of the k-mers that the code will use, and the larger this value is, the less noisy the graph will be.

So the code begins by reading the file ‘EPECgenes.txt’ and storing the names of the genes in one list called ‘listofnames’ and the sequences of the promoter regions in another list, ‘listofstrings’. Then, two functions will be called, ‘makeATD()’ and ‘makeindexD’. The first of these utilizes another function, ‘calcAT()’, to get the AT-content for each successive kmer in a string and store them in a sequential list. This list, called the ‘ATlist’, is used to make the dictionary ‘ATD’, in which the key is each gene name in ‘listofnames’, and the value for each genes is the ‘ATlist’. The makeindexD() function also makes a dictionary ‘indexD’ in which the key is the name of each gene, but instead of lists of AT-content as values, they are instead lists called the ‘iList’ in which

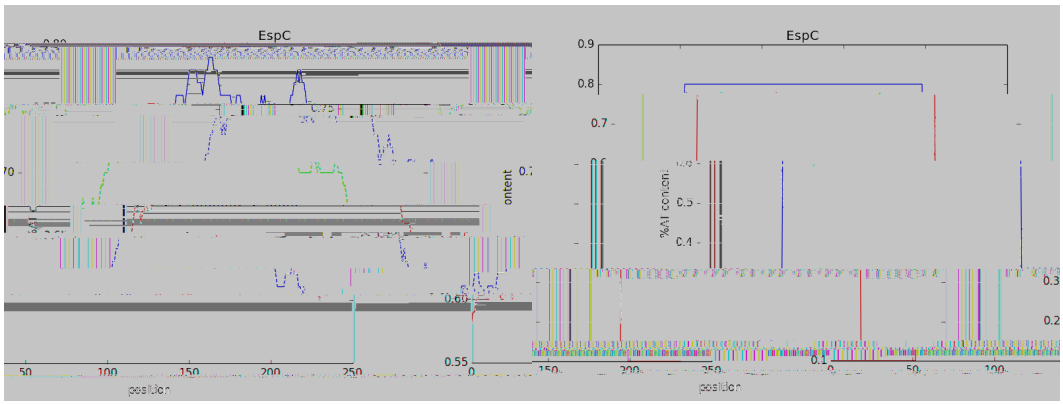

Figure 2. Graphs showing (left) the %AT-content by position in the promoter region of EspC, and (right) the same data, but with a peak only where the %AT exceeds a threshold value of 0.7. The one on the right is certainly more convincing, although it wasn't necessary in order to interpret the first graph. The presence of a peak indicates an AT-rich region of the sequence, suggesting direct regulation of this gene by Ler.

The 'easyAT()' function also made it easier to pick out the genes that do not have a significant AT-rich region, such as in eaeA (Fig 3).

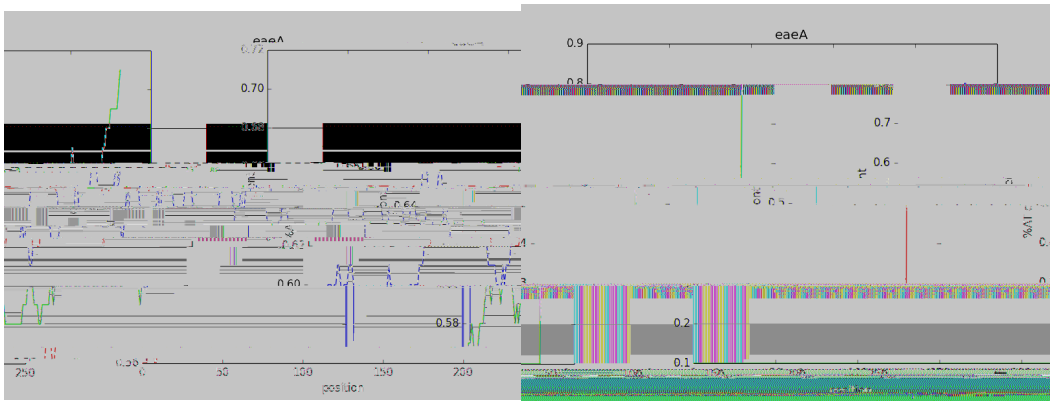

Figure 3. . Graphs showing (left) the %AT-content by position in the promoter region of EspC, and (right) the same data, but with a peak only where the %AT exceeds a threshold value of 0.7. . The left one is not easy to interpret, and it is not clear until one sees the right graph that there is no appreciable AT-rich region.

The main assumption that I made that may prove to be problematic is that the Ler binding site will be within the 300nt of the promoter. I based this off of the paper from which I got the original idea, however, there is no reason that it could not be farther. I also assumed that 0.7 was a reasonable threshold for AT-richness, but it is also possible that Ler may be able to bind with less than that. Finally, I assumed that the genes that showed the highest fold change in the microarray with the Ler-deletion strain would be those most likely to have Ler bind. However, the regulon of Ler is quite complex, and it is entirely possible that there are some that it binds to that simply have multiple regulatory inputs and are thus not showing a huge net change.