

22BIO211: Intelligence of Biological Systems - 2

Lab Sheet 2

1. Write a python program to create a directed overlap graph for the given K-mer set .

{CAT,ATG,TTC,TCC,TGC,GCA,ATT,CCA}

Also visualize the graph using python libraries.

2. Construct a weighted and directed overlap graph of all 6-mers from “GTACGTACGAT” where edges are overlaps of length ≥ 4 . Visualize the graph using ‘Networkx’.

Steps:

- Construct all k-mers
 - Represent each k-mer as a node
 - Connect two nodes with an edge if they overlap
 - Assign overlap length as the weight
3. Solve the following Rosalind Challenge : <https://rosalind.info/problems/ba3c/> and upload the ‘strand’ page.

Overlap Graph Problem

Construct the overlap graph of a collection of k-mers.

Given an arbitrary collection of k-mers Patterns, we form a **graph** having a node for each k-mer in Patterns and connect k-mers Pattern and Pattern' by a directed edge if $\text{Suffix}(\text{Pattern})$ is equal to $\text{Prefix}(\text{Pattern}')$. The resulting graph is called the overlap graph on these k-mers, denoted $\text{Overlap}(\text{Patterns})$.

we use the terms prefix and suffix to refer to the first $k - 1$ nucleotides and last $k - 1$ nucleotides of a k-mer, respectively.

Given: A collection Patterns of k-mers.

Return: The overlap graph $\text{Overlap}(\text{Patterns})$, in the form of an **adjacency list**.

Sample Dataset

ATGCG

GCATG

CATGC

AGGCA

GGCAT

Sample Output

AGGCA -> GGCAT

CATGC -> ATGCG

GCATG -> CATGC

GGCAT -> GCATG

4.