Genomes Comparision via de Bruijn graphs

Student: Ilya Minkin
Advisor: Son Pham

St. Petersburg Academic University

August 31, 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 $\Rightarrow$ no non-branching paths
- We need exact shared k-mers, so *directly* our approach can be applied to closely related species (different strains of a bacteria, etc)
- 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'

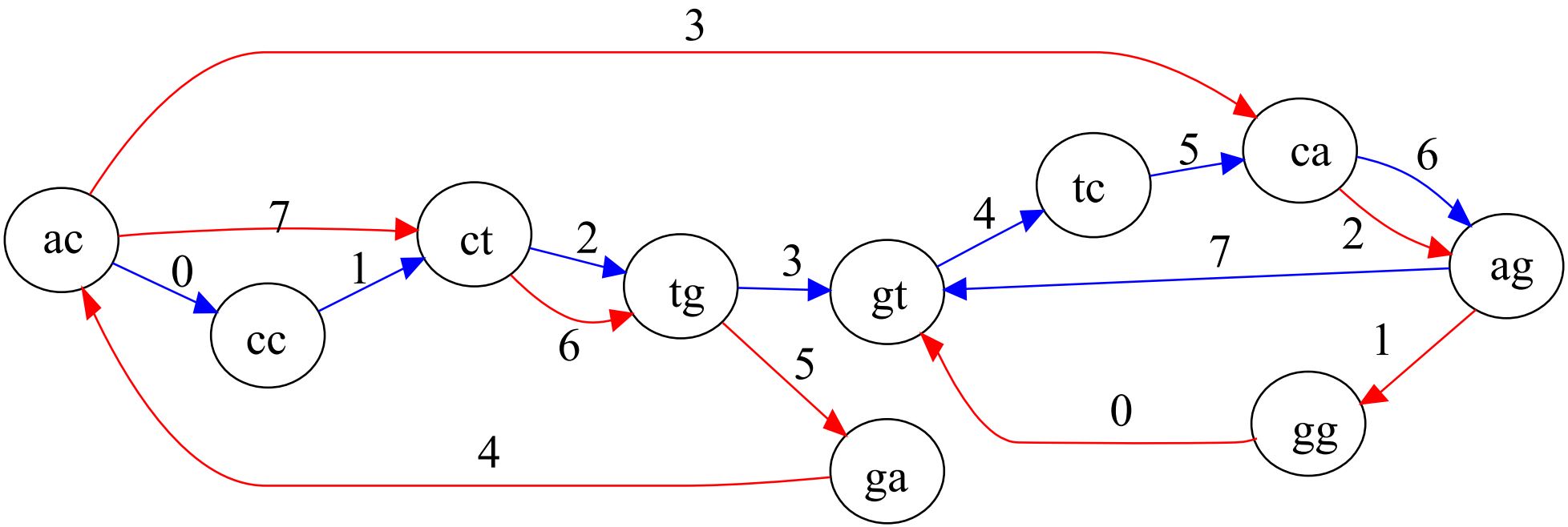


Figure 1: Colored de Bruijn graph built from two strands

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 $|W_2| \leq \delta$

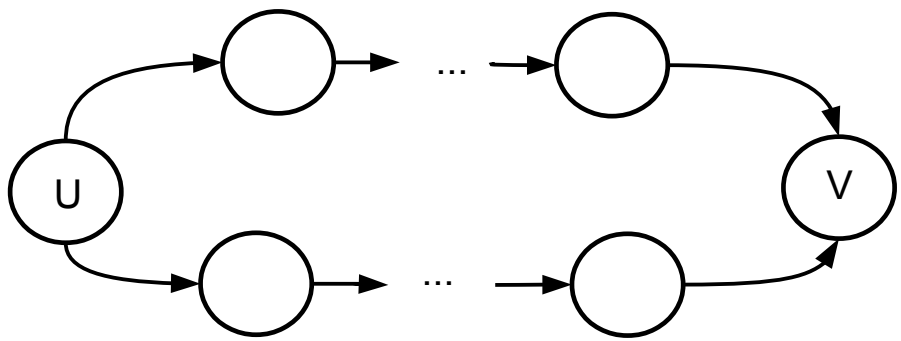


Figure 2: A bulge

General pipeline

- Build de Bruijn graph from the genome
- Remove bulges
- Bulges are removed by replacing one branch with another
- Output non-branching paths

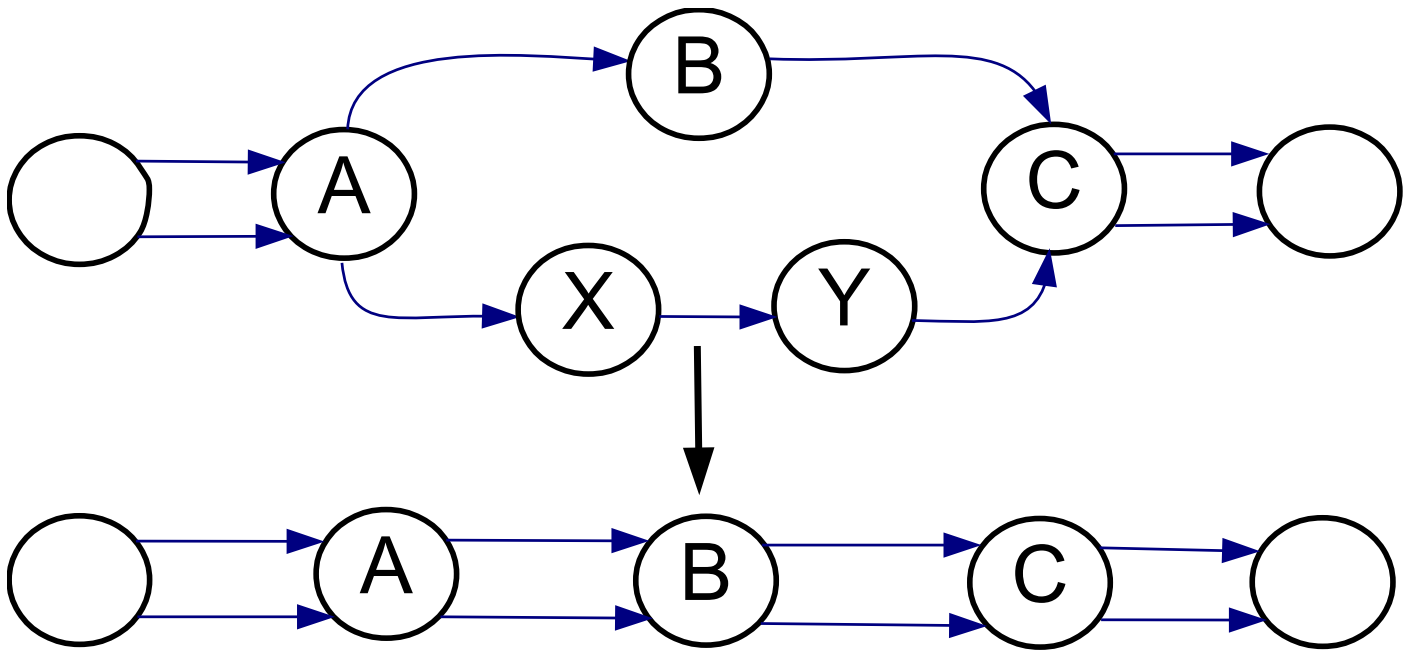
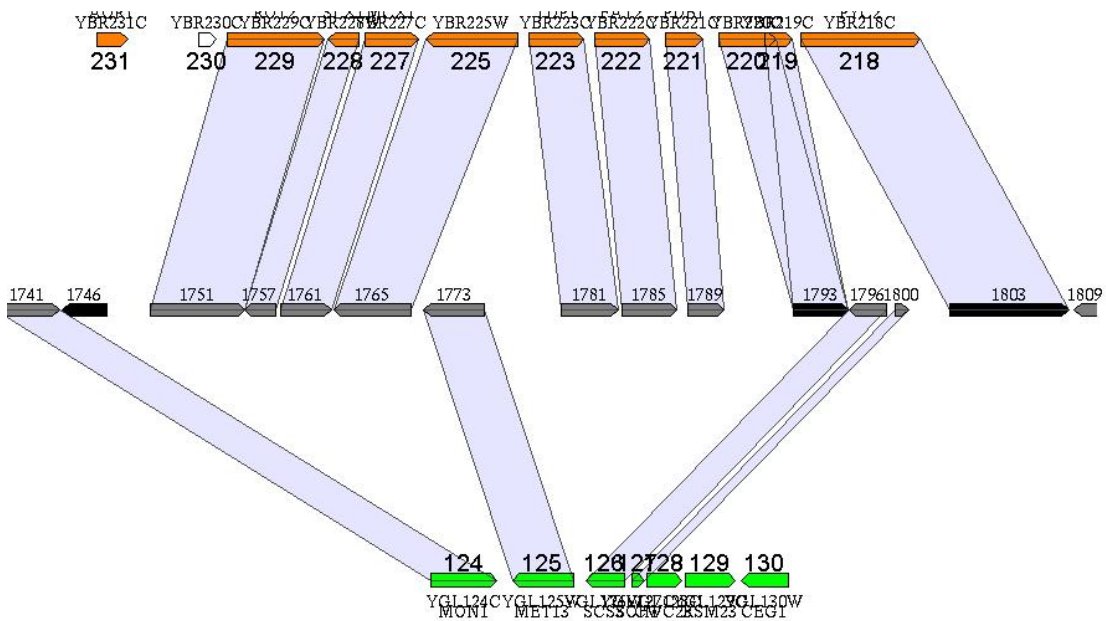


Figure 3: Bulge removal illustration

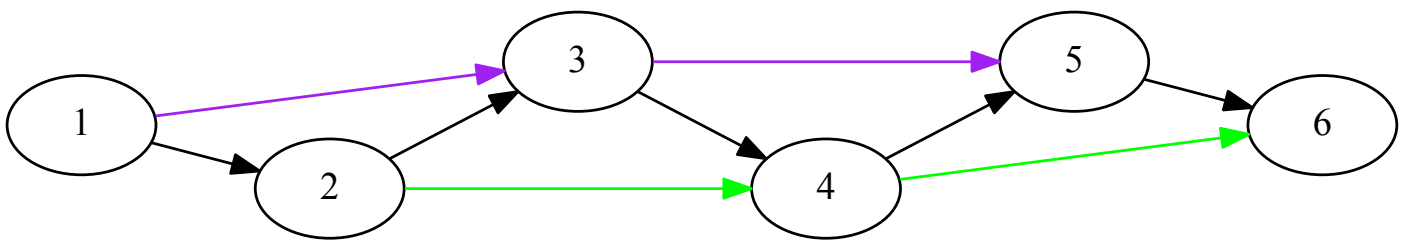
Bulge removal strategy

- It does matter which bulge's branch we replace
- Synteny blocks with muliplicity > 2 can have no k-mers shared across all instances of a same block
- Example: a synteny block from yeasts [Kellis et al., 2004]



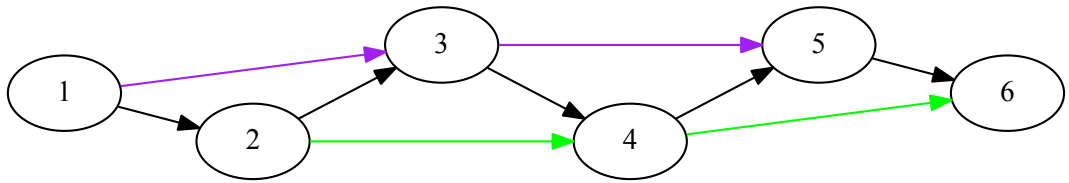
Bulge removal strategy

- Suppose that we have synteny 3 regions, k-mers are denoted by integers:
  2 4 6
  1 2 3 4 5 6
  1 3 5
- Let's build de Bruijn graph for this situation:

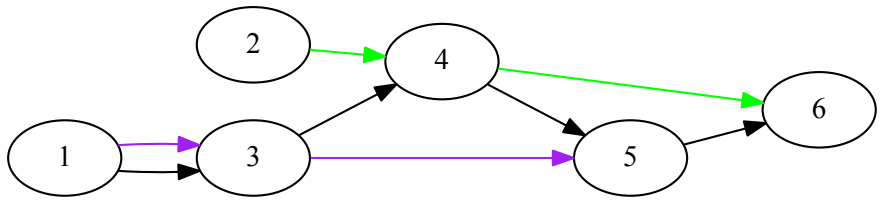


Wrong strategy

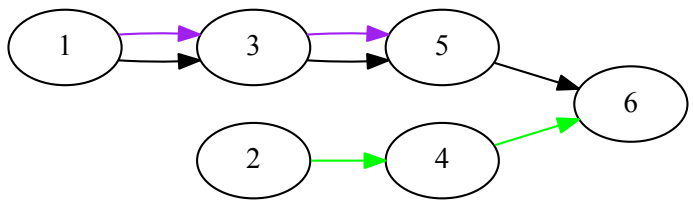
- Initial situation:



- Replace 1 3 by 1 2 3:

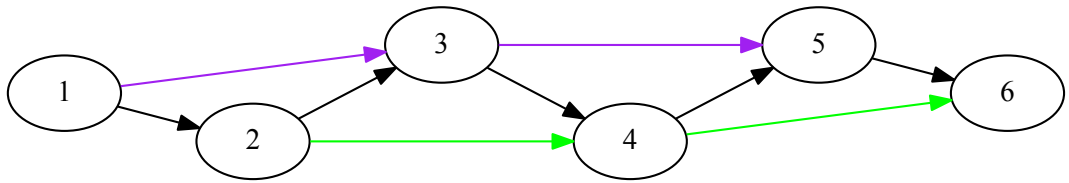


- Replace 3 4 5 by 3 5:

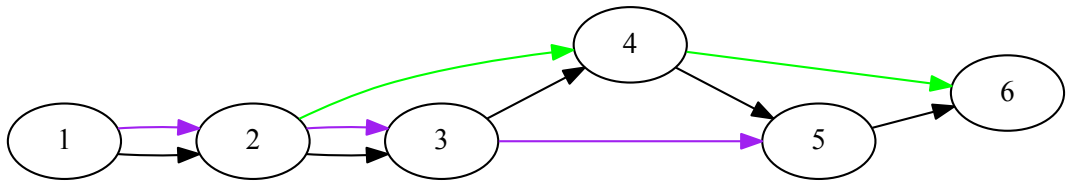


Proper strategy

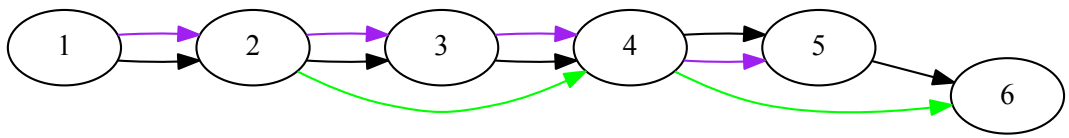
- Initial situation:



- Replace 1 3 by 1 2 3:

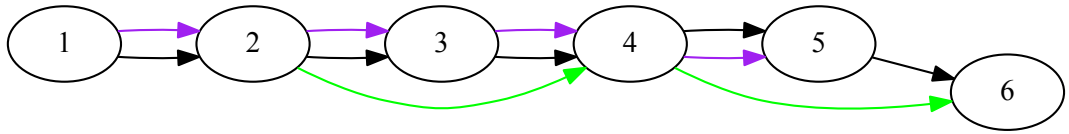


- Replace 4 6 by 4 5 6:

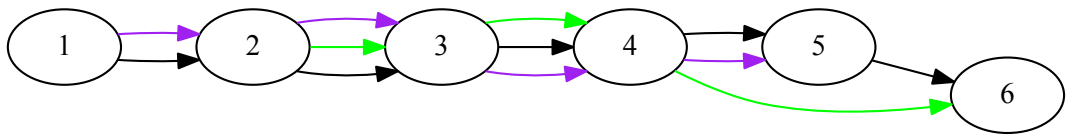


Proper strategy

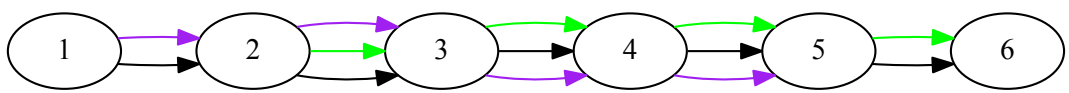
- From previous slide:



- Replace 2 4 by 2 3 4:



- Replace 4 5 by 4 5 6:



Proper strategy

- We incapsulate this intuition into a heuristic
- A vertex v in the graph is called *bifurcation* if there are at least two ingoing (outgoing) edges incident v that spell different k-mers.
- Let's denote by $MaxBif(p)$ maximum degree of bifurcation that lies on path p
- If two paths p_1 and p_2 form a bulge, then we replace p_1 by p_2 iff $MaxBif(p_1) > MaxBif(p_2)$, otherwise we replace p_2 by p_1
- And it seems to work

Results: a simple example

- We took two bacteria from *Pseudomonas aeruginosa* group and dot plot them:
  Pseudomonas aeruginosa PAO1
  Pseudomonas aeruginosa UCBPP-PA14

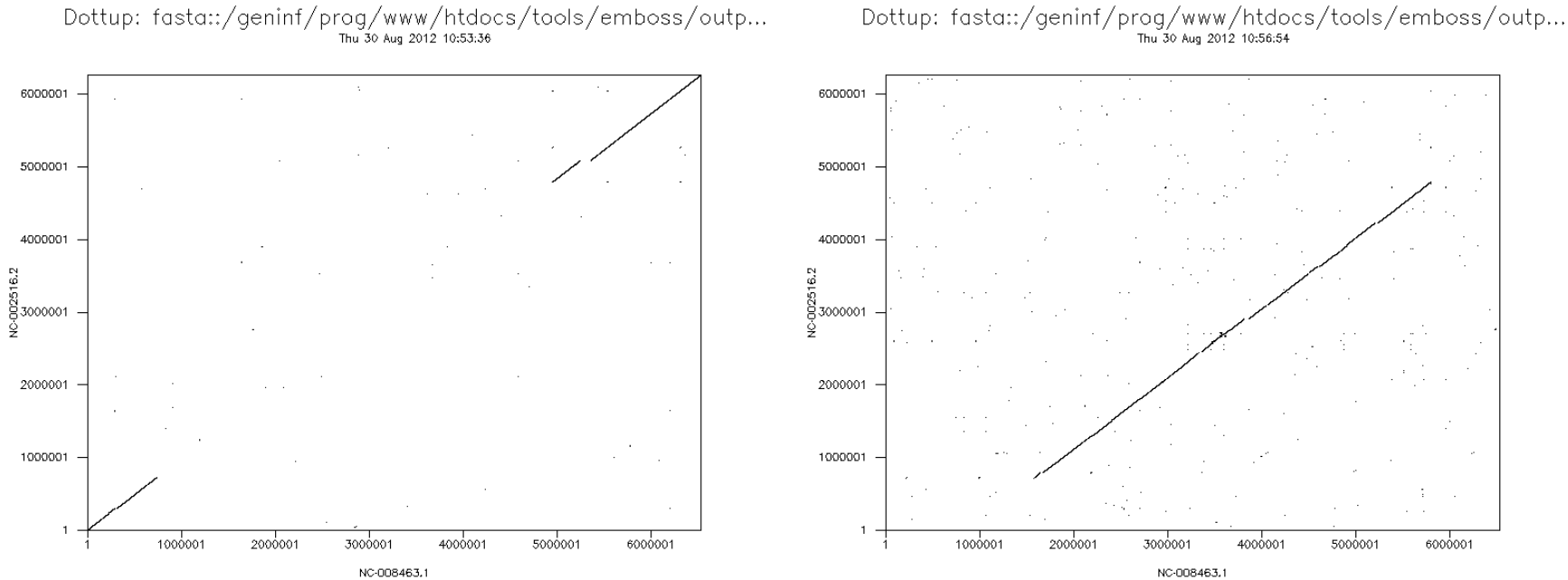


Figure 5: Left plot corresponds to PA01 against PA14, right plot corresponds to reverse-complementary of PA01 against PA14

Results: a simple example

- Then we constructed the blocks using our program
- $K = 5000, \delta = 25000$
- Results:
  +9 -0 -1 -2 -3 -4 -5 -6 +7 +10 +8
  +9 -7 +6 +5 +4 +3 +2 +1 +0 +10 +8

Results: three bacteria dataset

- Three strains of *Mycobacterium tuberculosis H37Rv*:
  Laboratory reference strain *H37Rv*
  CCDC5180
  CCDC5079
- One of this strains has multiple drugs resistance
- We use $K = 1000$ and $\delta = 5000$
- Blocks with multiplicity 3 cover 96% of the genome
- And Son see an evidence of rearrangements there

Results: three bacteria dataset

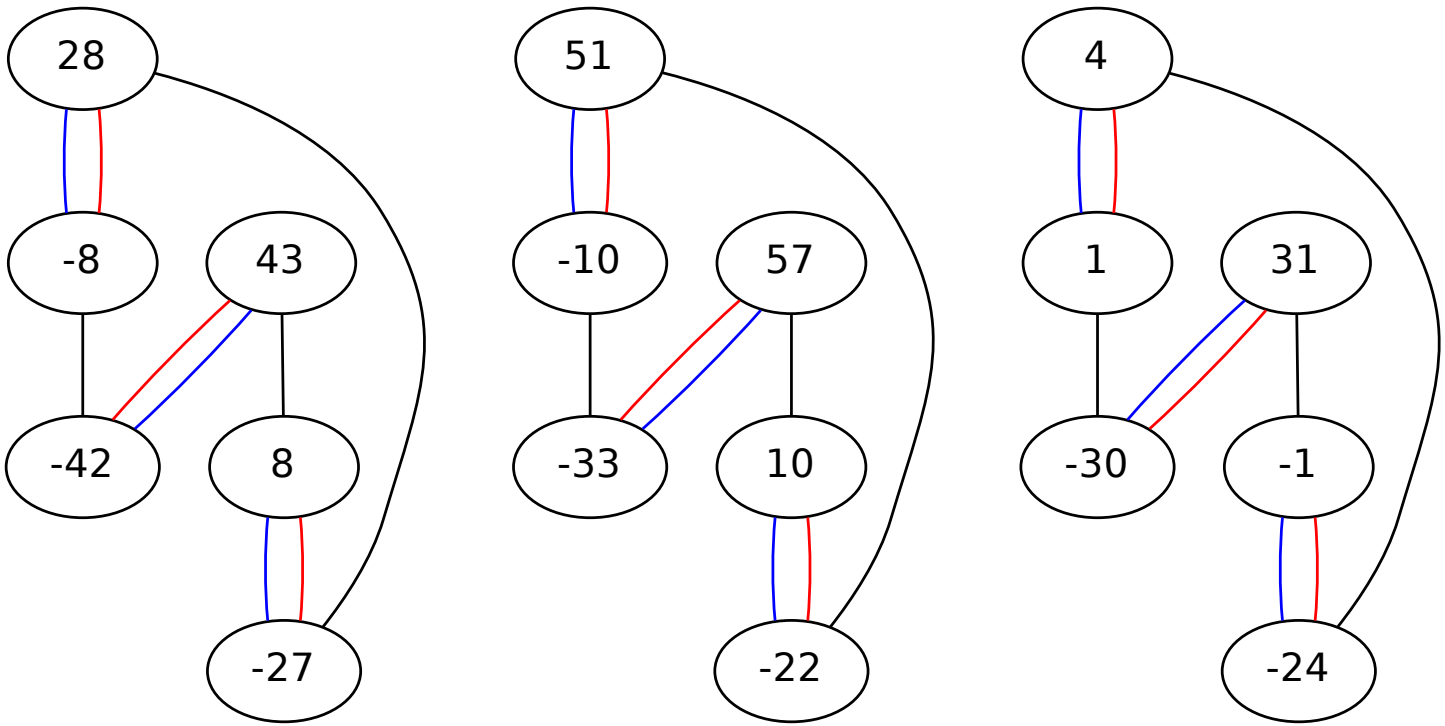


Figure 6

Results: yeasts dataset

- Well known research about yeasts *Saccharomyces cerevisiae* and *Kluyveromyces waltii* [Kellis et al., 2004]
- This paper shows evidence of so called *double conserved synteny* regions: one region (block) in k.waltii corresponds to two regions in cerevisiae
- We can't apply our method directly. But we can use alignments of ORFs from the paper to enrich number of k mers
- With $k = 1000$ and $\delta = 20000$ we cover 67% of the genome by blocks with multiplicity 3. Our blocks match 212 blocks from overall 252 in Kellis paper

Conclusions

- Our method is applicable for reconstructing synteny blocks in closely related species
- It can be extendede to handle more complicated cases

Future plans:

- Perform additional tests and evaluations
- Release software for finding synteny blocks in closely related species (end of Setptember)
- Write a paper
- Incorporate a local alignment tool and extend range of use to more complicated cases

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
- 5. Kellis M, W. Birren B, S. Lander E, (2004)

Thank you!