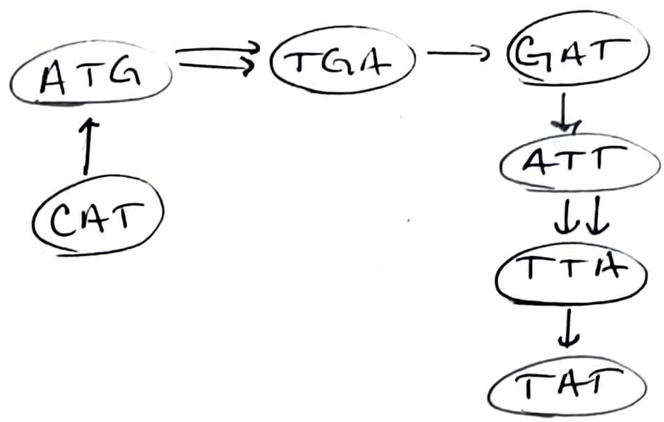
1. Construct a De Bruijn Graph for the following set of reads with $k = 4$:
$R = \{ATGAT, GATTA, ATTAT, CATGA\}$. Make sure to clearly indicate edge multiplicities.




CATGA
ATGAT
GATTA
ATTAT

* Every read is a subpath in the graph.

What do you think the original sequence was?
**CATGATTAT**

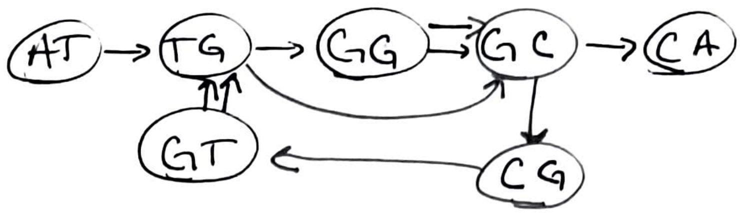
2. Sequencing coverage (the number of times each base was sequenced) can affect what a de Bruijn graph looks like. Suppose you have "perfect sequencing" data (no errors and reads uniformly cover the genome) and have an average sequencing coverage of $c$. What effect does this have on the resulting de Bruijn graph?

- **Every edge will have a multiplicity that is a multiple of c.**
- **Ex. If c = 30, edge multiplicities are 30, 60, 90. Why?**
- **Ex. 30 for a k-mer that appears once, 60 for twice, etc.**
- **Non uniform data will cause variation in multiplicities.**

3. Sequencing coverage (the number of times each base was sequenced) can affect what a de Bruijn graph looks like. Suppose you have gaps in your sequencing - that is portions of the genome that are not covered. What effect does this have on the resulting de Bruijn graph?

- **Graph may become disconnected.**
- **If reads have overlap of $< k - 2$, will also become disconnected.**

4. Eulerization is the process of turning a graph (or a multi-graph) into a Eulerian graph (one that contains an Eulerian Path). Construct the De Bruijn graph for the following set of sequences with $k = 3$. $R = \{ATGGC, GTGCA, GGCGTG\}$. (The fact that this set of sequences has different lengths should not be of a concern to you in this problem).




What is the fewest number of edge additions or removals to make the resulting graph Eulerian?
**2**

Recall: a connected, directed graph G has a Eulerian path if and only if it contains at most two semi-balanced vertices and all other vertices are balanced.
TG and GC are semi-balanced, but must be balanced.

5. What effects can choosing smaller or larger values of $k$ have upon the resulting de Bruijn graph?
**Smaller:**

- **Fewer vertices (limit is $4^{k-1}$). Also, a limit on number of "unique" edges. (Pro)**
- **Graph can become more interconnected. Fewer unique paths. (Con)**

**Larger**

- **We can span repeats. (Pro)**
- **Graph can become disconnected. (Con)**

6. Can you think of any graph simplification that would be helpful for being able to find Eulerian paths (or a set of such paths) in a de Bruijn graph?

- **Collapse paths**
- **Unzip edges**

**Summary - Use of Eulerian Path to do assembly is appealing (fast), but has lots of practical issues.**