On the Gradual Evolution of Complexity and the Sudden Emergence of Complex Features

Charles Ofria*,**
Michigan State University

Wei Huang**
Michigan State University

Eric Torng**
Michigan State University

Abstract Evolutionary theory explains the origin of complex organismal features through a combination of reusing and extending information from less-complex traits, and by needing to exploit only one of many unlikely pathways to a viable solution. While the appearance of a new trait may seem sudden, we show that the underlying information associated with each trait evolves gradually. We study this process using digital organisms, self-replicating computer programs that mutate and evolve novel traits, including complex logic operations. When a new complex trait first appears, its proper function immediately requires the coordinated operation of many genomic positions. As the information associated with a trait increases, the probability of its simultaneous introduction drops exponentially, so it is nearly impossible for a significantly complex trait to appear without reusing existing information. We show that the total information stored in the genome increases only marginally when a trait first appears. Furthermore, most of the information associated with a new trait is either correlated with existing traits or co-opted from traits that were lost in conjunction with the appearance of the new trait. Thus, while total genomic information increases incrementally, traits that require much more information can still arise during the evolutionary process.

Keywords
Avida, biocomplexity, digital evolution, experimental evolution, information theory, self-replication

1 Introduction

When Darwin released his theory of evolution by natural selection, some biologists objected on the grounds that it could not explain the origin of entirely new traits, arguing that incipient forms on the way to producing useful structures served no purpose and hence would not be supported by natural selection [10]. Darwin had anticipated this difficulty, and in his first edition of *On the Origin of Species* [3] noted that "In considering transitions of organs, it is so important to bear in mind the probability of conversion from one function to another" and furthermore, that "Different kinds of modification would, also, serve for the same general purpose." In other words, Darwin proposed that new traits arise due to functional shifts from previously existing traits, and that even though

* Contact author
** Department of Computer Science & Engineering, Michigan State University, East Lansing, MI 48824, USA. E-mail: ofria@cse.msu.edu (C.O.); huangw10@cse.msu.edu (W.H.); torng@cse.msu.edu (E.T.)

any specific modification may be unlikely to evolve, many different pathways may all lead toward the same goal; even if the outcome we witness seems incredibly unlikely, it was only one of many possibilities.

Substantial evidence has been collected indicating that complex traits can be produced through the evolutionary process, including such examples as the evolution of the eye [6, 12, 18], exoskeletons [21], the Krebs cycle [9], insecticide resistance [11], nutritive "milk" in the cockroach [23], and many others [2, 4, 5, 17]. In terms of artificial life studies, a detailed demonstration of the evolution of complex traits in digital organisms was performed using the Avida system [8].

The question remains as to how information flows into the genome to generate new complexity. If a new trait must arise from nothing, then the probability for this trait to appear drops exponentially as the number of genomic positions it requires increases. Even considering that portions of the information required to express a new trait come from existing traits (which may be destroyed in the process), all of the unique information associated with the new trait must arise without the benefit of selection for the incipient forms. Thus a new complex trait can only arise when the majority of its information has already made it into the genome through the presence of preexisting traits, typically of lesser complexity.

2 Measuring Genomic Information

We can use information theory to frame our analysis of biological complexity by measuring the amount of information associated with each trait, and how much information is shared between traits. In general, an organism can be thought of as an information channel [14] that passes a message (its genome) to a recipient (its offspring) in the presence of noise (mutations). The key difference here from traditional studies of information theory is a feedback loop: The message being passed will be used to build the next organism, which, in turn, will pass the message on. Any flaws in the information transmitted may reduce the capacity of the subsequent channel. However, errors also have a small probability of being beneficial and improving the quality of the offspring.

In a typical population, organisms will be subject to a uniform mutation rate across their genomes. Positions that contain no information can mutate freely, while those that store information important to the organism's survival will typically be perfectly conserved. We can use the distribution of symbols at each genomic position to estimate the amount of information stored at that position, and then sum these values to approximate the genomic complexity [1, 7].

In Shannon information theory [19, 20], the entropy

$$H(X) = -\sum_{x \in X} p_x \log p_x \tag{1}$$

is used to measure the expected number of bits required to specify the state of a system, and is maximized when all possible states are equally likely. In the case of genomic sequences, any reduction from maximal entropy at a locus indicates that information is being stored at that locus. A site that encodes no information can take on all possible symbols without affecting fitness; with all symbols equally probable, entropy is maximized and information content is zero. Likewise, a site that cannot be altered without reducing fitness encodes the maximum amount of information about its environment.

Given D possible symbols at each genomic position, we can take our logarithms to base D to normalize the entropy at that position to be between 0 and 1. To determine the amount of information at this position, we can then simply subtract its entropy from the maximal entropy of 1. Next, we approximate the information content of the whole system by summing the per-site information. This is only an approximation, because it ignores interactions between sites. For example, if two

positions maintain redundant information, they may be miscounted. However, in practice this approximation proves to be sufficiently accurate.

3 Experimental System

To study what happens to information in genomes during the evolutionary acquisition of a complex trait, we must use a system where we can isolate organisms associated with the adaptive event and then fully manipulate them to measure the information content of their genomes. This requires a level of knowledge about the state of the system that is unattainable in natural systems, but can be easily imagined in a computational one. In this article, we use digital organisms in the artificial life system Avida [16].

Digital organisms are self-replicating computer programs that are subject to mutations and will evolve to best survive in their virtual environment. Evolution progresses very naturally in these systems, and they are not subject to artificial fitness functions. Indeed, organisms will often evolve novel and even surprising survival strategies [8, 15, 22].

In Avida, each digital organism consists of a virtual CPU that processes a sequential program (the genome) made up from a genetic language consisting of 26 possible commands [13]. The commands in the language are simple, atomic operations that can be strung together to produce programs to perform any possible computation; that is, the genetic language is Turing complete. The organisms exist in an environment where resources are present that they can metabolize into extra CPU cycles by performing Boolean-logic based operations. In all of the experiments presented here, we provide unlimited resources to the organisms, and space (due to a finite population size) is the only limiting factor that they must compete over. Additional CPU cycles allow an organism to execute their genome more rapidly, and therefore increase its replication rate. In the default Avida environment used here, nine resources are available, each associated with a different Boolean-logic operation. The most complex of these is EQU, which is the focus of this study, as it has been for previous studies of complex traits [8]. To perform the EQU operation, an organism must input two 32-bit sequences and output a third sequence where each bit is set to one if the corresponding bits in the other sequences are the same (that is, both one or both zero) and is set to zero otherwise.

We can perform tests on digital organisms without influencing the course of evolution in the population by first copying them into a separate, isolated test environment. In particular, if we want to measure the amount of information contained at a genomic position, we can perform every possible point mutation at that site and measure the fitness relative to the unmutated wild type. If all modifications at a position are lethal, that position clearly contains the maximum possible information, which we will term 1 *complexity unit* (CU). If all mutations at that site are neutral, the site clearly contains no information (0 CU). For intermediate values, we can use the fitness values of the mutants to calculate their expected relative abundance (as described in [7]) and use this abundance to calculate the information content.

We can also measure genomic information in an altered environment to tease out what traits the information is associated with. For example, if we wish to measure the amount of unique information associated the EQU trait, we can calculate an organism's information content in the default environment, and then recalculate it in an environment that lacks the EQU resource. It will not be harmful for an organism in the latter environment to lose the ability to perform EQU, and hence mutations at sites associated with this trait will no longer be detrimental unless that site's information was shared with a secondary trait. The difference between these two measures is the amount of information uniquely associated with the EQU-metabolizing trait.

4 Experiments and Results

We performed a set of 50 Avida runs using a default configuration. Specifically, we used population sizes of 3,600 organisms, a per-site mutation rate of 0.0025, and a genome-level insertion and

deletion rate of 0.05. We used an ancestor with a length 100 genome, which was capable of self-replication but had no other functional traits. The organisms evolved in an environment with unlimited resources associated with all nine basic bitwise logic operations. We then measured the complexity of organisms in four different environments.

4.1 Three Sources of Complexity

We tested the complexity of the organisms in four different environments to identify which complex traits each instruction's information is associated with. *Environment I* is the default environment with unlimited resources for all nine logic operations. Measuring the information content in this environment gives us the total complexity for an organism. *Environment II* is identical to environment I, but without the resource associated with the complex trait EQU. The difference between an organism's complexity in environments I and II provides us with the amount of information uniquely associated with performance of the EQU trait. *Environment III* contains only the EQU resource, but none of the others, and *environment IV* contains no resources at all. The difference between the complexity measures in environments III and IV yields the total amount of information used to perform EQU, even if this information is also associated with other traits.

Figure 1 shows these information measures for a typical Avida run. In this example, the first organism in the population to possess the EQU trait appeared at update 24,891. While this organism did gain EQU (and 4.9 CU of information over its parent), it simultaneously lost the ability to perform the logical operation AND. When we tested it in environment II, where EQU is not rewarded, the information content of the genome actually dropped by 5.0 CU worth of information that was converted to be used solely by EQU. At this point, the genome contained a total of 9.9 CU of information unique to EQU (4.9 CU new + 5.0 CU converted) and another 23.0 CU of information shared with other traits, for a total of 32.9 CU required for EQU.

A total of 24 out of the 50 trials obtained the EQU trait. Figure 2 displays the average information content of genomes from these lineages, centered on update zero as the point where EQU is first acquired. From this figure, it is clear that the information associated with the EQU trait comes

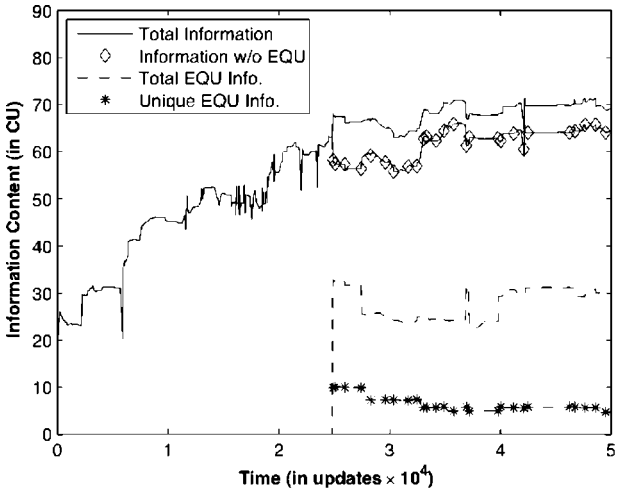

Figure 1. Information content over the course of a typical Avida experiment. We selected the most abundant genotype from an evolved population, and measured the information content of all of its ancestors along its line of descent to track complexity over time. The solid line displays the full complexity of the organisms, including a small jump up when EQU was first evolved. The diamond line shows the complexity of the organisms when EQU is ignored; a tradeoff during the evolution of EQU causes this measure to drop slightly. The dashed line tracks the total information associated with the EQU trait, and the starred line shows how much of that complexity is unique to EQU, indicating that most of its information is shared with other traits.

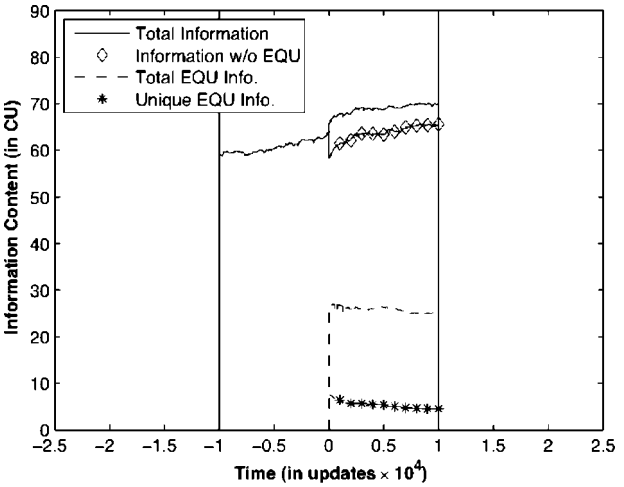

Figure 2. Average information during the acquisition of the complex trait EQU. We centered 24 runs on their acquisition of the EQU trait and averaged their information values to create this graph. The average focuses on 10,000 updates before and after the time point where EQU is first acquired, centered on that event as time zero. As with the sample run displayed in Figure 1, the total complexity (solid line) tends to increase due to EQU, but tradeoffs cause the complexity excluding EQU (diamond line) to drop. The total complexity associated with just EQU (dashed line) makes up almost half of the complexity in the genome, but most of this is shared, as indicated by the unique information associated with EQU (starred line), which is much smaller.

from three different sources. The majority of the information (72.3%) is shared with other traits. The remainder is split between information used that was once part of now-defunct traits (22.0%) and newly incorporated information that has appeared for the first time due to this trait (a mere 5.7%). The total information associated with the newly evolved EQU task is, on average, 26.8 CU; the mean newly incorporated information is 1.5 CU, and the mean information coming from now-defunct traits is 5.9 CU.

The relative importance of these information sources varies widely from one run to the next. To better understand the breakdown of where information comes from, we have created a histogram isolating the sources, as seen in Figure 3. Surprisingly, the change in organism complexity when EQU first arises was negative in three of the 24 cases (Figure 3a). This occurs when the traits lost during the acquisition of EQU actually had a greater combined complexity than the complexity of the EQU trait itself. In all three cases, the complexity was restored (through the reacquisition of lost traits) shortly after the rise of EQU. It is equally unexpected that the complexity of the organisms can go up when EQU first appears in the environment where EQU is not rewarded (Figure 3c). This happens if other traits arise simultaneously with EQU, or existing traits reorganize due to EQU's appearance. Figure 3b shows the unique information associated with the EQU task. The range of this unique information is large, from 0 to 16.7 CU. The total information associated with EQU differs among organisms, ranging from 12.8 to 35.3 CU (Figure 3d).

4.2 The Distribution of Complexity Changes from Beneficial Mutations

To further examine the amount of information that can appear *ex nihilo* in association with a beneficial mutation, we performed a set of experiments at three different time points in our 50 evolved populations (1% of the way into the run, 10% of the way in, and at the end of the run). At each time point, we chose 2,000 random organisms to mutate (with replacement) from each of the 50 populations. In each case we examined the complexity both before and after the mutation. Figures 4a to 4c show the distribution of the complexity changes associated with beneficial mutations at each time point. While a total of 100,000 mutations were examined for each graph (50 runs × 2,000 mutations tested per run), only a small fraction of them were beneficial. At 1,000 updates there were 18,675

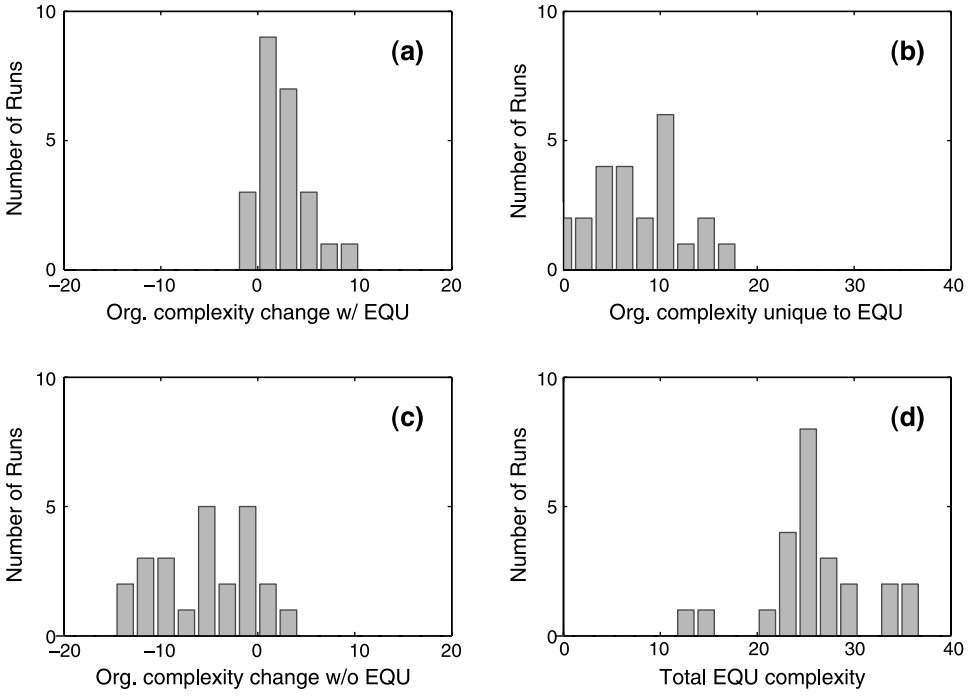

Figure 3. Histograms indicating the distribution of information in the 24 trials where organisms developed the EQU trait: (a) the change in complexity for the whole organism when EQU first arises; (b) the complexity unique to EQU, not shared by other traits; (c) the change in complexity of all traits other than EQU at the time EQU arises; and (d) the total complexity of the EQU trait, including information shared with other traits.

beneficial mutations, at 10,000 updates there were 5,500 beneficial mutations, and at 100,000 updates there were 2,473 beneficial mutations. Over time, populations became better adapted to the environment, reducing the amount of information left to incorporate, explaining the decline in the number of beneficial mutations.

The distribution of complexity increases (i.e., where complexity change is greater than zero) exhibit a clear exponential distribution at all three time points, as shown in Figure 4a–c. On average, every instruction that adds to complexity should have the same probability of mutating into the genome, meaning that more complex structures are exponentially less likely to occur by random chance alone. This leads to the exponential distribution observed, as long as a sufficiently rich environment exists. Figure 4c has a more uneven distribution of complexity increases than the earlier time points. This is largely due to the reduced availability of beneficial mutations as the organisms become so well adapted to the environment that there is little room for improvement; the mean number of tasks evolved over 50 runs at 1,000 updates is 0.1; at 10,000 updates it is 5.6; and at 100,000 updates it is 7.9 out of the possible 9. In other words, at 100,000 updates very few traits are still available to be acquired, leading to this effect.

To account for the effects of simpler tasks on the continued evolution of complexity, we performed a set of control runs where the organisms evolve in an environment without any resources associated with logic operations. We then analyzed the complexity changes from mutations (as above procedure) when these organisms and their mutants were moved into the nine-resource environment. Figure 4d shows the distribution of these complexity effects at update 10,000. We again see a similar exponential distribution of positive complexity changes. This indicates that the absence of building blocks does not limit the amount of new complexity that comes into the genome; however, without building blocks it becomes nearly impossible to evolve any of the more complex traits, such as EQU.

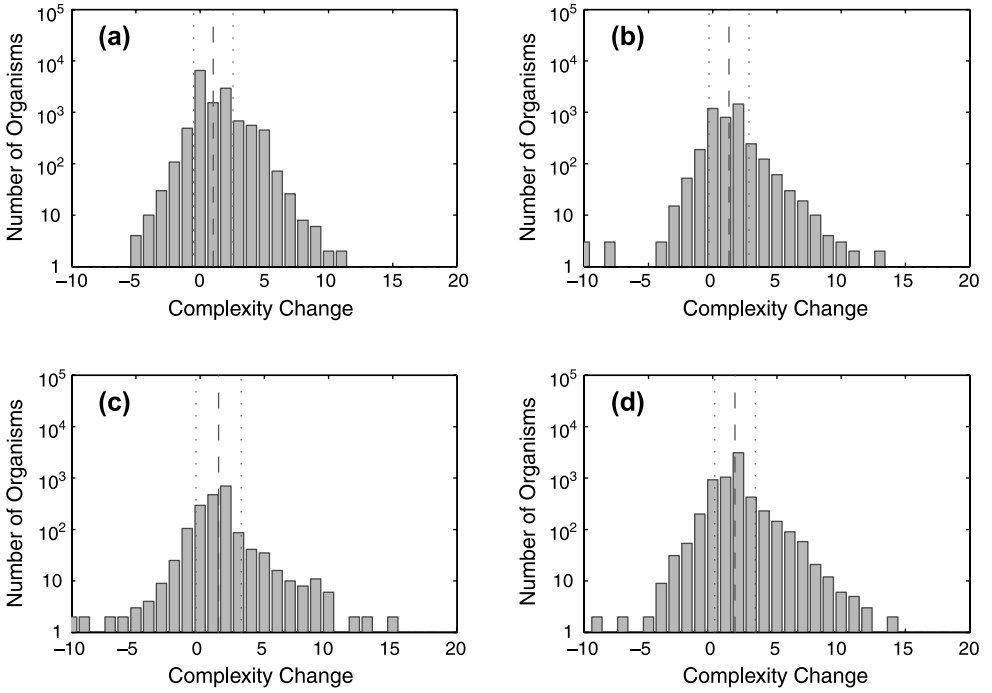

Figure 4. The distribution of complexity changes with beneficial mutations from 100,000 random sampled organisms in the populations (a) at 1,000 updates, (b) at 10,000 updates, (c) at 100,000 updates under a nine-task environment, and (d) at 10,000 updates of control runs. The dashed line represents the mean complexity changes, and the dotted lines are one standard deviation from the mean.

To improve our understanding of how complexity enters the genome, we need to focus on those mutations that provide a selective advantage and persist over evolutionary time scales. Clearly, mutations that lead to beneficial traits will be selected for. Thus we expect mutations that have become fixed in the population to provide more complexity, on average, than random mutations (even random beneficial mutations). This effect is clearly shown in Figure 5, where we examine the distribution of complexity changes in lineages in the nine-resource environment.

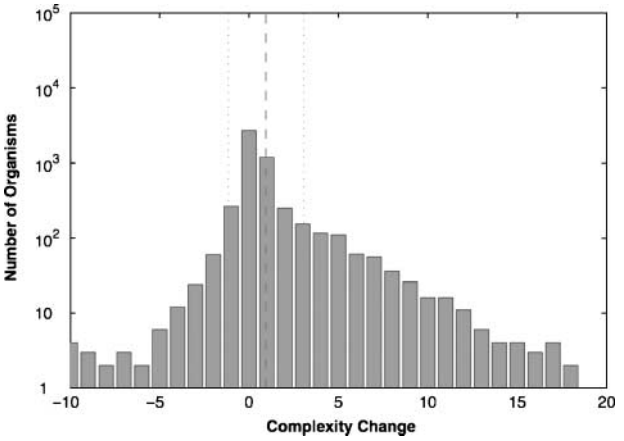

Figure 5. The distribution of complexity changes from a beneficial mutation along the lineages of 50 runs in the nine-task environment. The dashed line represents the mean complexity changes, and the dotted lines are one standard deviation from the mean.

5 Conclusions

We have demonstrated that when a new trait is introduced into a genome, its complexity comes from three sources: complexity shared with other traits; complexity from once functional but now defunct traits; and complexity belonging to newly incorporated information. We performed random mutation tests on populations at multiple time points and evolutionary stages. Based on the distribution of complexity changes due to beneficial mutations, we see that mutations leading to large complexity changes are exponentially rare. In particular, we see that the complex trait EQU requires on average 26.2 CU, but we never see more than half of this complexity appear in a single mutational step. If a complex trait such as EQU is to evolve, it must utilize preexisting complexity.

The experiments presented here were performed in digital organisms that are simple compared to life in the natural world, where there are far more than just nine resources for the organisms to interact with. Note that every new trait that appears provides new building blocks to work with, which, in turn, increase the probability for even more complex adaptation. We expect dramatically more complex traits to emerge in the natural world under such gradual increases in organismal complexity.

Acknowledgments

Thanks are due to Chris Adami, Dehua Hang, Richard Lenski, Matt Rupp, and Tom Schmidt for fruitful discussions about this research, as well as three anonymous reviewers for their useful comments. This research was funded by NSF grants CCF-0523449 and EIA-0219229 and the Quantitative Biology and Modeling Initiative at Michigan State University.

References

1. Adami, C., Ofria, C., & Collier, T. C. (2000). Evolution of biological complexity. *Proceedings of the National Academy of Sciences of the USA*, *97*, 4463–4468.
2. Chen, L., DeVries, A. L., & Cheng, C.-H. C. (1997). Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. *Proceedings of the National Academy of Sciences of the USA*, *94*, 3811–3816.
3. Darwin, C. (1859). *On the origin of species by means of natural selection*. London: Murray.
4. Dean, A. M., & Golding, G. B. (1997). Protein engineering reveals ancient adaptive replacements in isocitrate dehydrogenase. *Proceedings of the National Academy of Sciences of the USA*, *94*, 3104–3109.
5. Ganfornina, M. D., & Sanchez, D. (1999). Generation of evolutionary novelty by functional shift. *BioEssays*, *21*, 432–439.
6. Goldsmith, T. (1990). Optimization, constraint, and history in the evolution of eyes. *Quarterly Review of Biology*, *65*, 281–322.
7. Huang, W., Ofria, C., & Torng, E. (2004). Measuring biological complexity in digital organisms. In *Proceedings of the Ninth International Conference on Artificial Life*, 315–321.
8. Lenski, R. E., Ofria, C., Pennock, R. T., & Adami, C. (2003). The evolutionary origin of complex features. *Nature*, *423*, 139–144.
9. Meléndez-Hevia, E., Waddell, T. G., & Cascante, M. (1996). The puzzle of the Krebs citric acid cycle: Assembling the pieces of chemically feasible reactions, and opportunism in the design of metabolic pathways during evolution. *Journal of Molecular Evolution*, *43*, 293–303.
10. Mivart, S. G. J. (1871). *On the genesis of species*. London: D. Appleton and Co., Macmillan.
11. Newcomb, R. D., Campbell, P. M., Ollis, D. L., Cheah, E., Russell, R. J., & Oakeshott, J. G. (1997). A single amino acid substitution converts a carboxylesterase to an organophosphorus hydrolase and confers insecticide resistance on a blowfly. *Proceedings of the National Academy of Sciences of the USA*, *94*, 7464–7468.
12. Nilsson, D., & Pelger, S. (1994). A pessimistic estimate of the time required for an eye to evolve. *Proceedings of the Royal Society of London*, *256*, 53–58.
13. Ofria, C., Adami, C., & Collier, T. C. (2002). Design of evolvable computer languages. *IEEE Transactions in Evolutionary Computation*, *17*, 528–532.

14. Ofria, C., Adami, C., & Collier, T. C. (2003). Selective pressures on genomes in molecular evolution. *Journal of Theoretical Biology*, *222*, 477–483.

15. Ofria, C., Adami, C., Collier, T. C., & Hsu, G. K. (1999). Evolution of differentiated expression patterns in digital organisms. *Lecture Notes in Artificial Intelligence*, *1674*, 129–138.

16. Ofria, C., & Wilke, C. O. (2004). Avida: A software platform for research in computational evolutionary biology. *Artificial Life*, *10*, 191–229.

17. Piatigorsky, J., & Wistow, G. (1991). The recruitment of crystallins: New functions precede gene duplication. *Science*, *252*, 1078–1079.

18. Salvini-Plawen, L., & Mayr, E. (1977). On the evolution of photoreceptors and eyes. *Evolutionary Biology*, *10*, 207–263.

19. Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423, 623–656.

20. Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana, IL: University of Illinois Press.

21. Terwilliger, N. B., Ryan, M. C., & Towle, D. (2005). Evolution of novel functions: Cryptocyanin helps build new exoskeleton in *Cancer magister*. *Journal of Experimental Biology*, *208*, 2467–2474.

22. Wilke, C. O., Wang, J., Ofria, C., Adami, C., & Lenski, R. E. (2001). Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, *412*, 331–333.

23. Williford, A., Stay, B., & Bhattacharya, D. (2004). Evolution of a novel function: Nutritive milk in the viviparous cockroach, *Diploptera punctata*. *Evolution & Development*, *6*, 67–77.