Growth of Complexity in Molecular Evolution

Christoph Adami[1,2]

[1]Digital Life Laboratory 136-93
California Institute of Technology
Pasadena, CA 91125

[2]Jet Propulsion Laboratory 126-347
California Institute of Technology
Pasadena, CA 91109

Summary

Arguments for or against a trend in the evolution of complexity are weakened by the lack of an unambiguous definition of complexity. "Physical complexity", a measure based on automata theory and information theory, is a simple and intuitive measure of the amount of information that an organism stores, in its genome, *about* the environment it evolves in. It can be shown, both theoretically and experimentally, that physical complexity *must* increase in molecular evolution of asexual organisms in a single niche if the environment does not change. Responsible for this law of increasing complexity is natural selection, which acts as a natural Maxwell Demon in simple evolving systems. This law can be violated in co-evolving systems as well as at high mutation rates, in sexual populations, and in time-dependent landscapes. However, it turns out that these factors usually help, rather than hinder, the evolution of complexity.

Keywords: Evolution, Complexity, Entropy, Information, Digital Life

Introduction

Whether or not complexity increases in evolution is one of the central questions of evolutionary biology. Opinions about this subject vary, but generally belong to one of three camps. To the first it is obvious that complexity has increased, while the second claims there is not enough evidence to argue for or against an increase, and the third denies 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. Most agree however that, by and large, nobody knows precisely what is meant by the word "complexity" when referring to a biological organism. Indeed, while 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 and function, or to the sequence that codes for it. Functional (or structural) complexity is generally what we mean when we consider animals, but this seems to be the hardest measure to define. McShea[3] has studied several measures of structural and functional 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, while a trend can often be observed in the maximum, it tends to erode in the mean. Finally, complexity is often erroneously equated with "evolutionary success,"[1] a misconception that has led to many controversies.

The argument against a universal increase in complexity usually involves either a reference to the inordinate success of prokaryotes (and the fact that they likely represent the largest percentage of biomass on earth), or else examples in which adaptation has given rise to organisms that appear to have more complex ancestors in their direct line of descent. Neither of these observations runs contrary to the law of increasing complexity we shall be considering. In fact, as we shall see below, they are in perfect harmony with it, as long as it is understood that complexity is a relative, rather than universal, concept.

It is hard to imagine that a universal measure for structural or functional complexity can be devised, given that organisms differ so greatly in form and function. However, all these differences are sidestepped when we consider the nucleid acid sequences from whence all creatures come. Of course, we understand that the difficulty of biology lies precisely in the intricacy of this map from sequence to function. Nevertheless, it is very likely that a properly defined *sequence complexity* should mirror the complexity of the organism that the sequence gives rise to. If this is so (and at this juncture this is only a conjecture) then the problem of defining structural complexity can be demoted to the problem of defining sequence complexity, which is naturally much simpler because sequences are amenable to a mathematical characterization. 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., Ref.[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 a biologist, because random sequences do not give rise to organisms.

Recently, I have introduced the concept of "physical complexity" into the literature.[4] This complexity measure is carefully defined from an automata-theoretic point of view (just as Kolmogorov complexity was), but it has a very simple relationship to information theory, and turns out to be very intuitive. Furthermore, it appears to correspond exactly to that which biologists think is increasing when "self-organizing systems organize themselves". 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[5], due to the actions of natural selection in the guise of Maxwell's Demon. This will be illustrated 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 the wild, 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 multi-cellular organisms.

Physical Complexity

The physical complexity of a sequence refers to the amount of information that is stored in that sequence *about* a particular environment. For a genome, this environment is the one in which it replicates and in which its host lives, a concept roughly equivalent to what we call a *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, and not of the physical world in which such a sequence may mean something. Information, on the other hand, is always *about something*. Consequently, a sequence may embody information about one environment (niche) while being essentially random with respect to another. This makes the measure *relative*, or conditional on the environment, and it is precisely this feature that brings a number of important observations that are incompatible with a universal increase in complexity in line with a law of increasing physical complexity.

"Randomness" is in some ways the "flip side" of information, and is called "entropy" in information theory[6]. Entropy is a measure of how much there is to know potentially, or if applied to a sequence, a measure of how much information a sequence *could* hold, and thus quantifies our uncertainty about the genetic identity of a randomly selected individual from a pool. It is useful to think of sequence entropy as the *length* of a tape, while information is the length of tape containing recordings. Measurement (i.e., recording) turns empty tape into filled tape; entropy into information. As we shall see,

this is what happens during adaptation, and it is the force that drives the increase of complexity.

Information is a statistical form of correlation, and thus requires, mathematically and intuitively, a reference to the system that the information is *about.* The sequence on your information-filled tape allows you to make predictions about the state of the system the sequence is information about. This predictive capability implies that your sequence and the system have "something in common," that they are correlated. Your sequence will most likely *not* make predictions about any other system (unless the systems are very similar). If you do not know which system your sequence refers to, then whatever is on it *cannot* be considered information. Instead, it is *potential information* (a.k.a. entropy). This is the fundamental difference between entropy and information, often misrepresented in the literature[7].

Information-theoretic measures of complexity have been considered before, only to be discarded because of incorrect applications of the concept. Most often, *entropy* is used as a candidate for information-theoretic complexity. From the previous discussion, we realize that the entropy of a sequence is the amount of information it could possibly carry. Of course, this is just the length of the sequence. But it was recognized early on that sequence length is not a good predictor of organism complexity (the C-paradox), an observation that has discredited information-theoretic approaches to complexity. Physical complexity, a true measure of information, does not suffer from this handicap.

Nonmathematical, intuitive descriptions of complexity often make use of a concept very much akin to the one presented here. Most often, this is described as "genes embody knowledge about their niche" (Deutsch[8]) or, as put eloquently by Wilson: "(Organisms) encoded the predictable occurrence of nature's storms in the letters of their genes[9]." This is precisely what physical complexity measures, since physical complexity *is* information about the environment, which is used to make predictions about it. Being able to predict the environment allows an organism to exploit it for differential survival. Thus physical complexity translates into fitness for the organism.

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. I will refer to previous articles[4,5] for technical points not immediately relevant for the present non-technical discussion.

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 of relevance to us here, in particular because this definition is not practical, since it does not allow the unambiguous determination of sequence complexity from available data. However, it is worth mentioning that it is an instance of "effective complexity," a concept independently developed by Gell-Mann and Lloyd[10]. 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. Needless to say, 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.

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. In the section below, I give a technical exposition of the complexity measure. Readers who are satisfied with the intuitive description can skip this section without loss.

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

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

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

$$H_{\max}(X) = -\sum_{i=1}^{N} (1/N)\log(1/N) = \log N \ . \quad (2)$$

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.,

$$N = D^L \ . \qquad (3)$$

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:

$$H_{\max}(X) = L \quad . \qquad (4)$$

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 non-uniform: most sequences do not appear (because either they simply never occur, or because their fitness in the particular environment vanishes), while a few sequences are overrepresented. 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:

$$I(X:E) = H_{\max} - H(X \mid E) = L + \sum_{i=1} p_i \log p_i \ . \quad (5)$$

Here, I have introduced 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 while X in the above formulae represents an ensemble of sequences, E stands for one particular environment, not an ensemble of environments[1].

The probabilities p_i that go into the calculation of the conditional entropy in (5) 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[11,12]. 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

$$H(j) = -\sum_{i=G,C,A,T} p_i(j) \log p_i(j) \qquad (6)$$

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

---

[1] 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[6], but I will use the term information anyway.

$$H(X) \approx \sum_{j=1}^{L} H(j) \ , \qquad (7)$$

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

$$C_1(X) = L - H(X) \ , \qquad (8)$$

with $H(X)$ given by Eq. (7) 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.[13] for a review), in particular in populations that are not well equilibrated. Fortunately, as described in the appendix of Ref.[5], it is possible to correct for this using information about the strength of directional epistasis in the gene under consideration. In the following, we are going to assume that epistatic effects are sufficiently weak that the corrections can be ignored[2].

The Natural Maxwell Demon

Darwinian evolution is often described as a mechanism that increases the fitness of a population. Such a portrayal is problematic because the fitness of a population can depend on many parameters and is difficult to measure. Here, I show that Darwinian evolution increases the amount of *information* a population harbors about its niche (and therefore, its complexity). The mechanism by which evolution achieves this is best illustrated with the metaphor of Maxwell's Demon, a hypothetical creature invented by James Maxwell[14] to exemplify a possible threat to the second law of thermodynamics. This law guarantees that all isolated systems evolve from order towards disorder, that is, from a state of low entropy towards higher entropy. Incidentally, Darwinian evolution was long thought to violate precisely this theorem, since it appeared to produce more ordered states from less ordered ones, and protected the ordered ones from the decay ordained by the second law. However, evolving populations are not isolated (as stipulated in the conditions under which the second law holds), but rather are in contact with the sun that provides the power to keep them at low entropy.

---

[2] Epistasis is more problematic in asexual 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.[5].

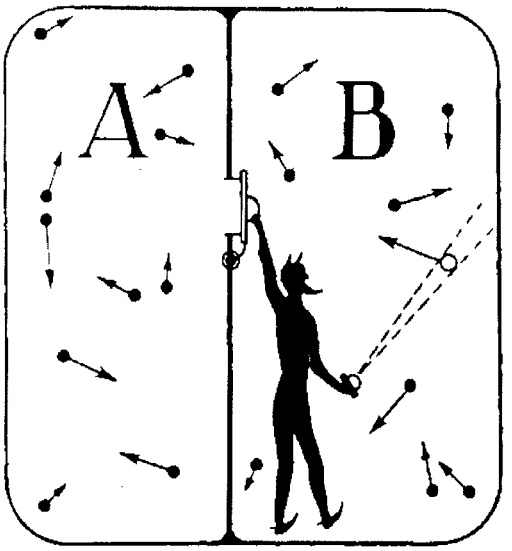

FIG. 1. Maxwell's Demon at work

Maxwell's Demon, on the contrary, seemed to be able to create ordered states without expenditure of energy. Consider the two halves of a vessel in the image in Fig. 1, which is a period depiction of the Demon at work. Initially, the halves are at the same temperature and pressure, a condition known as "thermodynamic equilibrium" in the lingo of physics. Imagine that there is a trap door separating the two halves, cunningly operated by the Demon. Equipped with a device that can measure the velocity of molecules headed for the door, he can make "intelligent" decisions about whether or not to open the door for the molecule in question. Should he open the door only for the fast molecules, say, and keep it closed for the slow ones, we can imagine that he can indeed create a disequilibrium between the two halves, such that the half with the slower molecules is in a more ordered state, in violation of the second law. While it was shown later that the operations of the demon do require energy after all[15], it does represent a convenient metaphor to illustrate the dynamics of natural selection.

In the case of Maxwell's Demon, *measurements* allowed the Demon to reduce the entropy of one half of the vessel. If we substitute information-theoretic uncertainty for the thermodynamical entropy, we can view evolution as a process that reduces the randomness (or uncertainty) inherent in a non-adapted genome by making judicious use of *measurements* performed on the environment, the niche. These measurements are not of the ordinary kind, though. Imagine each mutation, instead, as a candidate measurement (much like the molecule flying toward the trap door). A perfect natural Maxwell Demon then evaluates each mutation with respect to the immediate benefit for the population, and either *rejects it* (if it did not provide a fitness advantage), or *accepts it* (if it did). Note, however, that natural selection does not operate this perfectly: According to standard population genetics, beneficial mutations are fixed within the population with a probability of twice the selective advantage,[16] while neutral mutations also have a probability of "passing through the trap door". For finite population sizes, even deleterious mutations have a probability of becoming fixed (see below).

If the natural Maxwell Demon, i.e., natural selection, would operate perfectly, the complexity of a population could never decrease. As we have seen, the real Demon is imperfect, his measurements are imprecise and his actions probabilistic. Nevertheless, under normal conditions this dynamics must lead to an information increase on average. There are some situations, however, in which the Demon's handiwork crumbles altogether. We shall examine each of these important exceptions in detail below. Let us however first observe the Demon at work, in a population of digital organisms adapting to an artificial world, inside a computer.

Evolution of Complexity in Digital Organisms

Because evolution is an exceedingly slow process, it is difficult to witness the emergence of novelty and the concomitant 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 much shorter generation time. Digital organisms are just such a form of life: they are computer programs that self-replicate, mutate, and compete for resources[17-24]. 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 creating this virtual world, we do not specify a target sequence that represents the pinnacle of success. Instead, rewards (in the form of bonus execution time for the programs that reap them) are specified for *phenotypes* only, and thus natural selection acts on those. Because the underlying genetic space (the space of computer programs written in this particular language) is so high-dimensional, a large number of genotypes usually map to any particular phenotype, making the identification of a global genotypic optimum practically impossible. Phenotypes in this computational world are computational in nature, as we shall see presently.

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, no digital organism can survive, since they need to copy themselves to survive, and without the code being executed, no copying takes place. Fig. 2 below shows a sketch of the world that is created inside of a standard computer by running the Avida software[19], which is used for all the experiments described here.

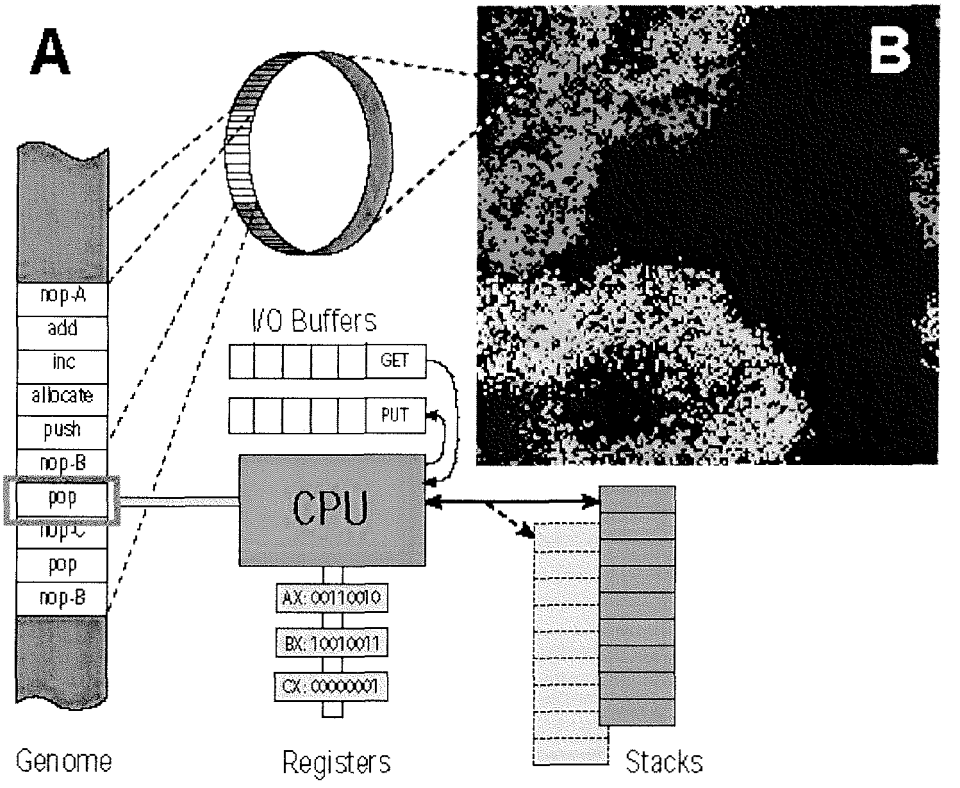

FIG. 2 (A) Each organism is executed on its own virtual CPU, which consists of an instruction pointer, registers (blue), two stacks (green), as well as input/output buffers (yellow). The genome of each organism is circular, like those of most bacteria and some biochemical viruses. (B) Population of digital organisms living in a two-dimensional artificial world with periodic boundary conditions, coloured according to their genotype. Because newborn programs are placed next to their progenitors by the Avida program, clones of identical organisms spread in roughly circular fashion.

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" are evolved by the digital organisms, and 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 the CPU described above. For these experiments[5], the

alphabet has 28 possible instructions, one of which is a logical primitive: NAND (the "not-and" operation).

Consider the behaviour of fitness over time (depicted here is the replication rate of the fastest replicator in a population of 3,600 adapting programs whose sequence length is kept fixed at 100, and seeded with a single simple replicator) in Fig. 3. Time is here measured in arbitrary units called "updates", where one update is the time it takes to execute about 30 instructions for each of the 3,600 programs in the population. One generation corresponds to between 10-100 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 much superior to all others. Following our discussion, we expect this increase in fitness to be associated with an increase in information, so that this genotype is a good candidate to inspect for an increase in complexity.

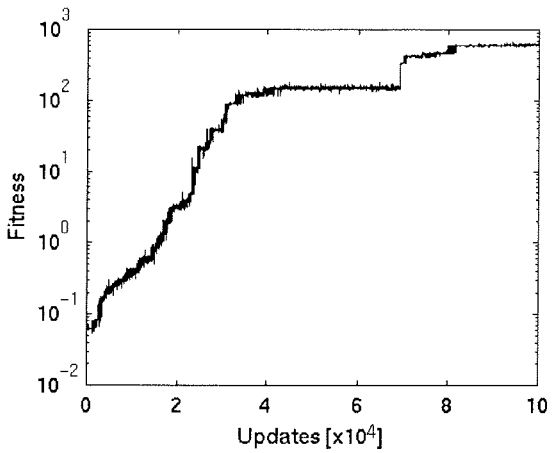

FIG. 3 Replication rate of fastest replicator in a population of 3,600 adapting digital organisms.

A plot of the approximate complexity (calculated according to Eq. (8)) can be seen in Fig. 4, 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 because mutations have not had sufficient time to randomize the "non-coding instructions". Also, evolution may struggle with a genome (hand-written by the experimenters) that is extremely ill-suited to the environment, but also difficult to "re-code." It may simply be "badly compressed", and evolution takes a while to find a better way to represent the same information. 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 Fig. 4, to which we now turn our attention.

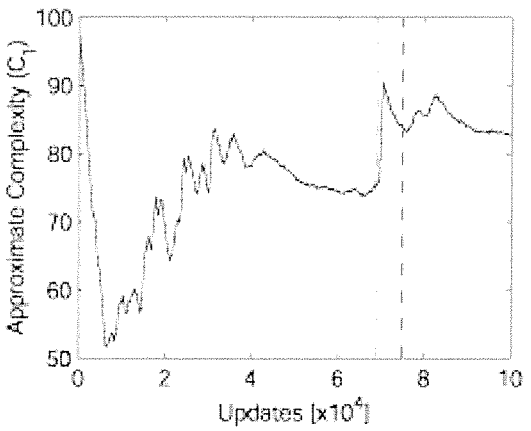

FIG. 4 Approximate complexity according to Eq. (8) for a population adapting to a complex world. The red-dashed lines indicate the times chosen as "pre" and "post"-transition, at which the genotypes analyzed in Fig. 5 were extracted.

Because of the hitchhiking effect mentioned earlier, the amount of information gained in the transition highlighted in Fig. 4 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. (6), we can extract dominating genotypes at just before and after the transition. In order to determine whether an instruction is entropy or information, we create all one-point mutants of the 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, while 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

$$H(x_i) \approx \log_{\mathrm{D}}(N_{\mathrm{viable}}) \quad , \qquad (9)$$

where N_{viable} is the number of neutral or beneficial substitutions at that locus. In equation (9), the logarithm is taken to the base of the alphabet size, thus ensuring that our measure for the "randomness" at each location is normalized to lie between zero and one. If we do this for two organisms before and after the transition, we obtain the per-site entropies of Figure 5. 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 Maxwell-Demon mechanism pointed out above: 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[5].

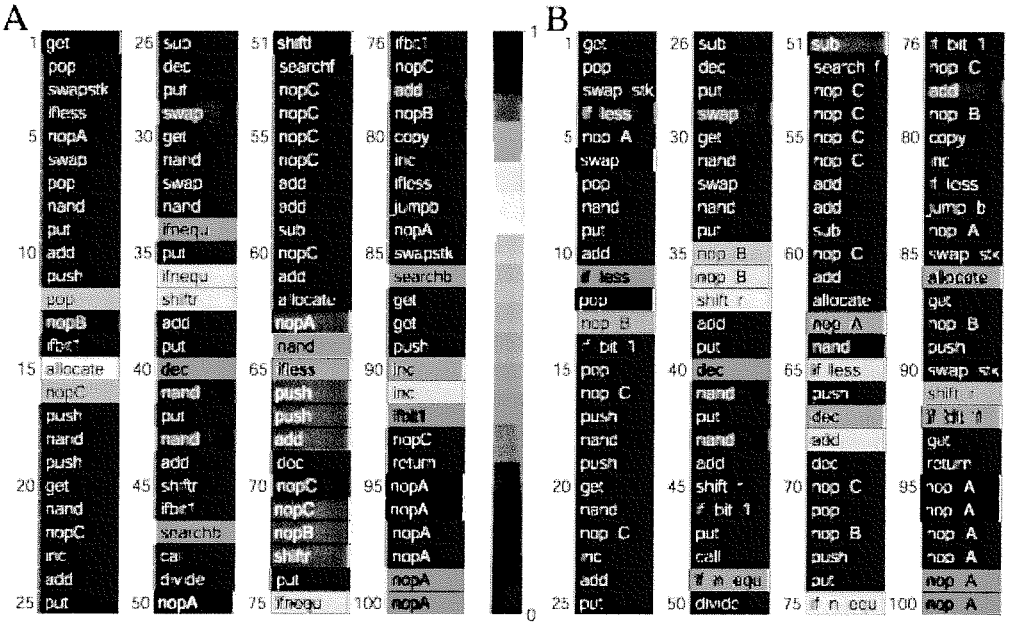

FIG. 5 Each instruction in the two genomes in (A) and (B) are coloured according to their per-site entropy (scale in the middle). An instruction that is fixed in the population has entropy close to zero (blue), implying that a mutation of that locus produces a non-functional organism. On the contrary, loci that can be mutated with impunity have entropy one (red). The genome (A) was extracted from the population after 2,991 generations (the left of the two red-dashed lines in Fig. 4), while genome (B) was extracted just after the transition at 3,194 generations (right dashed line in Fig. 4).

Causes for Complexity Declines

In this section I discuss the mechanisms by which complexity can fail to increase, or even crash. 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, not only will the organism be maladapted to this new environment, but also its measurable physical complexity will have decreased commensurately. High mutation rates can also lead to a loss of complexity. We can imagine high rates of mutations within the Maxwell's Demon metaphor as molecules flying towards the door separating the two halves of the vessel at a very high rate, too high for the Demon to control the door

accurately enough to prevent some molecules from escaping the vault he attempts to protect. In evolutionary theory, this process is known as the "hitchhiking" of deleterious mutations on beneficial ones. In small populations, high mutation rates are even more problematical, because the Demon becomes sloppy and information can leak through the trap door. In the extreme case (critically high mutation rates) the Demon can be paralyzed, leaving the door open (selection becomes inactive), a phenomenon known as the "error catastrophe" in the molecular evolution literature[24].

As is well known, sexual recombination can also lead to an accumulation of deleterious mutations, which is a signature of the Demon's malfunction. While 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[25]. 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 the Demon. In this case, 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. I discuss complexity growth in ecosystems briefly below.

Evolution of Ecosystem Complexity

With the present tools we cannot, strictly speaking, make any prediction about a trend in the complexity of entire ecosystems of interacting niches, since the concept of physical complexity only makes sense within an organism's own niche. An increase in complexity can only be observed in any particular niche, for the amount of time that this niche exists unchanged. Furthermore, the complexity of an organism can never exceed the potential complexity of the niche. Because niches do 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), while another 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 reduce an organism's complexity. In general, a change in niche will almost always 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, because the potential complexity of the niche has increased, an organism's amount of information about a niche can never exceed the amount of potential information in it.

Thus, we have to look at the process of co-evolution and its capacity to create more complicated environments as the possible unifying process that could give rise to an overall trend. Unfortunately, the mathematics of information in co-evolving environments appears as yet too daunting 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, which reduce the environment's complexity, 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, a formula that is difficult to write down, and a quantity even more difficult to measure.

Conclusions

In order to be able to speak about complexity, we must define it. In this review, 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 is equivalent to the dynamics of Maxwell's Demon, it possible to show that, within that niche, physical complexity must increase if the environment does not change.

While the Maxwell-Demon mechanism can fail in just about all those ways in which we are accustomed to see natural selection fail, 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 in a trend in the overall ("total") complexity of an ecosystem. Physical complexity tracks the performance of Maxwell's Demon perfectly (it increases if the Demon functions accurately, and decreases if he 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.

Acknowledgements

I thank Charles Ofria and Travis Collier for collaboration in the experimental work reported here, as well as Richard Lenski for valuable discussions. I am further indebted to Murray Gell-Mann for discussions on complexity, and for pointing out the relation between physical and effective complexity. Thanks are also due to Allan Drummond for comments on the manuscript. 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 SJ. Full House. New York: Harmony Books. 1996. p 3
2. Badii R, Politi A. Complexity–Hierarchical Structures and Scaling in Physics. Cambridge: Cambridge University Press. 1997.
3. McShea DW. Metazoan complexity and evolution: Is there a trend? Evolution 1996, 50:477-492.
4. Adami C, Cerf NJ. Physical complexity of symbolic sequences. Physica D 2000; 137:62-69.
5. Adami C, Ofria C, Collier TC. Evolution of biological complexity. Proc. Natl. Acad. Sci. USA 2000; 97:4463-4468.
6. Shannon CE, Weaver W. The Mathematical Theory of Communication. Urbana: University of Illinois Press. 1949.
7. Adami C. Information theory in molecular biology. 2002, in review.
8. Deutsch D, The Fabric of Reality. New York: The Penguin Press. 1997. p 179
9. Wilson EO, The Diversity of Life. Cambridge: Harvard University Press. 1992. p 9
10. Gell-Mann M, Lloyd S. Information measures, effective complexity, and total information, Complexity 1996; 2:44-52.
11. Basharin GP. On a statistical estimate for the entropy of a sequence of independent random variables. Theory Probability Appl. 1959; 4:333-336.
12. Schneider TD, Stormo GD, Gold L, Ehrenfeucht A. Information content of binding sites on nucleotide sequences. J. Mol. Biol. 1986; 188:415-431.
13. Wolf J, Brodie E, Wade, M. Epistasis and the Evolutionary Process. Oxford: Oxford University Press. 2000.
14. Maxwell JC. Theory of Heat. London: Longmans. 1871.
15. Landauer R. Irreversibility and heat generation in the computing process. IBM J. Res. Dev. 1961; 5:183.
16. Haldane JBS. A mathematical theory of natural and artificial selection. V: Selection and mutation. Proc. Camb. Phil. Soc. 1927; 23:838-844.
17. Ray TS. An approach to the synthesis of life. In Langton CG, Taylor C, Farmer JD, Rasmussen S. ed;Proc. Artificial Life II. Redwood City: Addison Wesley. 1991.

18. Adami C. Learning and complexity in genetic auto-adaptive systems. Physica D 1995; 80:154-170.
19. Adami C. Introduction to Artificial Life. New York: Springer Verlag. 1998.
20. Lenski RE, Ofria C, Collier TC, and Adami C. Genome complexity, robustness, and genetic interactions in digital organisms. Nature 1999; 400:661-663.
21. Wagenaar D, Adami C. Influence of chance, history, and adaptation on evolution in *Digitalia*. In Bedau MA, McCaskill JS, Packard NH, Rasmussen S. ed; Proc. Artificial Life VII. Cambridge: MIT Press. 2000. p 216-220
22. Yedid G, Bell G. Microevolution in an electronic microcosm. Am. Nat. 2001; 157:465-487.
23. Wilke CO, Wang JL, Ofria C, Lenski RE, Adami C. Evolution of digital organisms at high mutation rate leads to survival of the flattest. Nature 2001; 412:331-333.
24. Eigen M. Natural selection: a phase transition? Biophys Chem 2000;85:101-123.
25. Kondrashov AS. Deleterious mutations and the evolution of sexual reproduction. Nature 1988; 336:435-440.