Jul 8 2016 Jonathan Arndt

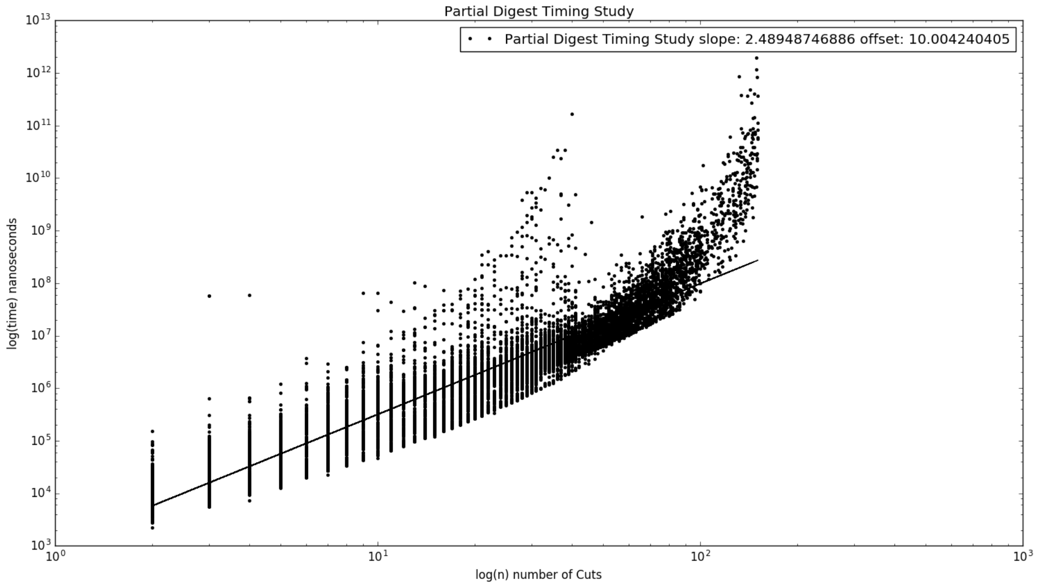
Assignment Two: Partial digest problem

Biologist often times would like to know how species and genes are related so experiments can be correlated. They can determine this by looking at the related genes in DNA. Experimentally, biologists can use enzymes to cut DNA at specific points where the enzymes will bind to proteins on the DNA. Unfortunately, though, even with filtering techniques, biologists struggle to determine the original placement of the cuts.

This is where the partial digest problem plays a big role. This algorithm sorts through all possible cuts and looks at relative sizes to determine if they are possible cuts. With a branch and bound algorithm this can be done quite efficiently. However, this isn't always the case, in some circumstances the algorithm can be very slow.

From my studies, the number of cuts isn't the only factor determining speed because I've noticed that problems with the same number of cuts and same max size result in drastic differences in timing.

By randomly running a large number of simulations, a guess at the average running time may be obtained.



This image is the loglog graph of the number of cuts with max size held constant at 1000 vs time to complete the algorithm in nanoseconds. As you can see from the chart, for lower values you can assume it is polynomial with only slight digression in the middle. A curve fitting algorithm determined a linear slope of 2.48948747, which means it’s just over an n^2 algorithm. This is good until you try to fit to the curve at the end which appears to be polynomial even on a loglog graph.