Assignment 7: Motif Finding

Your colleague Dr. Glenn calls you in a panic. She has been studying the action of a transcription factor (TF) on two differentially transcribed promoters in a new strain of bacteria she isolated last year. The postdoc who was working on the project quit and she's discovered that he didn't finish writing the code. Knowing your Python expertise, she is hoping you can help her figure out what he's been doing and continue the coding part of the project for her. After reassuring her that you can help out, she explains where the project is stuck. Dr. Glenn has created a scoring matrix (`polymerase_scoring_matrix.txt`) for the polymerase that transcribes her promoters. She also discovered a homolog of the TF she thinks regulates the region. Its scoring matrix (`tf_score_matrix.txt`) is also provided to you. In the scoring matrices, the columns are positions and the rows correspond to A, C, G, and T.

All of the files for this assignment are located in `/home/assignments/assignment7`.

Question 1
For both the TF and the polymerase matrices, determine the highest affinity binding site. You may find the highest affinity binding site from inspection or write a script. What are the scores for these sites? What sequence(s) produce these scores? Explain how you arrived at your answer.

Dr. Glenn's postdoc was supposed to be writing a Python script to scan the two promoters (`promoter1.txt` and `promoter2.txt`) to find native TF and polymerase binding sites to test in binding experiments. Dr. Glenn has given you the postdoc's unfinished code (`scan_sequence.py`).

The usage of `scan_sequence.py` is:

```
$ python3 scan_sequence.py <scoring_matrix> <sequence_file>
```

Run `scan_sequence.py` with one of the scoring matricies and promoter sequences to see what kind of output the script gives.

The postdoc did not write many comments. Add comments for the code blocks. (These areas have been annotated with "TODO: explain what this code does.") Additionally, document the user-defined functions by adding docstrings for the functions. (These areas have been annotated with "TODO: write function docstring.")

Astutely, you notice that the postdoc's code only scans the input sequence in the forward direction. Modify `scan_sequence.py` so that it also scans the reverse complement of the input sequence.

You also realize it would be helpful to have the program only report hits that are a good match to the input scoring matrix. After talking it over with Dr. Glenn, you decide that a score threshold of 40 for the TF and 45 for the polymerase should work well. Modify `scan_sequence.py` so that only matches above the threshold are printed. Add a command line argument for the score threshold. The suggested usage will now be:

```
$ python3 scan_sequence.py <scoring_matrix> <sequence_file> <score_threshold>
```

TIP: Remember to update your script's usage and command line argument checking!

Since Dr. Glenn needs the sequence of the putative binding sites as well, modify `scan_sequence.py` to print out the sequence of each hit as well as the score and position. The output should be tab-delimited with the following columns:

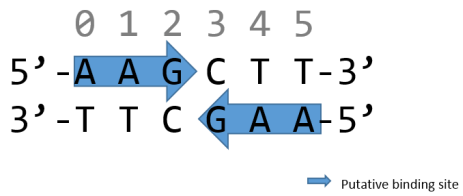
Column	Description	Example
0	Strand the putative binding site was found on (forward or reverse)	forward
1	Sequence of the putative binding site. Use the strand that the putative binding site was found on.	GGTCAG
2	Leftmost position of the putative binding site (0-based and **relative to the forward strand**)	15
3	Score of the putative binding site	43.6

Here's an example of what this would look like:

Consensus sequence

A A G

Schematic of putative binding sites



Output table

strand	sequence	position	score
forward	AAG	0	12.8
reverse	AAG	3	15.6

Scan both promoters with both matrices using the appropriate thresholds given above. Paste the commands to do this and their output in your README.

Question 2
Knowing that the genes are differentially expressed, which promoter would you expect to be repressed by this TF? Which would be activated? Why?

What to turn in

- Your modified and commented `scan_sequence.py`
- A `README.txt` with the answers to the questions and the commands you used to answer the questions