Algorithms for Bioinformatics (Autumn 2015)

Exercise 3 (Tue 22.09., 10-12, B222)

If you cannot make it to the exercise session, please e-mail your solutions and the reason why you cannot attend to daniel.valenzuela@cs.helsinki.fi before the exercise session to get credit.

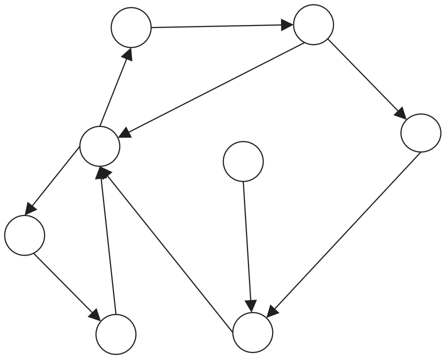
Some of the problems below are programming exercises on the Rosalind platform at `http://rosalind.info/problems/list-view/?location=bioinformatics-textbook-track`

1. Solve the Rosalind problem NEWBA3C: *Construct the Overlap Graph of a Collection of k-mers.*
   Note: There seems to be a problem with the Rosalind datasets. Some example datasets and results will be available on the course homepage.

2. Solve the Rosalind problem NEWBA3D: *Construct the De Bruijn Graph of a String.*
   Note: There seems to be a problem with the Rosalind datasets. Some example datasets and results will be available on the course homepage.

3. A de Bruijn sequence of order k is a string that contains every possible k-mer exactly once. For example, ABAACBBCCA is a de Bruijn sequence of order 2 over the alphabet {A,B,C}. Construct a de Bruijn sequence of order 3 over the alphabet {A,B,C}.
   Hint: The original use of de Bruijn graphs was for constructing de Bruijn sequences.

4. Insert and delete minimum number of edges to/from the graph below so that it has an Eulerian path.



5. An experiment trying to determine the 3-mer composition $\text{Composition}_3(\textit{Text})$ of a string *Text* has returned the answer {`GAG`,`GAT`,`TAG`,`ATA`,`ATA`,`AGA`,`TAC`}, but one 3-mer seems to be missing. Reconstruct *Text* by the Eulerian path approach, taking into account the missing 3-mer.
   Hint: Note that because the composition contains the 3-mer `ATA` twice, the graph should have two edges from `AT` to `TA`.