AMRITA VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY

22BIO211: Intelligence of Biological Systems - 2

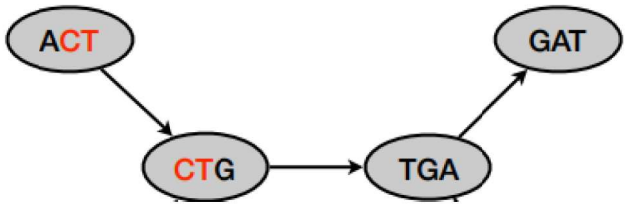
GENOME RECONSTRUCTION USING OVERLAP GRAPH

Dr. Manjusha Nair M
Amrita School of Computing, Amritapuri

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

Overlap Graphs

- *A graph is an overlap*
  - *if its vertices may be put into a one-to-one correspondence with intervals on a line,*
  - *such that two vertices are adjacent iff there intervals partially overlap,*
  - *that is, they have non-empty intersection, but neither contains the other.*
- Eg : $[ACTGAT]_3$ = **{ACT,CTG,TGA,GAT}**



Overlap Graphs

- Overlap graphs highlight situations in which there are interaction effects.
- Overlaps between reads can be determined by constructing an **overlap graph**.
- Edges represent overlap relationships between nodes.
- In Overlap Graph data structure, overlapping reads are connected using arrows.
- Overlap graph is a weighted directed graph, the weight of each edge being the length of overlap

Overlap Graph Problem

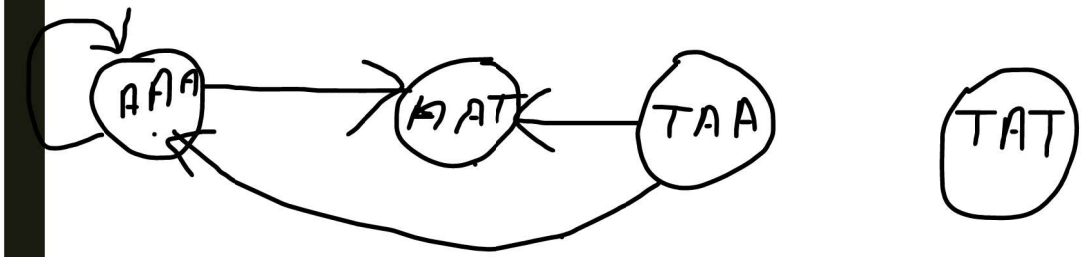
Overlap Graph Problem:
Construct the overlap graph of a collection of k-mers.

Input: A collection *Patterns* of *k*-mers.
Output: The overlap graph OVERLAP(*Patterns*).

Constructing Overlap Graphs

- The overlap graph Overlap(Patterns) is a directed graph that Contains
  - *One node for each k-mer in Patterns*
  - *An edge from Pattern $_i$ to Pattern $_j$ whenever the same (k -1)-mer is a suffix of Pattern $_i$ and a prefix of Pattern $_j$.*
  - *(E.g., AA is a suffic of TAA and a prefix of AAA)*

Example:
Patterns = {AAA,AAT,TAA,TAT}

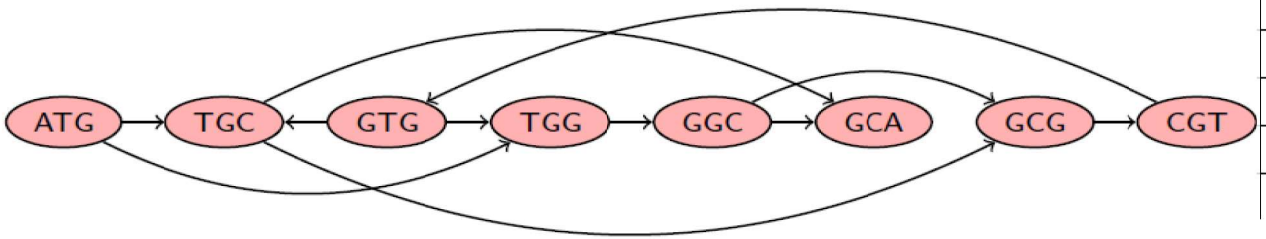


Node	Suffix	Prefix	Connected Nodes Suffix= Prefix
AAA	AA	AA	AAT, AAA
AAT	AT	AA	
TAA	AA	TA	AAT,AAA
TAT	AT	TA	

Constructing Overlap Graphs

- No of Nodes – Number of k- mers
- No of edges – number of overlapping connections
  - *One node for each k-mer in Patterns*
  - *Connect two nodes with an edge, if there is an overlapping relation*
    - ($suffix_{(source)}$ =$prefix_{(destination)}$)

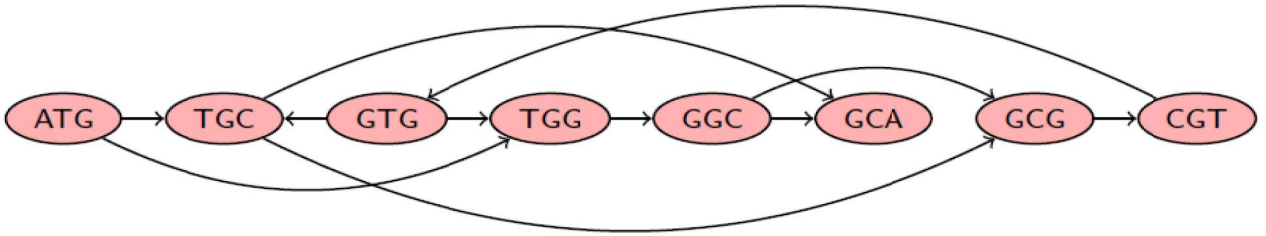
Example:
Patterns = {ATG;TGC;GTG;TGG;GGC;GCA;GCG;CGT}



Node	Suffix	Prefix	Connected Nodes Suffix= Prefix
ATG	TG	AT	TGC,TGG
TGC	GC	TG	GCA,GCG
GTG	TG	GT	TGC,TGG
TGG	GG	TG	GGC
GGC	GC	GG	GCA,GCG
GCA	CA	GC	
GCG	CG	GC	CGT
CGT	GT	CG	GTG

Constructing Overlap Graphs – Python Dictionary

Node	Suffix	Prefix	Connected Nodes Suffix= Prefix
ATG	TG	AT	TGC,TGG
TGC	GC	TG	GCA,GCG
GTG	TG	GT	TGC,TGG
TGG	GG	TG	GGC
GGC	GC	GG	GCA,GCG
GCA	CA	GC	
GCG	CG	GC	CGT
CGT	GT	CG	GTG




```
graph = { "ATG" : ["TGC","TGG"],
          "TGC" : ["GCA", "GCG"],
          "GTG" : ["TGC", "TGG"],
          "TGG" : ["GGC"],
          "GGC" : ["GCA", "GCG"],
          "GCA" : [],
          "GCG" : ["CGT"],
          "CGT" : ["GTG"]
        }
```

Constructing Overlap Graph : Another Example

- Construct an overlap graph of all 6-mers from GTACGTACGAT"
  - where edges are overlaps of length >= 4

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

Constructing Overlap Graph : Another Example

- Overlap Length = 4

GTACGTACGAT

Node	Suffix	Prefix	Connected Nodes Suffix= Prefix
GTACGT	ACGT	GTAC	ACGTAC
TACGTA	CGTA	TACG	CGTACG
ACGTAC	GTAC	ACGT	GTACGT,GTACGA
CGTACG	TACG	CGTA	TACGTA, TACGAT
GTACGA	ACGA	GTAC	
TACGAT	CGAT	TACG	

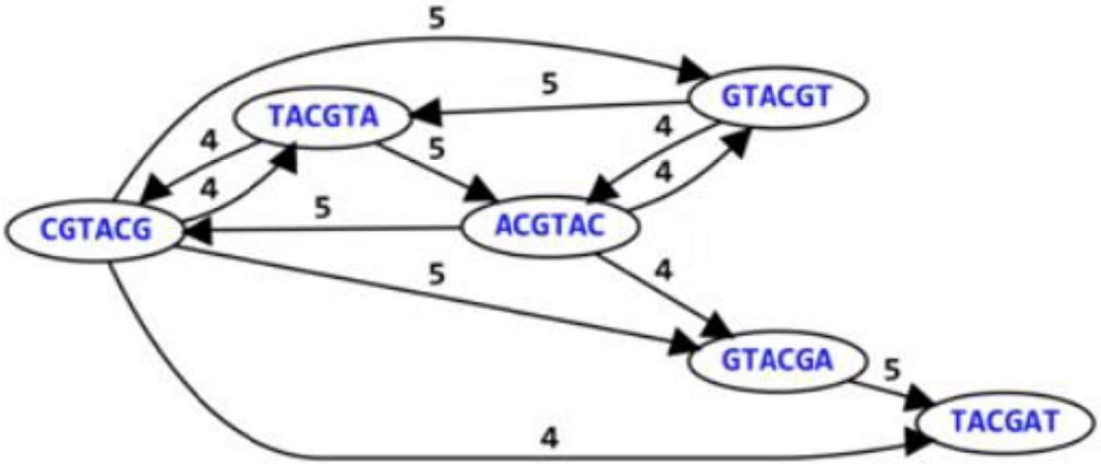
- Overlap Length = 5

GTACGTACGAT

Node	Suffix	Prefix	Connected Nodes Suffix= Prefix
GTACGT	TACGT	GTACG	TACGTA
TACGTA	ACGTA	TACGT	ACGTAC
ACGTAC	CGTAC	ACGTA	CGTACG
CGTACG	GTACG	CGTAC	GTACGA, GTACGT
GTACGA	TACGA	GTACG	~~GTACGA~~
TACGAT	ACGAT	TACGA	

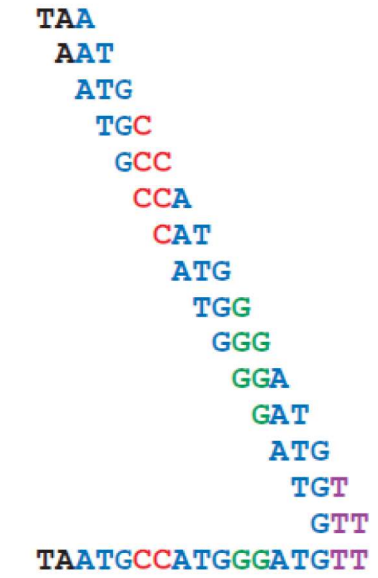
TACGAT

Constructing Overlap Graph : Another Example



Genome Path.

- Consecutive k-mers in a genome, are linked together to form genome path.
- Every k-mer has a relation



Constructing a genome path from its k-mer composition

If each k-mer is represented as a node in a graph, then the genome is a path through all the nodes

PREFIX(**TAA**) = **TA**

SUFFIX(**TAA**) = **AA**

The suffix of a 3-mer in the genome path is equal to the prefix of the following 3-mer in the path

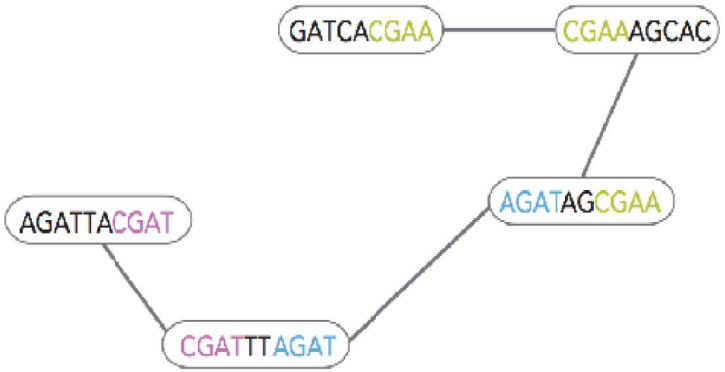
TAATGCCATGGGATGTT

SUFFIX(**TAA**) = PREFIX(**AAT**) = **AA**

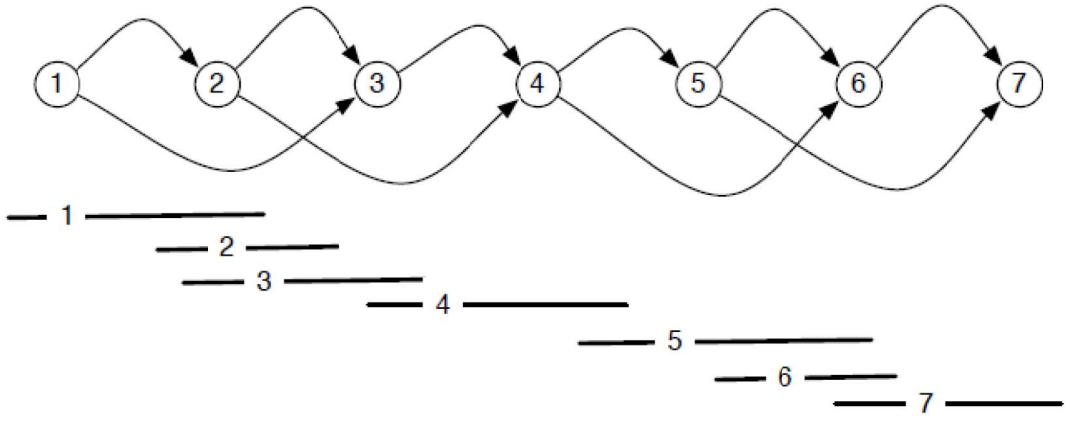
Use an arrow to connect any k-mer pattern to another k-mer pattern, if the suffix of first pattern is equal to the prefix of second pattern.

Genome Path and Overlapping Graph

Reads with overlapping sequence probably originate from overlapping portions of the subject genome



The full genome sequence is a "tour" of the graph



Given overlap graph, how can we find a good candidate assembly?

String Spelled by a Genome Path Problem

String Spelled by a Genome Path Problem:
Reconstruct a string from its genome path.

Input: A sequence of k-mers $Pattern_1, \ldots, Pattern_n$ such that the last $k-1$ symbols of $Pattern_i$ are equal to the first $k-1$ symbols of $Pattern_{i+1}$ for $1 \leq n-1$.
Output: A string $Text$ of length $k+n-1$ such that the i-th k-mer in $Text$ is equal to $Pattern_i$ (for $1 \leq i \leq n$).

Reconstructing a genome from its genome path is easy: adding one new symbol to the genome at each new k-mer
Already Solved in one of the previous labs – using string

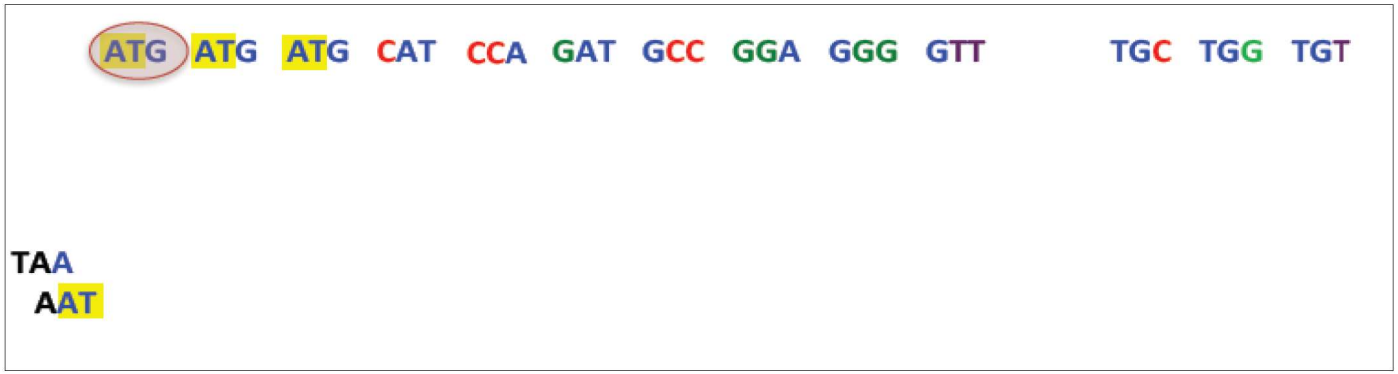
Unfortunately, constructing this string's genome path requires us to know the genome in advance.

Previous Example Revisited...

Step: 1

Step: 2

Step: 3



Previous Example Revisited...

Step: 4

Step: 5

What's Next?

ATG ATG CAT CCA GAT GCC GGA GGG

TGC TGG

TAA
AAT
ATG
TGT
GTT

Extending ATG by TGC instead of TGT

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

Constructing a genome from its k-mer composition - using Overlap Graph and Genome path

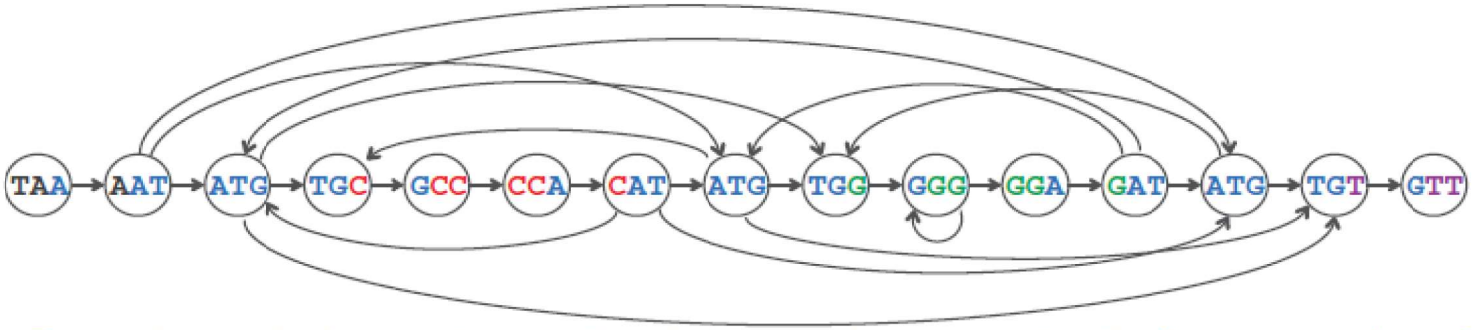
Input

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

Output

TAATGCCATGGGATGTT

No. of Nodes : 15 (No of K-mers)
No. of Edges : 25 (no of possible connections)



The overlap graph showing all connections between nodes representing the 3-mer composition of TAATGCCATGGGATGTT.

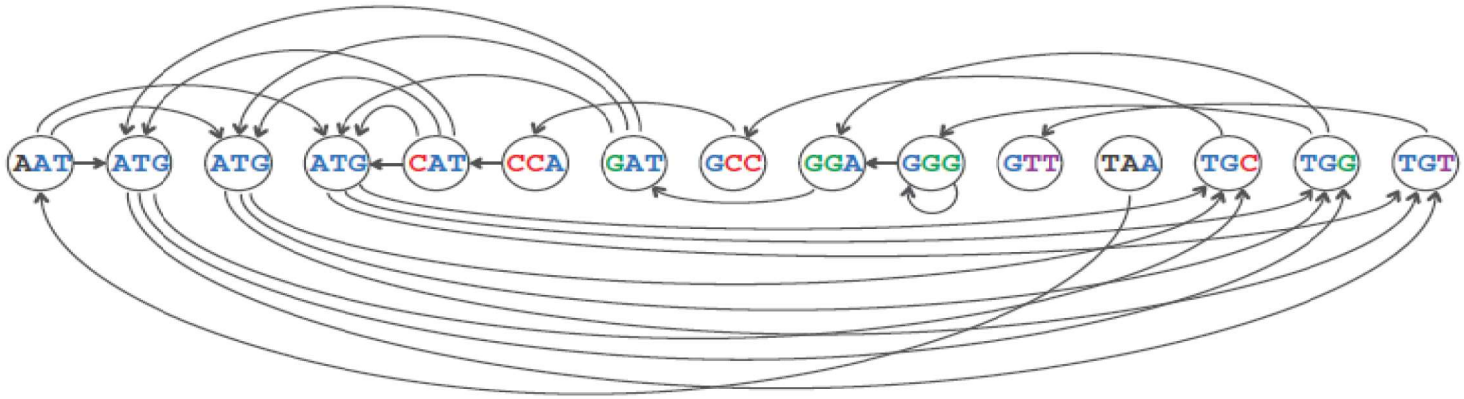
Constructing a genome from its overlap graph

Genome is a walk through the graph, connecting all the nodes exactly once

When Genome path exist- Note that the genome can still be spelled out by walking along the horizontal edges from **TAA** to **GTT**

Reconstructing Genome when there is no genome path

- In genome sequencing, we do not know in advance how to correctly order reads.
- Therefore we will arrange the 3-mers lexicographically
  - *Resulting in the following overlap graph*



Reconstructing Genome when there is no genome path

- The genome path has disappeared!
- The path through the graph representing the correct assembly is now harder
- Needed to find a path through the graph visiting each node exactly once;
  - *such a path "explains" all the 3-mers in the 3-mer composition of the genome.*

Summary

- Overlap Graph
- Constructing Overlap Graph
- Genome Path and Overlap Graphs
- Reconstructing Genome from its k-mer composition- using overlap graph and genome path
- Reconstructing Genome when there is no genome path