Integrative Molecular and Cellular Bioengineering, SPRING '07
16:125:584:01

<u>Assignment 1: Due Tuesday, January 30, 2007</u>

1. Find a) the mRNA sequence and b) the protein sequence (in the latter case, one-letter abbreviation is sufficient) for the human stat3 gene.

2. a) based on purely probabilistic considerations, do you expect the following sequence to have an exact match in the human genome?
b) Analyze the homologies of the following sequence with the human genome.

5'- gctgactaag catcgatcgaa -3'

3. Below is a position frequency matrix (PFM) for the transcription factor HNF1. Using this information, determine

a) the position weight matrix (PWM)
b) the consensus sequence
c) the site score for the following target sequence (which is an actual site for the humain HNF1 target on the albumin gene): 5'- agttaataatctaca-3'
d) (Bonus) Write a code to generate 1000 random sequences of length 15, calculate their site scores against this HNF1 matrix, determine the mean and median scores, and determine the fraction that score higher than the actual sequence from part c. How do you interpret the result?

Note: Use 0.375 as the "pseudocount" for all bases. Also, assume that the background probabilities are all equal to 0.25.

PO	A	C	G	T
1	6	1	16	3
2	2	0	24	0
3	1	0	0	25
4	2	0	0	24
5	25	0	0	1
6	20	2	1	3
7	2	0	0	24
8	8	3	8	7
9	16	0	1	9
10	3	1	1	21
11	1	4	0	21
12	16	1	6	3
13	9	11	3	3
14	7	14	0	5
15	11	7	4	3