Hands-On Thirteen
Detecting Motifs

Problem 1

We will explore the TATA-box position weight matrix (PWM) at JASPAR:

- Go to http://jaspar.binf.ku.dk/
- If you get a "Welcome" message, click on "OK".
- Scroll down to "Browse sorted by" and click on "Structural class".
- Click on "Browse" to the right of that window to go to the "JASPAR matrix models:" page.
- Go to "TATA-binding" under "family" with "ID" equal to "MA0108.2".
- Click on the logo graph once to get a new page: "Summary page for ID: MA0108.2".

a) Explain how the sequence logo depicted under "SEQUENCE LOGO" represents the information in the table under "FREQUENCY MATRIX".

b) Explain in your own words how the table under "FREQUENCY MARIX" was constructed.

- Click on "Reverse complement" to get a new logo and a new table.

c) Explain the relationship between the new logo and table and the previous logo and table.

Problem 2

Return to the JASPAR page.

- Scroll down to "Browse sorted by" and click on "ID".
- Click on "Browse" to the right of that window to go to the "JASPAR matrix models:" page. You will get the same page as above in Problem 1, but the entries are now sorted according to "ID" and not "class".
- Scroll down and choose the entry with ID number MA0040.1, Click on the logo graph once to get a new page: "Summary page for ID: MA0040.1".
- Repeat the above process twice to obtain the sequence logos of MA0041.1 and MA0047.2.

By looking at the sequence logos, explain which of these three transcription factors in rat (and mouse) is most likely to bind to DNA containing the motif TGTTTA.

Problem 3

We are going to use the spreadsheet pwm.xls to compute the total TATA PWM score of the following three sequences, and determine which one is most likely to be a true TATA box:

i) ATATATATAGGCTGG ii) CTATATATATGCTGG iii) CTATAAATAGGCCGG

To use "pwm.xls" with the first sequence:

- Copy "pwm.xls" and save it on your computer.
- Go to the file and double click on the "pwm.xls" icon.
- Enter "ATATATATAGGCTGG" in Excel cell 24A, and hit the return key.
- Read the total PWM score for that sequence in Excel cell 40Q and also in Excel cell 41Q since the TATA box might be on the other strand (the reverse complement of ATATATATAGGCTGG).

Repeat the above steps for the remaining two sequences and compare the three results to determine which sequences is most likely to be a true TATA box.

a) Results of using the TATA box PWM matrix:

First sequence: ATATATATAGGCTGG produces a total PWM score of _____.

Second sequence: CTATATATATGCTGG produces a total PWM score of _____.

Third sequence: CTATAAATAGGCCGG produces a total PWM score of _____.

b) Because it has the highest score, the _____ sequence is the most likely to be a true TATA box.

Problem 4

Use the spreadsheet pwm.xls to compute the total TATA PWM score of CCGGCCTATTTATAG. Explain why the score is so high, even though this sequence does not look like a true TATA box.

Problem 5

We are going to see how the different tables in the spreadsheet were constructed. Note that the table in Excel cells 10A, 11A, 12A, 13A to 10P, 11P, 12P, 13P is identical to the FREQUENCY MATRIX - MA0108.2 of Problem 1.

a) Explain how the table that is found in cells 16A, 17A, 18A, 19A to 16P, 17P, 18P, 19P, of Excel, was constructed.

b) Explain how the table that is found in cells 35A, 36A, 37A, 38A to 35P, 36P, 37P, 38P, of Excel, was constructed.