

Validation of Evolutionary Activity Metrics for Long-Term Evolutionary Dynamics

Andrew Stout
Department of Computer Science
University of Massachusetts, Amherst
Amherst, MA, USA
stout@cs.umass.edu

Lee Spector
Cognitive Science
Hampshire College
Amherst, MA, USA
lspector@hampshire.edu

ABSTRACT

As artificial life systems grow in number and sophistication, it is becoming increasingly important that the field agree on principled metrics for evaluating them. This report describes a series of experiments validating the evolutionary activity statistics developed by Bedau and his colleagues [2, 3, 4]. The work described herein was motivated by a feeling that the ‘null hypothesis’—that is, that the evolutionary activity statistics fail to exclude intuitively unlikelike systems from Class 3 dynamics [3]—had not been sufficiently disproved in the existing literature. We conducted a series of experiments applying the statistics to such systems, attempting to ‘break’ the scheme by measuring Class 3 dynamics in an intuitively unlikelike system. The evolutionary activity measurement scheme has so far proved robust to our attempts to break it, but we believe that this work is still valuable in advancing the validity of the scheme, and that this does not mean the scheme is without shortcomings.

Categories and Subject Descriptors

I.6 [Simulation and Modeling]: Simulation Theory—*Model classification*; J.3 [Computer Applications]: Life and Medical Sciences—*biology and genetics*

General Terms

Measurement, Experimentation, Theory

Keywords

evolutionary activity metrics, long-term evolutionary dynamics, artificial life

1. INTRODUCTION AND MOTIVATION

As artificial life systems grow in number and sophistication, it is becoming increasingly important that the field

agree on principled metrics for evaluating them. Of particular importance is a means of evaluating claims of “open-ended evolution”. How can the long-term evolutionary dynamics of an artificial system be quantified, and how can they be compared to those of the natural biosphere? As Bedau et al observe [4], such a comparison necessarily implies a classification scheme for the long term behavior of natural and artificial evolutionary systems.

One approach to classification is the suite of *evolutionary activity* statistics developed by Bedau and his colleagues [2, 3, 4]. Evolutionary activity focuses on trends involving *adaptation*, as opposed to trends involving *complexity*, and attempts to capture the intuition that evolution can be characterized by

- constant introduction of new adaptive “innovations” into the system, and
- persistence of those innovations.

Section 2 discusses the evolutionary activity statistics in detail.

The work described herein was intended as a ‘validation’ of the evolutionary activity statistics. Motivated by the observation that populations under no selective pressure can display surprisingly sophisticated dynamics [1], we felt that the ‘null hypothesis’ had not been sufficiently disproved in the published literature. That is, if an artificial evolutionary system under no selective pressure, or under ‘uninteresting’ selective pressure, nonetheless displays “unbounded evolutionary activity” (defined below), then these statistics do not adequately capture the intuitive criteria we expect of a lifelike system. The null hypothesis, which must be disproved if the evolutionary activity measurement scheme is to be considered valid, is that the scheme fails to exclude such intuitively unlikelike systems. If the statistics stand up to our attempts at a ‘reductio’, as indeed has been the case so far, then we will have gained a better understanding of them, and in rigorously testing the null hypothesis we will have advanced the validity of these measures.

The rest of this report is organized as follows. The next section gives a detailed summary of the evolutionary activity metrics. Section 3 describes the experiments we carried out to validate the metrics, and the results we obtained. Section 4 draws conclusions, and finally section 5 discusses ideas for future work.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

GECCO’05, June 25–29, 2005, Washington, DC, USA.
Copyright 2005 ACM 1-59593-010-8/05/0006 ...\$5.00.

2. BACKGROUND

The most mature and most widely applied measure of long-term evolutionary dynamics in the artificial life literature has been developed by Bedau and his colleagues over more than a decade [2, 3, 4]. Here we will first give an overview of the approach and important improvements that have recently been made to it, finally giving the specific formulas used in this work (which reflect those improvements).

The scheme is predicated on the view that “an evolving system consists of a population of components, all of which participate in a cycle of birth, life and death, with each component largely determined by inherited traits” [3]. Birth allows for potential innovations—that is, for new components to enter the system through genetic operators—and adaptive innovations are likely to persist. Based on long-term trends in the introduction of new adaptive innovations, the persistence of innovations, and the system’s diversity (number of distinct components), the system’s dynamics may be classified as having unbounded, bounded, or no evolutionary activity.

The classification is general enough that it can be applied to any system in which a record of the components’ existence times is available. Much of this generality comes from the general nature of the notion of a ‘component’, which is why no mention of what constitutes a component has been made thus far. The evolutionary activity statistics were originally formulated (in [2]) using alleles as components. When applied to the fossil record of the natural biosphere, taxonomic families have been used as components [3]. The most popular ‘unit’ for use as components in artificial systems, particularly in more recent work, has been a full genotype. We follow that trend and use genotypes as components in the work presented here, and discuss the some of the ramifications of this choice in section 5.

To measure a system’s evolutionary activity, an activity counter is kept for each component. At each time step the counter for each component is incremented according to an activity incrementation function. Different functions have been used, but a popular one appropriate for genotype or taxonomic family components is based on *existence*: simply increment a component’s activity counter by one if it exists at that time step. However, what we really want here is a measurement of *adaptive* activity—not simply existence—so it is necessary to normalize with respect to a “neutral shadow” of the system, which is identical to the real system except that selective pressure is entirely random. For the natural biosphere, this normalization is often assumed to be *de facto* accomplished (as in [3]). That is, any component (e.g. taxonomic family) that persists in the fossil record is assumed to be adaptive, although this assumption is not always valid, as in cases of exaptation (see, e.g. [7]). From these counts, then, several statistics can be computed:

Diversity is simply the number of distinct components in the system.

Mean cumulative activity is the sum of activity counts over all components in the system (*total cumulative activity*) divided by diversity.

New activity measures the introduction of new adaptive components (see details below).

Formulas for all three are given below. Based on these statistics, a system may display one of three fundamental classes

of long-term evolutionary dynamics:

Class 1 No adaptive evolutionary activity. Diversity is bounded, mean cumulative activity is zero, and new activity is zero.

Class 2 Bounded evolutionary activity. Diversity is bounded, mean cumulative activity is bounded, and new activity is positive.

Class 3 Unbounded evolutionary activity. Diversity is unbounded and mean cumulative activity is bounded (Class 3a), diversity is bounded and mean cumulative activity is unbounded (Class 3b), or both are unbounded (Class 3c). New activity is positive. [6]

Obviously, evolution cannot be literally unbounded on a computer with finite resources. The concern here is with *trends*—for example, that mean cumulative evolutionary activity continues to rise throughout the course of the experiment, and shows no indication of plateauing. This demands, of course, that experiments be run for long enough to exceed any transient behavior.

Bedau *et al* measure the fossil record of the natural biosphere as displaying Class 3a dynamics [3]. Thus this classification scheme constitutes an ALife test—if one of the major goals of Artificial Life is to develop systems which display dynamics similar to natural life, it is clear that Class 3 dynamics are desirable.

Channon ([5]) makes two important criticisms of the evolutionary activity measures, which he then addresses in [6]. One is that in a system with bounded diversity, retention forever of a single component results in unbounded evolutionary activity (Class 3b). Channon suggests examining *median* cumulative evolutionary activity instead of *mean* cumulative evolutionary activity. His main criticism, however, is of the method of normalization:

The test relies on normalization (or validation) from a shadow that can drift away from core aspects of the real run that it is intended to shadow. [...] Once the real and shadow populations have been allowed to evolve, we are no longer comparing the real run with a true shadow. The longer the period since the shadow was initialized to match the real run, the less relevant the shadow is to the real run. [5]

Channon’s solution is to ‘reset’ the shadow population to match the real population immediately after each snapshot (when an entry in the record of component existences is made—due to computational resource constraints, typically evolutionary activity statistics are collected by taking snapshots at regular intervals, every 100 timesteps/generations, for instance). Channon then normalizes activity at the level of components.

We now (finally) present the formulas for the evolutionary statistics, following Channon’s ([6]) improvements. The component activity counter increment function for the real and shadow populations is simply 1 if the component exists in the population at that time step, and 0 otherwise:

$$\Delta_i^R(t) = \begin{cases} 1 & \text{if comp. } i \text{ exists in real run at time } t \\ 0 & \text{otherwise} \end{cases},$$

$$\Delta_i^S(t) = \begin{cases} 1 & \text{if comp. } i \text{ exists in shadow run at time } t \\ 0 & \text{otherwise} \end{cases}.$$

Then the normalized evolutionary activity increment function for a component is computed by subtracting off the shadow activity from the real activity for that component:

$$\Delta_i^N(t) = \Delta_i^R(t) - \Delta_i^S(t).$$

This increment is then used to compute the normalized evolutionary activity for a component:

$$a_i^N = \begin{cases} \sum_{\tau=0}^t \Delta_i^N(\tau) & \text{if comp. } i \text{ exists in real run at time } t \\ 0 & \text{otherwise} \end{cases}$$

The normalized total cumulative evolutionary activity is the sum of the activity counts for every component in the real population:

$$A_{cum}^N(t) = \sum_{i \in R} a_i^N(t),$$

where $i \in R$ indicates that component i exists in the real population at time t . Then the normalized mean cumulative evolutionary activity is simply

$$\bar{A}_{cum}^N(t) = \frac{A_{cum}^N(t)}{D^R(t)},$$

where $D^R(t)$ is the diversity of the real population. The median cumulative evolutionary activity, $\bar{A}_{cum}^N(t)$, is simply the median of all $a_i^N(t)$ such that component i exists in the real population at time t .

Channon’s improvements to the statistics invalidate the prior method for determining when a component can be counted as ‘new’, which involved checking whether a component’s activity fell in a small ‘window’ that could be considered minimally adaptive (defined to be just above the highest activity of the shadow). As component-normalized activities are no longer monotonic, a new method is needed. The solution is simple but requires slightly more bookkeeping: a component is considered ‘new’ (that is, newly adaptively significant) the *first* time its activity exceeds a threshold a_0^N , and never afterward. This leaves the matter of determining a_0^N . Channon argues,

If the presence or absence of a component confers no adaptive advantage or disadvantage, then the real and shadow systems are equivalent for this component. [...] So the (normalised) activity distribution for this class of components will be symmetric about the origin. Therefore, provided we can make the assumption that the most negative normalised activity encountered during a run is from such a component, we can negate this value to find a level at which normalised activity can be considered adaptively significant. [6]

Then normalized new activity per component is:

$$A_{new}^N(t) = \frac{1}{D^R(t)} \sum_{i: \text{component } i \text{ is 'new'}} a_i^N(t).$$

This completes the definitions of the statistics needed for the classification given above.

3. EXPERIMENTS AND RESULTS

3.1 Experimental System Overview

Recall that the motivation for these experiments was to test the ‘null hypothesis’, that is, to see if the evolutionary activity statistics would classify as Class 3 something which is intuitively not lifelike. Thus it is important that the experimental system be intuitively not lifelike, that is, it must be simple. In our case, adding complication or subtlety to the evolutionary system being measured risks defeating the purpose of the experiment.

With this in mind, we designed a fairly standard genetic algorithm evolutionary system. Individuals consisted of genotypes ranging in length from 1 to 500 genes, each of which could take on one of four values. Genotypes were initialized randomly with length uniformly distributed between 1 and an upper limit (30, 50, or 150, depending on the experiment). Reproduction was asexual, with mutation rates dictating the probability of the length of the genome growing or shrinking by a gene and the probability of any given gene changing. Selection (for the privilege of reproducing) was done through tournament selection with a tournament size of two for most experiments; some early experiments used a ‘roulette wheel’ fitness-proportional selection. Enough selections for reproduction were made from the population at a given generation to fill the population for the next generation, the size of which was held constant over each experiment. The fitness function used varied, and will be described for each experiment below. An overview of the experimental parameters can be found in Table 1.

Collecting the evolutionary activity statistics required implementing a shadow population. This consisted of a GA identical in most respects to the selective system. The difference was that while the selective system selected individuals for reproduction based on fitness, the shadow population selected individuals to reproduce strictly at random. As is common practice, the statistics were calculated and collected based on ‘snapshots’ of the two populations taken at regular intervals. (This sampling is necessary due to computational constraints, specifically memory.) The number of generations between snapshots varied, and will be noted for each experiment below. At each snapshot, the activity counters for all components in the system were incremented based on existence as described above. Immediately after each snapshot the shadow population was ‘synchronized’ to be identical to the real population, again as described above. For all experiments, a_0^N , the activity threshold above which a component may be considered adaptive, was found as suggested by Channon ([6], quoted above), either by running an identical pre-experiment with the same random number generator seed, or (for the earlier experiments) by averaging the a_0^N s computed for several pre-experiments identical in all parameters except for random number generator seed.

3.2 Experiments

3.2.1 Genetic drift

As mentioned above, one of the motivations of this work was the observation that evolving systems can often display surprisingly sophisticated dynamics even without any selective pressure. Thus we first experimented measuring the evolutionary activity of a population with no selective pressure (random selection), merely undergoing genetic drift. In

Fitness function	None (Genetic Drift)	Simple Static	Reset	Moving	Factored
Selection mechanism	random	fit.-prop., tourney (2)	tourney (2)	tourney (2)	tourney (2)
Max. init. genome size	30	50	50	50	50
Population size	1000	500, 100	100	100	100
Genome size mut. rate	0.001	0.001	0.001, 0.0001, 0.00001	0.001, 0.0001, 0.00001	0.001, 0.0001, 0.00001
Gene mut. rate	0.002	0.001	0.001, 0.0001, 0.00001	0.001, 0.0001, 0.00001	0.001, 0.0001, 0.00001
Generations per snapshot	1000	1000, 500	500	20	20 40
# of snapshots	100	1000, 500	500, 1000, 1000,	10000	10000
a_0^N	7.7	0	0	1 2 1	105 72 15 250 142

Table 1: Parameters for the various experiments described in section 3.

this case the maximum initial size of the genotype was 30. The population size was 1000, the genome size mutation rate was 0.001 and the gene mutation rate was 0.002. There were 100 snapshots of 1000 generations each.

a_0^N was measured as 7.7 by averaging four runs (meaning that a component would need to persist for 8 snapshots, or 8000 generations, to be considered adaptive). The system in this experiment exhibited no *adaptive* evolutionary activity. That is, the activity for the ‘real’ population (which was, in this case, functionally identical to the shadow population, as both had random selection) never exceeded the a_0^N threshold. Thus the new evolutionary activity statistic was zero, resulting in a Class 1 classification: no adaptive evolutionary activity. This is, in fact, the *raison d’être* of the shadow—to filter out the activity present in a drifting population, and thus no matter how complex the dynamics of the drifting population, it will be canceled out by the shadow (provided the shadow’s dynamics are complex in a statistically similar way).

3.2.2 Simple static fitness function

The genetic drift experiment made it clear that we needed some selective pressure to achieve adaptive evolution, according to the evolutionary activity metrics. This is unsurprising, but a worthwhile sanity-check for this classification scheme. We started with a simple static fitness function: fitness was measured as proximity of the sum of the genes (which had values 0 through 3) to an arbitrary value chosen uniformly in the range [0, 1500) at initialization of the system. We ran two experiments, one with roulette-wheel selection, a maximum initial genome size of 50, gene and genome length mutation rates of 0.001, and a population size of 500, which was run for 1000 snapshots of 1000 generations each, and another with tournament selection, a population size of 100, and other parameters the same, which was run for 500 snapshots of 500 generations each. In the

former a_0^N was found by averaging four runs, while in the latter a_0^N was found directly for the population in question (i.e. using one run with the same random number generator seed).

In both cases a_0^N was measured at 0—any component which persisted for more than a single snapshot was considered adaptive. The results of both experiments were the same: new activity was positive, diversity was bounded, and mean and median cumulative activity were both bounded, indicating Class 2 dynamics—bounded evolutionary activity.

The classification actually masks some subtlety which only became clear in later experiments. The upper bound for the mean and median cumulative evolutionary activity of the real populations in the static fitness function experiments were approximately and exactly (respectively) one—that is, the typical *adaptive* component only persisted for one snapshot. This was due to the mutation rates, which were perhaps higher than they really should have been. The new activity in the system was due almost entirely to neutral mutations, which were soon replaced by other neutral mutations. Had the mutation rates been significantly lower, persistence would have been greater, leading to a higher bound for mean and median cumulative activity, but the new activity would have been much lower—probably zero most of the time, as would be expected of a population which had converged upon a solution.

3.2.3 Reset fitness function

In an effort to avoid the behavior of such convergence, as well as to address the questionable (and arbitrary) choice of mutation rates, we next ran a set of experiments with a changing fitness function, at three different rates of mutation. The fitness function was as in the previous experiment; however, the target function was changed with a 0.1 probability after each snapshot. All three experiments used a

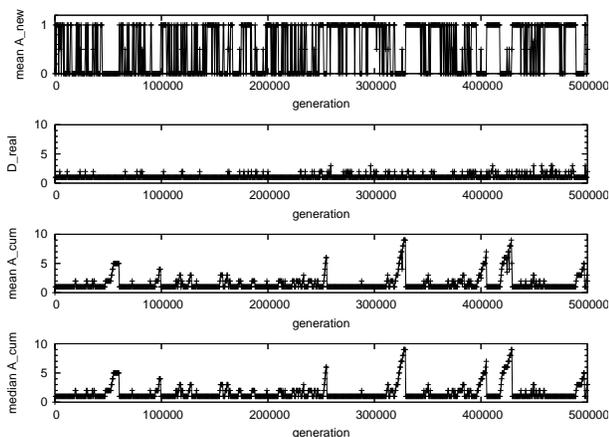


Figure 1: Mean new evolutionary activity, diversity, mean and median cumulative evolutionary activity for the resetting fitness function with a mutation rate of 0.00001. Note the peaks in the cumulative evolutionary activity, corresponding to periods of persistence (presumably due to stasis in the fitness function).

maximum initial genotype length of 50, a population size of 100, and tournament selection with a tournament size of 2. All three were run with 500 generations per snapshot; the experiment with mutation rates of 0.001 was run for 500 snapshots, and the experiments with mutation rates of 0.0001 and of 0.00001 were run for 1000 snapshots.

Figure 1 displays graphically the results for the mutation rate 0.00001 case. There are several features to note. The most obvious is the waves or spikes in the mean (and median, which is nearly identical) cumulative evolutionary activity, which indicate periods of persistence, during which a component (that is, genotype) persists over many generations. These periods presumably correspond to periods in which the fitness function is not changing. This conclusion is further supported by the observation that the peaks in mean cumulative evolutionary activity often correspond to periods of no new evolutionary activity, which, as suggested above, we would expect of a population which had converged. Then, when the fitness function is reset, the level of new activity is again positive, but persistence (measured by mean and median cumulative evolutionary activity) again drops to a low level. The result is that the system appears to trade persistence and innovation, but largely fails to achieve both simultaneously. Thus mean and median activity remain bounded, with bounded (and, indeed, minimal) diversity, and new activity positive on the whole, resulting again in a Class 2 classification of dynamics—bounded evolutionary activity.

Similar results were observed in the experiment with a mutation rate of 0.0001, although the features described were less pronounced, masked to some extent by neutral mutations as suggested above. This trend continued for the 0.001 mutation rate case, where the features seen in the other cases were obscured to a greater extent by neutral mutations.

3.2.4 Moving fitness function with short snapshots

We also ran several experiments using a much shorter snapshot of 20 or 40 generations per snapshot and a ‘moving’ fitness function which changed (by a random amount within a small range) at fixed intervals. The motivation for all of these experiments was to capture the evolutionary dynamics of the system with fine enough granularity that there would be less turnover between snapshots, but that there must still be selective pressure in order for adaptive evolutionary activity to be detected.

The first of these sets of experiments used a simple fitness function like that described in 3.2.2 above, except that every five generations the target of the fitness function was changed by a random amount drawn uniformly from a small range centered around 0, so that the target moved up or down by a small amount every five generations. Snapshots were taken every 20 generations, so that four movements of the fitness function occurred between each snapshot. Experiments were run with mutation rates of 0.001, 0.0001, and 0.00001. There was no adaptive persistence exceeding a_0^N in the experiments with higher mutation rates. In the 0.00001 mutation rate experiment there was some but very little new activity. In all cases the median cumulative activity was bounded at one, leading to a classification of Class 1 and weakly Class 2—no or little adaptive evolutionary activity.

We also experimented with a ‘factored’ fitness function, which had three different components: a broad selective pressure, such as the sum of the alleles mod four; a narrow selective pressure, such as the distance of the sum from some target (applicable only within a small radius of the target); and a flat ‘bonus’ for achieving a sum within some larger radius of the target. The target moved as in the simple moving fitness function described above. We tried applying this fitness function to systems with 5 generations between fitness function movements and 20 generations between snapshots as above, and with 40 generations between both fitness function movements and snapshots, in both cases at several different rates of mutation. In all cases a_0^N was measured at high levels compared to the earlier experiments described above, and components rarely if ever achieved the persistence necessary to exceed a_0^N and be considered newly adaptively significant, resulting in Class 1 dynamics—no adaptive evolutionary activity.

A summary of the experimental results can be found in Table 2.

4. CONCLUSIONS

It is clear that the Evolutionary Activity scheme for classifying the evolutionary dynamics of a system has so far withstood our attempts to break it. Two features appear to be of particular importance to the scheme’s robustness. The first is the use of the neutral shadow to effectively subtract out all non-adaptive activity in the system. As discussed in section 3.2.1, no matter how complex the dynamics of a drifting population, this complexity will be canceled out by the shadow. The second important feature is the method of component normalization introduced by Channon ([6]), which assures that the neutral shadow remains a faithful shadow of the real population, canceling out the potential for spurious results arising from the (random) divergence of the real and shadow populations.

Experiment	D	A_{new}	median A_{cum}	Evolutionary Dynamics
Genetic Drift	bounded	zero	zero	Class 1: None
Simple Static Fitness	bounded	positive	bounded	Class 2: Bounded
Reset Fitness	bounded	positive	bounded	Class 2: Bounded
Moving Fit./Short Snapshots	bounded	zero	zero	Class 1: None

Table 2: A summary of the results of our experiments validating the evolutionary activity statistics.

The dominant trend observable from the experiments described in this paper is that with simple fitness functions there appears to be a trade-off between new activity (innovation) and persistence. When the fitness function is stable components persist, but there is little new evolutionary activity, as the population is essentially already converged on a solution. Change in the fitness function promotes new innovation, but there is little retention of components until the population again converges on a stable solution. In order to achieve sustained innovation and persistence simultaneously the fitness function must be such that new innovation is always necessary (and possible), which implies a function that is always changing, and yet also such that adaptive innovations remain adaptive. This is a tall order, and probably not achievable by any simple, “uninteresting” fitness function.

5. DISCUSSION AND FUTURE WORK

Immediate future work might focus on further validation or challenges of the evolutionary activity metric. For example, there are problems with considering a full genotype as a component. Genotype granularity leaves no room for neutral variation within a component, as exists in, say, a phenotype or taxonomic family. When studying artificial systems which allow easy examination of the genotype, genes or alleles are a less problematic choice for the notion of component, and it is possible that at that level of granularity it might be easier to “break” the evolutionary activity metric—that is, to exhibit an intuitively unlikelike artificial evolutionary system that nonetheless achieves a Class 3 classification.

The results presented in this paper suggest, however, that the evolutionary activity metrics are indeed robust. The larger problem is their limited applicability: they apply only to systems where a clear notion of uniform component exists. Indeed, Bedau and Packard [2] acknowledge that the scheme will measure false positives if usage counters are attached to non-functional micro-level units (such as introns), and that false negatives occur if the “micro-level units fail to reflect some aspect of genuine evolutionary activity”, such as “a genetic system in which combinations of genes can have adaptive significance over and above the adaptive significance of their individual gene components”—which is true, we would argue, of almost every nontrivial genetic system. We believe this is a more severe shortcoming than the existing literature would suggest, and one which has not been adequately addressed. Adapting the existing metrics or developing new ones suited to combinatorial systems is a crucial challenge, and one we hope to undertake in the future. We are particularly interested in metrics that can be applied to genetic programming systems, where adaptive significance comes almost entirely from the combination of micro-level units (functions and terminals).

6. ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under Grant No. 0308540 and Grant No. 0216344, and by the Defense Advanced Research Projects Agency (DARPA) and Air Force Research Laboratory, Air Force Materiel Command, USAF, under agreement number F30502-00-2-0611. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation. A. Stout also gratefully acknowledges a Swarthmore College Lockwood Fellowship.

7. REFERENCES

- [1] J. C. Avise. *Phylogeography: The History and Formation of Species*. Harvard University Press, 2000.
- [2] M. A. Bedau and N. H. Packard. Measurement of evolutionary activity, teleology, and life. In C. Langton, C. Taylor, D. Farmer, and S. Rasmussen, editors, *Artificial Life II, Santa Fe Institute Studies in the Sciences of Complexity, Volume X*, pages 431–461, 1992.
- [3] M. A. Bedau, E. Snyder, C. T. Brown, and N. H. Packard. A comparison of evolutionary activity in artificial evolving systems and in the biosphere. In P. Husbands and I. Harvey, editors, *Proceedings of the Fourth European Conference on Artificial Life*, pages 125–134. MIT Press, 1997.
- [4] M. A. Bedau, E. Snyder, and N. H. Packard. A classification of long-term evolutionary dynamics. In C. Adami, R. Belew, H. Kitano, and C. Taylor, editors, *Artificial Life VI*, pages 228–237. MIT Press, 1998.
- [5] A. Channon. Passing the ALife test: Activity statistics classify evolution in Geb as unbounded. In *Proceedings of the European Conference on Artificial Life*, 2001.
- [6] A. Channon. Improving and Still Passing the ALife Test: Component-Normalised Activity Statistics Classify Evolution in Geb as Unbounded. In Standish, Abbass, and Bedau, editors, *Artificial Life VIII*, pages 173–181. MIT Press, 2002.
- [7] S. J. Gould. *The Structure of Evolutionary Theory*. Belknap Press, 2002.