SGN-6156 Computational Systems Biology II

Exercise 3, April 23, 2008, at 12:15-13:45, in class TC415

This exercise will familiarize you with multiple sequence alignment algorithms and transcription factor binding site prediction methods using Matlab. Exercises can be done in class (during the exercise session).

1. Assume a transcription factor X has been measured to bind a set of 10 sequences: 'AGTAGCCA', 'CGTTCCTACA', 'GTTGGTACC', 'GTTGCCA', 'TGTCGCCATG', 'CGTTGTCAT', 'AGTTACCA', 'GTTAGCACA', 'GTTTTTATG', 'GGTTGGTA'.
   (a) Use a multiple sequence alignment algorithm to align these 10 sequences. Try e.g. the multiple sequence alignment algorithm function 'multialign' that is implemented in Matlab's Bioinformatics toolbox.
   (b) Identify the core 7 residue part of the alignment that can be aligned without gaps. Use e.g. an alignment viewer 'multialignviewer'.
   (c) Display the consensus sequence and sequence logos for the 7 residue multiple alignment (Matlab functions 'seqconsensus' and 'seqlogo'). This should correspond roughly to the first 7 residues of the RFX1 binding site shown during the lectures.
   (d) Convert this 7-long multiple alignment into a position specific frequency matrix $f_{b,i}$, where $b \in \{A, C, G, T\}$ and $i = 1, \ldots, 7$, using the maximum-likelihood method (i.e., normalized counts for each column of the matrix). Note that the ith column of $f_{b,i}$ represents the probability of seeing any of the residues in the i position of the binding site.
2. Analyze the promoter sequences of 'M22326' and 'X04724' genes for putative locations of RFX1 binding site.

   First, download the following m-files 'PSWM_MotifLocator.m', 'PSWM_Scan.m', 'BM_Scan.m' and 'basepairs2num.m' from
   http://www.cs.tut.fi/∼harrila/teaching/CSBII2008/.

   Promoter sequences are shown on the next page. (You should be able to copy and paste those into Matlab.) Define the sequences as char vectors/strings and convert them into an integer representation using 'basepairs2num'. Then use 'PSWM_MotifLocator' to find possible binding sites on both sequences (try W = PSWM_MotifLocator(S,F,0.25*ones(1,4),0,[],[]); where S is your sequence in integer representation and F is your $f_{b,i}$ matrix). You can assume that in the background model all residues {A,C,G,T} are equally likely, i.e., occur with probability 0.25.

   Which one of the sequences, 'M22326' and 'X04724', is more likely to contain a binding site for RFX1? What is the most likely location of the binding site?

- M22326
  CGCGGGCGTCCCCGACTCCCGCGCGCGCTC
  AGGCTCCCAGTTGGGAACCAAGGAGGGGGA
  GGATGGGGGGGGGGGGTGTGCGCCGACCCG
  GAAACGCCATATAAGGAGCAGGAAGGATCC
  CCCGCCGGAACAGACCTTATTTGGGCAGCG
  CCTTATATGGAGTGGCCCAATATGGCCCTG
  CCGCTTCCGGCTCTGGGAGGAGGGGCGAGC
  GGGGGTTGGGGCGGGGGCAAGCTGGGAACT
  CCAGGCGCCTGGCCCGGGAGGCCACTGCTG
  CTGTTCCAATACTAGGCTTTCCAGGAGCCT
  GAGCGCTCGCGATGCCGGAGCGGGTCGCAG
  GGTGGAGGTGCCCACCACTCTTGGATGGGA
  GGGCTTCACGTCACTCCGGGTCCTCCCGGC
  CGGTCCTTCCATATTAGGGCTTCCTGCTTC
  CCATATATGGCCATGTACGTCACGGCGGAG
  GCGGGCCCGTGCTGTTCCAGACCCTTGAAA
  TAGAGGCCGATTCGGGGAGTC

- X04724
  ACTGGGTCCCCACTACCTTTATAGACCAAA
  GCACCTCCTCTCTGCCCCCTGGACTTTGCT
  GTTTGACCCATTAAGGGCTCCAGGTGGGGT
  AGGTCAGCAGATGGCCAGAGGGGCTGAAGC
  TGCAGTTTCCAAACACTTCCCTGGTGCTAG
  GTCTGCAGAAAGCGCTCATTGGACGTCAAC
  ACCTCTACTTAGTCCTAGGGTAATTAGAGT
  CTTAACAAGGGGCCCTGATGGCCTGATGAA
  CCAGTTCACAGCAGGGGACATTGTTCCAGA
  GTGGGTGATGCTTTCTTTGTCCTTGGCTGA
  GCATTTTTCCACATCATTCCCCAGGAAGTT
  GCAGTTGGGGACCTAGTATCTTTGTCCCTT
  GGACTGTTTCACAGTTCCAATTGATAGCTG
  GGTTCTTAACTCAGCCGAGTCCCCAGCTCT
  CTCTCAGAGGTAGAAGGAAAGCAGAATTCA
  GGCAGCAAGGCACTTAATGGTCCCTCCTCC
  TCTATCTCTCTTTCCATATC