Replaying the Tape:
An Investigation into the Role of Contingency in Evolution

Tim Taylor and **John Hallam**
Department of Artificial Intelligence, University of Edinburgh
5 Forrest Hill, Edinburgh EH1 2QL, U.K.
timt@dai.ed.ac.uk

Abstract

The role of contingency (random events) in an artificial evolutionary system is investigated by running the system a number of times under exactly the same conditions except for the seed used to initialize the random number generator at the beginning of each run. Twelve different measures were used to track the course of evolution in each run, and "activity wave diagrams" were also produced (Bedau & Brown 1997). The results of 19 runs are presented and analyzed. The performance of every run was compared with each of the others using a non-parametric test (a randomization version of the paired-sample t test). When comparing absolute values of the measures between the runs, some significant differences were found. However, looking at the *difference* in values between adjacent sample points for a run, no run was significantly different to any other for any of the measures. This suggests that the general behaviour is the same in all runs, but the accumulation of differences results in significantly different outcomes. The results lead us to propose a rule of thumb for future experiments with the system: *to check whether the outcome of any particular experiment is robust to contingency in the evolutionary process, at least nine runs should be conducted using different seeds for the random number generator, to be confident of seeing a variety of results.* The results are likely to be applicable to other A-Life platforms of self-replicating computer programs, but at this stage can probably tell us little about the role of contingency in biological evolution.

Introduction

There is much debate in the field of evolutionary biology over the role of contingency ("historical accidents") in determining the course of evolution (see, for example, (Gould 1989), and, for a flavour of the ensuing debate, (Ridley 1993; Gould 1993; McShea 1993)). If evolution were to be re-run on Earth, starting from the same initial conditions and proceeding for another 4 billion years, encountering the same sorts of perturbations from the physical environment that it encountered the first time around, what sort of a world would exist today? Would *homo sapiens* evolve again, or might life not even make the transition from prokaryotic to eukaryotic cells, or maybe not even reach the cellular stage at all? What, in other words, would happen if "the tape were played twice"?

The same question arises when considering artificial evolutionary systems, where we have the advantage of being able to "replay" evolution under experimental control. Indeed, in considering the performance of *any* evolutionary system, we generally wish to disentangle the relative influence of three factors: (1) contingency, (2) performance due to the particular design of the system, and (3) performance which may be general to a wide class of evolutionary systems (Taylor & Hallam 1997). However, considering the importance of these questions, very little has been published to date on the role of contingency in artificial systems. Fontana and Buss have done some excellent work on the subject, choosing to focus on self-maintaining organizations in an artificial chemistry, rather than presupposing the existence of self-replicating entities (Fontana & Buss 1994b; 1994a). Their results suggest that a number of generic organizational features may be expected to emerge in any comparable system.

Fontana and Buss have not, as yet, witnessed the emergence of high-level self-reproducing entities in their work (and that was not their primary goal). There do, however, exist a growing number of A-Life systems which *presuppose* the existence of self-replicators (e.g. Ray's Tierra (Ray 1991), Adami et al.'s Avida (Adami & Brown 1994), Skipper's Computer Zoo (Skipper 1992), and our own platform, Cosmos (Taylor & Hallam 1997; Taylor 1997)). Most publications relating to these systems mention in passing that the results being presented were typical of a large number of runs, but details are rarely given, and, to our knowledge, no systematic study of the role of contingency in such systems has yet been published. One factor that may have contributed to this omission is the difficulty of dealing sensibly with the huge amounts of data that such simulations can produce, which can make it difficult to usefully compare one run with another. However, Bedau et al. have recently been developing a number of techniques for visualizing evolutionary activity, and have also proposed some quantitative measures of evolution (Bedau & Packard 1991; Bedau *et al.* 1997; Bedau & Brown 1997). These analysis tools provide some fairly straightforward ways of comparing the results of a number of evolutionary runs, both qualitatively and quantitatively.

The purpose of this paper is twofold: (1) to report an experiment that runs an artificial life system a number of times, varying just the random number seed between runs, in order to compare how each run evolves and therefore get some idea of the role of contingency in the system; and (2) to use a variety of measures and visualization techniques to compare the runs, and hopefully to ascertain which are the most useful measures for such comparisons. The paper ends with a discussion of the results, including the extent to which they may be generalized to other evolutionary systems.

The A-Life System

Cosmos is a Tierra-like platform that supports a population of self-replicating computer programs living in an environment. Its design differs from Tierra in a number of ways, the most relevant of which, for the present discussion, are described below. For more details about Cosmos, refer to (Taylor 1997; Taylor & Hallam 1997), or look on the worldwide web at `http://www.dai.ed.ac.uk/daidb/people/homes/timt/research.html`. The source code is available from the authors.

Spatial Organization For the runs reported in this paper, the environment was configured as a two-dimensional toroidal grid. There is evidence that such spatial organization, where interactions between programs are restricted to a program's local neighbourhood, can promote heterogeneity and prevent premature convergence (Adami & Brown 1994).

Energy Collection At each time step, energy is distributed throughout the grid. Programs must collect energy from the environment in order to execute their instructions. If a program's internal energy level falls below a certain threshold, it dies. In addition, a maximum population size can be specified for the system. If this is the case, when the population maximum is reached, a fraction of the programs are killed off stochastically, but those with low internal energy have a higher probability of being killed. Programs therefore have to concern themselves with energy collection as well as reproduction, and thus have some degree of control over their own lifespans (i.e. those that collect more energy are less likely to be killed).

Communication Unlike in Tierra, programs in Cosmos can *not* directly read the code of other programs. However, any program can compose an arbitrary message (a string of bits) and transmit it to the local environment, and any program can issue instructions to receive such messages from the environment and interpret them how it wishes. However, in the experiments reported here, such communication did not evolve, so the programs generally had fewer ecological interactions than, for example, Tierran parasites that execute the code of other programs.

Mutations and Flaws As a run proceeds, variation may begin to appear amongst the programs in the environment, caused by the action of two different mechanisms: (1) *Mutations* can affect any program, by the random flipping of one or more bits in the program's code or associated structures. The mutation rate is a system-wide parameter, and does not vary throughout the run; (2) *Flaws*. While a program in running, a flaw may occur in its execution. If this happens, the instruction which was about to be executed will, with equal probability, either be executed *twice* consecutively, or not at all. The rate at which flaws occur is determined by a parameter owned by each individual program. Being a part of the program, it is therefore possible for the flaw rate to evolve over time (by being changed by mutations) in a lineage of individuals.

On a technical note, as this paper is concerned with the role of chance events in evolution, the choice of random number generator (RNG) is particularly relevant, as different types of RNG have different properties. Cosmos uses the `bsd_random()` RNG, which uses the linear feedback shift register generation technique. `bsd_random()` does not suffer from some of the deficiencies of many versions of the standard `random()` RNG.

Measurement Techniques

In any population of self-replicating entities which are competing against each other for resources required for replication (e.g. energy and materials), there are three factors which determine the rate at which any particular type of replicator will spread throughout the population (Dawkins 1989). These are the life-span or *longevity* of the replicator, the rate at which it replicates (its *fecundity*), and the number of errors in makes while producing copies of itself (its *copy-fidelity*). A number of measures were chosen to track changes in each of these three factors through an evolutionary run.

For *longevity*, we looked at the age at death of each program. Specifically, for time slice windows of equal width from the start to the end of the run, we plotted the age at death of each program that died within that time slice window. Example plots are shown in Figure 2. The plots for measures of fecundity and copy-fidelity, described below, also used this windowing technique. For the plots for all three of these factors, the data is pruned by only plotting values for individual programs of types which achieved a concentration of at least two individuals at some time during the run. In the plots, the darkness displayed at any point reflects the number of individual programs taking that particular value at that particular time.

For *fecundity*, we looked at two measures: the number of time slices between the first and second successful replication of each program (the *replication period*) (this could obviously only be applied to programs that successfully replicated at least twice in their lifetime), and the length of programs. Example plots for replication period are shown in Figure 9.

For *copy-fidelity*, we looked at three measures: the flaw rate, the number of faithful (error-free) replications made by individual programs over their lifetime, and the number of unfaithful replications. Example plots of these three measures are shown in Figures 3 and 4.

In addition to these six measures, the population size throughout the run was also recorded, as was the population diversity (the number of *different types* of program in the population).

Four measures suggested by Bedau et al. were used: the Activity (presence), Mean Activity (presence), Activity (concentration), and Mean Activity (concentration), along with their visualization technique of plotting "activity distribution functions" (also referred to as "activity waves"). The basic idea behind all of these techniques is the same, involving the notion of the *evolutionary activity* of each genotype (type of program) in the population:

> "the *evolutionary activity* $a_i(t)$ of the i^{th} genotype at time t [is] its concentration integrated over the time period from its origin up to t, provided it exists:
>
> $$a_i(t) = \begin{cases} \int_0^t c_i(t)dt & \text{if genotype } i \text{ exists at } t \\ 0 & \text{otherwise} \end{cases}$$
>
> where $c_i(t)$ is the concentration of the i^{th} genotype at t. A genotype's evolutionary activity ... reflects its adaptedness (relative to the other genotypes in the population) throughout its history in the system." (Bedau & Brown 1997)

Activity (concentration) is defined at time t as $\sum_i a_i(t)$. *Activity (presence)* is defined similarly, but with $c_i(t)$ defined to simply reflect whether genotype i *exists* at time t, rather than being a measure of concentration (i.e. $c_i(t)$ is 1 if genotype i exists at t, and 0 otherwise). *Mean Activity (concentration)* and *Mean Activity (presence)* are defined as their respective Activity measures divided by the diversity (number of different genotypes) of the population at t.

For a fuller explanation of these measures and the reasons they are defined as they are, refer to (Bedau *et al.* 1997; Bedau & Brown 1997; Bedau & Packard 1991).

To end this section, we acknowledge that paleobiologists have developed their own suite of measures of biological evolution. Daniel McShea has recently published some particularly interesting work on tests for evolutionary trends (McShea 1994), and definitions of complexity (McShea 1996; 1991). Ideally, we would like to be able to use the same set of measures for studying both natural and artificial evolution. Unfortunately, the amount of evolutionary change occurring in Cosmos in the runs reported here is really *very* small compared to the sorts of macroscopic trends that McShea's measures were designed to track, so it is not clear that these measures can usefully be applied to artificial evolutionary systems (or at least to Cosmos) at present.

Method

Nineteen runs of Cosmos were initialized, each with exactly the same ancestor programs, and exactly the same parameter values except for the seed for the random number generator.

Most of the parameters took on the system's default values; those that did not are listed in the Appendix. The most salient of these are `grid_size`, set to 40 (i.e. a 40 x 40 square environment), `max_cells_per_process`, set to 800, and `number_of_timeslices`, set to 300,000.

For each completed run, the measures described in the previous section were investigated. To recap, these measures were as follows:

1. Program age at death
2. Replication period (time between 1st and 2nd faithful replication)
3. Program length
4. Flaw rate
5. Number of faithful replications per program
6. Number of unfaithful replications per program
7. Population size
8. Population diversity
9. Activity (presence)

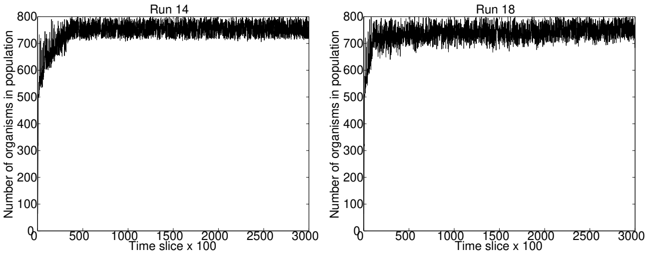

Figure 1: Population size, Runs 14 (*left*) and 18 (*right*)

10. Mean activity (presence)
11. Activity (concentration)
12. Mean activity (concentration)
13. Activity waves

Results

For each measure, the results from each of the 19 runs were compared. (In the following, the pairs of run results displayed in Figures 1–9 and Figure 15 were generally chosen because they illustrate noticeably different results.)

Population Size, Age at Death, Flaw Rate, Number of Faithful Replications, Number of Unfaithful Replications.

In each run, the *population size* rose rapidly from the initial value (64 ancestors) up to 800, the maximum number allowed. Whenever this ceiling was reached, 10% of the population was killed off stochastically, but according to each program's internal energy levels (as described earlier). After the ceiling had first been reached, the population size fluctuated in the region of around 700-800 programs for the rest of the run. Typical population size graphs are shown in Figure 1.

No trends were found for program *age at death*, *flaw rate*, *number of faithful replications per program*, and *number of unfaithful replications per program*. That is, for each of these measures, the distribution of values across the population showed no change right through the run. In addition to showing no trends, the *absolute* values of the measures were generally very similar in different runs. Example graphs for these measures are shown in Figures 2 (age at death), 3 (flaw rate), and 4 (faithful and unfaithful replications per program). In Figures 2 and 3, the plot on the left hand side shows a representative graph of the measure, as observed in the majority of the runs. The plots on the right hand side of Figures 2 and 3 show slightly unusual or noteworthy cases.

For *Age at Death* (Figure 2), there are a couple of points to note. Most obviously, there is considerable structure in the distribution of ages at which organisms die. This is interpreted as indicating that the cycle of births and deaths in the population is well synchronized throughout the run. The figure shows that the majority of programs live for some multiple of a little over 130 time slices, with fewer programs surviving for each successive multiple. This figure of 130 time slices corresponds very well with the time it takes the programs to replicate (see Figure 9). The obvious explanation is that each time the population size reaches the ceiling of 800 programs, a number of programs die, creating space for the remaining programs to reproduce. Once this reproduction stage occurs, the population size is soon at the ceiling again, so the cycle repeats. The extinctions triggered by the population size hitting the ceiling are therefore periodic, resulting in the observed distribution of ages, with most organisms surviving for an integral multiple of the period of this cycle. The second point about the *Age at Death* plots is that, in some runs, a slight kink in seen in them (e.g. in the middle section of the plot for Run 10, on the right hand side of Figure 2). Having just discovered that age of death is related to the replication period of the programs, it is not surprising to see that these kinks are associated with times of significant change in the replication period of the programs. For the graph of replication period for run 10, corresponding to the *Age at Death* plot on the right hand side of Figure 2, see the right hand side of Figure 9.

For flaw rates (Figure 3), in 16 out of the 19 runs, very few programs with flaw rates different to that of the ancestor programs appeared throughout the run. However, in three runs (3, 11 and 19), the whole population moved to a higher rate during the run (the figure effectively shows the reciprocal of the flaw rate, so the increase in flaw rate appears as a downward trend). If these changes in flaw rate were adaptive, one might expect to see corresponding changes in other measures, particularly the number of faithful and unfaithful reproductions per organism. However, no such trends were observed (the graph of number of unfaithful reproductions per organism for Run 3, for example, is shown on the right hand side of Figure 4). It therefore appears that these changes in flaw rate were the result of random (genetic) drift.

Activity (presence), Mean Activity (presence), Activity (concentration), Mean Activity (concentration), Diversity, Program Length, Replication Period.

To recap, the measures just discussed generally showed no trends, and their absolute values were very similar across different runs. In contrast, trends *were* ob-

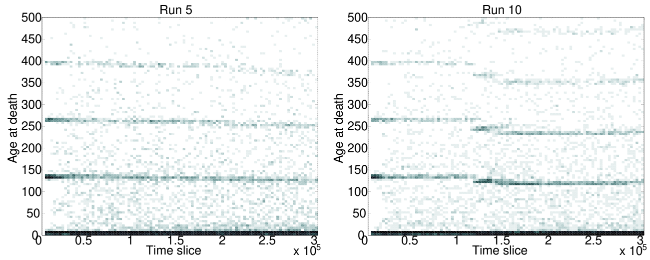

Figure 2: Age at Death, Runs 5 (*left*) and 10 (*right*)

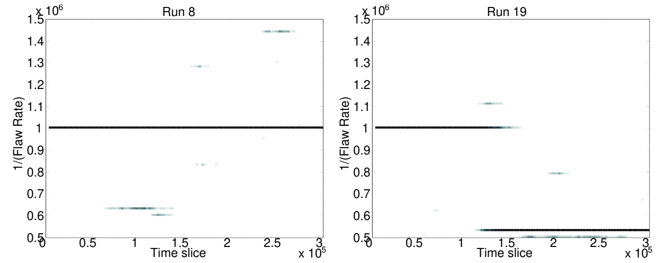

Figure 3: $\frac{1}{FlawRate}$, Runs 8 (*left*) and 19 (*right*). The vertical axis is scaled by a factor of 10^6.

served in seven of the other measures (i.e. Activity (presence), Mean Activity (presence), Activity (concentration), Mean Activity (concentration), Diversity, Program Length and Program Replication Period—discussion of the wave plots will be left until the end of the section), with noticeable differences between some of the runs. Plots for some of these measures are presented for two example runs (17 and 10) in Figures 5–9.

Ideally, we would like to know whether the differences in these measures between any of the runs are statistically significant. Such differences would indicate that evolution might genuinely be treading a different path, for no other reason than the different seed used for the random number generator when the runs commenced. The choice of a statistical test for this task was not immediately obvious. We wished to avoid parametric tests, as we did not want to make assumptions about the population parameters (for example, there is no reason to suspect that any of the measures we are looking at are normally distributed across all possible evolutionary runs).

We therefore chose a non-parametric method—a randomization version of the paired sample t test (see, for example, (Cohen 1995)). For each measure of interest, this test will tell us, for each run, which other runs produced significantly different results. The test can indicate whether two samples are related without any reference to population parameters. The procedure used was as follows:

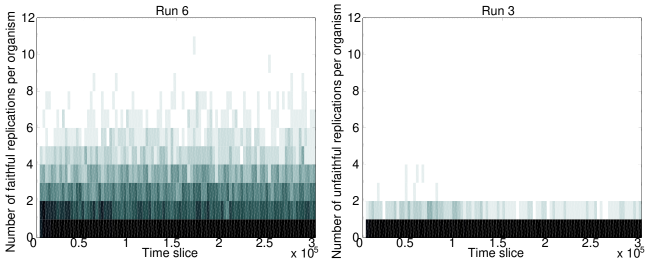

Figure 4: Number of Faithful Replications per Program, Run 6 (*left*). Number of Unfaithful Replications per Program, Run 3 (*right*)

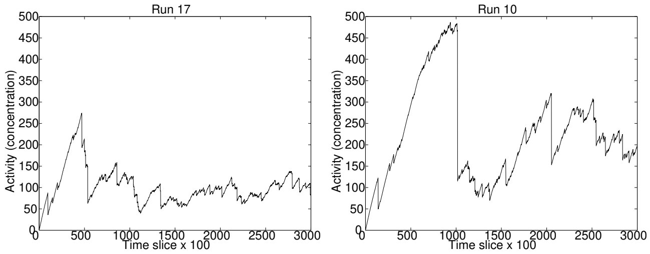

Figure 5: Activity (concentration), Runs 17 (*left*) and 10 (*right*)

Procedure: Randomization Version of the Paired-Sample t Test For each run, 10 sample data points were extracted, each one representing the value of the measure in question at one of 10 equally spaced times throughout the run.

The basic idea of the paired sample t test in this case is to consider the 10 sample points for pairs of runs in turn. By doing pairwise tests at 10 sample points we are comparing the measures at a number of points through the run, with no point having more significance than any other. For each pair of runs, the difference between corresponding samples is calculated, together with the mean value for the 10 differences. We then ask what the likelihood is of achieving this mean difference under the null hypothesis that the two runs are statistically equivalent. The method by which this is done will be explained shortly.

Obtaining Raw Sample Points In the case of measures which are already statistics of the whole population at any given time (i.e. both forms of the Activity measure, both forms of the Mean Activity measure, and Diversity), these 10 sample points could be taken directly from the value of the measure at the appropriate time. However, to prevent high-frequency

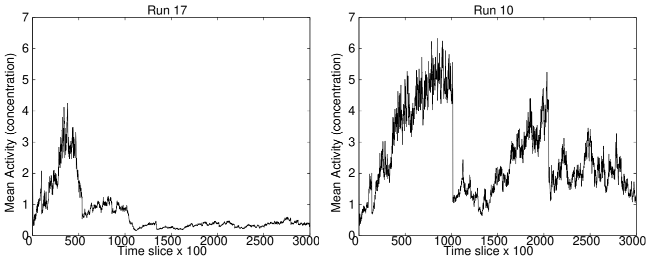

Figure 6: Mean Activity (concentration), Runs 17 (*left*) and 10 (*right*)

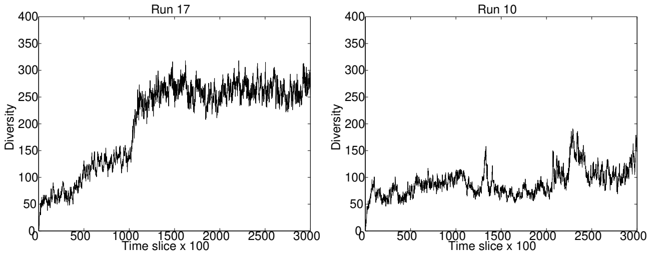

Figure 7: Diversity, Runs 17 (*left*) and 10 (*right*)

changes in these measures from producing aberrant results, the measures were first smoothed before the samples were taken (using median-smoothing with a window of 10,000 time slices).

In the case of the measures where the existing data consisted of multiple values at each time slice, each representing individual programs (i.e. the Program Length and Replication Period measures), each of the 10 sample points was produced by taking the median value of all values lying within a window of 1000 time slices around the time slice being sampled.

Obtaining Differenced Sample Points Because of the cumulative nature of evolution, it is possible that a small difference in the sampled value of a measure early on in a pair of runs will be magnified into a large difference later on, even if the two runs are actually proceeding in a fairly similar fashion. In order to gauge the magnitude of this effect, a duplicate set of tests was run, which used the *difference* in value between adjacent sample points as the figure to compare between runs, rather than the *absolute* value of the sample points. Using differenced data should reduce the influence of any cumulative disparity between runs.

Testing for Significance We are considering the difference in values between corresponding sample points in a pair of runs. Under the null hypothesis that

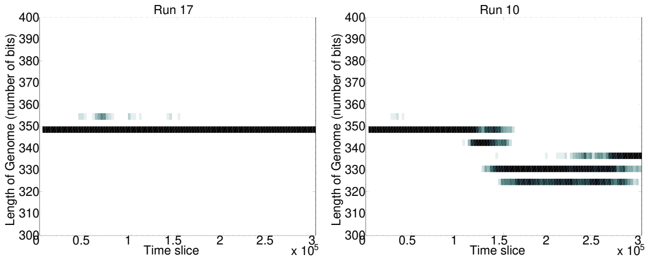

Figure 8: Program Length, Runs 17 (*left*) and 10 (*right*)

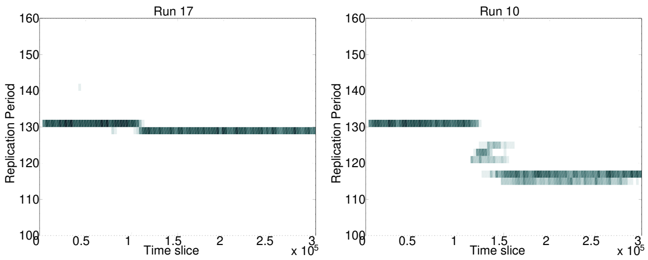

Figure 9: Replication Period (interval between first and second faithful replications), Runs 17 (*left*) and 10 (*right*)

the two runs are equal, however, it is equally likely that these values would be reversed (i.e. for sample point n for runs A and B, the null hypothesis is that the values A_n from run A and B_n from run B are just as likely to have come from the other run—A_n from run B and B_n from run A). If this were the case, the difference between the values would be the same as before, but with the sign reversed. We can test for the significance of the observed mean difference by constructing the distribution of all mean differences obtained from looking at each possible combination of each of the paired samples into one or other of the runs. As there are 10 paired samples, there are 2^{10} (1024) such combinations. The exact procedure is listed below (adapted from (Cohen 1995)), which may make things clearer:

1. For run I and J, if S_I and S_J are lists of the 10 sample data points for each run, construct a list D of the differences between these values, $D = S_I - S_J$. Denote the mean of these differences $\bar{x}_D$.

2. *if* $\bar{x}_D = 0$

   $p = 0.5$

   else

   (a) Set a counter C to zero.

   (b) for $i = 0..1023$

   - Construct a list D^* such that $D^*_j = D_j$ if $b_{ij} = 0$, or $D^*_j = -D_j$ if $b_{ij} = 1$, for $j = 1..10$, where b_{ij}

is the j^{th} digit of i in base 2.

- denote the mean of the new list $\bar{x}_{D^*}$
- *if* $\bar{x}_D > 0$
  if $\bar{x}_{D^*} \geq \bar{x}_D$, then increment C by one
  else if $\bar{x}_D < 0$
  if $\bar{x}_{D^*} \leq \bar{x}_D$, then increment C by one
  endif

(c) $p = (C/1024)$

p is the (one-tailed) probability of achieving a result greater than or equal to $\bar{x}_D$ (or less than or equal to $\bar{x}_D$ if $\bar{x}_D < 0$) by chance under the null hypothesis. That is, p is the probability of incorrectly rejecting the null hypothesis that systems I and J have equal population mean scores for the measure in question.

For each of the seven measures being considered (Activity (presence), Mean Activity (presence), Activity (concentration), Mean Activity (concentration), Diversity, Program Length and Replication Period), this procedure was followed for each of the $19(19-1)/2 = 171$ pairwise comparisons between runs, for both the raw sample data and the differenced sample data.

The p values for each pairwise comparison are shown graphically in Figures 10–14. These figures show one histogram for p values obtained using raw sample data, and another for p values obtained using differenced sample data. In all of the histograms, any p value less than 0.05 is plotted as zero. Bars of non-zero height on the histograms therefore represent pairs of runs which are not significantly different from each other for the measure in question at the $p = 0.05$ level.

(Note that, in order to emphasize the formation of various clusters of runs in these histograms, the runs in each histogram are arranged along the x and y axes in increasing order according to the mean of their 10 sample values. While this emphasizes clusters in any one histogram, it means that clusters occurring in similar positions in the histograms of different measures do not necessarily represent the same runs.)

The randomization version of the paired-sampled t test has some advantages over other methods of investigating pairwise comparisons (e.g. it is non-parametric), but it has the disadvantage that it is "virtually certain to produce some spurious pairwise comparisons" (Cohen 1995) (p.203). Cohen suggests one way, not to get around this problem, but at least to have some idea of the reliability of a particular set of pairwise comparisons (Cohen 1995) (p.204). The idea is to first calculate, at the 0.05 level, how many runs, on average, each run differed from (call this $\bar{n}_{0.05}$). Then calculate a similar figure at a much more stringent level. As we have 1024 numbers in our distribution of mean differences, the 0.001 level is appropriate. Finally, calculate the *criterion differential*, $C.D. = \bar{n}_{0.05} - \bar{n}_{0.001}$. If $C.D.$ is large, this indicates that many significant differences at the 0.05 level did not hold up at the 0.001 level. A small $C.D.$ value indicates that the experiment differentiates runs unequivocally, therefore lending more weight to the validity of the results at the 0.05 level. Table 1 shows $\bar{n}_{0.05}$, $\bar{n}_{0.001}$ and $C.D.$ for each measure, and for both raw and differenced sample data.

Table 1 reveals a number of interesting results. The most striking is the difference in the results of using raw sample points compared with differenced sample points.

Using raw data, the average number of runs that any particular run was significantly different to at the 0.05 level ranged from 3.89 for Activity (presence) to 13.26 for Diversity. However, the criterion differential for all of these measures is high (ranging from 3.68 for Activity (presence) to 12.32 for Program Length). This suggests that the validity of the figures at the 0.05 level are questionable, and the true figures are probably somewhat lower than those calculated. Having said this, the average number of runs that any particular run was significantly different to even at the 0.001 level was non-zero for the five measures suggested by Bedau et al. (ranging from 0.21 for Activity (presence) to 6.32 for Diversity).

Using differenced data, the results have a very different look. In only two measures were any runs significantly different from any others even at the 0.05 level (0.11 for Activity (concentration) and 0.42 for Diversity), and both of these vanished at the 0.001 level. In other words, these figures suggest that, for *all* of these measures, starting off at any point during any of the runs, the amount the measure *changed* over a given period was not significantly different compared to any of the other runs.

Activity Wave Diagrams

Whereas the Activity and Mean Activity measures produce a summary figure for a whole population of genotypes at time t, activity wave diagrams plot the success of every genotype in the population at every stage of the run (Bedau & Brown 1997). They are therefore a useful visualization technique for competition between genotypes, and the shape of an individual wave can also suggest the level of adaptive value of the corresponding genotype relative to its competitors.

The activity wave diagrams for most of the runs looked surprisingly different, although it is hard to quantify these differences (the Activity and Mean Activity measures do quantify some aspects of them,

Measure	Data Type	$\bar{n}_{0.05}$	$\bar{n}_{0.001}$	$C.D.$
Activity (presence)	raw	3.89	0.21	3.68
	differenced	0.00	0.00	0.00
Mean Activity (presence)	raw	12.00	4.53	7.47
	differenced	0.00	0.00	0.00
Activity (concentration)	raw	8.42	2.11	6.32
	differenced	0.11	0.00	0.11
Mean Activity (concentration)	raw	10.32	4.11	6.21
	differenced	0.00	0.00	0.00
Diversity	raw	13.26	6.32	6.95
	differenced	0.42	0.00	0.42
Program Length	raw	12.32	0.00	12.32
	differenced	0.00	0.00	0.00
Replication Period	raw	10.21	0.00	10.21
	differenced	0.00	0.00	0.00

Table 1: Mean number of runs that each run is significantly different from at the 0.05 level ($\bar{n}_{0.05}$) and 0.001 level ($\bar{n}_{0.001}$), and the criterion differential ($C.D.$). See text for details.

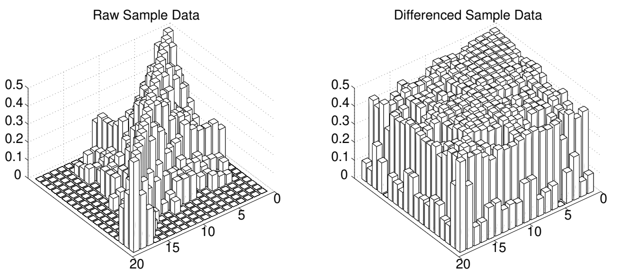

Figure 10: **Activity (concentration)**: Pairwise comparisons (p values) between runs. Raw Sample Data (*left*). Differenced Sample Data (*right*). p values below 0.05 are plotted as zero, so bars of non-zero height indicate pairs of runs that are not significantly different at the 0.05 level. See text for details.

but no single measure captures all of the important information that the diagrams can tell us). Example activity wave diagrams (for runs 17 and 10) are presented in Figure 15.

One way in which the activity wave diagrams can be very useful is in evaluating the effectiveness of different measures of evolution at highlighting the important adaptive events during a run. In particular, in the runs reported here it was observed that the Activity and Mean Activity measures based purely upon the *presence* of genotypes in the population bear little resemblance to the salient features of the wave diagrams. Indeed, these measures were introduced mainly so that they could be applied to fossil data as well as to data from artificial systems (the concentration data for fossil taxa being unknown) (Bedau *et al.* 1997). The meas-

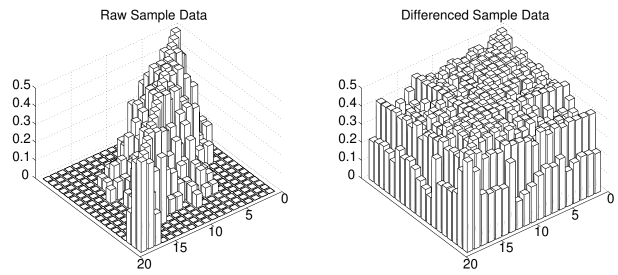

Figure 11: **Mean Activity (concentration)**: Pairwise comparisons between runs. See text and caption of Figure 10 for details.

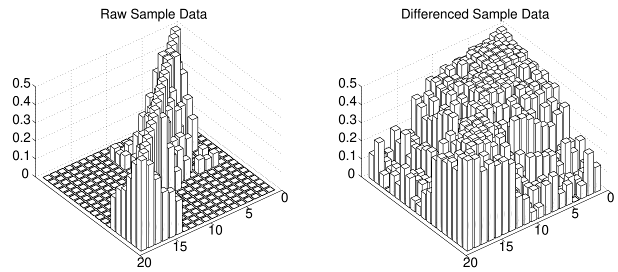

Figure 12: **Diversity**: Pairwise comparisons between runs. See text and caption of Figure 10 for details.

ures based upon the *concentrations* of genotypes should be better, and the results of these runs indicate that this is indeed the case. Activity (concentration) usually seems to give a better reflection of the wave diagram than does Mean Activity (concentration). This is possibly because the latter measure is defined as

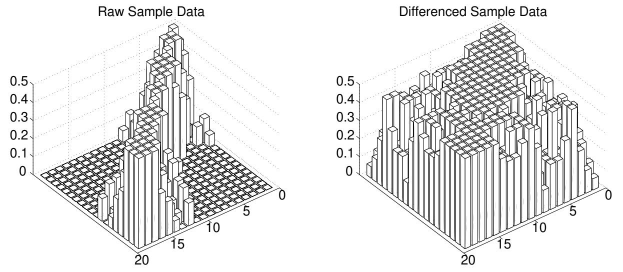

Figure 13: **Program Length**: Pairwise comparisons between runs. See text and caption of Figure 10 for details.

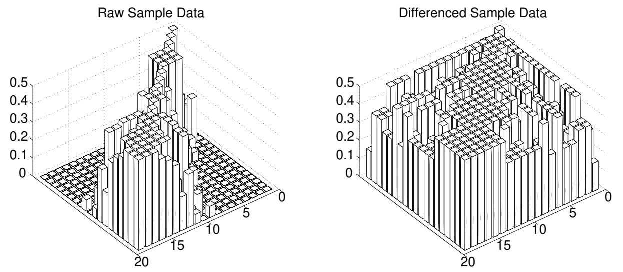

Figure 14: **Replication Period**: Pairwise comparisons between runs. See text and caption of Figure 10 for details.

Activity divided by Diversity, but diversity, by its very nature, does not take account of the *concentrations* of different genotypes, but merely their presence.

Discussion

As discussed earlier in the paper, the three factors that are fundamental to the success of genotypes in an evolving population are the longevity, fecundity and copy-fidelity of the individuals. The measures chosen to track these factors in the runs reported here were *Age at Death*, *Replication Period*, *Program Length*, *Flaw Rate*, *Number of Faithful Replications* and *Number of Unfaithful Replications*. Very little change was observed in any of these measures except Program Length and Replication Period throughout the course of any of the runs. It therefore appears that, under the set of parameters used in these runs, the programs are only able to evolve along one of the three axes (fecundity) theoretically available to them. Studying some of the programs that evolved during the runs suggests that most adaptive events involved either making the program shorter by removing (what turned out to be) redundant instructions, or by adding energy collection instructions to reduce the chance of the program being culled.

For Program Length and Replication Period, signi-

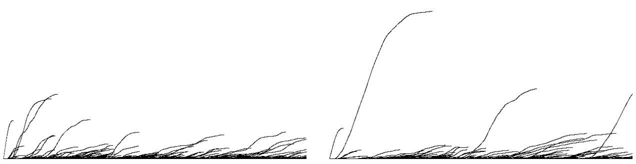

Figure 15: Activity Wave Diagram, Runs 17 (*left*) and 10 (*right*)

ficant differences (at the 0.05 level) were observed in the raw data values between some runs. For these measures, the mean number of runs that each run is significantly different from at this level was calculated as 12.3 for Program Length and 10.2 for Replication Period, but the high criterion differential on these scores suggests that the true value should be somewhat lower (looking at Figures 13 and 14, probably somewhere in the range of 6 to 10).

Looking at the derived measures suggested by Bedau et al. (Activity (presence), Mean Activity (presence), Activity (concentration), Mean Activity (concentration) and Diversity), significant differences were found between runs which did hold up even at the 0.001 level. Again, the true value of each of these differences probably lay in the range of roughly 6 to 10.

These results indicate that each run, on average, performed significantly differently to between a third and a half of the other runs. One of the main reasons for doing these experiments was to understand how we should deal with contingency when conducting further experiments with Cosmos. If we assume that at least the finding that each run is statistically different to more than a third of the others is a general result, then we can use the following rule of thumb: For each re-run of a trial with a different seed for the RNG, the probability of its outcome being statistically equivalent (at the $p = 0.05$ level) to the original one is, at most, about $\frac{2}{3}$. Therefore, the number of *re*-runs that should be conducted to be confident (at the 95% level) of at least seeing one statistically different type of behaviour is n, where $(\frac{2}{3})^n \leq 0.05$, i.e. $n \geq 7.388$, or, in round figures, $n \geq 8$. This is the number of *re*-runs *after* the original, so, finally, we can say that *any trial should be conducted nine times with different seeds for the RNG.*

Having said that each run performed significantly differently to at least a third of the other runs, precisely *which* runs were significantly different depended upon the particular measure being looked at. This emphasizes the fact that one should be clear about exactly what measure is being used when talking about comparisons between evolutionary runs.

The fact that *no* significant differences were found between any of the runs for any of the measures when looking at *differenced* sample data is of great interest. It suggests that the significant differences observed in *raw* sample data may be caused (at least in part) by the cumulative magnification of initially small differences as a run proceeds. If this effect is controlled for (which was the purpose of using differenced data), the behaviour of the runs in terms of the *change* in values of the measures over a given time period would seem to be very similar in all of the runs. However, because of the cumulative magnification of small differences, the *absolute* outcomes of the runs *do* differ significantly in some cases, so contingency *does* play a big role.

Finally, we can ask to what extent these results can be generalized to other evolutionary systems. Considering biological evolution first, it is clear that even just in terms of population size and the length of runs, the system is completely trivial. Also, the role of contingency may be different in systems which have rich ecological interactions (of which Cosmos programs have very little). It would therefore be unwise to claim that these results can tell us much about the role of contingency in biological evolution, but they may be relevant in specific cases. As for other artificial evolutionary systems, Cosmos is of comparable design, so the results, and the rule of thumb about the number of trials that should be run, should be broadly applicable to these platforms as well. The extent to which ecological interactions affect the results may be investigated by running similar trials on systems that display stronger interactions of this kind (such as Tierra).

Acknowledgements

Thanks to Chris Adami and four anonymous reviewers for helpful comments on a draft of this paper, and also to Mark Bedau and Emile Snyder for supplying software for producing evolutionary activity data from the raw data of a run. One of the authors [TT] is supported financially by EPSRC grant number 95306471. The facilities used for this work were provided by the University of Edinburgh.

Appendix:Non-default parameter values

ancestor=user_defined number=64 rng_seed=[*variable*]
limited_run=yes number_of_timeslices=300000 grid_size=40
horizontal_wrap=yes vertical_wrap=yes max_cells_per_process=800
x_delta=0.025 et_value_constant=0.025 et_value_power=1.0
max_energy_tokens_per_cell=50 apply_flaws=yes
max_energy_tokens_per_grid_pos=25 mutation_period=1000000
mutation_application_period=1 default_flaw_period=1000000
neighbouring_genomes_readable=yes

References

Adami, C., and Brown, C. 1994. Evolutionary learning in the 2D artificial life system 'Avida'. In Brooks, R., and Maes, P., eds., *Artificial Life IV.* The MIT Press. 377–381.

Bedau, M. A., and Brown, C. T. 1997. Visualizing evolutionary activity of genotypes. (preprint).

Bedau, M. A., and Packard, N. H. 1991. Measurement of evolutionary activity, teleology and life. In Langton; Taylor; Farmer; and Rasmussen., eds., *Artificial Life II.* Redwood City, CA: Addison-Wesley. 431–461.

Bedau, M. A.; Snyder, E.; Brown, C. T.; and Packard, N. H. 1997. A comparison of evolutionary activity in artificial evolving systems and in the biosphere. In Husbands, P., and Harvey, I., eds., *Fourth European Conference on Artificial Life*, 124–134. MIT Press/Bradford Books.

Cohen, P. 1995. *Empirical Methods for Artificial Intelligence.* MIT Press.

Dawkins, R. 1989. *The Selfish Gene.* Oxford: Oxford University Press, 2nd edition.

Fontana, W., and Buss, L. W. 1994a. 'The arrival of the fittest': Toward a theory of biological organization. *Bull. Math. Biol.* 56:1–64.

Fontana, W., and Buss, L. 1994b. What would be conserved if "the tape were played twice"? *Proc. Natl. Acad. Sci. USA* 91:757–761.

Gould, S. 1989. *Wonderful Life: The Burgess Shale and the Nature of History.* Penguin Books.

Gould, S. J. 1993. How to analyze the Burgess disparity—a reply to Ridley. *Paleobiology* 19(4):522–523.

McShea, D. W. 1991. Complexity and evolution: What everybody knows. *Biology and Philosophy* 6:303–324.

McShea, D. W. 1993. Arguments, tests, and the Burgess Shale—a commentary on the debate. *Paleobiology* 19(4):399–402.

McShea, D. W. 1994. Mechanisms of large-scale evolutionary trends. *Evolution* 48(6):1747–1763.

McShea, D. 1996. Metazoan complexity and evolution: Is there a trend? *Evolution* 50(2):477–492.

Ray, T. 1991. An approach to the synthesis of life. In Langton; Taylor; Farmer; and Rasmussen., eds., *Artificial Life II.* Redwood City, CA: Addison-Wesley. 371–408.

Ridley, M. 1993. Analysis of the Burgess Shale. *Paleobiology* 19(4):519–521.

Skipper, J. 1992. The computer zoo—evolution in a box. In Varela, F., and Bourgine, P., eds., *Toward a Practice of Autonomous Systems: Proceedings of the First European Conference on Artificial Life*, 355–364. Cambridge, MA: MIT Press.

Taylor, T., and Hallam, J. 1997. Studying evolution with self-replicating computer programs. In Husbands, P., and Harvey, I., eds., *Fourth European Conference on Artificial Life*, 550–559. MIT Press/Bradford Books.

Taylor, T. 1997. The COSMOS artificial life system. Working Paper 263, Department of Artificial Intelligence, University of Edinburgh. Available from `http://www.dai.ed.ac.uk/daidb/people/homes/timt/papers/`.