Bioinformatics Research Center
Aarhus University

Advanced Statistical Methods in Bioinformatics
Asger Hobolth
August 28, 2018

Handout 2

Lectures in Week 36

Monday, September 3 and Wednesday, September 5 (in room 1110-214):

Finite Markov Chain Imbedding (FMCI)
I will cover selected parts of the paper:
Fu and Koutras (1994). Distribution theory of runs: A Markov chain approach. *Journal of the American Statistical Association*, **427**, 1050–1058.
I will start by discussing why the paper is of interest to us.

Lectures in Week 37

Hidden Markov Models and the EM-algorithm.

Exercises in Week 36

1. Write a computer program in R where you reproduce the numbers for $N_{5,2}$ in Table 1 in Fu and Koutras (1994).

2. Write a computer program in R where you reproduce the numbers for $N_{15,2}$ in Table 2 in Fu and Koutras (1994).

3. Write a computer program in R that simulates a homogeneous finite state Markov chain.
   Input should be an initial distribution, a transition matrix, and the length of the desired chain.
   Output should be a simulation from the Markov chain with the desired length.

4. We are interested in the word 'CTT' in a number of DNA sequences. The null model is that letters A,C,G,T are independent and occur with frequencies (0.3,0.2,0.2,0.3).
   Write a computer program in R for calculating the probability for $0, 1, 2, \geq 3$ number of occurences of the word 'CTT' in a DNA sequence of arbitrary length.
   Hint: Use similar ideas as in Fu and Koutras (1994) to define an appropriate Markov chain.
   Verify your calculations by a simulation study. (E.g. use the computer program from the previous exercise.)