Digital evolution of genomic complexity [1]

I General ideas of this research

Complexity has not been rigorously defined or easily measurable. Neither does natural selection guarantee that organisms will increase in 'complexity' as they evolve. But we all have an intuitive feeling that our biological system gets more and more complexities through the process. In order to make a case for or against this trend in the evolution, a recent information-theoretic definition identifies genomic complexity with the amount of information a sequence stores about its environment. Through digital evolution of genomic complexity, this research suggests that as the natural selection forces genomes to behave as a natural "Maxwell Demon", within a fixed environment the genomic complexity is forced to increase.

By skirting the issue of structural and functional complexity, they examined genomic complexity. The assumption is that genomic complexity is mirrored in functional complexity and vice versa. The new perspective is, on the one hand, genomic complexity can be defined in a consistent information-theoretic manner, as the *physical complexity*. "Physical" implies that it should measure the amount of information coded in the sequence about its environment. (This is a very good point). On the other hand, it has been shown that evolution can be observed in an artificial medium, providing a unique glimpse at universal aspects of the evolutionary process in a computational world.

In this system, the symbolic sequences subject to evolution are computer programs that have the ability to self-replicate via the execution of their own code. In this respect, they are computational analogs of *catalytically active RNA sequences* that serve as the templates of their own reproduction. In populations of such sequences that adapt to their world (inside of a computer's memory), noisy self-replication coupled with finite resources and an information-rich environment leads to a growth in sequence length as the digital organisms incorporate more and more information about their environment into their genome. Through these populations it is available to observe the digital organism incorporate more and more information about their environment into their genome, and also to distinguish distinct evolutionary pressures acting on the genome, and the analysis by mathematical framework.

II Using information theory to derive the genomic complexity

By using information theory, those parts of the genome that contain information should correspond in fact to the environment the genome lives in. This environment is extremely complex, consisting of the ribosomes the messages are translated in, other chemical and the abundance of nutrients inside and outside the cell, the environment of the organism proper (e.g., the oxygen abundance in the air as well as ambient temperatures)…

It is known that not all the symbols in an organism's DNA correspond to information. These are referred as "junk DNA", usually consist of portions of the code that are unexpressed or untranslated. In the absence of a complete map of the function of them, they could only correspond to *potential information*, that is, entropy. To distinguish whether a sequence is informative, the "fitness" is used, which implies that a sequence's information content is *conditional* on the environment it is to be interpreted within. And, a genetic locus that codes for information essential to an organism's survival will be fixed in an adapting population because all mutations of the locus result in the organism's inability to promulgate the tainted genome, whereas inconsequential (neutral) sites will be randomized by the constant mutational load. Thus, examining an ensemble of sequences large enough to obtain statistically significant substitution probabilities would thus be sufficient to separate information from entropy in genetic codes. The neutral sections that contribute only to the entropy turn out to be exceedingly important for evolution to proceed.

In a genome, for a site *i* that can take on four nucleotides with probabilities

$$\{ p_C(i), p_G(i), p_A(i), p_T(i) \} \qquad (1)$$

the entropy of this site is

$$H_i = -\sum_j^{C,G,A,T} p_j(i) \log p_j(i) \qquad (2)$$

The maximal entropy per-site is 1 (if take logarithms to base 4, i.e., the size of the alphabet), which occurs if all the probabilities are all equal to ¼. If the entropy is measured in bits (take logarithms to base 2) the maximal amount of information per site is two bits. A site stores maximal information if, in DNA, it is perfectly conserved across an equilibrated ensemble. Then, assign the probability p=1 to one of the bases and zero to all others, rendering H_i =0 for that site according to (2). The amount of information per site is thus,

$$I(i) = H_{max} - H_i \qquad (3)$$

Thus, for an organism of l base pairs the complexity is

$$C = l - \sum_i H(i) \qquad (4)$$

This is an approximation to the *true* physical complexity of an organism's genome. In reality, sites are not independent and the probability to find a certain base at one position may be conditional on the probability to find another base at another position. Such *correlations* between sites can render the entropy per molecule significantly different from the sum of the per-site entropies. We will explain later, that this is an important part to be taken into account actually. Thus,

$$H = -\sum_{g} p(g \mid E) \log p(g \mid E) \qquad (5)$$

this entropy per molecule involves an average over the logarithm of the conditional probabilities *p(g| E)* to find genotype *g* given the current environment E. In every finite population, estimating *p(g| E)* using the actual frequencies of the genotypes in the population results in corrections to (5) larger than the quantity itself, rendering the estimate useless.

Another avenue for estimating the entropy per molecules in (5) is the creation of mutational clones at several positions at the same time [2] to measure the correlation effects. This approach is feasible within experiments with simple ecosystem of digital organisms. The main idea is, for a length *l* with instruction taken from an alphabet of size *D*, a multi-site entropy, reflecting the average entropy of a sequence, can be defined as

$$H_l = \log_D [\omega(l) D^l] \qquad (6)$$

Where D^l is the total number of different sequences of length *l.* ω is the fraction of neutral mutants. $\omega(l)\ D^l$ is the number of neutral sequences, in other words all those sequences that carry the same information as the wild-type.

This value of ω can be got from testing all possible mutants of the wild-type for fitness and sampling the *n*-mutants to obtain $\omega(n)$, which is well fit by a two-parameter ansatz:

$$\omega(n) = D^{-\alpha n^{\beta}} \qquad (7)$$

where, 1-α measures the degree of neutrality in the code ($0<\alpha<1$), and β reflects the degree of correlations ($\beta>1$ for synergistic deleterious mutations, $\beta<1$ for antagonistic ones). It turns out that in most cases the constant α and β can be estimated from the first few n. Thus, the complexity of the wild-type can be approximated, by (4), (6) and (7) as follow:

$$C_l = \alpha\, l^{\beta} \qquad (8)$$

III Results for digital evolution and progression of complexity

By developing a tool to study evolution in a computational medium – the Avida platform [3], which hosts populations of self-replicating computer programs in a complex and noisy environment, within a computer's memory. The evolution is limited in speed by the computers used, with generations (for populations of the order $10^3 – 10^4$ programs) in a typical trial taking only a few seconds. Despite the apparent simplicity of the single-niche environment and the limited interactions between digital organisms, very rich dynamics can be observed. In this world, a new species can obtain a significant abundance only if it has a competitive advantage thanks to a beneficial mutation. The new species will gradually exert dominance over the population, bring the previously dominant species to extinction.

By tracking the entropy of each site in the genome, it is possible to measure the difference in complexity between the pair of genomes in Fig 1, separated by only 203 generations and a powerful evolutionary transition. New gene emerged in the transition at which sites the entropy has been drastically reduced. Continually surveying the entropies of each site gives Fig 2. The trend toward more conserved sites is obvious, and the evolutionary transitions can be identified by vertical darkened “band” because the genome instigating the transition replicates faster than its competitors thus driving them into extinction.

By plotting the sum of per-site entropies for the population (as an approximation for the entropy of the genome), Fig 3A shows it as a function of evolutionary time while Fig 3B shows the fitness of the most abundant genotype as a function of time. Fig 4 gives the complexity estimated by (4) (that means, no correlations were included in entropy – it is a pity! I don’t know why they don’t use (5) even though they provided the method to achieve this better estimation) as a function of time. It increases monotonically except for the periods just after evolutionary transitions, when complexity estimate settles down according to thermodynamics’ second law. Such a typical evolutionary history documents that the physical complexity, measuring the amount of information coded in the sequence about its environment, indeed steadily increases.

Note that the evolutionary transition is not of the equilibrium type so the drop in entropy is commensurate with the second law. In fact, each such transition is best described as a measurement, and evolution as a series of random measurements on the environment. Darwinian selection is a filter, allowing only informative measurements to be preserved. As a sequence, only mutations that reduce the entropy are kept while mutations that increase it are purged. Thus the mutations can be viewed as measurements and this is the classical behavior of the Maxwell Demon.

In this neat way, it appears that genomic complexity can be well measured during the evolution. However, to take into account other factors which play important roles in generating genomic complexities, such as transcriptional regulation factors, alternative splicing and posttranscriptional modification of RNA transcripts (RNA editing), I do not regard the digital evolution research above is credible enough now to reflect the real genomic complexity. Anyway, in the framework they developed, as the using of information theory, the considerations of correlations, we may find ways future.

[1] C. Adami, *et. al.*, *Evolution of Biological Complexity*, Proc. Nat. Acad. Sci. USA 97 (2000), 4463-3368
[2] R. E. Lenski, *et. al.*, *Genome complexity, robustness and genetic interactions in digital organisms*, Nature 400 (1999), 661-664
[3] C. Adami, *et. al.*, *Introduction to Artificial Life* (1998), Springer, New York

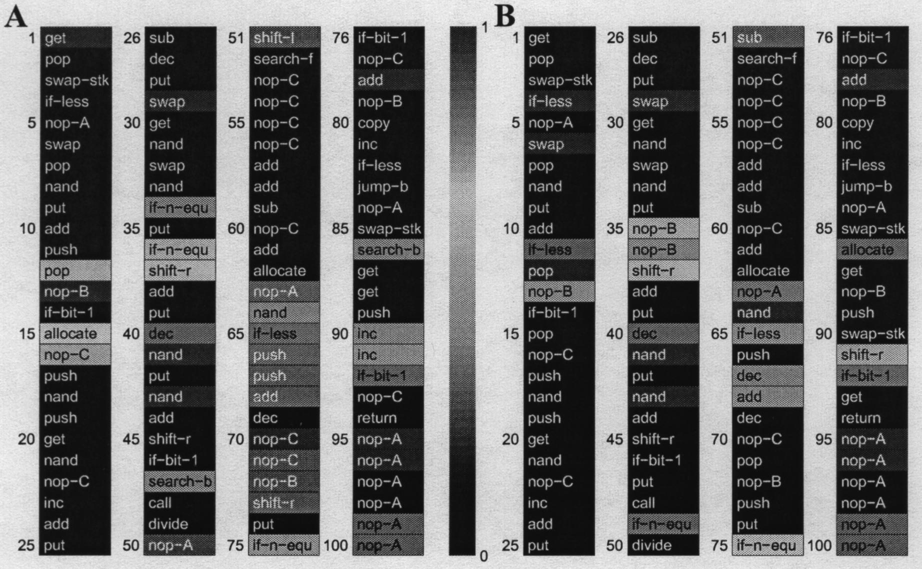

Figure 1: Typical **Avida** organisms, extracted at 2,991 (A) and 3,194 (B) generations respectively into an evolutionary experiment. Each site is color-coded according to the entropy of that site (see color bar). Red sites are highly variable whereas blue sites are conserved. The organisms have been extracted just before and after a major evolutionary transition.

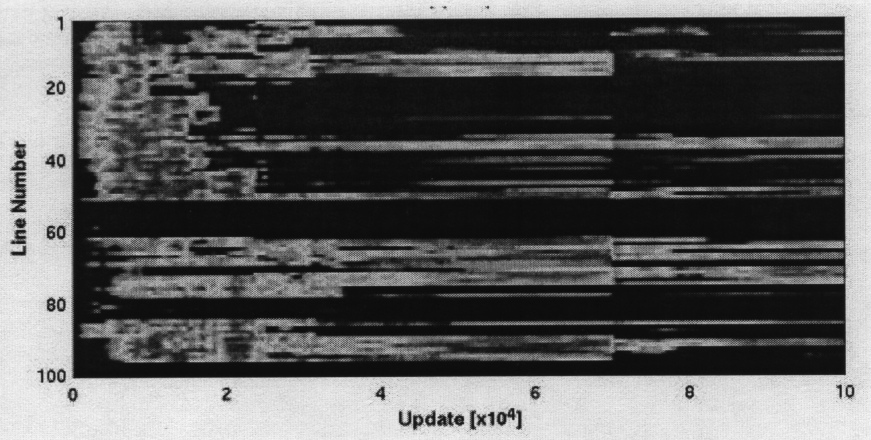

Figure 2: Progression of per-site entropy for all 100 sites throughout an **Avida** experiment, with time measured in "updates" (see *Methods*). A generation corresponds to between 5 and 10 updates, depending on the gestation time of the organism.

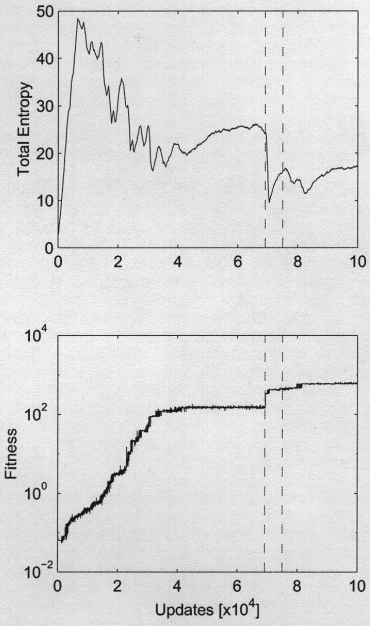

Figure 3: (A) Total entropy per program as a function of evolutionary time. (B) Fitness of the most abundant genotype as a function of time. Evolutionary transitions are identified with short periods in which the entropy drops sharply, and fitness jumps. Vertical dashed lines indicate the moments at which the genomes in Fig. 1 A and B were dominant.

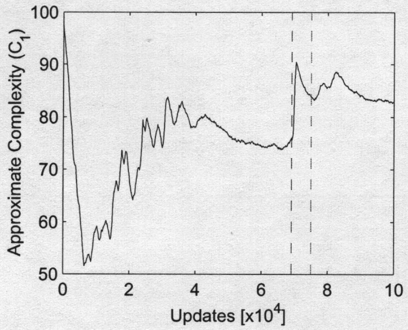

Figure 4: Complexity as a function of time, calculated according to Eq. (4). Vertical dashed lines as in Fig. 3.