Computational Biology: Assignment #6

Due on Monday, Apr 28, 2014

Jianyang Zeng 1:30pm

Weiyi Chen

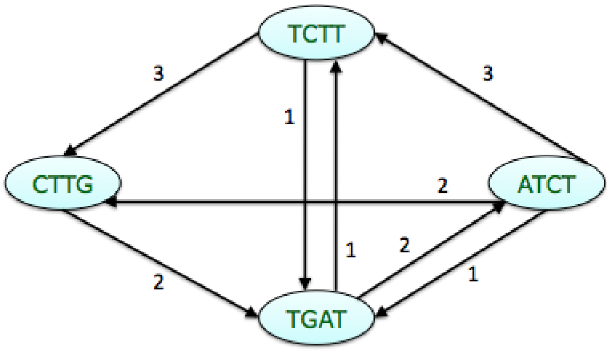
Problem 1

Overlap Graphs and Sequence Assembly

(a)

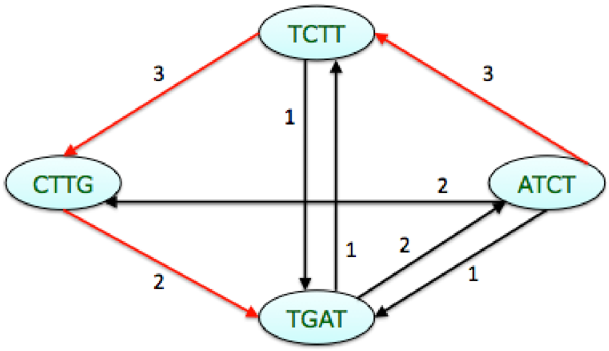
The assembled sequence computed by greedy algorithm is ATCTTGAT or TGATCTTG.

(b)



(c)

One Hamiltonian path is colored red as follows.



The corresponded assembled sequence is ATCTTGAT.

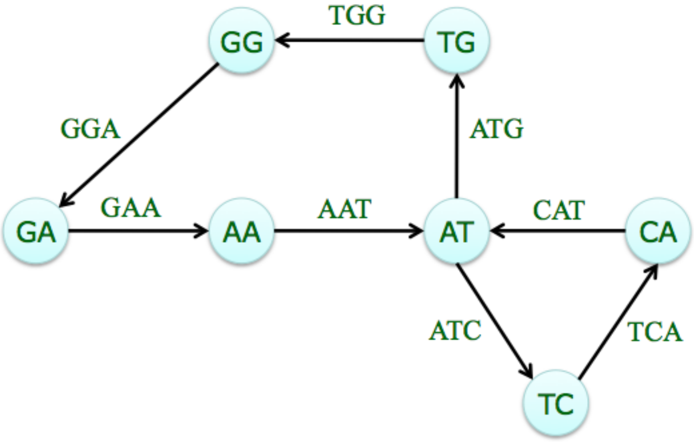
Problem 2

Eulerian Graphs and Sequence Assembly

Problem 2 continued on next page. . .

(a)

The de Bruijn graph:



(b)

Since all vertices are balanced, we can start at any vertex. So there are 8 possible Eulerian paths, since there are 7 vertices and 2 possible start path for vertex AD. All Eulerian paths are listed as follows.

- Eulerian path: AT-TG-GG-GA-AA-AT-TC-CA-AT
  Assembled sequence: ATGGAATCAT
- Eulerian path: AT-TC-CA-AT-TG-GG-GA-AA-AT
  Assembled sequence: ATCATGGAAT
- Eulerian path: TC-CA-AT-TG-GG-GA-AA-AT-TC
  Assembled sequence: TCATGGAATC
- Eulerian path: CA-AT-TG-GG-GA-AA-AT-TC-CA
  Assembled sequence: CATGGAATCA
- Eulerian path: TG-GG-GA-AA-AT-TC-CA-AT-TG
  Assembled sequence: TGGAATCATG
- Eulerian path: GG-GA-AA-AT-TC-CA-AT-TG-GG
  Assembled sequence: GGAATCATGG
- Eulerian path: GA-AA-AT-TC-CA-AT-TG-GG-GA
  Assembled sequence: GAATCATGGA
- Eulerian path: AA-AT-TC-CA-AT-TG-GG-GA-AA
  Assembled sequence: AATCATGGAA

Problem 3

Sorry I didn't work out a solution to improve further though I worked hard to read some paper like Linear Approximation of Shortest Superstrings, which only achieve a constant factor approximation, proving an upper bound of 4.
If possible, I hope you could share us the paper with upper bound of 3.5 stated. I feel interested in reading it.