Problem Set 4: Assembly

Handed out Monday, March 23. Due at the start of class Friday, March 27.

Homework Information: Please upload a PDF of your solutions to Moodle by 2pm central time. If you write your solutions by hand, use an app like Adobe Scan to take a picture of it and turn it into a PDF.

1. (4 pts) In a Sequencing By Hybridization study the following spectrum of 3-mers is collected: $S = \{$'aca', 'att', 'cat', 'ctg', 'ctt', 'gct', 'tct', 'tga', 'tgc', 'ttc', 'ttg', 'ttt'$\}$

(a) Draw the Hamiltonian (overlap) graph for S.

(b) Draw the Eulerian (De Bruijn) graph for S.

(c) Give all possible DNA sequences whose spectrum is S. (Hint: Find all Eulerian paths in the De Bruijn graph.)

2. (4 pts) A shortest superstring is the minimum length string that contains, as a substring, all strings from a given input set. A k-mer is a k-length substring (k consecutive elements) of a larger string. In the following questions "digit" refers to an alphabet of the base 10 integers $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$.

(a.) What is the minimum length of the shortest superstring of a set of n, 2-digit integers?

(b.) What is the minimum length of the shortest superstring of a set of n, 3-digit integers?

(c.) What is the maximum number of unique k-mers in an n-digit string?

3. (2 pts) Given the following genome: CATACCGCATAC and let $k = 5$.

(a) List all k-mers of the genome.

(b) Draw the De Bruijn graph.