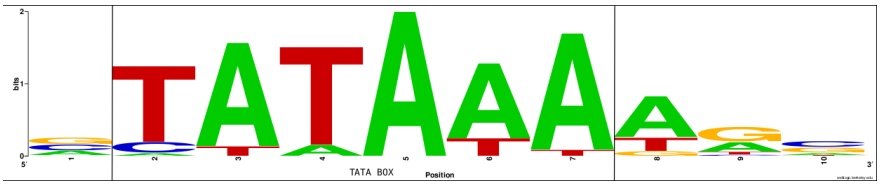
October 24, 2015



Motif Finding Within Genes Via Python Programming

by

Nathan Morse

Introduction

The python program included (morse_a4.py) investigates the DNA sequences upstream of a given microbial organism. Much of the program's analysis is based on finding a TATA-box, which simply consists of the nucleotides TATA followed by either an A or T, then an A, then an A or T, then either an A or G. The program finds various motifs through use of regular expressions within the upstream region of the gene prior to the TATA-box.

Methods

Among the 6 FASTA formatted test files provided, there is also a FASTA formatted file of the microbial genome of what is known as the *Sunflower chlorotic mottle virus*. If the program is run on this genome, the reader achieves extensive data on the upstream region prior to the first found TATA-box. The python program finds various motifs such as mirror repeats in addition to each motif's location within the upstream region and percentage of which it makes up. Both motifs and the TATA-box are found through use of regular expressions (see below):

The following python code is used to find the TATA-box. The highlighted portion indicates the regex used to define TATA (TATA followed by either an A or T, then an A, then an A or T, then either an A or G). The brackets [] indicate the option to have only one of the letters within:

```
TATAregex = re.compile(r"tata[at]a[at][ag]")
```

The next portion of python code is what finds the motif for mirror-repeated motif that are 6 characters in total length (6 base pairs in this case). Theoretically speaking, mirror-repeated motifs of 6 characters would be similar to a palindrome if working with English words, such as the noun "Hannah":

```
MRregex = re.compile(r"(.)(.)(.)\3\2\1")
```

"(.)" indicates any given character and says "remember this character", while the "\3\2\1" recalls the three found characters backwards, thus making a mirror image.

Results and Discussion

When the python program runs on the microbial genome provided, the results included the location of the TATA-box in accordance with the upstream and downstream regions. The total length of the genome was 9965 base pairs and the upstream region alone was 723 base pairs.

The upstream region starts at base pair 1 and goes up to base pair 723

The TATA-box is found at base pair 724 and goes up to base pair 731

The downstream region starts at base pair 732 and up to base pair 9965

Many direct repeat, mirror repeat, and AT/TA/AA/TT motif runs were found in the upstream region. The direct repeat motifs made up 27.9% of the upstream region while the mirror repeat motifs made up only 10% of the upstream region. AT/TA/AA/TT motif runs made up a large 40.7% of the upstream region.