Ben Siranosian
February 8, 2014
CSCI1820 hw2

Problem 1

1) Transcription factors are cellular proteins that bind to specific DNA sequences, typically upstream of a gene, and control the rate of transcription from DNA to mRNA. Transcription factors work alone or as members of a complex. When bound to DNA, they can enhance or reduce the binding of RNA polymerase, therefore increasing or decreasing transcription of the sequence.
2) Proposed by Francis Crick in 1966, the wobble hypothesis states that the last base in a codon (the 3' base on mRNA or 5' base on tRNA) is more likely than the other two to undergo non-standard base pairing. The wobble hypothesis accounts for the fact that most organisms don't encode tRNAs for every possible codon: non-standard binding enables every codon to be translated with a smaller set of tRNAs.
3) Primary structure: the sequence of amino acids that make up the polypeptide chain.
   Secondary Structure: the folding of the amino acid chain into alpha helix and beta sheet structures.
   Tertiary structure: the globular, three-dimensional structure after secondary structures have folded.
   Quaternary structure: the three-dimensional structure of multiple protein subunits coming together to make a finalized protein.
4) The central dogma of biology explains the way proteins are made from the genetic code. Genetic information starts in DNA, is transcribed into mRNA, and then translated into a protein.

Problem 2

Similarity Matrix 1

	A	**T**	**C**	**G**
A	10	-10	-10	0
T	-10	10	0	-10
C	-10	0	10	-10
G	0	-10	-10	10

Similarity Matrix 2

	A	**T**	**C**	**G**
A	10	0	0	0
T	0	10	-10	-10
C	0	-10	10	-10
G	0	-10	-10	10

Similarity matrix 1 prioritizes matches over mismatches but reflects the idea "purines and pyrimidines are more similar to each other than members of the other class." This has the effect of scoring alignments where mismatched bases of the same class line up higher than alignments where mismatched bases of opposing classes line up.
Similarity matrix 2 prioritizes matches over mismatches but reflects the idea "A frequently mutates into one of the other three bases." With this matrix, a mismatch with an A will have no effect on the score.

Problem 3

1) The algorithm runs in O(m * log(n)) time.

2) A[6] = 8 after the code runs with N=5 and M=6.

Problem 4

Code for the warmup problem is found in *repeated_substrings.py*
Code for the inexact repeated substrings problem is found in *repeated_substrings_mismatches.py*
README.txt contains instructions on how to run each program.

The inexact repeated substrings of 'AATTCAAT' of length at least 2 are:
AAAT,AGT,CCA,TAT,TTAA,TAA,GT,GA,ACT,TTT,TTA,TCC,ATC,ATT,ATTT,CC,CA,CT,AATC,AAG,AAA,AAC,CAT,
AAT,TT,GAT,TG,TC,TA,AA,AC,AG,AT,CATT,TCCA

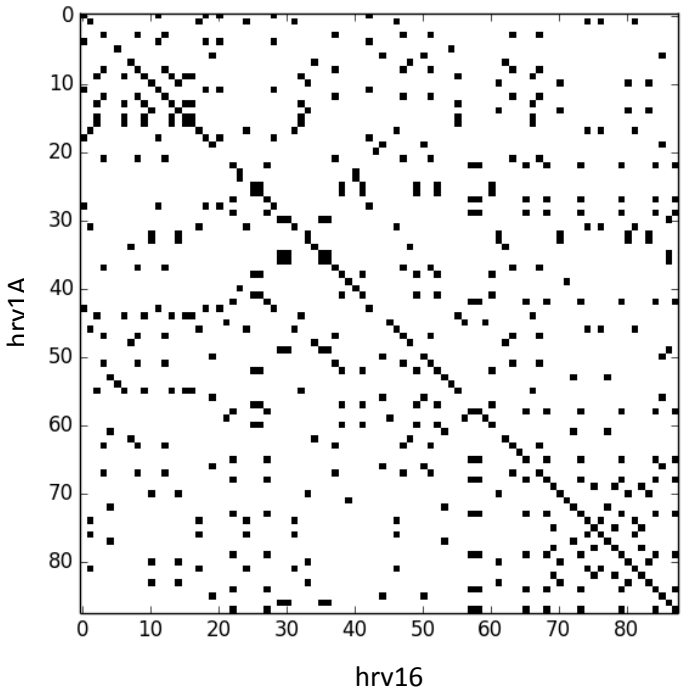
Problem 5

The Needleman-Wunsch algorithm is used to conduct global alignment between two sequences. The global alignment between the two capsid proteins will highlight segments that have remained constant through evolutionary time and place gaps in regions that could represent insertions or deletions.
The Smith-Waterman algorithm is used for local alignment between two sequences. The local alignment between the two capsid proteins would highlight the short regions that are *most* similar, but wouldn't be best for seeing overall evolutionary trends.

Problem 6

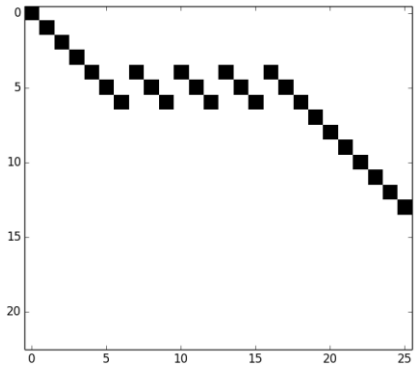
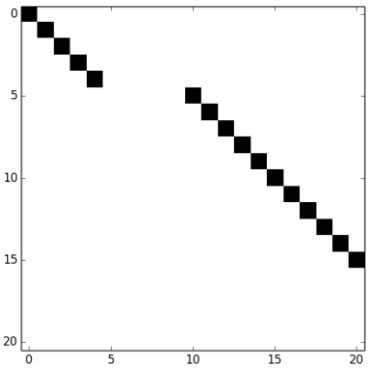
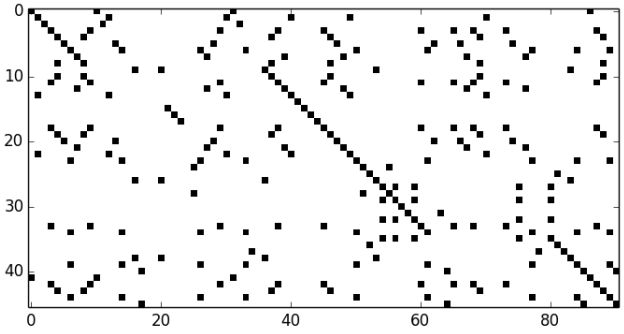
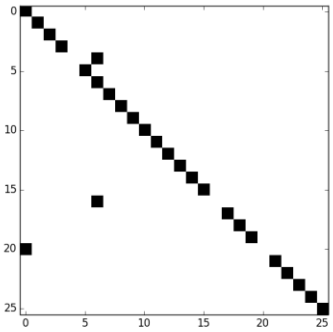
The code for producing dot-plots can be found in dot_plot.py

Dot-plot of the two sequences:



Ben Siranosian
February 8, 2014
CSCI1820 hw2

Example plots:

Repeats	Indels
	
Exons/Introns	**SNPs**
	

Problem 7

There is only one optimal global alignment of the two sequences. The edit graph below shows this.

	A	**C**	**G**	**T**
A	1	0	0	0
G	0	1	2	1
T	0	1	1	3

Optimal Global Alignment:

```
A C G T
|   | |
A - G T
```