Samantha Rack
Cloud Computing - A2

Part 1:

Approximately **12 comparisons** could be made in one minute (5 sec / comparison) when it was run locally (student machines ran about 50% more slowly).

With 250 sequences to compare to one another, these will be a total of **30,876 comparisons** required if no comparisons are repeated (n*(n-1)/2 where n = 249). Running this sequentially would require **42 hours and 53 minutes**. To finish the comparisons of *agamiae.small.fasta* in one hour, **43 machines** would be needed. For the complete data set with 100,000 sequences, there will be about 500 billion comparisons which would take about **79,274 years** for all the comparisons to be performed sequentially.

However, to simplify my approach to the problem, I allowed for repeated comparisons and comparisons of the same sequence against itself. So, the speedup calculations below are based on **62,500 comparisons** being performed, which would be completed sequentially in **86 hours and 49 minutes (312,500 seconds)**.

Part 2:

Top Ten Matches:

```
seq 1102140176186 matches seq 1102140143903 with a score of 1539
seq 1102140177183 matches seq 1102140177182 with a score of 874
seq 1102140172678 matches seq 1102140167168 with a score of 818
seq 1101555423227 matches seq 1102140171813 with a score of 777
seq 1102140164074 matches seq 1101555423815 with a score of 767
seq 1102140171192 matches seq 1101671610328 with a score of 764
seq 1102140171192 matches seq 1102140170611 with a score of 744
seq 1102140178449 matches seq 1102140171192 with a score of 742
seq 1102140171192 matches seq 1101555423803 with a score of 741
seq 1102140178846 matches seq 1101671610328 with a score of 736
```

Part 3:

Table 1. Results of running swalign tool on *agamiae.small.fasta* with a varying number of workers.

Number of Workers	Run time	Speedup	Efficiency
50	1 hr 41 min 40 sec (6100 sec)	51.23	102%
100	1 hr 1 min 46 sec (3706 sec)	84.32	84%
150	1 hr 11 min 28 sec (4288 sec)	72.88	49%
200	54 min 18 sec (3258 sec)	95.92	48%

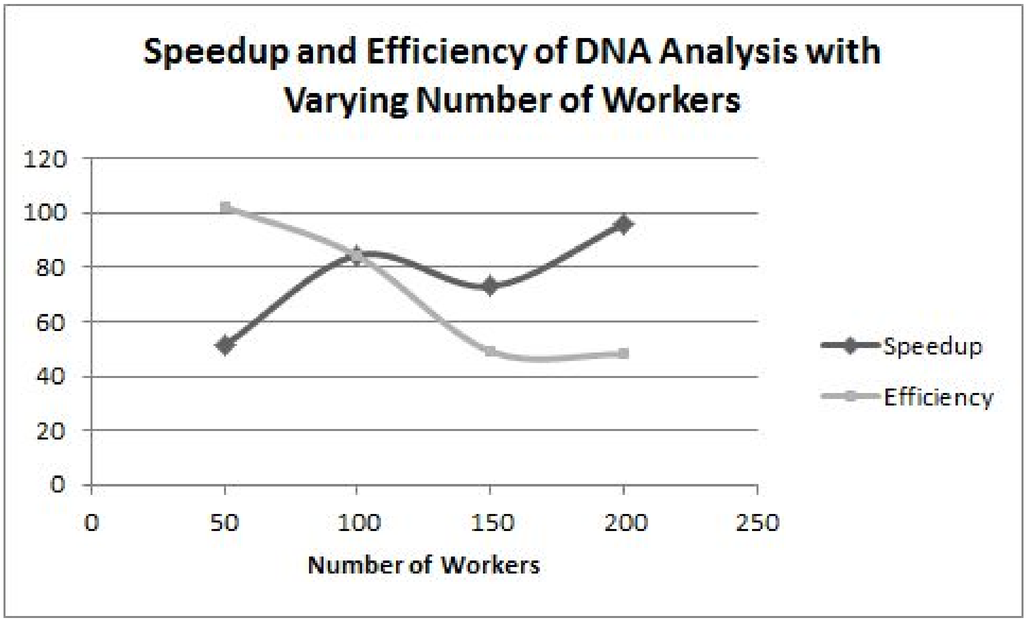

Figure 1. Graph of speedup and efficiency values calculated for the analyses using 50, 100, 150, and 200 workers.

The results displayed above are mostly what was anticipated. The speedup in general increases as the number of workers increases, but the rate of the increase reduces. This is evident with the efficiency values found for each run. The diminishing returns are expected; when a larger number of workers are employed the total run time is more affected by stragglers.

The one unexpected run was with 150 workers; the run time exceeded that of the run with 100 workers, and the speedup and efficiency were not what would be expected with the trend of the remaining runs. This likely is a result of Condor having a large number of jobs queued at that time as compared to the other runs, or the result of a long tail that caused the run to last much longer than it would on average.

The only other unanticipated result from these tests was the efficiency of greater than 100% for the run with 50 workers. This could be a result of better machines available via Condor than the local machine used for the sequential test which allowed the 50 workers to compute overall more quickly.