Ancestral Genome Construction And Multiple Genomes Comparison

Sergey Knyazev

Finding Synteny Blocks Algorithm

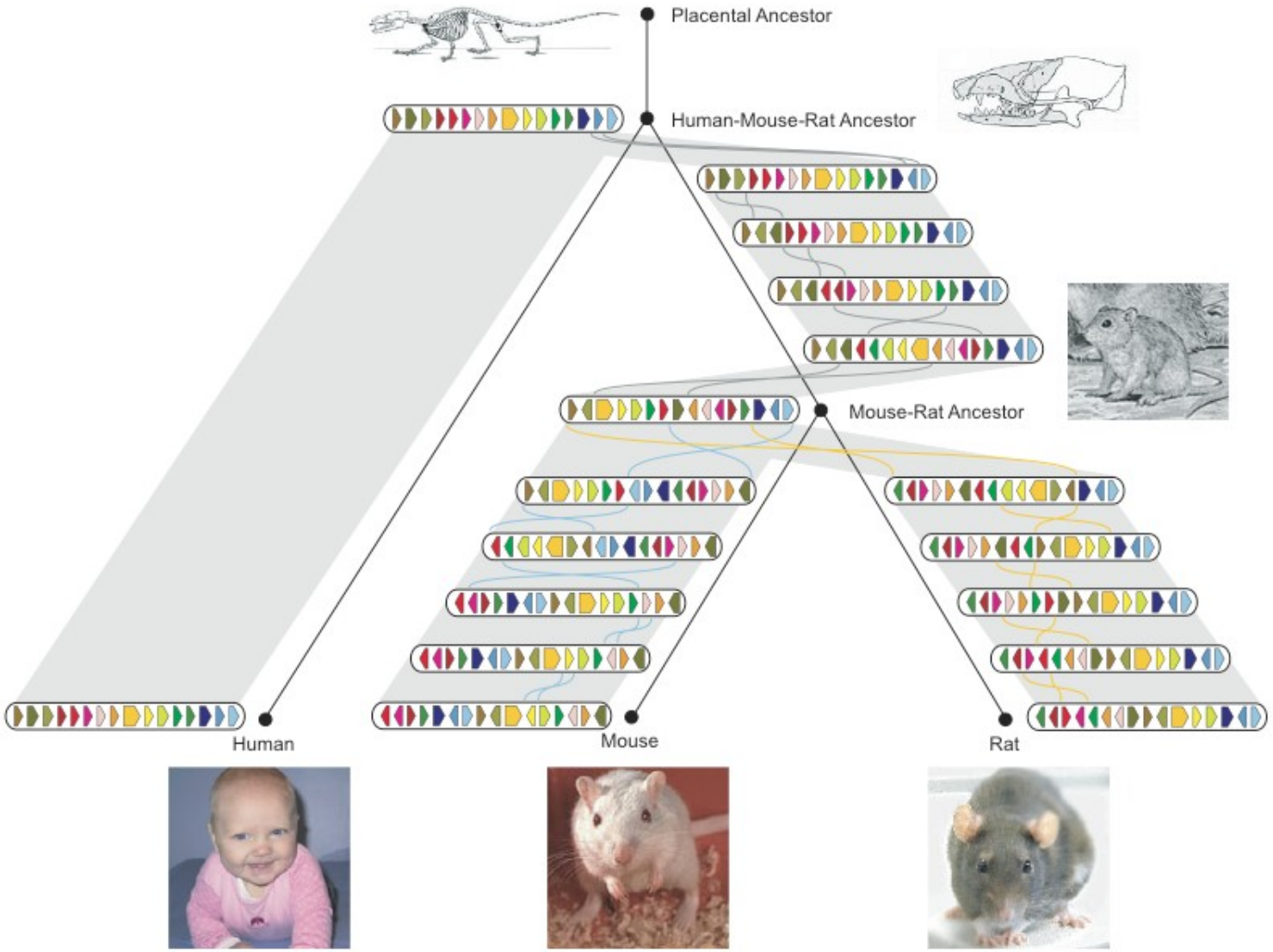
- Construct DeBruijnGraph
- Simlify DeBruijnGraph
- Discover synteny blocks

Biological Motivations

- Genome 10K Project's Goal: Assemble genomes of 10000 vetebrate species
- Inspired by Genome 10K, the i5k initiative to sequence 5000 insect genomes.
- Questions: How are these species related?

- *None of the current comparative genomics tools can work with such a high number of genomes*.

History of Chromosome X



Rat Consortium, *Nature*, 2004

Previous Work

- Breakpoint Graph and Ancestral Genome Reconstruction (Max Alekseyev and Pavel Pevzner) Genome Research 2009
- Reconstructing contiguous regions of an ancestral genome (Jian Ma et. al.) Genome Research 2006
- DRIMM-Synteny: decomposing genomes into evolutionary conserved segments (Son K. Pham and Pavel Pevzner) Bioinformatics 2010

Problem formulation

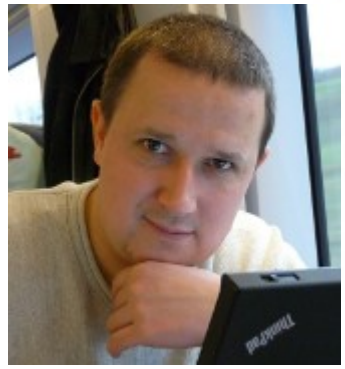
- Input: Given a set of current genomes (in the alphabet of genes), and a phylogenetic tree of these species.
- Output: Anscentral Genome and an order set of reliable rearrangements that transforms each genome to the ancestral genome.

Limitation

- Any methods requires to compute synteny blocks for all genomes as an initial step.
- Synteny blocks --- Highly Conserved Blocks.
- But as soon as more sequences are added, the current synteny blocks reconstruction algorithms will output synteny blocks with very small length with various multiplicity.
- Current methods for ancestral genome rearrangment cannot work with these ”low quality” synteny blocks.

Proposed Solution

- Combine genome rearrangments algorithm (Multiple Breakpoint graph) to synteny blocks reconstruction.



Academic Project

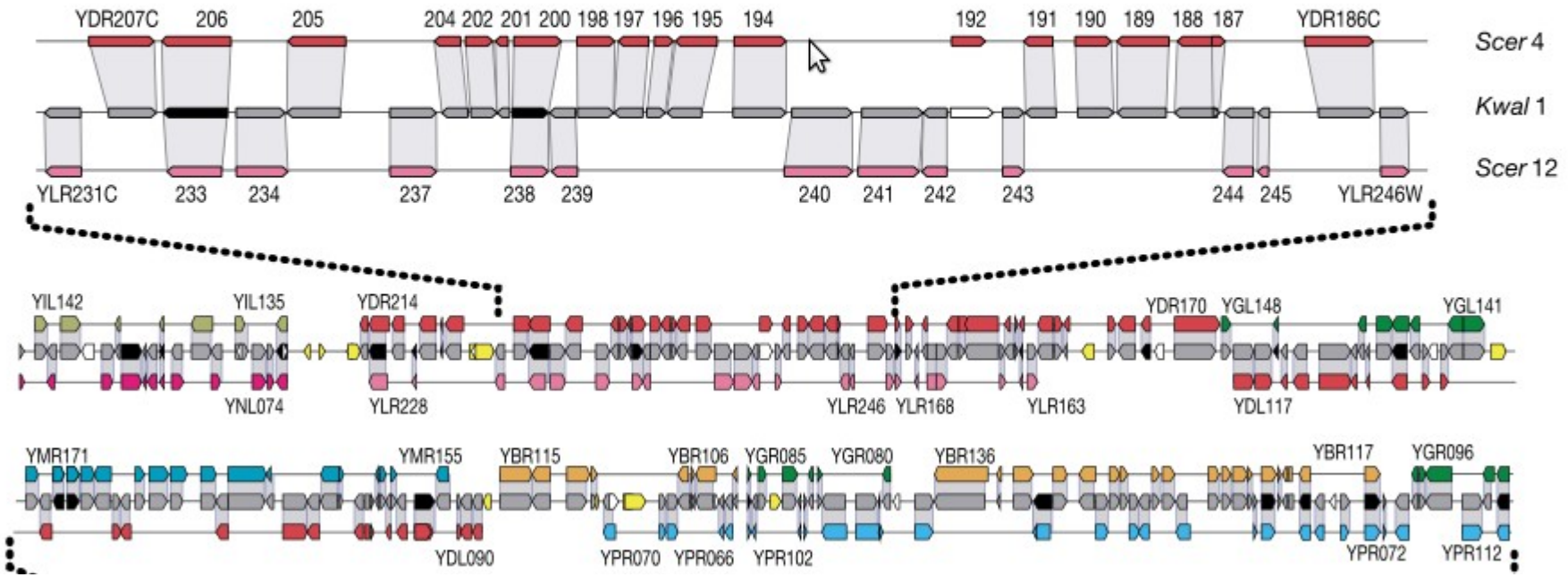
- Develop a Synteny Blocks software from sequences **in the alphabet of genes** that has ability to plug-in Genome Rearrangement Algorithm Module.

Synteny Blocks – Problem Formulation

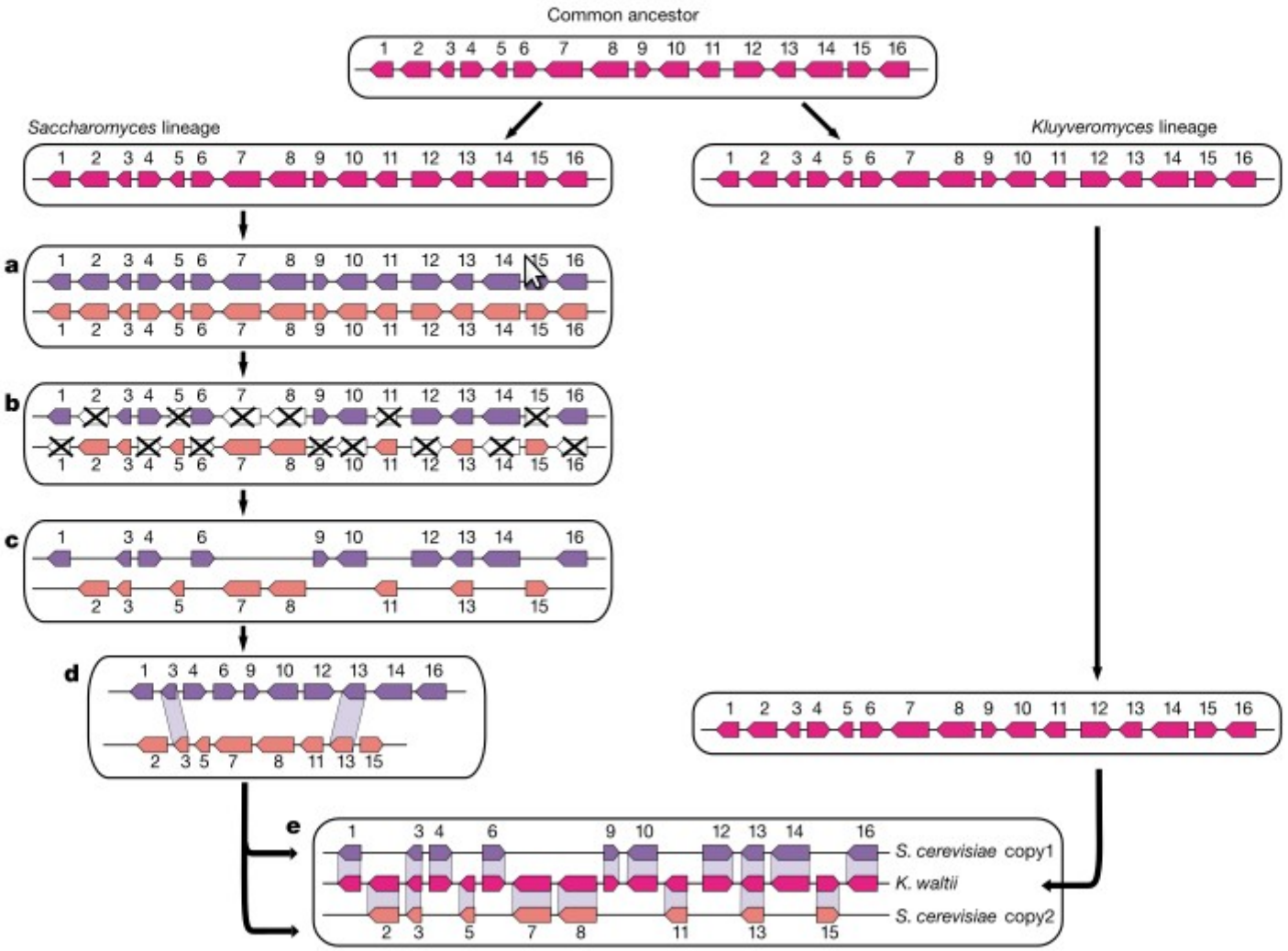
- Sequence modifying algorithm:
- Let d(S,S') be the minimum number of edit operations (e.g. insertions/deletions/substitutions of letters or short substrings) to transform a string S into a string S' . We define the SMP as follows:
- SMP: given a string S and a parameter girth, find a string S with minimum d(S,S') among all strings such that AB(S') has no cycles shorter than girth.
- Where AB(S) is the A-Bruijn graph constructed from sequence S.

Whole Genome Duplication

- Main problem - to find synteny blocks in genomes after extencive duplications followed by massive gene losses.

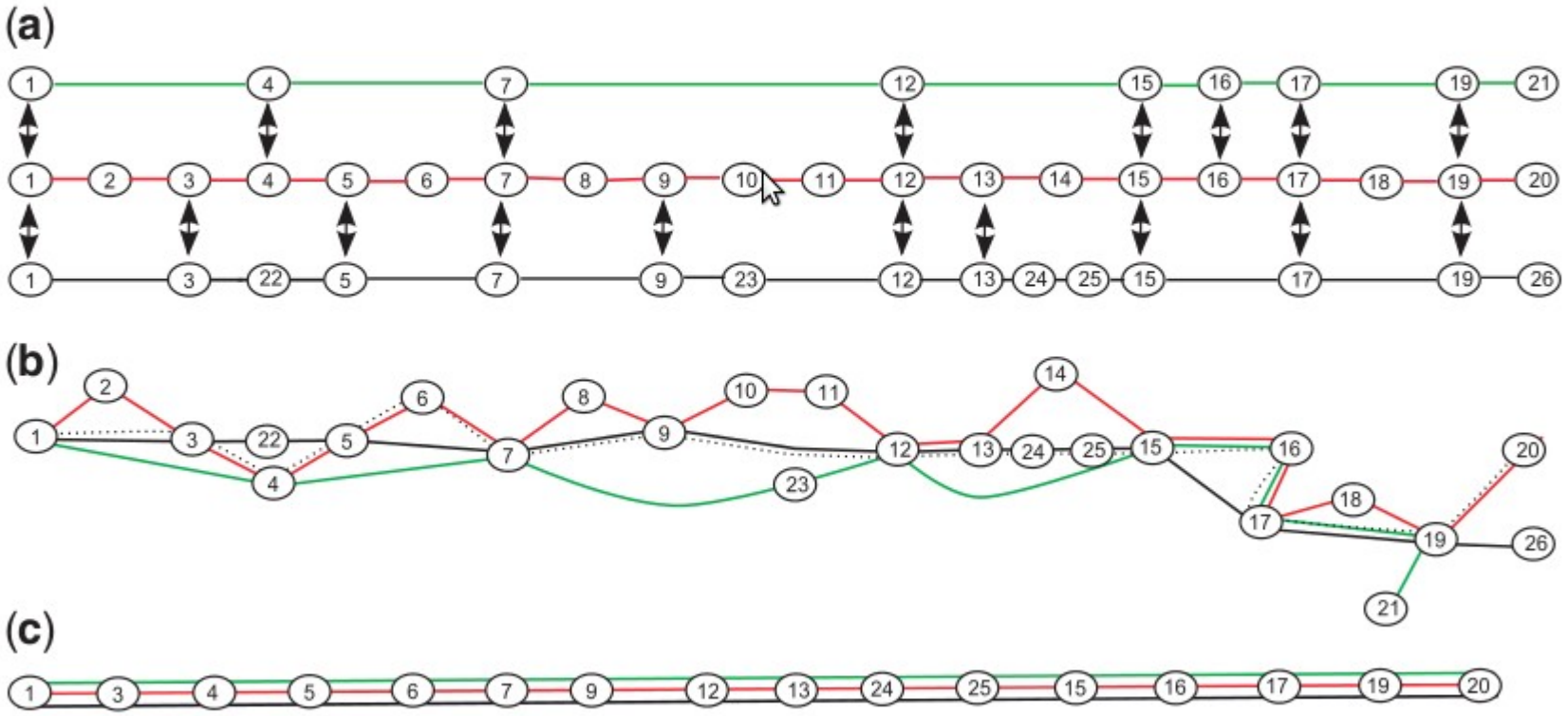


Synteny Blocks after Whole Genome Duplication

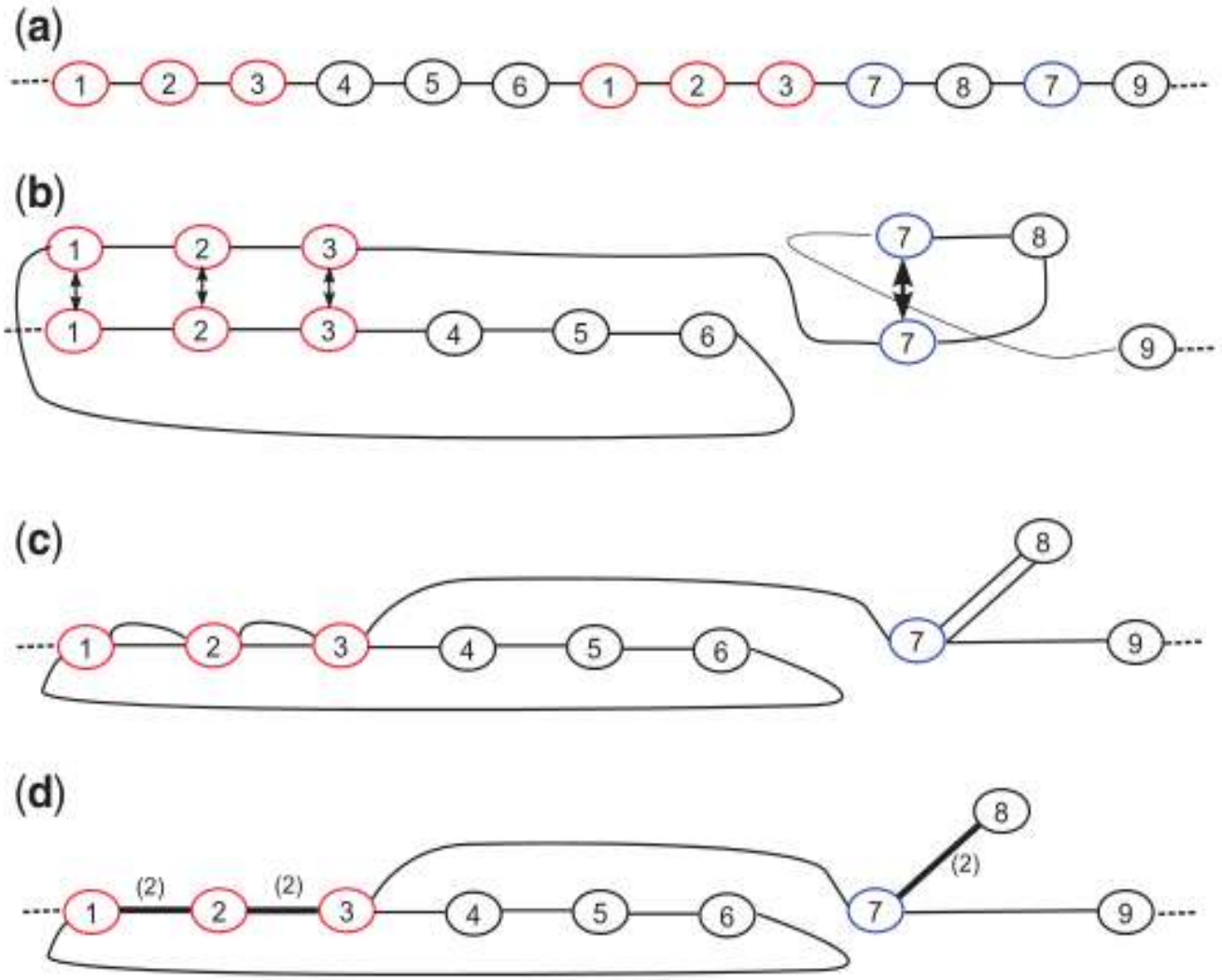


Solution

- Implement sequence modication to discover synteny blocks.

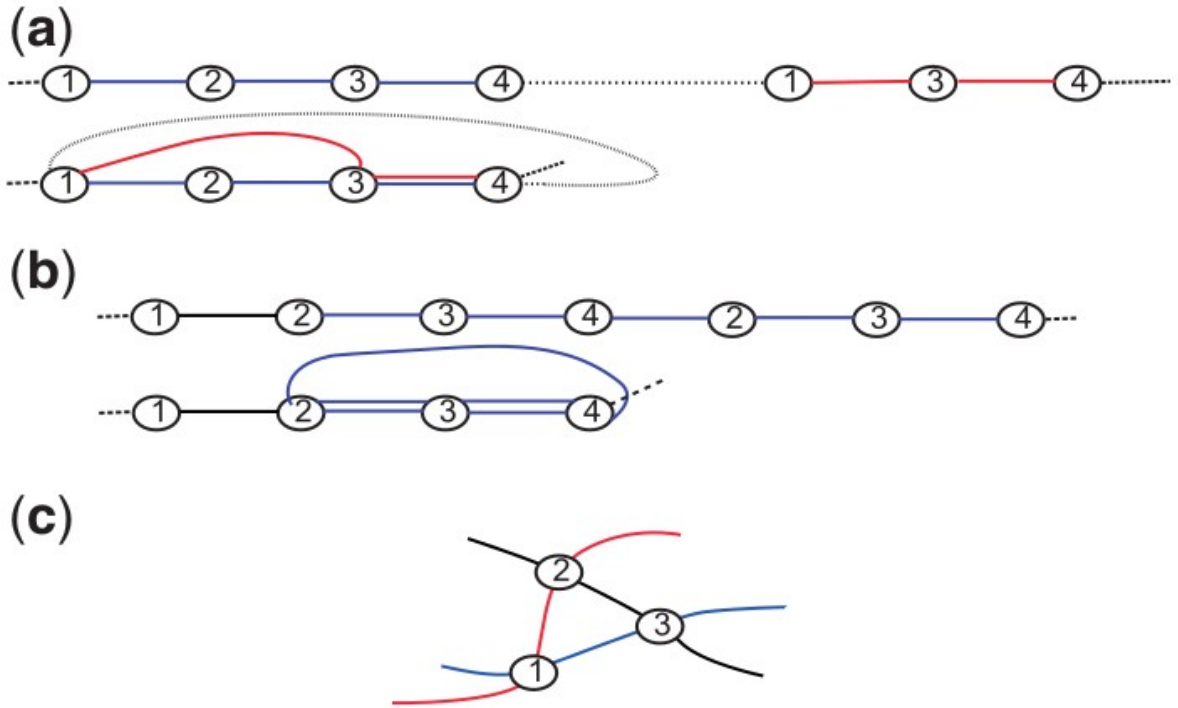


A-Bruijn Graph Construction



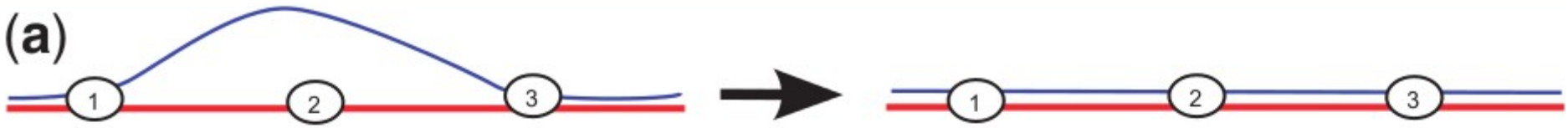
A-Bruijn Graph Simplification

- ABruijnGraph cycles
  - Two-way cycle is formed by two paths P' and P"
  - One-way cycle is formed by one path P
  - Composite cycle is formed by three and more paths



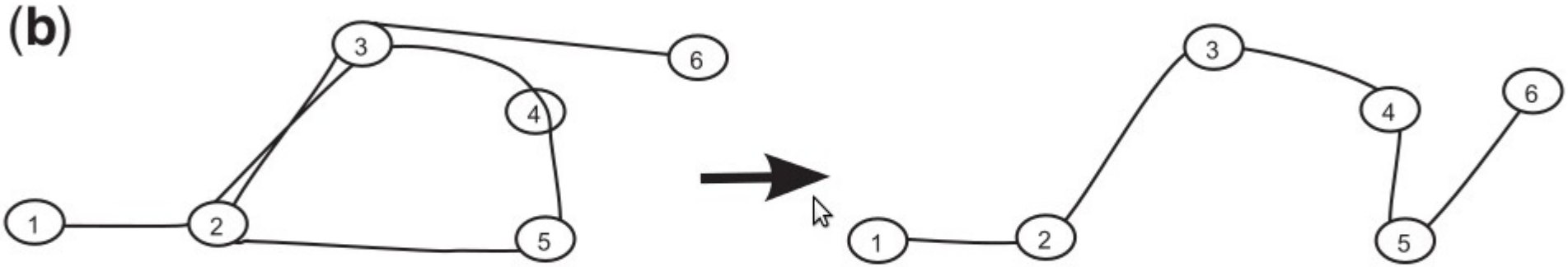
A-Bruijn Graph Simplification

- Detour two-way cycles



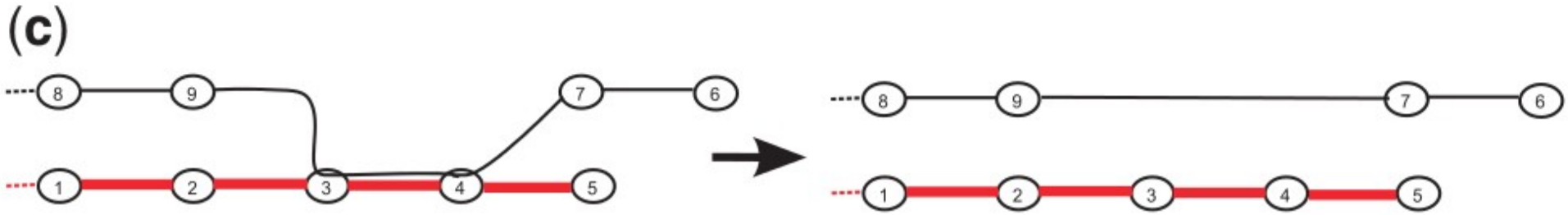
ABruijnGraph Simplification

- Shortcut a one-waycycle
- A one-way cycle formed by a path P = (vin ,...,u,vin ,...,vout), where vin and vout are the first and the last vertices of P. P-transformation substitutes every instance of path P by a shorter path (vin ,...,u)



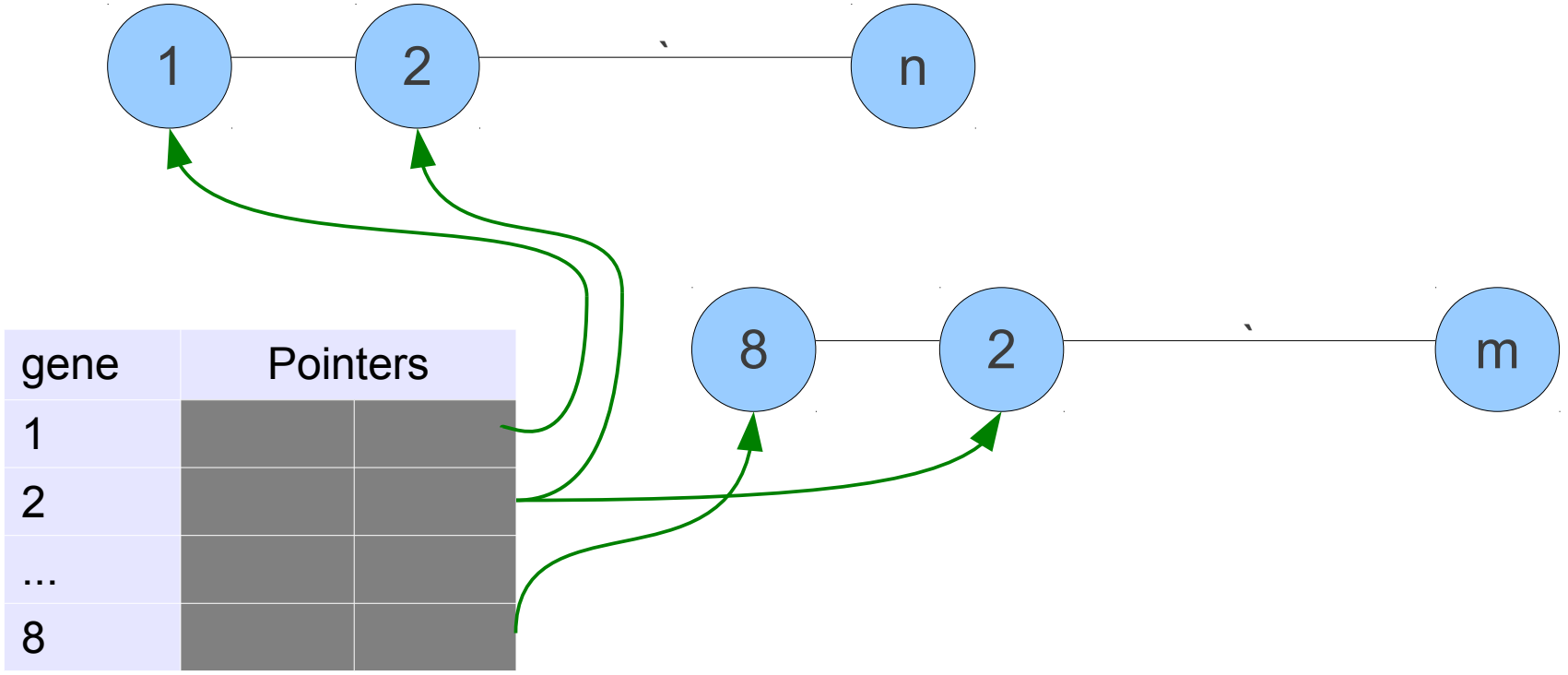
ABruijnGraph Simplification

- Path splitting eliminates spurious similarities



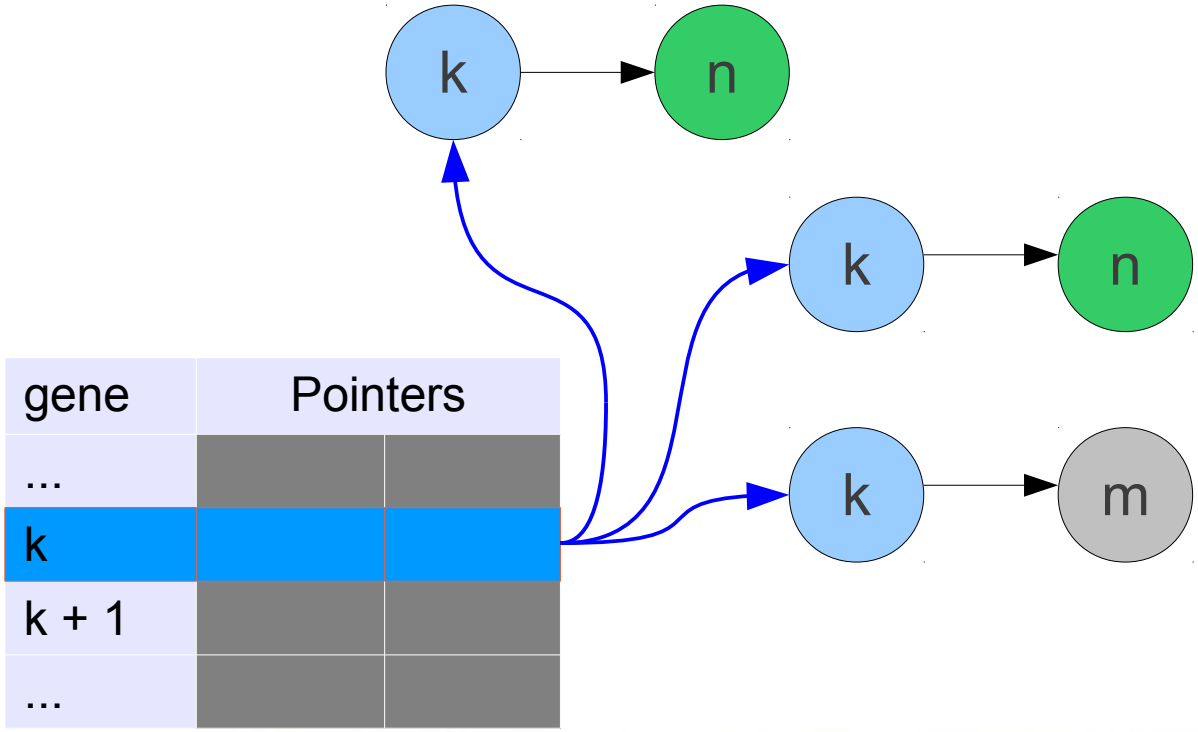
Simplification Algorithm

- Chromosomes are std::list of genes
- Graph is a google::dense_hash_map of gene pointers std::list



Simplification Algorithm

- Checking for branch
- Branch is a gene followed by different genes in different chromosomes.

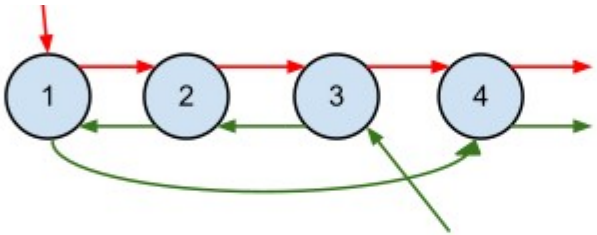


Simplification algorithm

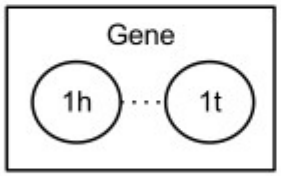
```
for all chromosomes:
        for all genes:
                if check_for_branch()
                        if find_cycle()
                                rerout_cycle()
```

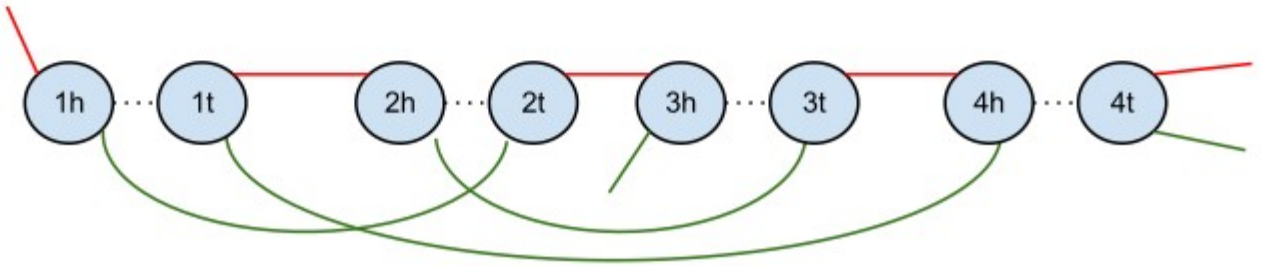
Simplification

- Graph for simple dataset.



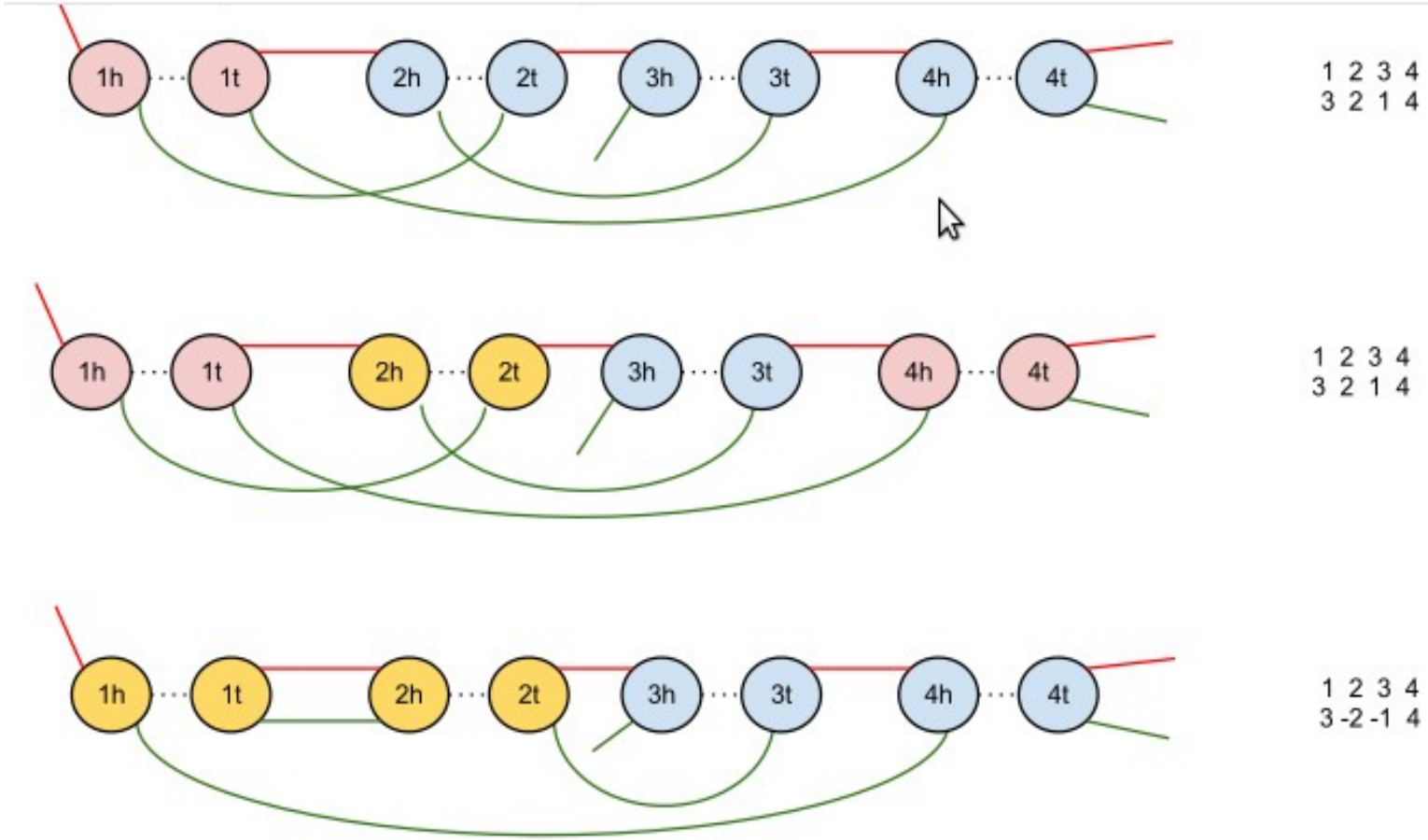
1 2 3 4
3 2 1 4





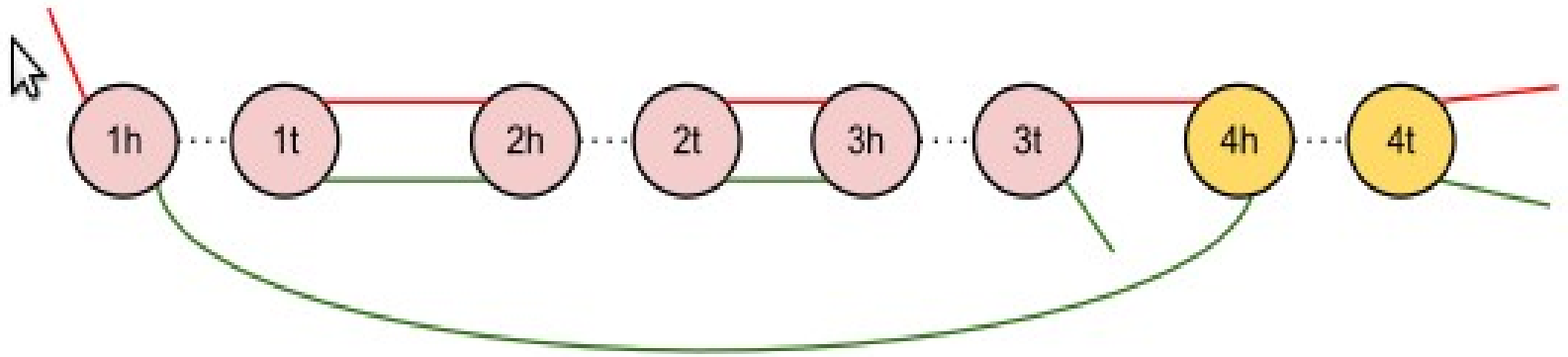
Simplification

- One step of simplification



Simplification

- Output



1h 1t 2h 2t 3h 3t 4h 4t
3t 3h 2t 2h 1t 1h 4h 4t

1 2 3 4
-3 -2 -1 4

Results

- Synteny blocks in K.waltii and S.cerevisiae:
  - The concatenation of S.cerevisiae (S) and K.waltii (K) results in a genome with 10686 genes (6240 unique genes).
  - After simplification concatinated genome has 9674 genes (6070 unique genes).
  - Coverage: 0.70829

Statistics

- Syntenies count (min synteny length = 2): 339

  Coverage: 0.70829
- Syntenies count (min synteny length = 3): 255

  Coverage: 0.660947
- Syntenies count (min synteny length = 4): 208

  Coverage: 0.618772
- Syntenies count (min synteny length = 7): 117

  Coverage: 0.493384
- Syntenies count (min synteny length = 11): 63

  Coverage: 0.356213

Thank you!