Project 3 Report (Followup)

Brian Satzinger

December 17, 2009

Contents

1 Introduction

In the conclusion to the first part of this report, I stated that the data could be improved by running more experiments and using a more complex graph. For this report, the size of the graph has been increased from 10 vertexes to 20 vertexes, and 6000 experiments were conducted instead of 24. This large increase in the number of experiments performed is due to the development of automated test software, which ran the genetic algorithm repeatedly and collected the data about it. These 6000 experiments generated 1.1GB of data summarizing the condition of each of the 16 populations at the end of each generation in every experimental run, and took approximately four hours to complete on my single-processor computer.

Ultimately, the convergence time data was too varied to draw any conclusions about the effectiveness of migration (directed or random). The effect of the mutation gradient was more pronounced on the results, but it does not also appear to be very beneficial.

2 Changes to Graph Coloring Problem

The graph was changed from a 10 vertex symmetric graph, to a 20 vertex randomly generated graph, which is shown in Figure 1 with a four-color solution (generated by the GA). This increase in size substantially increases the computational complexity of the problem. The figure below shows a correct coloring of the graph using only four colors. The GA developed this solution, but it did not reliably develop correct four-color solutions. For testing, the GA was terminated when it developed the first correct five-color solution, which it was able to do reliably.

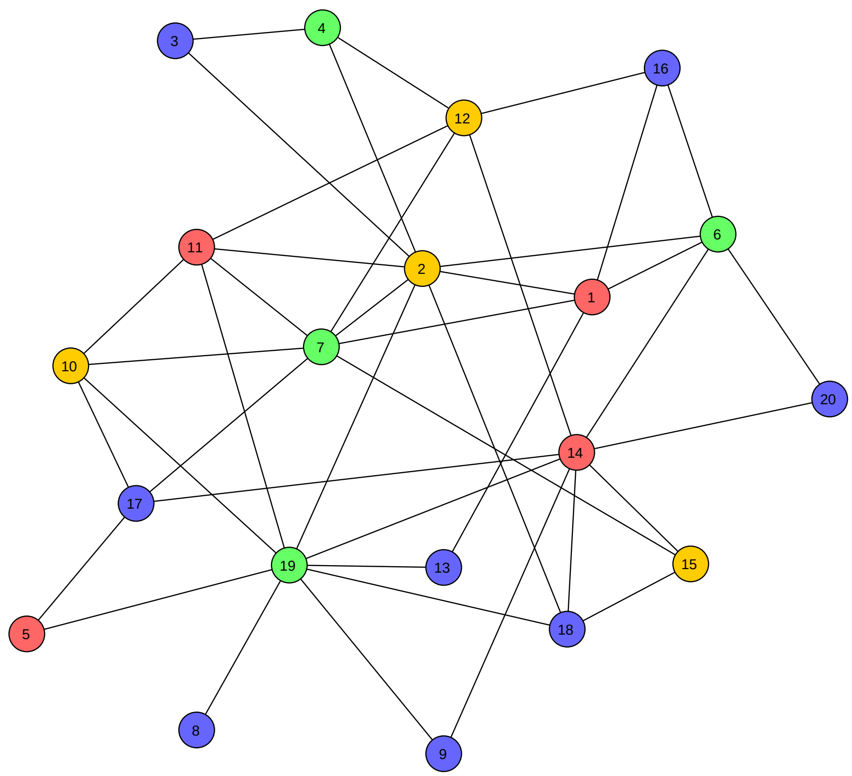

Figure 1: The new graph used in experiments (with 4 color solution)

3 Changes to Fitness Function

The fitness function was revised. The old fitness function (1) selected too much for the number of colors used, and did not penalize duplicated colors enough. This led to solutions with too few colors used. The new fitness function (2) penalizes incorrect colorings more, and divides by the square root of the number of colors used to reduce the effect of n_{colors} on the result.

$$fitness = \frac{n_{different} - 2 \cdot n_{same}}{n_{colors}} \tag{1}$$

$$fitness = \frac{n_{different} - 4 \cdot n_{same}}{\sqrt{n_{colors}}} \tag{2}$$

4 Results & Discussion

The 6 configurations used for testing in the first part of the report are also used here. They are summarized in table 1. For each of the 6 configurations, the GA was run 1000 times. The time to develop the first correct five-color solution to the graph in Figure 1 was recorded.

Run	Type of Migration	Mutation Gradient Used?
A	Directed	Yes
B	Directed	No
C	Random	Yes
D	Random	No
E	None	Yes
F	None	No

Table 1: Configurations used in experiments

4.1 Time to First 5-Color Result

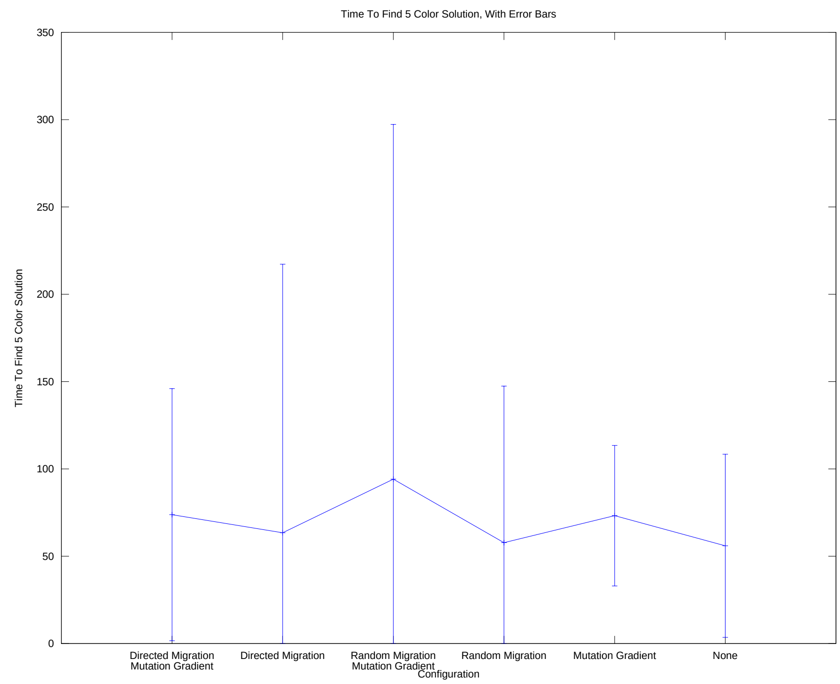

Figure 2: Time to produce first correct five-color result, with error bars

Run	Mean Time (Generations)	Std. Dev. (Generations)
A	73.97	72.2
B	63.42	153.77
C	94.13	203.15
D	57.78	89.61
E	73.21	40.22
F	55.99	52.44

Table 2: Time to produce first correct five-color result

These results are disappointing. There was no clear difference between the techniques, since the time to find the first correct five-color result varied so much between trials in the same run, possibly due to variations in the random initial populations. Although there were some differences in the mean time, the standard deviations are so large that I hesitate to draw any conclusion about the comparative effectiveness of the techniques. However, it should be noted that the technique with the lowest mean time was the one with neither migration nor a mutation gradient.

There does seem to be a relationship between the mean time and the use of the mutation gradient. Those runs which did not use a mutation gradient (B, D, and F) had lower mean times than those runs which did

(A, C, and E). The effect of the mutation gradient on the time to completion will be further explored in the next section, using histograms to better characterize the distribution of completion times.

4.1.1 Histograms

The data from runs B, D, and F were combined to give Figure 3, while the data from runs A, C, and E were combined to give Figure 4. These histograms show the distribution of the time required (in generations) to find the first correct five-color solution. Figure 3 shows the results from using no mutation gradient, while Figure 4 shows the results from using a mutation gradient. Each histogram represents the results from 3000 trials.

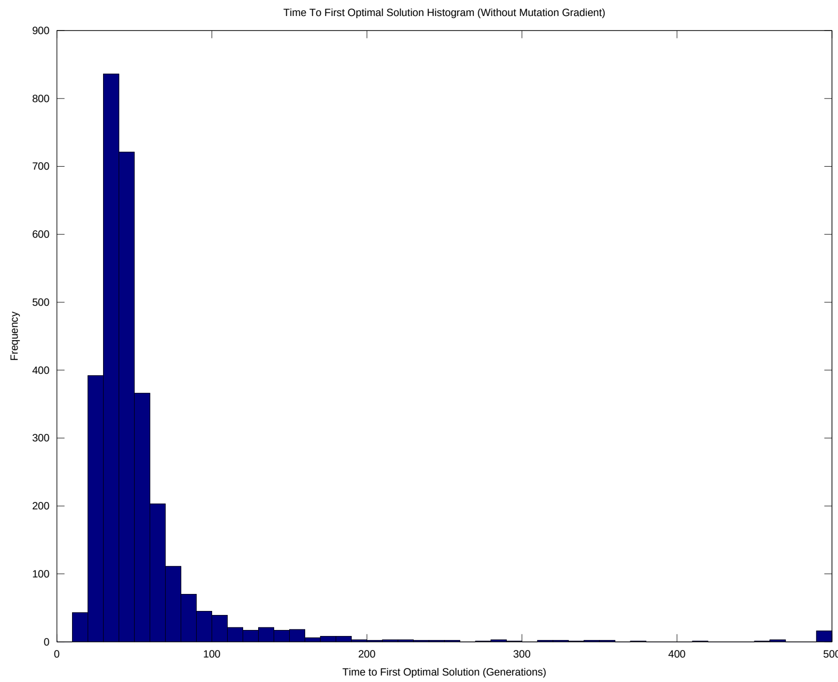

Figure 3: Histogram of completion time for runs using no mutation gradient

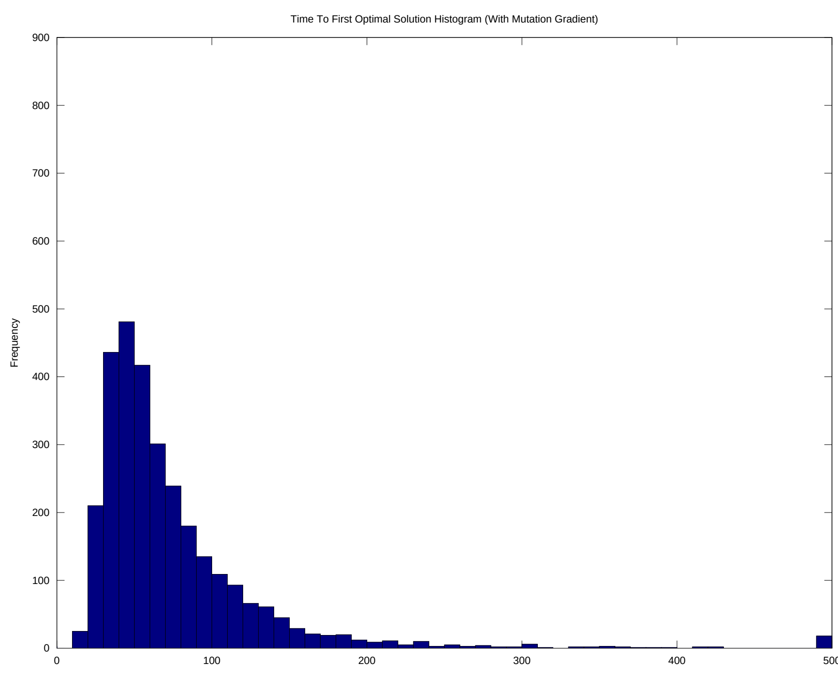

Figure 4: Histogram of completion time for runs using a mutation gradient

There is a clear difference between Figures 3 and 4. Figure 4 has a much lower peak, and is skewed toward longer times. Figure 3, which shows the results from not using a mutation gradient, has a sharper peak, and peaks at a lower (better) value. This shows that using mutation gradients in this experiment caused a greater variance in the completion time, and also reduced the mean completion time.

4.2 Location of First 5-Color Result

I was able to extract information about which populations tended to produce the first five-color result. These results are plotted as surface plots by the x and y coordinate of the population in the population grid, with the z coordinate showing how many times that population was the first to produce the first result. Each surface plot represents 1000 trials, although there may be more than 1000 data points represented, since there were some cases where different populations developed optimal solutions simultaneously. Runs B, E and F are compared to illustrate the spatial effects of mutation gradients and directed migration.

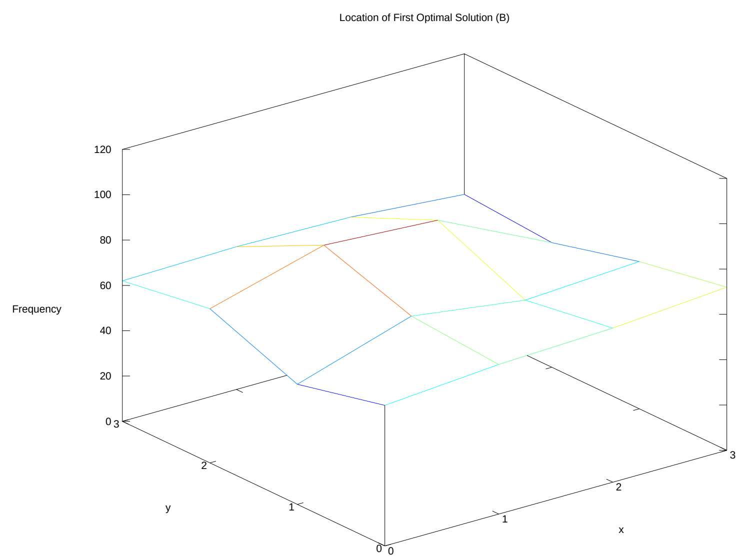

Figure 5: Location of first five-color solution for run B (using only directed migration)

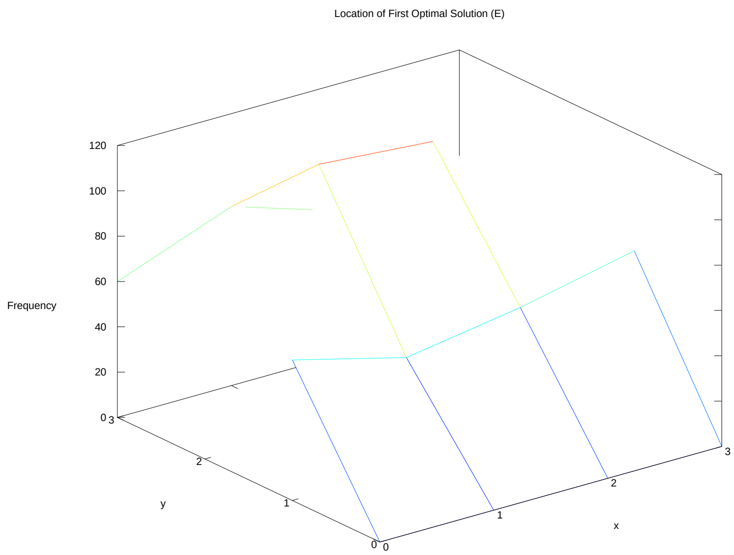

Figure 6: Location of first five-color solution for run E (using only a mutation gradient)

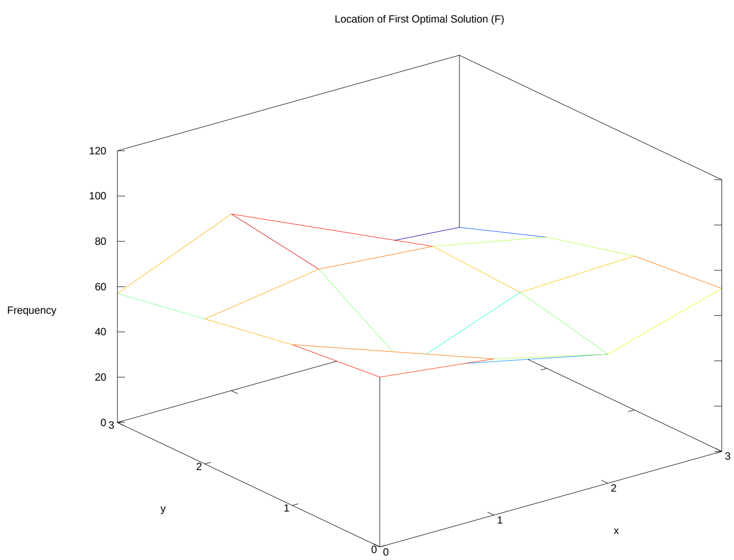

Figure 7: Location of first five-color solution for run F (using neither migration nor a mutation gradient)

Figure 6 shows that the mutation gradient dominates the success of each population. There is a large link between the y coordinate of the population and the ability of the population to generate the first correct five-color solution. In this case, the populations with y=0 were not the first to generate a correct solution in any of the 1000 trials. This can be attributed to the mutation gradient used, which causes a mutation rate of 0 in these populations. Without migration, there is no way for these populations to change once a candidate solution takes over. These populations likely converged too quickly on poor solutions. This may explain some of the poor performance of the mutation gradient technique. It would be interesting to reformulate the mutation gradient so that the mutation rate was non-zero for these populations and repeat the experiments.

For Figure 5, which shows the effect of directed migration, the populations in the middle of the grid (who have more neighbors than the populations on the edge of the grid) are successful more often. This might suggest that directed migration would be more beneficial if the grid was increased in size to reduce the proportion of edge-populations. The populations could also be connected around the edges of the grid to eliminate this problem.

Figure 7 shows the results from using neither migration nor a mutation gradient. In this case, each population is independent. Here, there is no clear trend between location and success. This is expected, since the location of the population does not have a mechanism for affecting its success.

5 Conclusion

A more detailed examination of the effectiveness of directed migration and mutation gradients show that there is not a clear benefit to using the techniques. The use of mutation gradients definitely reduced the ability for the GA to reliably find solutions. However, this may be due to the mutation gradient equation assigning a mutation rate of 0 to some of the populations. The poor performance of these populations might have reduced the overall ability of the GA to find a correct solution quickly. Something to investigate in the future would be an alternate formulation of the mutation gradient that assigned non-zero mutation rates to all populations.

In looking at the results, I have wondered if the way I formulated the graph coloring technique might have made the migration strategies less effective. In the chromosome representation, each color is represented as an integer (in this case between 0 and 19). When populations converged on solutions, these solutions used only a few of the 20 possible colors. The particular colors that were settled on tended to vary between different populations. For example, one population might develop a five-color solution using 17, 5, 15, 19, and 9. A different population might develop a solution using colors 1, 4, 8, 13, and 2. An individual traveling between these populations would not produce viable offspring through crossover because the "language" of the colors used was so different. The resulting organisms would inevitably have a much higher n_{colors} which would reduce their fitness. If this problem could be mitigated, it might improve the performance of the migration techniques. Alternatively, a different problem domain might be more suited to these techniques.

Perhaps more interesting than the lack-luster results are the techniques I developed for conducting a large number of trials and collecting the results. I developed a Linux shell script for starting the simulation repeatedly and recording the results from each run in a different file. This allowed the 1000 trials of each configuration to execute in an automated way. In the previous part of the report, the experiments were conducted only four times. This was because it was time consuming to run each simulation manually. I was also able to use command line utilities to analyze the results. For example, I searched for the string "FOUND" in the entire set of simulation results to extract the lines which corresponded to finding a correct solution. These results were further divided by searching by x and y coordinate to generate the surface plots seen in sub-section 4.2. It is apparent that such data-management techniques are very important to dealing with such a large amount of data in an efficient way.