Avida Project – Exploring the (μ,N) Landscape Student from Fall 2011

Introduction — Adaptation in biological populations is affected by many parameters, among them mutation rate (μ) and population size (N). The effects of extreme values of either μ or N are relatively well-established. At very small values of μ, adaptation is mutation-limited, and at very large values, there is insufficient fidelity of replication to preserve the advantage of beneficial mutations. At moderately large values of μ, clonal interference results in slower adaptation, because beneficial mutations in different lineages battle each other while trying to reach fixation (1). In small populations (small N), genetic drift, or the fixation of alleles by random chance, is a more powerful force; because drift acts without regard to the selective advantage of mutations, it tends to disproportionately amplify neutral and deleterious mutations, compared to the effect of selection alone. At large N, adaptation speeds up because a larger population increases the chance that a highly beneficial mutation will be present in any given generation, but clonal interference can also slow adaptation. High mutation supply rates tend to increase the rate of adaptation, with diminishing returns (2). However, combinations of μ and N may produce interesting dynamics; for instance, there is evidence that at very large values of μN, clonal interference decreases (3).

In this project, I explore the interaction between μ and N. We can express a population's mutation supply rate per generation (M) in terms of these factors, such that

=

At a given value of M, there is a range of values of μ, along with the corresponding values of N; as μ increases, N decreases (Fig. 1). I allow populations of digital organisms to evolve at different points on this M curve and examine the average fitness after 2000 generations in order to investigate the combined effects of small μ and large N, or large μ and small N. More generally, I explore different points on the (μ,N) landscape, specifically the effect of increasing N given constant μ. I expect that at points with small μ and large N, adaptation will be mutation-limited and thus slower, and that at large μ and small N, adaptation will be slower because it will be dominated by genetic drift. I also hypothesize that the optimal combination of μ and N has an intermediate value of μ and a large value of N. Thus, I expect that rate of adaption is not solely determined by the generational mutation supply rate, but also depends on the component values of μ and N.

Methods — I used the program Avida-ED to digitally simulate evolution (4). For 24 combinations of μ and N corresponding to five values of M, I ran ten identical trials for 2000 generations each, for a total of 240 trials (Table 1). The range of μ was chosen based on preliminary tests that suggested that the maximal values of average fitness resulted from μ between 0.01 and 0.02 (1 and 2 percent respectively; values of μ are given as a decimal rather than a percent throughout this report). Also, this range resulted in computationally feasible values of N for running repeated trials. Offspring were placed randomly in the dish, and the repeatability mode was set to experimental. Average fitness values were recorded at the end of each experiment, and the mean and median of the average fitness, were calculated for each selected point in the (μ,N) landscape. All plots and statistics were calculated in Mathematica.

Results — The primary value of M under study was M=12, and a set of trials was run at each of the μ-values 0.005, 0.01, 0.012, 0.015, 0.018, 0.02, and 0.03. Figure 2 shows the values of average fitness at the end of each trial, plotted at the appropriate point in the (μ,N) plane. There is a consistent dip in fitness across the higher μ-values, with significant variability in the lower range of μ (Table 1, Fig. 4a). The data takes on a Poisson-like distribution—fitness values cluster towards the lower end for each set of trials, with a few outliers in the upper range, consistent with the step-like increases in fitness seen during the trials. Because of this distribution, outliers have a disproportionate effect on the mean of fitness, so median and the natural logarithm of median (hereafter log(median); appropriate because of Avida's exponential valuation of metabolic functions) were used for analysis.

Christopher Marx 11/15/11 11:53 PM
Comment: This is perhaps a bit technical of a title. Can you think of something a bit more broad?

Primrose Boynton 11/15/11 7:03 PM
Comment: You are very clear in your introduction in setting up the problem, and I like how you explicitly phrase your expectations and hypotheses here.

Christopher Marx 11/15/11 11:55 PM
Comment: Agreed – this is a fantastic framing of the question.

Christopher Marx 11/15/11 11:56 PM
Comment: Wow – that's a ton!

Christopher Marx 11/15/11 11:59 PM
Comment: I think I understand this – and that it is quite a nice line of thought – but perhaps you can draw it out a bit more…

Primrose Boynton 11/15/11 7:14 PM
Comment: Hm…you have more experience with these data than I do by now, but it might be more appropriate to just log-transform them and then use the mean and standard deviation.

Christopher Marx 11/15/11 11:59 PM
Comment: Great point.

The data along the curve M=12 showed significant variability, again probably due to the Poisson-like data distribution. To better understand the (μ,N) landscape, I ran more trials along different M-curves rather than simply running more trials along the curve M=12, in hopes of finding more general trends that would persist in spite of the inherently large variability in the data.

Fig. 3 shows the mean, median, and log(median) of each set of trials, plotted at the appropriate point in the (μ,N) landscape. As expected, the plot of the mean is much noisier than that of the median. All plots, however, show a general increase in average fitness as N increases, though the effect of μ is less clear.

I plotted log(median) of average fitness with respect to only one parameter, μ or N, to better understand the individual contribution of each parameter to the rate of adaptation. Fig. 4 shows the plot of μ versus fitness; in Fig. 4a, points are colored in groups of M, and in Fig. 4b, the linear best-fit model is plotted alongside the data. This best-fit model had an R^2 value of -0.01, indicating that the data follow no consistently linear pattern, and Fig. 4a shows the large variability of fitness within a given value of M. Fig. 5a shows the plot of N versus fitness, with points colored in groups of μ, and Fig. 5b shows the linear best-fit model. For N versus fitness, the best-fit model had an R^2 value of 0.61, suggesting that N is the limiting factor of fitness increase in the sampled range of μ and N. Plotting M versus average fitness yields results very similar to the plot of N versus fitness, and the best-fit model has a similar R^2 value of 0.59, reinforcing the observation that N is the main parameter affecting fitness increase in the studied region of the (μ,N) plane.

Discussion — The results described above point to two dominant trends. Average fitness is highly variable across replicates under the same conditions, making statistical analysis difficult. Also, in the sampled region, N is the main parameter limiting fitness increase; increasing N causes adaptation to increase regardless of μ in the range of μ under study, and M, a product of μ and N, has the same effect on adaptation as N alone.

The large inherent variability in fitness increase means that the results should be viewed with a note of caution. Since the data is distributed more like a Poisson than a normal, taking the mean of a data set is unduly influenced by outliers. I would like to conduct more replicate trials at each sampled point in the (μ,N) plane, but taking more trials might not cause standard statistical measures like the mean to converge to useful measures of the data (though the median would be more likely to converge).

Other studies have shown that rate of adaptation increases with an increase in mutation supply rate, with diminishing returns (2). This project establishes that the range of (μ,N) space sampled, from μ=0.012 to μ=0.02 and N=400 to N=2400, falls in the range in which the increase in adaptation resulting from increasing mutation supply rate is a function of population size. For each value of μ tested, increasing N results in an increase in adaptation, at about the same rate for different values of μ (Table 2). Thus, genetic drift probably has a strong effect in slowing adaptation. This result is unsurprising given the small population size and high mutation rate of these Avida trials, relative to natural or laboratory microbial populations.

It would be interesting to study a greater region of (μ,N) space to identify which regions are dominated by μ and which ones by N. Based on current data, I would guess that the region where μ ranges from 0.02 to 0.03 would be dominated by the high μ, so that increases in N have little to no effect on rate of adaptation. Because of the computational cost of running many replicates, this project did not examine the region in which μ is very low and N is relatively high, and it is possible that the range of μ explored does not include the optimal value of μ for rapid adaptation in this simulation—another possibility for further study.

Let us return to the original question of whether the generational mutation supply rate M alone can act as a predictor of rate of adaption. The analysis suggests that along the portion of the M-curve under examination, variations in the μ-to-N ratio are not reflected in the rate of adaption, except insofar as this variation also affects N (Fig. 4, 5). Along a given M-curve, μ and N are hyperbolically related, so for certain ranges of μ, lower than the range currently under study, a small change in μ corresponds to a large change in N (Fig. 1). Thus, this project does not rule out the possibility that different dynamics are in play along

Primrose Boynton 11/15/11 7:18 PM
Comment: Nice!

Christopher Marx 11/16/11 12:06 AM
Comment: Again – VERY clear…

Primrose Boynton 11/15/11 7:21 PM
Comment: This is a good insight

Christopher Marx 11/16/11 12:08 AM
Comment: Very cool. If I understand correctly, you are saying that despite the fact that either factor gives more mutations, the real issue is to avoid drift (which only depends upon N). Very cool.

Christopher Marx 11/16/11 12:09 AM
Comment: Yes, and I'd also consider greater time. Would this still hold true after 20K generations instead of 2K?

different regions of the M-curve due to the component values of μ and N, and further analysis in other regions would be needed to better characterize the (μ ,N) landscape.

References

1. P.J. Gerrish and R.E. Lenski, The fate of competing beneficial mutations in an asexual population. *Genetica* **102/103**, 127-144 (1998).
2. J.A. de Visser et al., Diminishing returns from mutation supply rate in asexual populations. *Science* **283**, 404-406 (1999).
3. J.P. Bollback, J.P. Huelsenbeck, Clonal interference is alleviated by high mutation rates in large populations. *Mol. Biol. Evol.* **24**(6), 1397-1406 (2007).
4. C. Adami, Digital genetics: unraveling the genetic basis of evolution. *Nat. Rev. Gen.* **7**, 109-118 (2006).

Primrose Boynton 11/15/11 7:30 PM

Comment: I found your project very clear and well thought out. I liked the way you broke M down into its components and looked at both, and I appreciated your thorough analysis of the data. Good work!

Christopher Marx 11/16/11 12:18 AM

Comment: Overall: 80/80.
This is textbook *perfect*. I could not have described what I wanted better than you did. There is absolutely no need to do more on the written project, but simply to think of how to convey this orally & visually in 5 minutes. Great work.

Figures

M	μ	N	mean fitness	median fitness	log(med)
8	0.02	400	55.2	18.5	1.26717173
12	0.03	400	47	22.5	1.35218252
8	0.018	444	94.5	28	1.44715803
8	0.015	533	107	11	1.04139269
12	0.02	600	48	26	1.41497335
8	0.012	667	123.4	25.5	1.40654018
12	0.018	667	163	96.5	1.98452731
8	0.01	800	51	27	1.43136376
12	0.015	800	251	121	2.08278537
18	0.02	900	265.8	177	2.24797327
12	0.012	1000	101	30	1.47712125
18	0.018	1000	367.5	89.5	1.95182304
12	0.01	1200	169	121	2.08278537
18	0.015	1200	161	87.5	1.94200805
24	0.02	1200	263	31.5	1.49831055
24	0.018	1333	320.1	139.5	2.14457421
27	0.02	1350	205	187	2.27184161
18	0.012	1500	164.4	112	2.04921802
27	0.018	1500	212	200	2.30103
24	0.015	1600	849.1	176.5	2.24674471
27	0.015	1800	203	145	2.161368
24	0.012	2000	655	435.5	2.63898816
27	0.012	2250	334	287	2.4578819
12	0.005	2400	95	53.5	1.72835378

Table 1. The mean and median of the average fitness from each set of ten trials.

M	μ	N	slope	R^2
8		-	-10.29	-0.25
12		-	-19.2	0.08
18		-	-41.5	0.62
24		-	-60.09	0.42
27		-	-16.35	-0.17
	total		-15.45	-0.01
-	0.012		0.00081	0.89
-	0.015		0.00071	0.45
-	0.018		0.00065	0.74
-	0.02		0.00079	0.24
		total	0.00067	0.61

Table 2. In the upper part of the table, the slopes and R-squared values of the linear best-fit model for the relation between μ and average fitness are shown for each set of points with the same M. In the lower part, the slopes and R-squared values of the linear best-fit model for the relation between N and average fitness are shown for each set of points with the same μ. For more information, see Fig. 4,5.

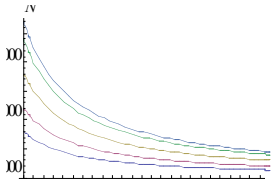


Figure 1. Curves are plotted for the points in the (μ ,N) landscape yielding M values of 8, 12, 18, 24, and 27, from bottom to top.

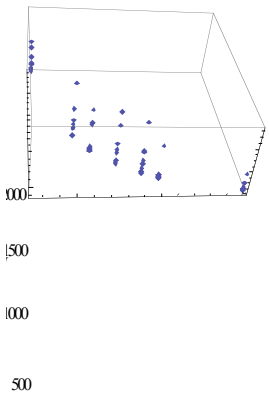
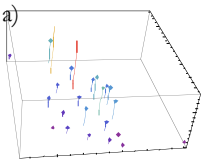


Figure 2. Average fitness at the end of each trial run at selected points in the (μ ,N) landscape corresponding to the curve M=12. In future analysis, the median of each set of trials was used for analysis.

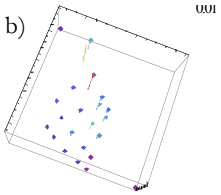
Christopher Marx 11/16/11 12:02 AM

Comment: I LOVE the idea of looking at it this way but I am having a bit of a hard time seeing the vertical dimension. Your use of color below seems to be an excellent solution that perhaps you could apply here, too.

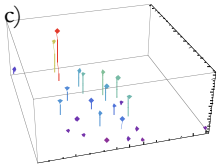
3. a)



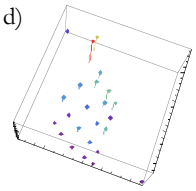
b)



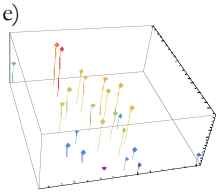
c)



d)



e)



f)

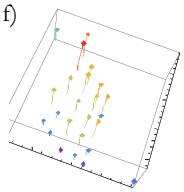
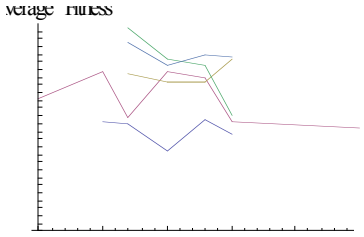
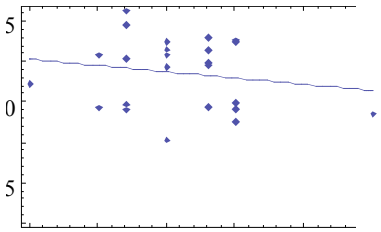


Figure 3. Each point in the plot represents the mean (a,b), median (c,d), or natural logarithm of the median (e,f) of the average fitness at the end of ten trials run under the conditions specified in "Methods" at the indicated point in the (μ ,N) plane. Each pair of figures represents two different views of the same plot.

4. a)



b)

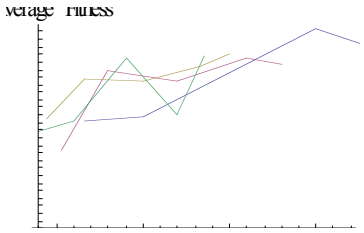


2.10418 _ 15.4535 x

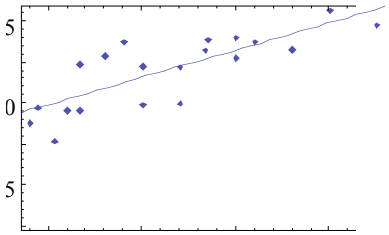
Figure 4. The natural logarithm of the median of the average fitness is plotted against the mutation rate μ only (a). Each line connects a set of points with the same value of M — 8 (dark blue), 12 (purple), 18 (yellow), 24 (green), and 27 (light blue). In figure 4b, all of the points are considered part of the same set, and are plotted with the linear best-fit line through them (R^2 = -0.01).

Primrose Boynton 11/15/11 7:17 PM
Comment: It was sometimes hard to see axis labels in your plots, and a legend next to the plot would make this easier to read.

5. a)



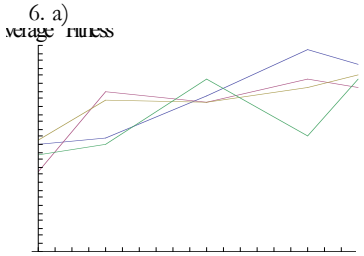
b)



1.14158 _ 0.000668414 x

Figure 5. The natural logarithm of the median of the average fitness is plotted against the population size N only (a). Each line connects a set of points with the same value of μ — 0.012 (dark blue), 0.015 (purple), 0.018 (yellow), and 0.02 (green). In figure 4b, all of the points are considered part of the same set, and are plotted with the linear best-fit line through them ($R^2 = 0.61$).

6. a)



b)

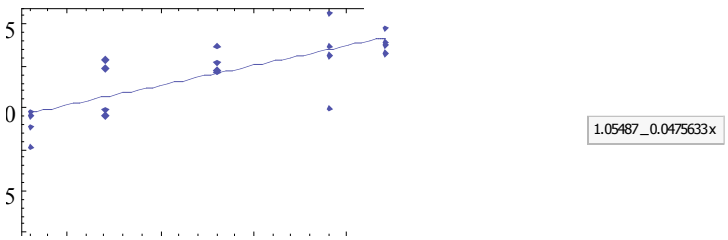


Figure 6. The natural logarithm of the median of the average fitness is plotted against the mutation supply rate M only (a). Each line connects a set of points with the same value of μ — 0.012 (dark blue), 0.015 (purple), 0.018 (yellow), and 0.02 (green). In figure 4b, all of the points are considered part of the same set, and are plotted with the linear best-fit line through them ($R^2 = 0.59$).