Genomes Comparision via de Bruijn graphs

Student: Ilya Minkin
Advisor: Son Pham

St. Petersburg Academic University

March 23, 2012

Biological Motivation

- Sequencing genomes is getting cheaper
- Probably genome assembly task will be easier with current development in sequencing machines (Nanopore)
  - 1000 Human Genomes
  - Genomes 10K: One genome for each vetebrate genus
  - 1001 Arabidopsis Genomes
  - Human Microbiome Project: sequence genomes of microbial communities at different sites on human body
- What can we do with these **thousands** of sequences?

Long Term Project

- None of the current comparative genomics tools were designed for a very high number of genomes.
- We aim to provide a tool for comparing multiple genomes that has the following functions (properties)
  - Find synteny blocks in (multiple) complicated genomes
  - Allocate insertions, deletions
  - Find other structure variations
  - Ability to work for incomplete genomes (contigs)
  - Provide a user friendly web interface for this tool.

Synteny Blocks: Algorithmic challenge

- Suppose that we are given two genomes
- The question is: how are they evolutionary related to each other?
- In order to do rearrangements analysis we must decompose genomes into synteny blocks
- Synteny blocks are evolutionary conserved segments of the genome
- These blocks cover most of the genome
- Occur in both genomes with possible variations

Academic Project

Project: Identify synteny blocks for duplicated genomes represented as sequences of **nucleotides**.

- **None** of the previous synteny blocks reconstruction software (DRIMM-Synteny (Pham And Pevzner 2010) included) can efficiently solve this problem.
- DRIMM-Synteny can find the synteny blocks for complicated genomes. But:

Academic Project

Project: Identify synteny blocks for duplicated genomes represented as sequences of **nucleotides**.

- **None** of the previous synteny blocks reconstruction software (DRIMM-Synteny (Pham And Pevzner 2010) included) can efficiently solve this problem.
- DRIMM-Synteny can find the synteny blocks for complicated genomes. But:
- It requires the genome to be represented as sequence of genes.

General Idea: de Bruijn Graph

- Create de Bruijn graph from the nucleotide sequence - no anchors
- Conserved regions will yield non-branching paths
- We are given an alphabet Σ and a string S over it, $|\Sigma| = m$
- A substring T, $|T| = k$ is called *k-mer*
- de Bruijn graph is a multigraph $G_k = (V, E)$, where
  $V = \Sigma^{k-1} = \{\text{all possible strings of length k-1}\}$
- If k-mer T is present in S then we add oriented edge $(T[1, k-1], T[2, k])$ to the graph

Challenges

- Variations in synteny blocks generate cycles, so we need to simplify graph
- Double strandness: conserved regions may occur on both strands. Example:
  5' AACCGGTT 3'
  3' TTGGCCAA 5'
- Such blocks are reversed complementary to each other $\Rightarrow$ no non-branching paths
- Memory efficiency

First Example: Ideal Situation

S = ACGTGGGACGTG
k = 3

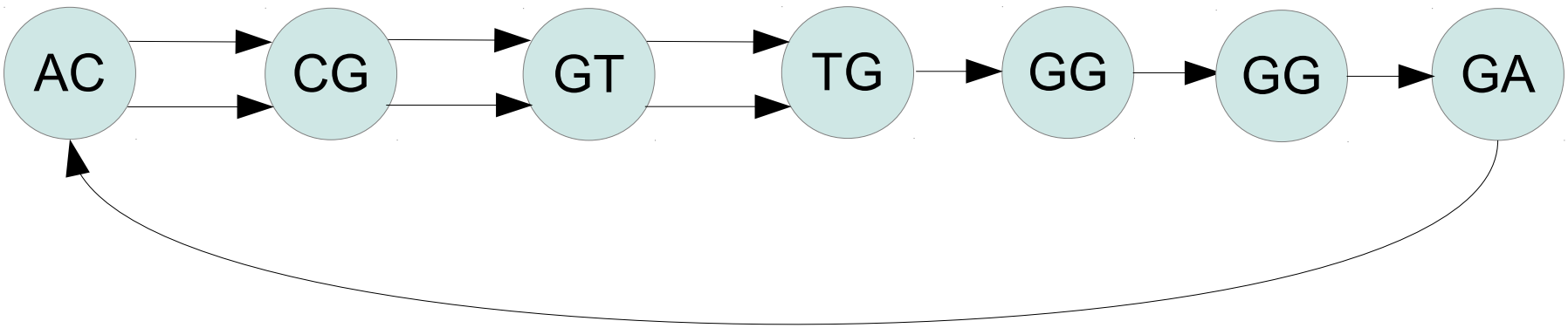

Figure 1: Here absolutely conserved region "ACGTG" generates clear non-branching path

Second Example: SNP

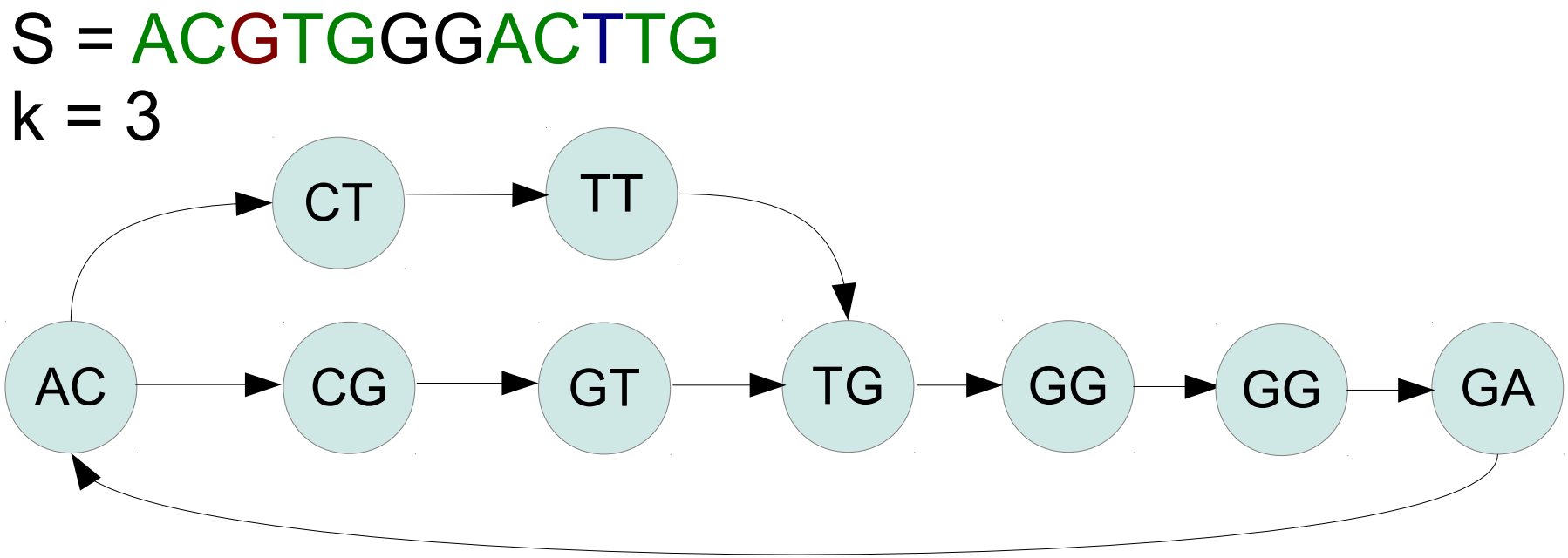

Figure 2: In this example SNP generates so-called "bulge" cycle

Double Strands: Possible Solutions

- Colored de Bruijn graph. Build two graphs for the direct and the reverse-complementary sequences. Color edges in each graph and merge graphs
- Bidirected graph. Each vertex has two parts – direct and reverse complementary. Every edge has two directions (one on each end) to indicate which part of the vertex we use
- Simplest possible solution – glue k-mers that are reverse complementary

Third Example: Double Strands

S_{dir} = 5' ACCTTAGGT 3'

S_{rev} = 3' TGGAATCCA 5'

k = 3

Glue complementary *k*-mers ACC/GGT & CCT/AGG

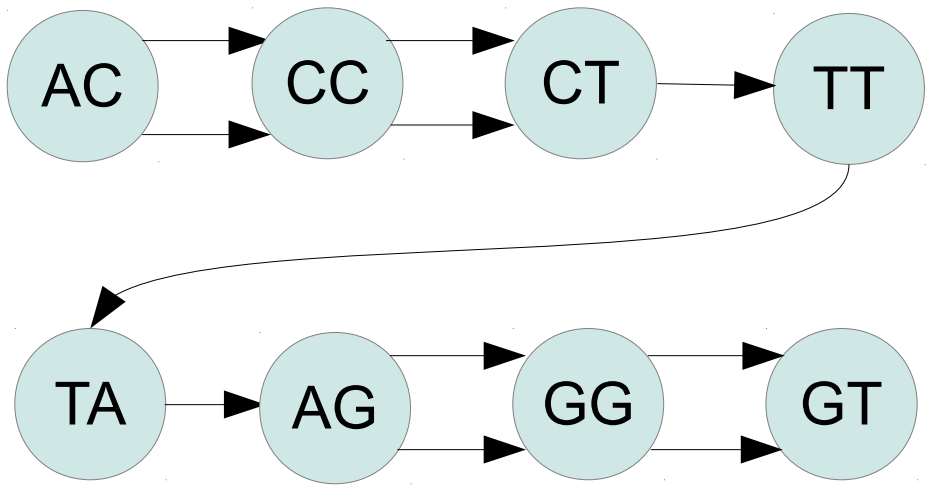

Figure 3: Gluing reverse complementary edges helps to resolve double strandness issue

Methods and expected results

- Glue complementary k-mers togetther
- Simplify graph by deleting short cycles (with size less than some Δ)
- Note that our graph simplification is different from the graph simplification in genome assemblers
- Find non branching paths $=$ synteny blocks
- Use the software to analyse repeats in Arabidopsis genome

Current Progress

Now:

- Program that can find absolutely conserved regions on one strand
- Handles 25 MB Arabidopsis chromosome with $\leq$ 500 MB RAM

Near future:

- Add graph simplification
- Resolve double strandness issue
- Get rid of hashtables, use suffix arrays $\Rightarrow$ reduced memory consumption

References

- 1. Pevzner P and Tesler G, (2003) Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution.
- 2. Pham S and Pevzner P, (2010) DRIMM-Synteny: Decomposing Genomes into Evolutionary Conserved Segments

Thank you!