Genomes Comparision via de Bruijn graphs

Student: Ilya Minkin
Advisor: Son Pham

St. Petersburg Academic University

June 4, 2012

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

- 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 $V = \Sigma^{k-1} = \{\text{all possible } (k-1)\text{-mers}\}$
- If k-mer T is presented in S, then we add an oriented edge $(T[1, k-1], T[2, k])$ to the graph
- Create de Bruijn graph from the nucleotide sequence
- Conserved regions will yield non-branching paths

Challenges

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

Colored graph

- We use colored de Bruijn graphs [Iqball et al., 2012] to handle double-strandness
- Suppose that S^+ and S^- are positive and negative strands of the chromosome
- Colored de Bruijn graph is a multigraph $G_k = (V, E)$ where $V = \Sigma^{k-1}$
- For each k-mer T^+ in S^+ add edge $(T^+[1, k-1], T^+[2, k])$ to G_k and mark it *blue*
- For each k-mer T^- in S^- add edge $(T^-[1, k-1], T^-[2, k])$ to G_k and mark it *red*

Edge labeling

- Note that our graph is built from a string, not set of reads
- Each walk in the graph represents a string
- We are interested only in walks that represent substrings of the source string
- Assign to each edge e label $L(e) =$ position of the corresponding k-mer on the positive strand
- Walk $W = (v_1\, e_1\, v_2\, e_2 \ldots)$ is considered valid iff:
  1. e_i and e_{i+1} are of the same color
  2. $|L(e_i) - L(e_{i+1})| = 1$

Example

5' ACCTGTCAGT 3'
3' TGGACAGTCA 5'

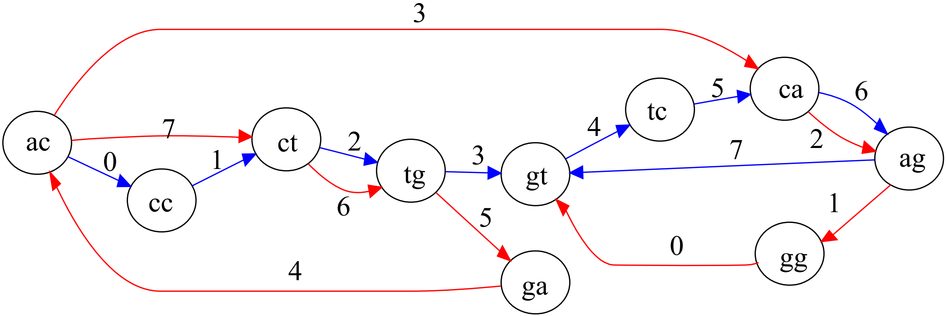

Figure 1: Colored de Bruijn graph built from two strands

Graph simplification

- Bulges spoil long non-branching paths and indicate indels/mismatches
- A pair of walks (W_1, W_2) is a bulge iff:
  1) Start and end vertices of W_1 and W_2 coincide
  2) W_1 and W_2 have exactly 2 common vertices
  3) There are no edges $u \in W_1$ and $v \in W_2$ such that $L(u) = L(v)$
  4) $|W_1| \leq \delta$ and $|W_2| \leq \delta$

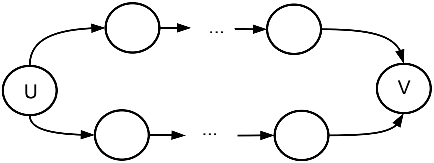

Figure 2: A bulge

General pipeline

- Build de Bruijn graph from the genome
- Remove bulges (BFS-like algorithm)
- Bulges are removed by replacing long branches with shorter ones
- Output non-branching paths

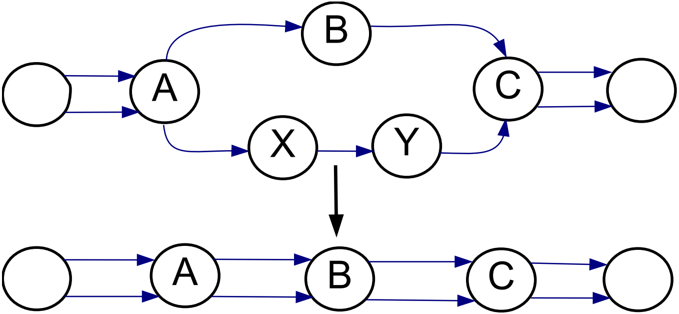

Figure 3: Bulge removal illustration

Parameters selection

- How should we choose K and δ?
- Duplicated genes can have no long ($K > 50$) shared K - mers
- Big $K \sim 50$ – we find only few synteny blocks
- Small $K \sim 10$ and small $\delta \sim 15$ – we find very short synteny blocks
- Small $K \sim 10$ and big $\delta \sim 200$ – the genome will be disrupted completely

Parameters selection

- How should we choose K and δ?
- Duplicated genes can have no long ($K > 50$) shared K - mers
- Big $K \sim 50$ – we find only few synteny blocks
- Small $K \sim 10$ and small $\delta \sim 15$ – we find very short synteny blocks
- Small $K \sim 10$ and big $\delta \sim 200$ – the genome will be disrupted completely
- Solution – do simplification in multiple stages

New pipeline

- General idea – "align" similar regions first, then glue them together into synteny blocks
- Start with small K and small δ to smooth duplicated regions and obtain long K-mers
- Rebuild and simplify the graph with higher K and δ
- Continue this process several times
- Final step can be done with $K \sim$ several hundreds

Experiment

- We have attempted to identify duplications in *Arabidopsis thaliana*
- Arabidopsis is known to be highly duplicated genome [Arabidopsis Genome Initiative]
- Size of the genome is $\sim 120\, Mbp$
- We used 4 stages and following parameters:

Stage number	K	δ
1	15	150
2	50	500
3	100	1000
4	500	5000

Computation results

- We have found 4722 synteny blocks in Arabidopsis
- These blocks cover 28 % of the genome
- Minimum length of the block is 1000 *bp*
- Largest block found has length $\sim$ 95 000 *bp*
- We tried to verify blocks by aligning instances of the same block
- At least 74 % of blocks have 70 % of exact matches

Computation results

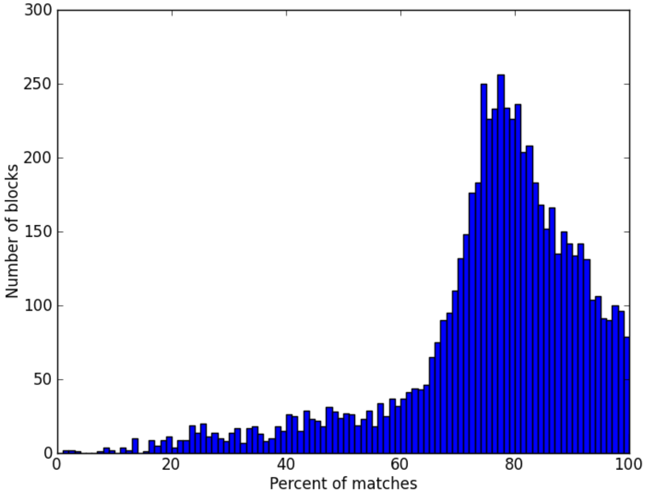

Figure 4: Matches percent vs. number of blocks plot

Computation results

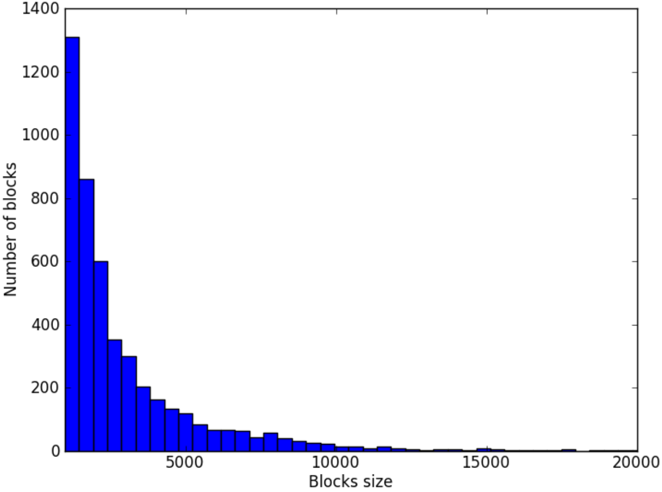

Figure 5: Synteny blocks length distribution

Future plans and summary

Summary

- We have covered 28 % of Arabidopsis genome with synteny blocks
- But we have missed some duplicated regions, described in [Arabidopsis Genome Initiative]
- Most of the blocks are short ($< 5000\,bp$)

Future plans

- Improve coverage
- Examine other genomes
- Optimize algorithms to handle larger genomes

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
- 3. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G, (2012) De novo assembly and genotyping of variants using colored de Bruijn graphs
- 4. Arabidopsis Genome Initiative, (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana

Thank you!