RESEARCH ARTICLE

Suboptimality and Complexity in Evolution

DAVID W. SNOKE,[1] JEFFREY COX,[2] AND DONALD PETCHER[2]

[1]*Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, Pennsylvania 15260; and*
[2]*Department of Physics, Covenant College, Lookout Mountain, Georgia 30750*

Received 4 March 2014; accepted 17 June 2014

A scalable model of biological evolution is presented which includes energy cost for building new elements and multiple paths for obtaining new functions. The model allows a population with a continual increase of complexity, but as time passes, detrimental mutations accumulate. This model shows the crucial importance of accounting for the energy cost of new structures in models of biological evolution. © *2014 Wiley Periodicals, Inc. Complexity 21: 322–327, 2015*

Key Words: numerical models of evolution; fitness collapse; growth of complexity; optimality; vestigiality

1. INTRODUCTION

A recent article by Guttenberg and Goldenfeld [1] (GG) presented some very helpful general conclusions about models of evolution. In particular, two conditions were presented for an ongoing increase of complexity. First, relative fitness must be scalable; in other words, fitness should be measured relative to the degree of fitness of other creatures in the environment at any point in time. Second, there should be no upper bound to the increase of complexity, which means that models which use a fixed fitness landscape [2,3] will not be realistic.

In all modeling, there is a balance between simplicity and generality versus detailed accuracy. The simplified GG model contained the above two key elements which allowed for continuous increase of complexity, something which had not been obtained for the previous models, such as TIERRA, AVIDA, and Webworld (for a survey, see discussion and references in Ref. 1). Conversely, the GG model neglected two aspects of real biological systems which have general impact. One of these is that there is an additional energy cost to increased complexity. In the model of Ref. 1, a fixed array of elements was considered, in which each element could be changed to another with no net increase (or decrease) of energy cost. In real systems, building new systems is costly, and the cost of carrying along useless or redundant systems is one of the arguments for the efficiency of existing living systems, as excess baggage is dropped as too costly.

A second simplification of the model of Ref. 1 was that there was effectively a direct reward for increased complexity, in the form of increased survivability for creatures with extra complexity. This was implemented by having an organism with a larger number of beneficial elements always defeat an organism with fewer such elements. The problem with this is that nature does not reward complexity per se, it rewards functions that enhance survival and reproduction. Thus, there may be many paths to the same function, some simpler and some more complex, and all

Correspondence to: David Snoke, E-mail:snoke@pitt.edu

 Vol. 21 No. 1
DOI 10.1002/cplx.21566
Published online 1 July 2014 in Wiley Online Library
(wileyonlinelibrary.com)

will be rewarded roughly the same whether or not the function is done elegantly or not; only the overall energy cost will deter some versions of obtaining the function.

In this article, a model is presented which includes the aspects of scalability and unlimited increase of complexity, but which incorporates two key features, namely (1) an energy cost for increasing number of elements produced and (2) multiple paths to beneficial functions. The primary issue addressed here can be termed vestigiality. In general, evolution is an undirected process; new functions are obtained as the system makes stabs in the dark, building new elements of the system, at some energy cost, and keeping them if they turn out to aid survivability and discarding them if not, through natural selection. For this process to work, it must be possible to make many stabs in the dark, assuming that the number of systems which lead to useful functions is a small fraction of the total of all possible systems which could be constructed. This assumption is eminently reasonable given the fact that most randomly generated strings of DNA do not lead to proteins which have a folded and compact form [4,5], and presumably also do not lead to proteins with useful function. Thus, at any point in time there must be some number of stabs in the dark going on, in the form of nonfunctional systems which might become functional with a few changes. We can call these at-present-useless systems "vestigial" although this term typically refers to elements of living systems which once had function; here it is generalized to include all nonfunctioning elements, both those which have had a function and lost it, and those which do not have a function but might eventually obtain one. In general, in a single snapshot in time, both types of elements would look the same.

A critical question is what fraction of all the systems in a creature may be expected to be vestigial at any point in time. There are two competing processes. On one hand, the energy cost of carrying vestigial systems makes them weakly deleterious, not neutral, which tends to reduce their number. Conversely, without stabs in the dark, that is, new systems which might eventually obtain new function but as yet have none, no novelty can ever occur, and no increase of complexity. Thus, if the energy cost of vestigial systems is too high, no evolution will occur. This leads one to expect, and has historically led evolutionary theorists to expect [6–8], that living systems carry a significant fraction of vestigial, or nonfunctional, elements, as well as quasi-vestigial elements which function with much less than optimal efficiency. By contrast, Bialek and coworkers (W. Bialek, private communication) [9,10] have recently argued that nearly all systems in existing living organisms function at near optimality, which implies high efficiency.

It is, therefore, valuable to create a model of evolutionary processes which incorporates the possibility of vestigial elements, and ask to what degree these vestigial elements are eliminated after a system successfully obtains new function, in an overall context in which complexity increases without limit as in the GG model.

2. NUMERICAL MODEL

The model presented here has an $n \times n$ grid of organisms with periodic boundary conditions, which compete with each other, as in the GG model. Each organism is replaced by one of its neighbors in the succeeding generation according to the following procedure. First, the fitness f of the organism and the average fitness $\bar{f}$ of the neighbors of an organism is found. Then, the probability, P, of replacing the organism is set equal to a quasi-Fermi–Dirac distribution:

$$P = \frac{1}{e^{(f-\bar{f})/T}+1}, \tag{1}$$

where T is an effective temperature which determines the "nonselectable range" of variations [11]. If the organism succumbs to replacement, one of its nearest neighbors is selected to replace it, with a probability given by

$$w_i = \frac{e^{(f_i - f_{\min})/T}}{\sum_{j \in n.n.} e^{(f_j - f_{\min})/T}}, \tag{2}$$

where $f_{\min}$ is the minimum fitness among the nearest neighbors.

The role of the Fermi–Dirac function here is just to give a smooth transition; no statement is being made about any element actually acting as a fermion. The Fermi–Dirac distribution (1) ensures that for $\bar{f}$ well above f, the organism is replaced with one of its neighbors with nearly unity probability, while for $\bar{f}$ well below f, the organism is almost never replaced. The middle range of random replacement is a key aspect of the model. If the nonselectable zone is too small, then all novelties are instantly eliminated because they have some energy cost. Conversely, if the nonselectable region is large, there will be effectively no energy cost for adding nonfunctional elements. Overall, this replacement scheme ensures that at high T, the replacement process is completely random, while at low T, the fitter organisms always replace the weaker. (High T also models the case of empty space between neighbors, which has the effect of allowing less fit individuals to have higher propagation probability).

The fitness of a single organism is determined by how many new functions it has obtained. New functions are represented as targets of unit radius randomly distributed in an unbounded d-dimensional space. The state of a single organism is represented as a set of paths starting at the origin in this d-dimensional space. Mutations consist of random additions and deletions of unit steps in these paths. If one of the paths reaches one of the target spots, then the organism is rewarded with an increased fitness

DOI 10.1002/cplx

value for each target that is hit. Near-hits can also be rewarded, based on the distance from the end of a path to the target spot.

There is no upper limit to the fitness of an organism, since there is assumed to be an infinite number of targets randomly distributed in the unbounded d-dimensional parameter space. Each organism can add new paths by mutations with some small probability, with no limit on the number of paths per organism other than the numerical resources. However, there is an energy cost proportional to the length of each path which is subtracted from the fitness of the organism. Thus, for example, an organism which has 50 paths with an average length of 50 units, out of which two paths have hit targets, might have a reward of 2×1000 points, minus 2500 points for the total length of all the paths. This model, therefore, realistically allows that the scoring of whether a single mutation is deleterious or beneficial is contextual: if a mutation is part of a path that allows a new function, it is beneficial, while the same path, not yet complete, is weakly deleterious. If multiple paths hit the same target, the reward is given only once; in other words, there is no reward for redundancy. Although a reward value for redundancy, namely extra rewards for multiple paths hitting the same target, is biologically realistic, when it was added to this model, it was found not to change the general conclusions given below. It is not fundamentally different from having two or more reward targets near to each other in the d-dimensional parameter space, which would just change the effective density of the targets.

Each element is coded for a different random direction in the d-dimensional space. Thus, each path, as it grows by preference of insertion of steps over deletion of steps in the mutation process, carries out a random walk in the d-dimensional space, and the net increase of the volume covered by the paths is diffusive. The unbounded nature of the d-dimensional space is modeled using a finite space which has extent larger than the distance covered by the random walk of any path produced during a simulation.

It should come as no surprise that growth of the paths only occurs if additions are more likely than deletions, on average. (Deletion of whole systems, which is common in real biological systems, was simulated in this model as deletion of whole segments of the paths used here, but has the same effect as making the overall deletion rate much higher.) In DNA in prokaryotic organisms, single deletions are more actually likely than single insertions [12], but eukaryotic organisms are assumed to have increasing number of expressed proteins over time, possibly through gene duplication. In the model described here, however, the paths in parameter space should not be equated with DNA sequences. Rather, they model an accumulation of many random, expressed structures which are built inside an organism and which collectively may lead to some new function. This may be taken to occur at the macroscopic, organ level, at the level of interactions between cells, or inside the cell at the protein, molecular machine level. The feature in the model that many different paths of the paths may lead to the same target allows the biologically realistic effect that more than one type of system may accomplish the same function, and only the function, not the structure, is rewarded; more inefficient structures (longer paths) are penalized only by their energy cost. Some functions can only be accomplished by many elements, however; this is represented in the model by the fact that some targets are further from the origin. Reaching targets further from the origin corresponds to increasing complexity.

For the particular results presented here, the d-dimensional space was assumed to be three-dimensional. This can be taken as generally representative of three physical parameters which can be varied, for example, the hydrophobicity, ionicity, and hydrogen bonding capability of the amino acids in a protein. The number of relevant parameters in living systems is likely much more than three, which would force a search in a high-dimensional space, which generally makes it harder to hit targets by random walk; we restricted the number of parameters to three to make the simulation numerically tractable. The mean distance between targets in this space was kept around five units, where a unit is defined as the length per step in one dimension per path element.

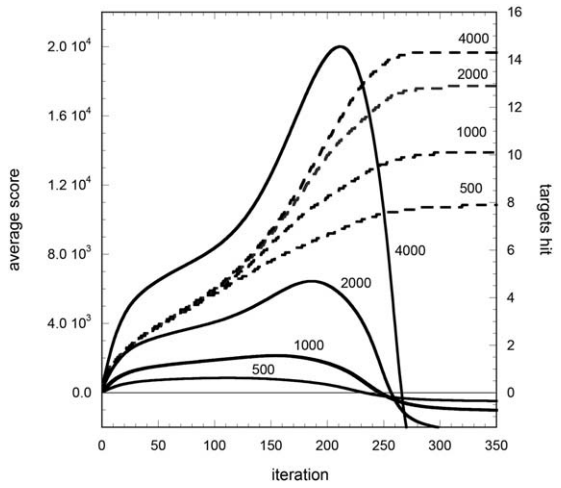
In this model, reasonable results, in which the average number of targets hit by paths increases monotonically over time, could be obtained if each organism was assumed to have one mutation per generation per path of length 10 units; additions were assumed to occur 65% of the time, deletions 34% of the time, and new paths created by splitting an old path 1% of the time. These are unrealistically high mutation rates per generation in real living systems, but a "generation" in this model may be taken as one mutation-time, which may be many physical generations.

3. RESULTS

The model was run for a 50×50 grid of organisms, and 2000 instantiations of this model were evolved and averaged for each set of parameters. As shown in the dashed curves Figure 1, the number of targets hit increases monotonically over time for several different reward values. A somewhat surprising general result, however, is that for a fixed target reward value, the population always undergoes a "fitness collapse" after some time, as shown in the solid lines of Figure 1. Even though the average number of hit targets increases, the overall fitness of the population eventually drops precipitously into the negative. This collapse of a population bears some similarity to other models of collapse, for example, the complexity catastrophe scenario due to increasing interaction of elements [13], error catastrophe analysis [14], and the

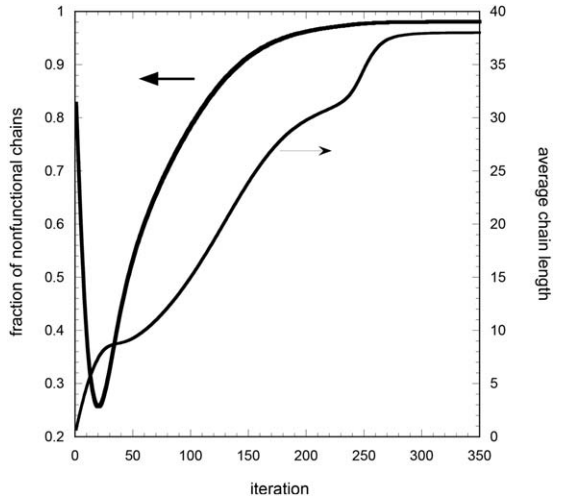
DOI 10.1002/cplx

FIGURE 1



Heavy lines, left axis: average population fitness as a function of time for the model described in the text, for a 50×50 grid, for four different target reward values. Fitness was determined by awarding the full point value for hitting each target exactly and half the full point value points for coming within one unit distance of a target. Energy cost was assessed as one point per path step. Thin lines, right axis: the average number of targets hit, for the same four cases.

FIGURE 2



Heavy line, left axis: the average degree of vestigiality as a function of time for the same model as Figure 1, for the case of a reward of 2000 per target hit. Thin line, right axis: the average path length per organism for the same model.

proposed idea of "genetic entropy" [15]. In each case, weakly deleterious mutations can accumulate, leading to overall less fitness of a population. In this model, as there is a nonselectable region, mutations with small energy cost, namely additional steps in paths that do not hit targets, are not selected against and, therefore, spread throughout the population. Over time, the entire population accumulates far more negative energy cost from vestigial paths than the reward for hitting function targets can overcome. It is not possible to overcome this collapse by varying the temperature T or by varying the reward value for hitting a target. As discussed above, if $T \to 0$, competition will remove individuals with nonfunctional paths before they can hit a target, while if $T \to \infty$, selection becomes completely random, and individuals which hit targets are not preserved. The natural unit for T is the energy cost per new path step. For a broad range of T around this value, making the reward value per target much greater than the step cost makes the cost of adding new steps negligible once a target has been hit, and so the length of the nonfunctional paths increases exponentially. Conversely, if the reward value is small, the paths diffuse in the d-dimensional space only up to some maximum value and do not continue to hit targets, because long paths are removed by selection due to their energy cost.

The collapse can be prevented if the reward value increases geometrically as more targets are hit (a linear or power-law increase is not enough). This may be biologically realistic, as having one function may make possible much greater functions. Simply allowing the possibility that hitting one target may open up new possible targets does not help prevent the collapse, unless those new targets have geometrically increasing reward value. This variation was explored in this study by altering the model described above to allow additional searches for a given organism whenever one of its paths hit a target; in effect, when a target was hit, new paths were allowed to start from the target instead of from the origin. This maps back to simply allowing a new origin with new paths in a new d-dimensional space, whenever a target is hit. Allowing these extra searches, with the same fixed reward value, actually reduced the average score compared to the case of a single origin for the paths. The additional paths in the extra searches just add to the total length without changing the overall probability of hitting targets.

Another general result is that in all cases where the model gives a continual increase of the number of targets hit, the vestigiality eventually approaches unity. This is true both in the case of constant reward value, which leads to fitness collapse as shown in Figure 1, and for the case of geometrically increasing reward value, which has ever-increasing average population fitness. Figure 2 shows the vestigiality as a function of time for a model with strong, fixed reward for hitting targets; the vestigiality is defined as the total length of paths which do not hit targets divided by the total length of all paths. The vestigiality does drop rapidly at first, as paths which hit targets are

DOI 10.1002/cplx

favored over nonfunctional paths. Eventually, however, the vestigiality increases even though the average number of targets hit by the organisms increases monotonically over time, because the reward for hitting new functions overcomes the cost of the large number of vestigial paths.

4. CONCLUSIONS

Although this model is simplified, it is reasonable to conclude that if evolution of biological systems is in a state of upward increase of complexity, then we should expect a large amount of vestigiality, that is, nonfunctional elements, both at the genomic level and at higher levels. This was historically the reason for treating a large part of the genome as "junk" DNA [16,17], although recently it has been argued that noncoding DNA has function [18]. In addition, vestigiality is natural to expect at the macroscopic level; this model indicates that the question is not why vestigial organs or structures exist, but why living systems are not full of them.

This conclusion is challenged by the growing popularity in recent years of the paradigm of optimality in systems biology as a general phenomenon (e.g., [19–22]). One option for adopting this viewpoint would be to assume that complexity is no longer increasing, that is, biological systems alive today have nearly reached an ultimate optimum. This might occur, in terms of this model, if the energy cost of extra path length relative to reward for new functions were to jump up strongly after some point in time, effectively stopping the scalable increase of complexity. This scenario would imply that earlier forms of living systems should have significantly greater vestigial matter than present ones.

Another alternative would be if there were no gaps between targets (set to around five unit steps in the present model); in other words, if there was a reward for every step along the way to a target, so that there was no need for stabs in the dark. This view seems to contradict the general observation that almost all single mutations are deleterious [23,24], but could be supported by a general study showing that those mutations which are not deleterious are almost always part of a successive path of progress toward new functions, an effect sometimes termed "channels in parameter space" [25].

In summary, the model presented here indicates that careful attention must be given to the energy cost of nonfunctional structures, which must have some energy penalty in order for selection to work to eliminate them. If the cost is too high, a scalable increase of complexity is not possible, while if the energy cost is too low, unrealistic, unchecked growth of nonfunctional structures will occur.

While this model is not biologically realistic because it predicts many population collapses, it points out the crucial need to account for the balance of these two pressures in realistic models of biological systems. In existing living systems, the fitness collapse seen in this model appears to be prevented by mechanisms which quickly eliminate nonfunctional elements, while leaving functional elements untouched. This type of mechanism would seem to prevent "stabs in the dark" of any great magnitude, and thus prevent ongoing increase of complexity. The model presented here has the flexibility to allow for new variations in the future, such as multiple species ("coevolution"), population bottlenecks, and changing environmental stresses, which may account for mechanisms to remove vestigial systems and prevent fitness collapse, while allowing for increasing complexity.

ACKNOWLEDGMENT

The authors thank W. Bialek and H. Salman for helpful conversations, and R. Jones for use of a computer cluster.

REFERENCES

1. Guttenberg, N.; Goldenfeld, N. Cascade of complexity in evolving predator-prey dynamics. Phys Rev Lett 2008, 100, 058102/1–4.
2. Gavrilets, S. Fitness Landscapes and the Origin of Species; Princeton University Press: Princeton, NJ, 2004.
3. Orr, H. The genetic theory of adaptation: A brief history. Nat Rev Genet 2005, 6, 119–127.
4. Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P. Molecular Biology of the Cell, 4th ed.; Garland Science: New York, 2002; p 141.
5. Keefe, A.D.; Szostak, J.W. Functional proteins from a random-sequence library. Nature 2001, 410, 715–718.
6. Scadding, S.R. Do vestigial organs provide evidence for evolution? Evol Theory 1981, 5, 173–176.
7. Naylor, B.G. Vestigial organs are evidence of evolution. Evol Theory 1982, 6, 91–96.
8. Scadding, S.R. Vestigial organs do not provide scientific evidence for evolution. Evol Theory 1982, 6, 171–173.
9. Tkacik, G.; Callan, C.G., Jr.; Bialek, W. Information flow and optimization in transcriptional regulation. Proc Natl Acad Sci USA 2008, 105, 12265–12270.
10. Bialek, W.; de Ruyter van Steveninck, R.R.; Tishby, N. Efficient representation as a design principle for neural coding and computation. In: IEEE International Symposium on Information Theory (IEEE Cat. No. 06TH8883C), New York, 2006; pp 659–663.
11. Kimura, M. Model of effectively neutral mutations in which selective constraint is incorporated. Proc Natl Acad Sci USA 1979, 6, 3440–3444.
12. Mira, A.; Ochman, H.; Moran N.A. Deletional bias and the evolution of bacterial genomes. Trends Genet 2001, 17, 589–596.

DOI 10.1002/cplx

13. Solow, D.; Burnetas, A.; Tsai, M.-C.; Greenspan, N.S. Understanding and attenuating thecomplexity catastrophe in Kauffmans N K model of genome evolution. Complexity 1999, 5, 53–66.
14. Eigen, M. Error catastrophe and antiviral strategy. Proc Natl Acad Sci USA 2002, 90, 13374–13376.
15. Sanford, J.C.; Baumgardner, J.; Brewer, W.; Gibson, P.; ReMine, W. Using computer simulation to understand mutation accumulation dynamics and genetic load, In: Proceedings of International Conference on Computational Science-ICCS 2007, Lecture Notes in Computer Science Vol. 4888, Springer, Berlin, 2007, pp 386–392.
16. Ohno, S. So much 'junk DNA' in our genome. Brookhaven Symp Biol 1972, 23, 366–370.
17. Orgel, L.E.; Crick, F.H.C. Selfish DNA: The ultimate parasite. Nature 1980, 284, 604–607.
18. Makalowski, W. Not junk after all. Science 2003, 300, 1246–1247.
19. Tanaka, R.; Csete M.; Doyle, J. Highly optimised global organisation of metabolic networks. IEEE Proc Syst Biol 2005, 152, 179–184.
20. Tyo, K.E.; Alper, H.S.; Stephanopoulos, G.N. Trends Biotechnol 2007, 25, 132.
21. Segré, D.; Vitkup, D.; Church, G.M. Analysis of optimality in natural and perturbed metabolic networks. Proc Natl Acad Sci USA 2002, 99, 15112–15117.
22. You L.; Yin, J. Evolutionary design on a budget: Robustness and optimality of bacteriophage T7. IEEE Proc Syst Biol 2006, 153, 46–52.
23. Bell, G. Selection: The Mechanism of Evolution; Chapman and Hall: New York, NY, 1997; p. 49.
24. Clune, J.; Misevic, D.; Ofria, C.; Lenski, R.E.; Elena, S.F.; Sanjuán, R. Natural selection fails to optimize mutation rates for long-term adaptation on rugged fitness landscapes. PLoS Comput Biol 2008, 4, e1000187.
25. Flatt, T. The evolutionary genetics of canalization. Q Rev Biol 2005, 80, 287–316.

DOI 10.1002/cplx