Passing the ALife test: Activity statistics classify evolution in Geb as unbounded

Alastair Channon

Department of Electrical and Electronic Engineering
Anglesea Road, University of Portsmouth, PO1 3DJ. UK
alastair@channon.net http://www.channon.net/alastair

Abstract. Bedau and Packard's evolutionary activity statistics [1, 2] are used to classify the evolutionary dynamics in Geb [3, 4], a system designed to verify and extend theories behind the generation of evolutionary emergent systems. The result is that, according to these statistics, Geb exhibits unbounded evolutionary activity, making it the first autonomous artificial system to pass this test. However, having passed it, the most prudent course of action is to look for weaknesses in the test. Two weaknesses are identified and approaches for overcoming them are proposed.

1 Introduction

Perhaps the most important goal in Artificial Life is the generation of an autonomous system (that is excluding systems such as the Internet, which owes its evolution to human endeavour) that exhibits unbounded evolution. "Life-as-it-could-be" [5], as opposed to models of life-as-we-know-it, can only be achieved within such a system, for life is a product of evolution. Thanks to Bedau and Packard's evolutionary activity statistics [1, 2], we have a test that a candidate system should pass before claims of unbounded evolution are taken seriously. The test is based on the following statistics: component activity increment Δ_i, component activity a_i, diversity D, total cumulative activity A_{cum}, mean cumulative activity $\bar{A}_{\mathrm{cum}}$ and new evolutionary activity per component A_{new}.

$$\Delta_i(t) = \begin{cases} 1 & \text{if component } i \text{ exists at } t \\ 0 & \text{otherwise} \end{cases} \qquad D(t) = \#\{i : a_i(t) > 0\}$$

$$a_i(t) = \begin{cases} \sum_{\tau=0}^{t} \Delta_i(\tau) & \text{if component } i \text{ exists at } t \\ 0 & \text{otherwise} \end{cases} \qquad A_{\mathrm{cum}}(t) = \sum_i a_i(t)$$

$$A_{\mathrm{new}}(t) = \frac{1}{D(t)} \sum_{i:a_i(t)\in[a_0,a_1]} a_i(t) \qquad \bar{A}_{\mathrm{cum}}(t) = \frac{A_{\mathrm{cum}}(t)}{D(t)}$$

For A_{new} to be a good measure of new activity, the range $[a_0, a_1]$ should be chosen such that component activities within it can be considered both adaptively significant (so a_0 should be high enough to screen out most non-adaptive activity) and not amongst the highest activities (so a_1 should be low enough that a good proportion of activities lie above it).

Table 1. Classes of evolutionary dynamics and their statistical signatures, based on table 1 from [2]. Rows 3b and 3c have been added to class 3 (see text).

| | | STATISTICAL SIGNATURE | | |
| | | D | A_new | $\bar{A}_\mathrm{cum}$ |
CLASS	EVOLUTIONARY DYNAMICS			
1	none	bounded	zero	zero
2	bounded	bounded	positive	bounded
3a	unbounded (D)	unbounded	positive	bounded
3b	unbounded ($\bar{A}_\mathrm{cum}$)	bounded	positive	unbounded
3c	unbounded (D & $\bar{A}_\mathrm{cum}$)	unbounded	positive	unbounded

For artificial systems, a "shadow" should be run, mirroring the real run in every detail except that whenever selection operates in the real system, random selection should be employed in the shadow. Shadow statistics can then be used to determine a_0 and levels of activity that can be considered significant.

After determining long-term trends in these statistics, the system being examined can be classified according to table 1. The hallmark of class 3 (unbounded evolutionary dynamics) is unbounded A_cum in combination with positive A_new.

Other possibilities exist with zero A_new, but these belong in class 1 (no evolutionary activity) because such cases have no significant new components. Table 1 in [2] only shows the first row (3a) for class 3, but footnote 1 in [2] mentions the other rows (3b and 3c). Note that table 1 includes all possibilities for positive A_new, because zero $\bar{A}_\mathrm{cum}$ implies zero A_new. So any system with unbounded evolutionary dynamics will belong to class 3 (one of 3a, 3b and 3c).

Previously, only the biosphere has passed this test,[1] although a number of Artificial Life systems have been evaluated. The following quote from their discussion section summarises Bedau, Snyder and Packard's conclusion.

> "We also suspect that no existing artificial evolving system has class 3 dynamics. In our opinion, creating such a system is among the very highest priorities of the field of artificial life. From one perspective, this is a negative result: Echo, and perhaps all other existing artificial evolutionary systems, apparently lack some important characteristics of the biosphere – whatever is responsible for its unbounded growth of adaptive activity. But at the same time this conclusion calls attention to the important constructive and creative challenge of devising an artificial model that succeeds where all others have failed." [2, p. 236]

Nehaniv [7] defined open-ended evolution as unbounded growth in a measure "cpx". However in [8] he proceeded to show that a trivial system exhibits open-ended evolution according to this definition.

---

[1] Maley [6] makes the claim that two of his models exhibit unbounded evolutionary activity. However, Urmodel 3 shows less new activity than its shadow (with no reason to think that it would become greater), Urmodel 4 shows a lower mean activity than its shadow and both are only examined during their initial growth stages.

2 Implementing the Statistics in Geb

Full details of Geb are available in [3, 4]. Here I include just enough for an explanation of the evolutionary components used in the statistics. Geb is a virtual world containing autonomous organisms, each controlled by a neural network. Each neuron has a bit-string label, or 'character', which is used during development and for matching the neural outputs of one organism with basic behaviours (turning, killing, etc.) and with inputs of other organisms. An organism is born with a simple axiom network that results in reproduction. This develops through the application of a genetically determined Lindenmayer system (L-system) [9]. Each L-system production rule has the following form:

$$\mathcal{P} \to \mathcal{S}_r, \mathcal{S}_n \; ; \; b_1, b_2, b_3, b_4, b_5, b_6$$

$\mathcal{P}$ Predecessor (initial bits of node's character)
$\mathcal{S}_r$ Successor 1: *replacement* node's character
$\mathcal{S}_n$ Successor 2: *new* node's character
bits: $b_1, b_2, b_3, b_4, b_5, b_6$ specify linkage details

The *successors* (1 and 2) are characters for the node(s) that replace the old node. If a successor has no character (0 length) then that node is not created. Thus, the predecessor node may be replaced by 0, 1 or 2 nodes. Necessary limits on the number of nodes and links are imposed.

An evolved genotype contains a large number of production rules (once decoded), but only the rules found to match neuron's characters most closely are used during development. In this way, increasingly specific production rules can evolve, with regressive rules existing as fall-back options should a rule be damaged by crossover or mutation, and as material for further evolutionary search.

When a new organism is 'born', all possible production rules are decoded from its genotype. Then the developmental process is part-simulated in advance of it truly taking place, as a means of filtering out all the production rules that would never be used, either because they would never match any possible node's character, or because more specific rules exist for each node that could develop. All rules that remain will be used if the organism lives long enough.

2.1 Choosing the Component Class

As Geb's genotypes both change length and contain a high degree of neutrality[2] the genotype is not a good choice of component class. Production rules, the alleles from a genotype, are a much better choice. It can be expected that if a production rule has an adaptive advantage, then it will persist. Better still though is the choice "production rules that survive the filtering process at birth", for these are actually used in the developmental process; the idea behind activity statistics is to measure the degree to which components both *persist* and are *used*.

---

[2] In order to avoid confusion, I only use the term *neutral* to refer to genetic variations that are phenotypically equivalent, and not in relation to shadow runs.

When mutation causes a component to not be expressed (currently present), the activity count of the original component is no longer included in the total activity of the system, even if the mutation is functionally neutral. At first I implemented the activity statistics on production rules directly. But there is often a high degree of neutrality in a production rule, especially when its 'successors' relate to neurons that are over-specified (have excess bits at the end of their characters) or development-terminal (not matched by any production rule). The predecessor and link-bits sections of production rules are more plastic. If a predecessor bit is mutated, then the rule will most likely either fail to match or be less specific to its target neuron than another rule. If a link detail bit is mutated, then the result will more often than not be a damaged network, and organisms with that production rule active (not filtered out) will be driven to extinction. So the choice of component used here is 'predecessor plus link details' $(\mathcal{P}, b_1, b_2, b_3, b_4, b_5, b_6)$. This can be thought of as a disjoint grouping of alleles, with each group being a component. Which individual a component is from is irrelevant: two identical production rules in two different organisms result in two instances of the same component.

This grouping does not completely remove the neutrality problem. As successor lengths increase, neural character lengths increase, and so the number of predecessors that can potentially match a typical neural character increases. If two rules have the same successors and link-details (or neutral variants), then it makes no difference to development which one is used. So, as component lengths increase, we can expect the level of neutrality to increase.

Having chosen this component class, there is a clear consequence for the possible classifications of evolutionary dynamics. Because the number of neurons that an organism can have is limited (for practical reasons), the number of production rules that can survive filtering is limited. And because the population size is small (a maximum of four hundred organisms), there is little room for more than a couple of species at a time. So diversity of these components will certainly be bounded, and we can rule out class 3a and 3c dynamics.

2.2 Implementation Details

Geb's shadow mirrors the real run in every detail except that selection is random. Whenever a real organism is killed, a randomly chosen shadow organism is also killed. Whenever a real organism is born (as the product of two real organisms), a new shadow organism is born as the product of two randomly chosen shadow organisms, using the same reproduction procedure with the same rates of crossover and mutation.

It is not feasible to gather the statistics at every timestep. So snapshot existence records are taken at regular intervals and the evolutionary statistics are calculated from these. In the results reported here, snapshots were taken every one thousand timesteps. To put this in context, the run reported lasted six million timesteps, during which time there were over five hundred and eighty million organism reproductions. Because activity is intended as a measure of how much a component both is used (already covered above) and *persists*, I

screen out (in each of the real and shadow populations) isolated occurrences: when a component occurs in the current snapshot but not the previous one.

In previously published work on Geb, total extinction (population size dropping to one individual) was not mentioned because it had not been encountered. However, during the long trial runs undertaken when experimenting with evolutionary statistics, I encountered occasional runs in which total extinction occurred. So for the set of runs from which the example reported here is taken, I set a minimum number of organisms to twenty. The fact that total extinction is so rare despite the population size being so small (a maximum of four hundred organisms) indicates that there is no serious problem here. Once population sizes can feasibly be increased, the problem should in practice disappear rapidly.

3 Results and Discussion

This section contains the results from a typical run, drawn from the full set of twenty runs. Atypical variations are discussed at the end of this section.

3.1 Activity Waves

In order to gain an understanding of the dynamics behind the higher level evolutionary statistics, it is a good idea to look first at the activity wave diagrams, which simply show all components' activities plotted against time. Figure 1 shows the activity wave diagrams for the real and shadow populations.
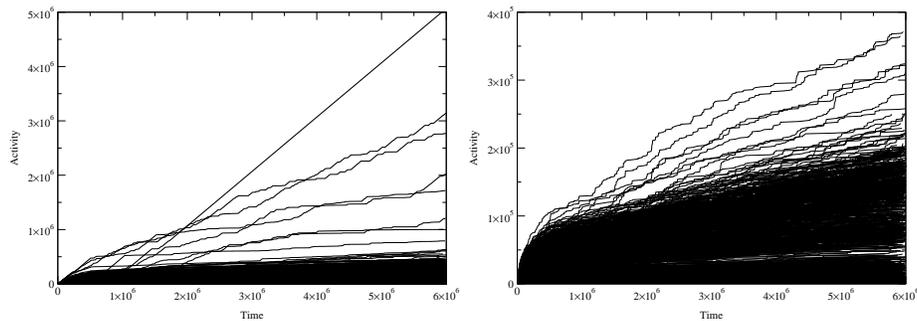

Fig. 1. Activity wave diagrams for the real (left) and shadow (right) runs. Note the different scales for real and shadow.

The most obvious feature of the real run's activity waves in figure 1 is that many of them keep increasing. This would also be true in a similar analysis of genes from the biosphere's evolution. Genes that are beneficial to life tend to remain in the population of genes and be used by many species: humans have a significant proportion of genes in common with mice, flies and even plants. So

because the components here are (groups of) genes, not whole genotypes, this feature does not imply a quasi-stable ecosystem.

In systems without neutrality new components initiate an activity (by presence) wave that increases with a constant slope and then stops when the component goes extinct. Here however, that increase is often shared between two or (perhaps many) more phenotypically equivalent components, with interchanging presence. If the population size was larger, then there would be greater scope for more than one of these neutral variations to be expressed in the population at any time. But with a population of less than four hundred, and short lifetimes, genetic variation spreads quickly through the population so wave transitions between neutral variants show up almost as either-or events.

We can also observe that as time goes by, the level of neutrality increases: the average rate of increase for a new component decreases and the number of components in each neutrality-group increases. This is consistent with the expectation from section 2.1 that as component lengths increase, we would see the level of neutrality increase.
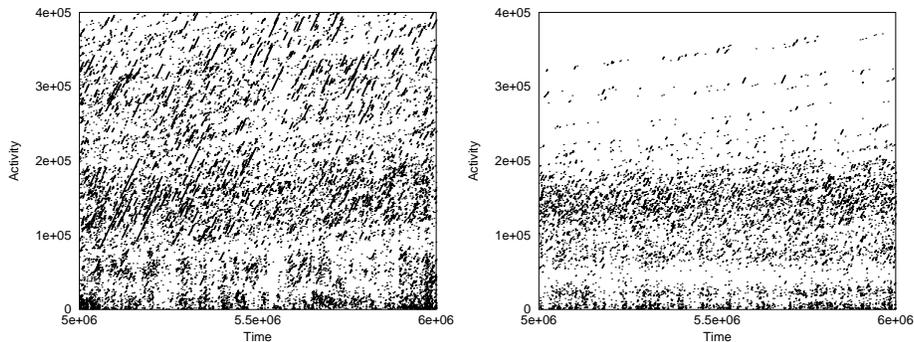
Fig. 2. Activity point-plots for the real (left) and shadow (right) runs in the last million timesteps, within the shadow's activity range.

Because of this increase in neutral-group size and decrease in average component activity rate of growth, most activity falls within the solid black regions at the bottom of each graph. So it is instructive to look in more detail at the bottom-right corner of the activity wave diagrams. Figure 2 shows the activity waves in the last one million timesteps, with just a point for each recorded activity value. Its scale covers the shadow run's full range only, so that the real and shadow data can be easily compared. Notice the long runs of consecutively increasing activity in the real run, and the lack of them in the shadow.

### 3.2 Determining the New-Activity Range

In order to measure new activity (A_{new}), we must first determine the range $[a_0, a_1]$ of component activity values that should be considered both adaptively

significant and not amongst the highest activities. The method given in [2] involves finding the activity value which is equally likely to have occurred in the real run as in the shadow run, and setting $[a_0, a_1]$ to be a narrow band that surrounds it. A log-log plot of the real and shadow component activity distributions for the run reported here shows that they cross at an activity of approximately 1.42×10^5. However, it is clear from figure 1 that the component activity distributions are far from constant over the run, and that this value increases during the course of the run. Looking again at figure 2 shows that by the end of the run, activities of approximately 1.42×10^5 are common in the shadow.

Here I have used a fixed range that screens out the majority of the shadow activity in the final million timesteps of the run. This results in artificially low values for A_{new} early on in the run, but the results are still positive despite this. In the last million timesteps, less than 3.5% of the shadow activity is above 2.8×10^5, and approximately 27% of the real activities are above 3.2×10^5. So the results that follow were calculated using a new-activity range of $[2.8 \times 10^5, 3.2 \times 10^5]$.

3.3 Evolutionary Statistics and Classification

Figure 3 shows both total and mean activity increasing rapidly in the real run, and much slower in the shadow run. New activity is positive in the real run, and much higher than in the shadow, which exhibits only occasional blips of new activity. Diversity is bounded in both the real and shadow, as expected (see section 2.1).
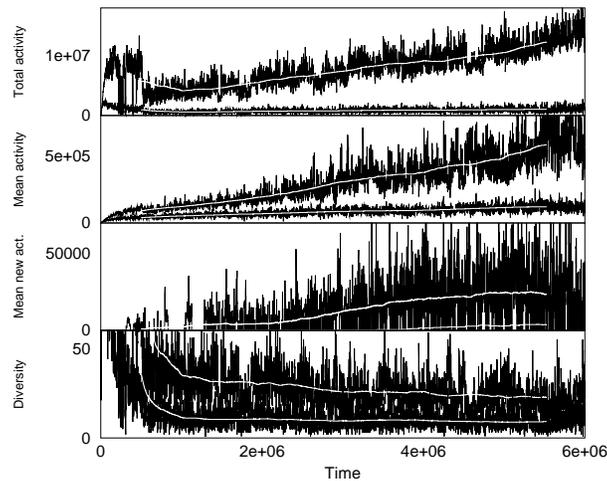
Fig. 3. Total activity, mean activity, new activity and diversity from a typical Geb run and its shadow. Running averages are shown in white.

Figure 4 shows the activity difference and "excess activity", for both total and mean activity, between the real and shadow statistics. Excess activity is defined in [10] as activity difference divided by shadow activity.
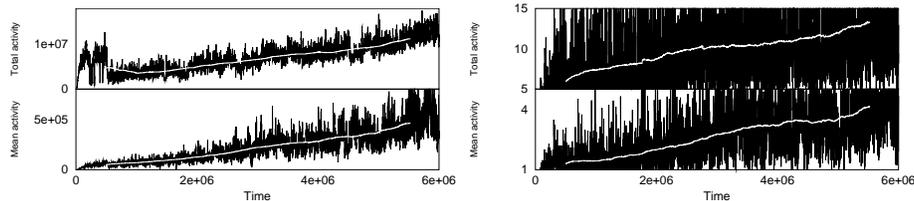
Fig. 4. Normalised total and mean activity: difference (left) and "excess" (right). Running averages are shown in white.

According to this classification system, these results clearly fall into class 3b: unbounded evolutionary activity.

3.4 Atypical Runs

These results are typical of the twenty runs that were carried out for this set of experiments. However, three of the runs encountered problems, causing their results to be atypical. Two of these effectively met total extinction. In section 2.2 I mentioned that I imposed a minimum limit on the number of organisms, in an attempt to avoid total extinction. However, if population size hits this limit and does not increase rapidly, then many reproductions may occur with selection effectively random. This causes evolutionary activity to plummet as adaptive traits are lost. Once lost, this activity cannot be regained, except by the evolution of new adaptive components. These results should not be a cause for concern, for the same reasons mentioned in section 2.2: once population sizes can feasibly be increased, the problem should in practice disappear rapidly.

In the third atypical run, it appears that a freak mutation has caused the only existing species to take on a behaviour of never reproducing or moving forward and always turning and trying to kill. Of course this would ordinarily be a very poor strategy. It is easy to imagine how the bad gene (production rule) could have spread through a population of just one species as fit individuals reproduced with the new unfit ones, causing their children to pick up the dominant bad gene. However, one would not expect this to pose a threat to a different species. This is easily verified: introducing just a few organisms from any of the other evolved populations (from the other runs), causes the old organisms to be rapidly displaced by the newcomers. So this result is also not a cause for concern, for the same reason: it is due to the small population size, which cannot support more than one or two species at a time.

4 Criticisms and Conclusions

Geb has demonstrated class 3 behaviour, and so passed the test. Does this mean that Geb truly exhibits unbounded evolution? Possibly, for it was designed to verify and extend theories of evolutionary emergent systems generation and so a number of potential pitfalls have been avoided. However, having passed the test the most prudent course of action is to look for weaknesses in the test.

The main concern that I have at this time is that the test relies on normalisation (or validation) from a shadow that can drift away from core aspects of the real run that it is intended to shadow. For example, the components that exist in the real population at any one time (well into evolution) are almost certainly more densely clustered than those in the shadow. So the mutation of a real component is more likely to produce another high-activity component than the mutation of a shadow component. Once the real and shadow populations have been allowed to evolve, we are no longer comparing the real run with a true shadow. One way around this problem would be to develop a method of comparing the real run with a shadow that is regularly reset (both components and activity history) to be identical to the real run but which evolves using random selection between resets. The normalised activity increment between resets would be determined by comparing the real and shadow increments.

My other criticism of the test as it stands is in its use of mean activity when looking for unbounded activity growth, especially when classifying a system as belonging to class 3b. When diversity is bounded, the retention (forever) of a single component results in unbounded mean activity. The test should not be so influenced by such components, and should rather look for trends in typical components. So it is median activity, not mean activity, that should be measured, and required to be unbounded for a system to be classified as within class 3b. The activity waves from Geb's runs indicate that median activity is also unbounded. However, when median activity is measured in both real runs and their shadows, it shows up as unbounded in both, and normalised median activity appears to be bounded (figure 5). However, in light of my main concern above, just as the positive results cannot be trusted, this negative result cannot be trusted either. The correct course of action is to proceed as outlined above, by developing a shadowing method that regularly resets the shadow state to the real state, and then look at the results again, including median activity.
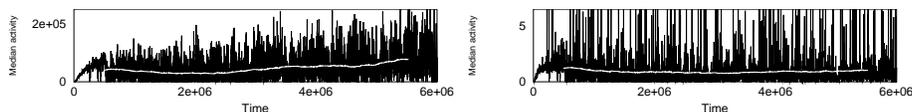
Fig. 5. Normalised median activity: difference (left) and "excess" (right). Running averages are shown in white.

The main conclusion of this paper is that these results are encouraging but, until the above concerns and any others that the Artificial Life community may raise are resolved, the test results cannot be considered conclusive. Whether or not Geb exhibits anything that deserves to be thought of as unbounded evolutionary activity remains to be determined. We do however now have an Artificial Life system with which to test the Artificial Life test.

References

[1] BEDAU, M. A., AND PACKARD, N. H. Measurement of evolutionary activity, teleology, and life. In *Proceedings of Artificial Life II* (Redwood City, CA, 1991), C. G. Langton, C. Taylor, J. D. Farmer, and S. Rasmussen, Eds., Addison-Wesley, pp. 431–461.

[2] BEDAU, M. A., SNYDER, E., AND PACKARD, N. H. A classification of long-term evolutionary dynamics. In *Proceedings of Artificial Life VI, Los Angeles* (Cambridge, MA, 1998), C. Adami, R. Belew, H. Kitano, and C. Taylor, Eds., MIT Press, pp. 228–237.

[3] CHANNON, A. D., AND DAMPER, R. I. Perpetuating evolutionary emergence. In *From Animals to Animats 5: Proceedings of the Fifth International Conference on Simulation of Adaptive Behavior (SAB98), Zurich* (Cambridge, MA, 1998), R. Pfeifer, B. Blumberg, J.-A. Meyer, and S. Wilson, Eds., MIT Press, pp. 534–539.

[4] CHANNON, A. D., AND DAMPER, R. I. Towards the evolutionary emergence of increasingly complex advantageous behaviours. *International Journal of Systems Science 31*, 7 (2000), 843–860.

[5] LANGTON, C. G. Artificial life. In *Santa Fe Institute Studies in the Sciences of Complexity, Vol VI: Proceedings of the Interdisciplinary Workshop on the Synthesis and Simulation of Living Systems (Artificial Life I)* (Redwood City, CA, 1989), C. G. Langton, Ed., Addison-Wesley, pp. 1–47.

[6] MALEY, C. C. Four steps toward open-ended evolution. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO99)* (San Francisco, CA, 1999), W. Banzhaf, J. Daida, A. E. Eiben, M. H. Garzon, V. Honavar, M. Jakiela, and R. E. Smith, Eds., Morgan Kaufmann, pp. 1336–1343. Volume 2.

[7] NEHANIV, C. L. Measuring evolvability as the rate of complexity increase. In *Artificial Life VII Workshop Proceedings* (2000), C. C. Maley and E. Boudreau, Eds., pp. 55–57.

[8] NEHANIV, C. L. Evolvability in biologically-inspired robotics: Solutions for achieving open-ended evolution. In *Sensor Fusion and Decentralized Control in Robotic Systems III, SPIE Proceedings 4196* (2000), P. McKee, G. T. Schenker, Ed., pp. 13–26.

[9] LINDENMAYER, A. Mathematical models for cellular interaction in development. *Journal of Theoretical Biology 18* (1968), 280–315. Parts I and II.

[10] RECHTSTEINER, A., AND BEDAU, M. A. A generic neutral model for measuring excess evolutionary activity of genotypes. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO99)* (San Francisco, CA, 1999), W. Banzhaf, J. Daida, A. E. Eiben, M. H. Garzon, V. Honavar, M. Jakiela, and R. E. Smith, Eds., Morgan Kaufmann, pp. 1366–1373. Volume 2.