Assignment 6: Motif Finding
Input Files

- Scoring matrix for polymerase binding site
  - polymerase_score_matrix.txt
  - Provided in /home/assignments/assignment6

- Scoring matrix for transcription factor binding site
  - tf_score_matrix.txt
  - Provided in /home/assignments/assignment6

- Format:
  - Rows are A, C, G, T (in that order)
  - Columns are scores (weights) for each base at that position in binding motif
Input Files

- Two promoter sequences
  - promoter1.txt, promoter2.txt
  - Provided in /home/assignments/assignment6
Step #1: Finding Highest Affinity Motif

- Determine highest affinity binding motif(s) based on scoring matrices
  - Can calculate by hand or with a script
  - Report sequence(s), score(s)
  - If calculated by hand, briefly explain method for calculation in your README
  - If calculated with script, provide command to run script in your README
Step #2: Test Run of Script, Commenting Script

- Run unmodified `scan_sequence.py` script to see output
  - Provided in `/home/assignments/assignment5`
  - Usage: `python3 scan_sequence.py <scoring_matrix> <sequence_file>`
  - Choose one of the scoring matrices, one of the promoters

- Add docstrings to functions
  - Marked by ‘TODO: write function docstring’

- Provide comments for code
  - Marked by ‘TODO: explain what this code does’
Step #3: Modify Script

- Search reverse complement of input promoter sequence
  - Reverse complement function already written, call it on input sequence
  - Scoring function already written, call it on reverse complement sequence

- Add argument for score threshold
  - Update usage statement in docstring
  - Update code for checking correct number of arguments

- Only report sequences with scores above threshold
  - Report strand (forward or reverse), binding site sequence, leftmost position of binding site, score
  - Output should be tab-delimited

- Don’t forget to comment code that you add
How to Report Position

- Leftmost position, 0-based, relative to the forward strand
  - Example on page 2 of assignment

- Another example:

<table>
<thead>
<tr>
<th>Forward</th>
<th>Reverse</th>
</tr>
</thead>
<tbody>
<tr>
<td>5'</td>
<td>3'</td>
</tr>
<tr>
<td>0</td>
<td>49</td>
</tr>
<tr>
<td>AGGTACCCAT</td>
<td>TCCATGGGTA</td>
</tr>
<tr>
<td>ATCGGTAACT</td>
<td>TAGCCATTGA</td>
</tr>
</tbody>
</table>

- What you should report:
  - forward, AGGTACCCAT, 0, score
  - reverse, AGTTACCGAT, 40, score
Step #4: Run Modified Script

- Run modified `scan_sequence.py`
  - Run using each matrix on each promoter
  - For `tf_score_matrix.txt`, use threshold of 40
  - For `polymerase_score_matrix.txt`, use threshold of 45
What to Turn In

- Modified, commented `scan_sequence.py`
- Completed `README.txt`