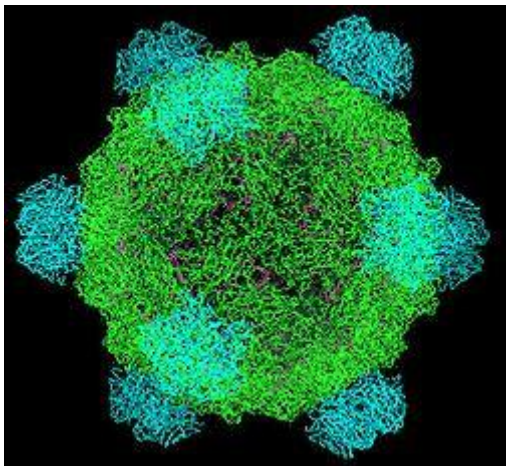
Assembled Phi-X174 genome using Overlap Graph, Kmer Composition and De-Bruijn Graph.

Phi X174

The phi X 174 (or ΦX174) bacteriophage is a single-stranded DNA (ssDNA) virus and the first DNA-based genome to be sequenced. This work was completed by Fred Sanger and his team in 1977. In 1962, Walter Fiers and Robert Sinsheimer had already demonstrated the physical, covalently closed circularity of ΦX174 DNA. Nobel prize winner Arthur Kornberg used ΦX174 as a model to first prove that DNA synthesized in a test tube by purified enzymes could produce all the features of a natural virus, ushering in the age of synthetic biology. In 2003, it was reported by Craig Venter's group that the genome of ΦX174 was the first to be completely assembled in vitro from synthesized oligonucleotides. The ΦX174 virus particle has also been successfully assembled in vitro. Recently, it was shown how its highly overlapping genome can be fully decompressed and still remain functional.



Problem Description.

- **Input**: A collection of Strings called reads of the original genome. Each read is a sub-string of the original genome(Genome can be circular also).
- **Output**: A string S of minimum length that cantains all the strings(reads) given in the input as its sub-strings.

Algorithms

- **Overlap Graph Algorithm**:
    - Construct an overlap graph. Two reads are joined by a directed edge of weight equal to the length of the maximum overlap of these two strings.
    - Then construct a Hamiltonian path in this graph in a greedy fashion.
    - Greedy Strategy : For each read select an outgoing edge of maximum weight. Why? Because the more the overlap between the reads shorter shorter will be the length of the combined string made of these reads.
    - Then read a string spelled by this path. i.e combine to form a super string.
    - Sometimes choosing the wrong first vertex may result in longer superstring. So you should generate random index probably 2-3 times and find minimum length super string.
    - Now in the last step since genome can be circular also so remove the overlap length between last and first read.

  Note: This greedy algorithm does not work with every genome as it might not give optimal solution every time.

- **K-Mer Composition Algorithm Using De-Bruijn Graph**:

  What is K-mer Composition? --> Given a String ACGTACTAT. Its 3-mer Composition is (ACG, CGT, GTA, TAC, ACT, CTA, TAT).

  Its De-Bruijn graph:

STEPS:

- Read the k-mer composition of the graph.
- Create the De-Bruijn graph from the k-mer composition.
- Find an eulerian cycle in the graph.
- Construct the genome from the found cycle.

- **Genome-Assembly using De-Bruijn graph from Error-Prone Reads.**:

  Algorithm:

  - Read the reads of the genome from the input.
  - Create De-Bruijn Graph from the k-mers - which are formed by spitting the reads into all substrings of length k.
  - Remove the tips from the graph.
  - Remove the bubbles from the graph.
  - Find an Eulerian cycle in the graph.
  - Form the genome from the eulerian cycle found in the graph.

  Note:

  - Do not remove bubbles of long lengths. Only remove bubbles of length less than k, which is the size of a k-mer.
  - You may also have to again remove tips after removing the bubbles.

  De-Bruijn Graph from reads:

  Bubbles:

  Tips:

  - In the above de-bruijn graph tips are:
    - ATGA-->TGAC-->GACC-->ACCA
    - TTGC
    - TGCT