AMRITA
VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY

22BIO211: Intelligence of Biological Systems - 2

ASSEMBLING GENOMES FROM READ-PAIRS

Dr. Manjusha Nair M
Amrita School of Computing, Amritapuri

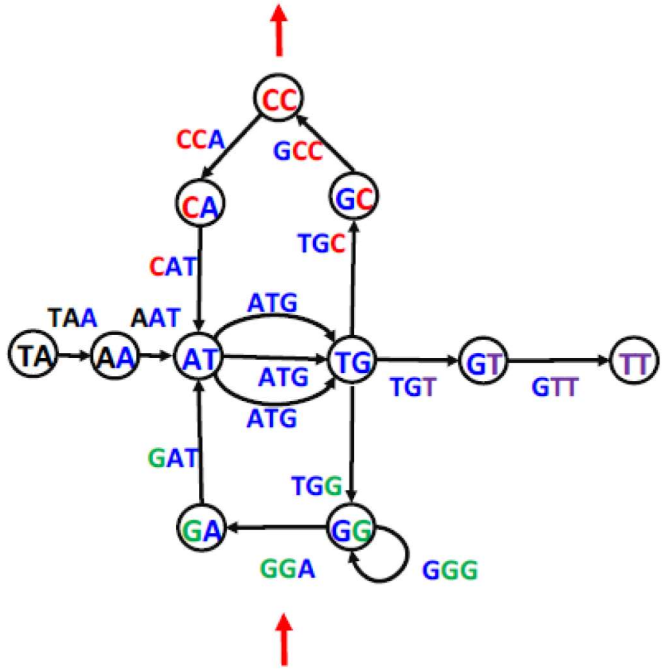
Email : manjushanair@am.amrita.edu
Contact No: 9447745519

From Reads to de Bruijn Graph to *Genome*

Genome

TAATGCCATGGGATGTT

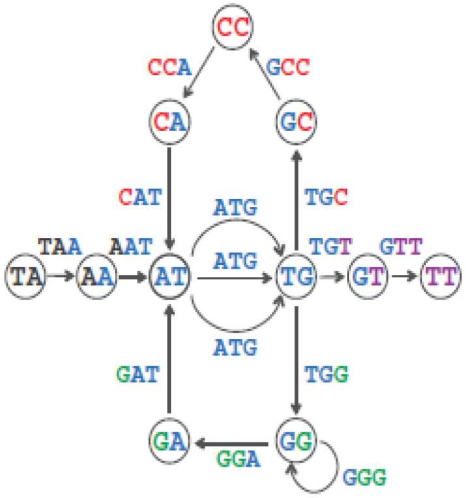
De- Bruijn graphs



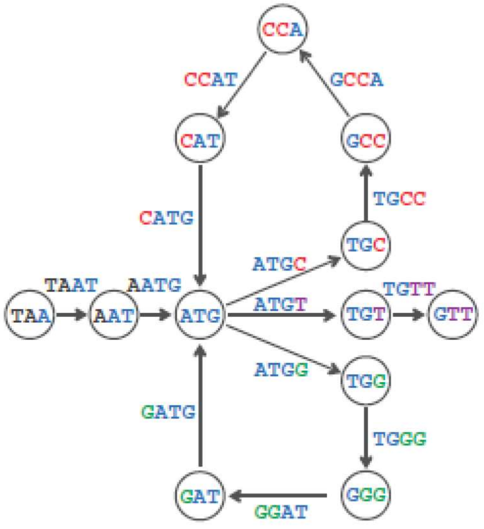
Reads

AAT ATG ATG ATG CAT CCA GAT GCC GGA GGG GTT TAA TGC TGG TGT

From Reads to de Bruijn Graph to *Genome*

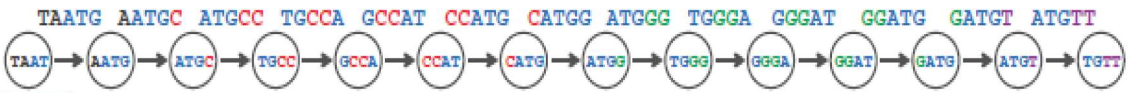


DEBRUIJN$_3$(TAATGCCATGGGATGTT)



DEBRUIJN$_4$(TAATGCCATGGGATGTT)

- de Bruijn graphs become less and less tangled when read length increases
- As soon as read length exceeds the length of all repeats in a genome, the de Bruijn graph turns into a path



DEBRUIJN$_5$(TAATGCCATGGGATGTT)

Repeats make the Graphs complicated

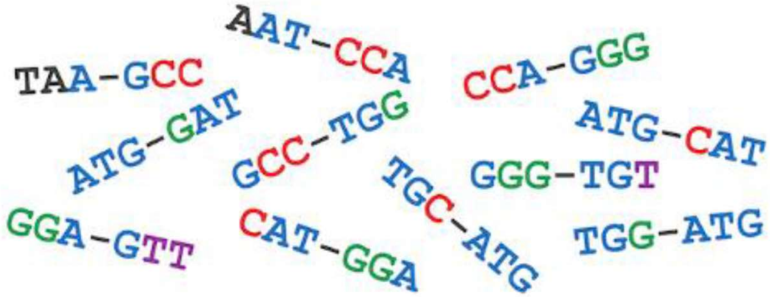
What is Ideal Read Length?

- Biologists have not yet figured out how to generate long and accurate reads.
  - *The most accurate sequencing technologies available today generate reads that are only about 300 nucleotides long*
    - which is too short to span most repeats, even in short bacterial genomes.
- Genome cannot be uniquely reconstructed from its k-mer composition

Biologists have suggested an indirect way of increasing read length by generating **read-pairs**

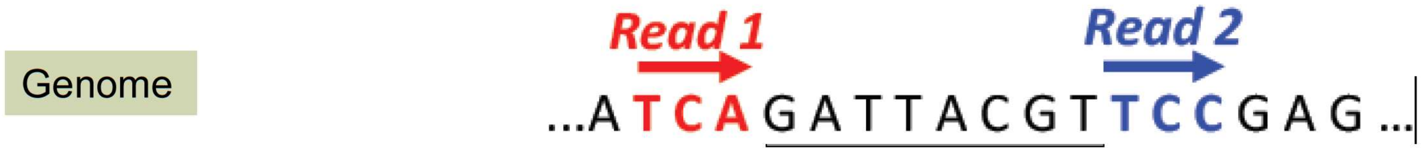
Reads to Read-pairs

- Read Pairs
  - *pairs of reads separated by a fixed distance d in the genome*
  - *long "gapped" read of length k + d + k whose first and last k-mers are known but whose middle segment of length d is*



Read-pairs sampled from **TAATGCCATGGGATGTT** (reads of length 3 separated by a gap of length 1)

From *k*-mers to Paired *k*-mers



A **paired *k*-mer** is a pair of *k*-mers at a fixed distance *d* apart in *Genome*.
E.g. **TCA** and **TCC** are at distance *d*=8 apart.

From *k*-mers to Paired *k*-mers

- Given a string Text, a (k, d)-mer is a pair of k-mers in Text separated by distance *d*
- We use the notation (Pattern1 | Pattern2) to refer to a (k, d)-mer whose k-mers are Pattern1 and Pattern2

(TAA|GCC) is a(3,1)-mer in TAATGCCATGGGATGTT

(**ATG** | **GGG**) is a (3, 4)-mer in **TAATGCCATGGGATGTT**

From Composition to Paired Composition

- PAIREDCOMPOSITION $_{k,\ d}$(Text), is the collection of all (k, d)-mers in Text (including repeated (k, d)-mers).

PairedComposition$_{3,1}$
(TAATGCCATGGGATGTT)

```
TAA GCC
 AAT CCA
  ATG CAT
   TGC ATG
    GCC TGG
     CCA GGG
      CAT GGA
       ATG GAT
        TGG ATG
         GGG TGT
          GGA GTT

TAATGCCATGGGATGTT
```

(AAT | CCA) (ATG | CAT) (ATG | GAT) (CAT | GGA) (CCA | GGG) (GCC | TGG) (GGA | GTT) (GGG | TGT) (TAA | GCC) (TGC | ATG) (TGG | ATG)

lexicographic order of the 6-mers formed by
their concatenated 3-mers:

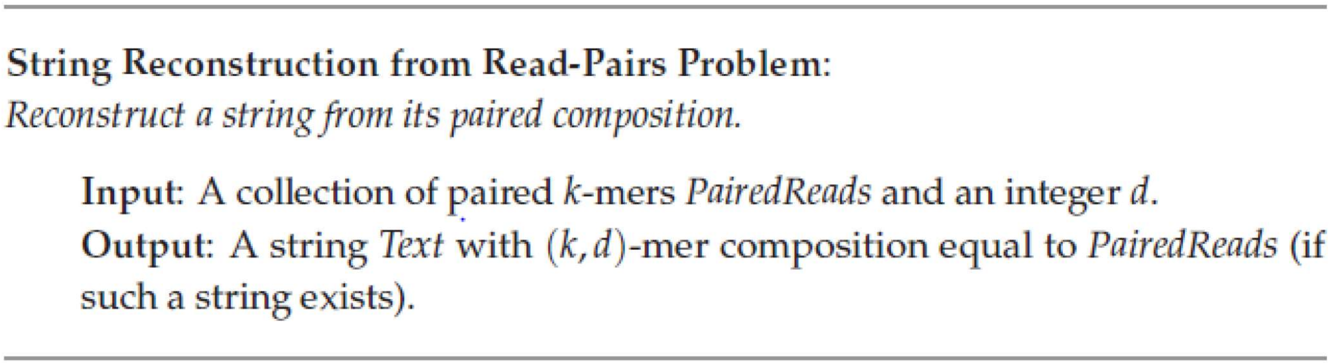
From Composition to Paired Composition

- **Advantages of Paired Composition**

1. **Reduced Repeats**
    - There were repeated 3-mers in the 3-mer composition
    - There are no repeated (3, 1)-mers in its paired composition

2.**Uniqueness**
    - Two different Genomes that can be reconstructed using previous example : **TAATGCCATGGGATGTT** and **TAAATGCCATGGGATGTT**
    - **They** have the same 3-mer composition
    - They have different (3, 1)-mer compositions.

String Reconstruction from Read-Pairs



String Reconstruction from Read-Pairs Problem:
Reconstruct a string from its paired composition.

Input: A collection of paired k-mers *PairedReads* and an integer d.
Output: A string *Text* with (k, d)-mer composition equal to *PairedReads* (if such a string exists).

Solution : Paired de Bruijn Graphs and Eulerian Paths

Summary

- Reads to Read-pairs
- From *k*-mers to Paired *k*-mers
- From Composition to Paired Composition