Whole Genome Comparison: Project Presentations

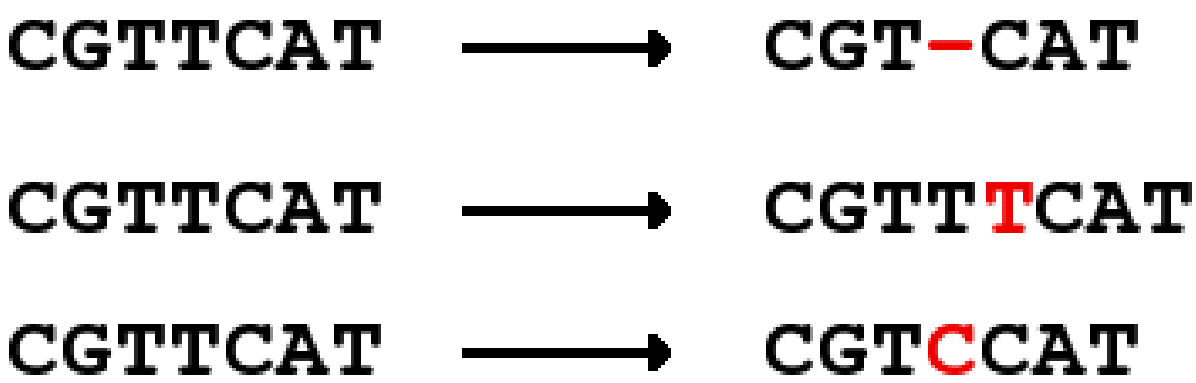
Felix Heeger, Max Homilius, Ivan Kel, Sabrina Krakau,
Svenja Specovius, John Wiedenhoeft

July 19, 2010

Outline

Evolutionary events

Nucleotide deletion, insertion and point mutation



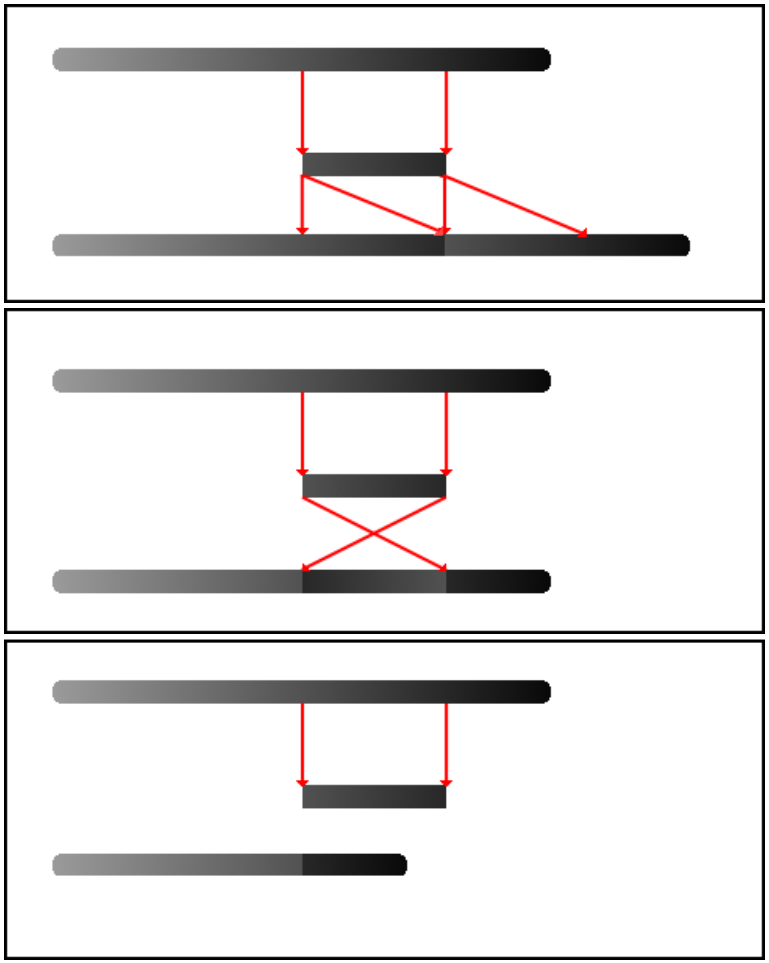
Collinear alignment

Columns of aligned sequences

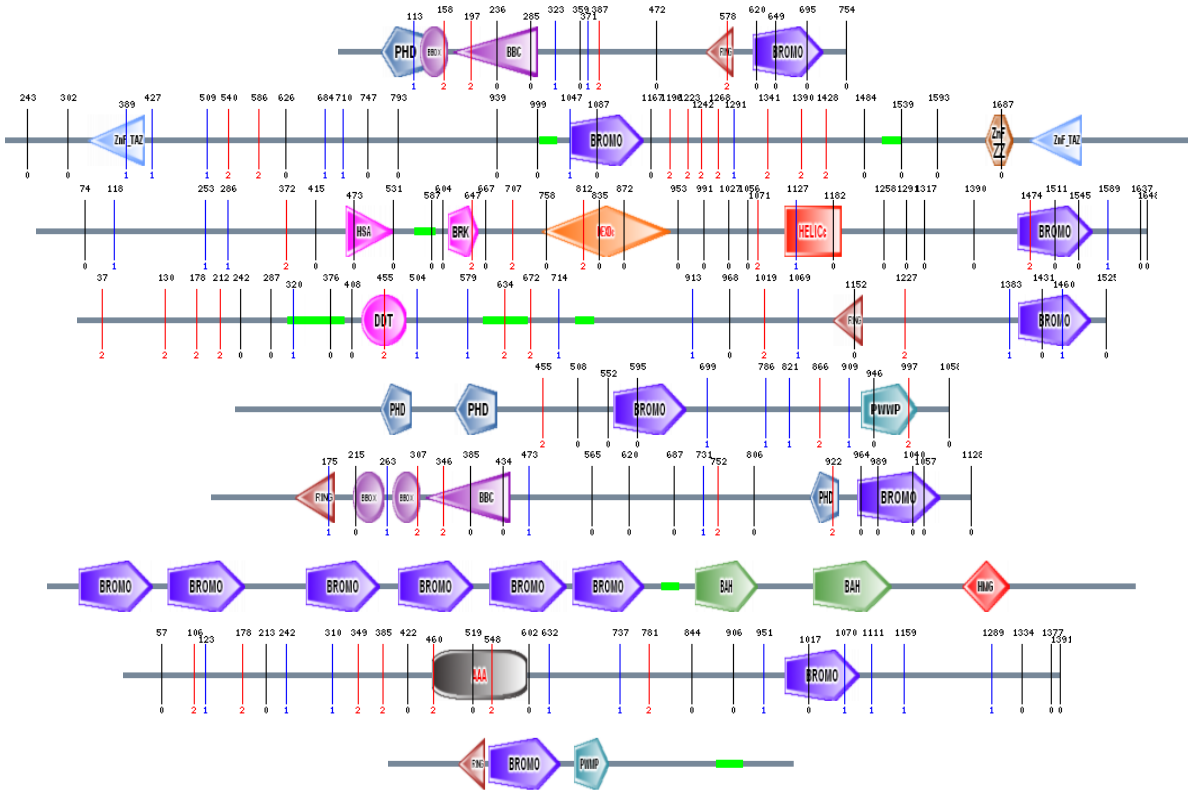
```
CONSENSUS       a.gttcctgc.tgcgtttgctggactgatgtactt.gtttgtgagg.caa
Hs#S663801      a.gttcctgc.tgcgtttgctggacttatgtactt.gtttgtgagg.caa
Hs#S337687      aagttcctgc.tgcgtttgctggactgatgtacttggtttgtgnaggcaa
Hs#S629177      a.gttcctgc.tgcgtttgctggactgatgtactt.gtttgtnagg.caa
Hs#S672957      a.gttcctgc.tgcgtttgct..........................
Hs#S672182      a.gttcctgc.tgcgtttgctggactgatgtactt.gttt.......
Hs#S674099      a.gttcctgc.tgcgtttgctggactgatgtactt.gtttgtgagg.caa
Hs#S196113      a.gttnctgn.tgngtttgctggactgatgtactt.gtttgtgagg.caa
Hs#S994400      ..........................gtacnt.gtttgtgagg.cta
Hs#S80460       a.gttcctgc.tgcgtttgctggactgatgtactt.gtttgtgagg.caa
Hs#S1988018     a.gttcctgc.tgcttttgctggactgatgtactt.gattgtgagg.caa
Hs#S1794113     a.gttcctgc.tgcgcttgctggactgatgtactt.gtttgtgagg.caa
Hs#S4698        a.gttcctgc.tgcgtttgctggactgatgtactt.gtttgtgcgg.caa
Hs#S813765      a.gt.cctgc.g.cgtttgc.ggacggatgtactt.gtt.gtgagg.caa
Hs#S1184845     .........................................g.caa
Hs#S1577463     ..........................................gg.caa
Hs#S914987      ......................ctgatgtactt.gtt.gtgagggcaa
Hs#S1985364     a.gttcctgc.tgcgtttgctggactgatgtactt.gtttgtgagg.caa
Hs#S1465644     ..gttc.tgcctgcgtttgctgaactgatgtactt.gttagt.aag.caa
Hs#S1850471     c.gttactgc.ggggtttgctggactcatg.actttgttngt.agg.caa
```

More evolutionary events

Genome rearrangements: duplication, reversal and deletion of segments

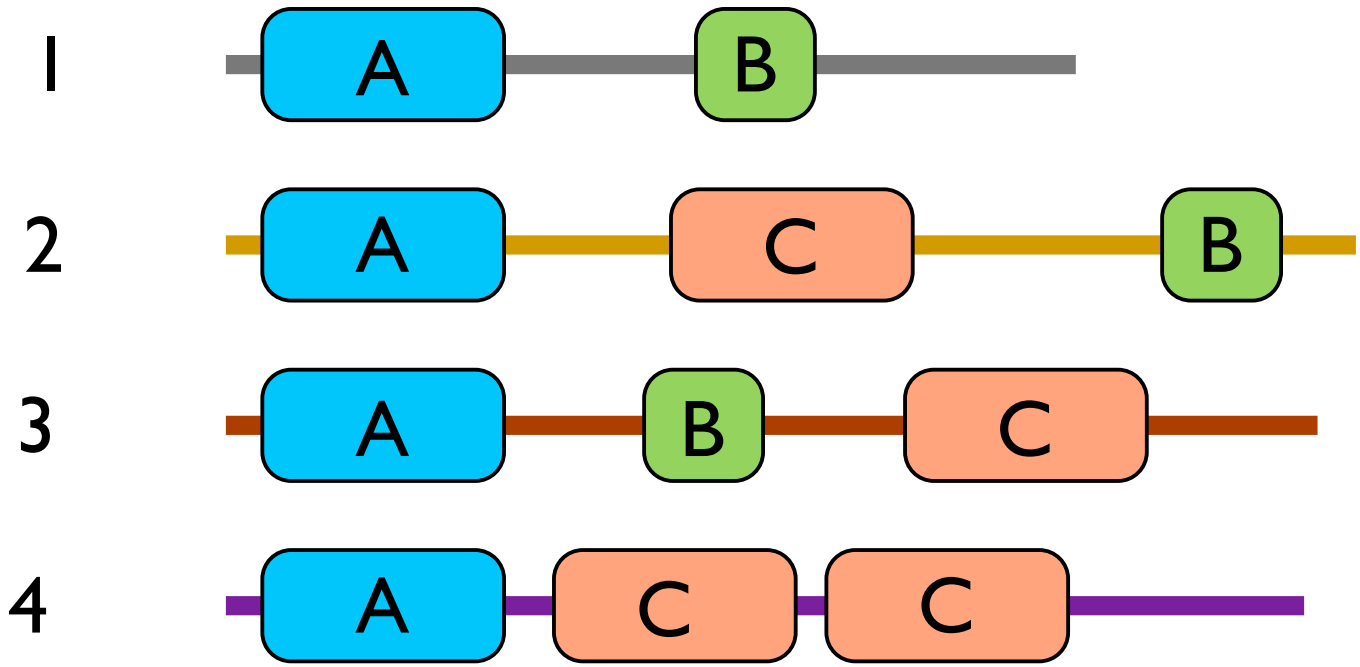


Multidomain proteins

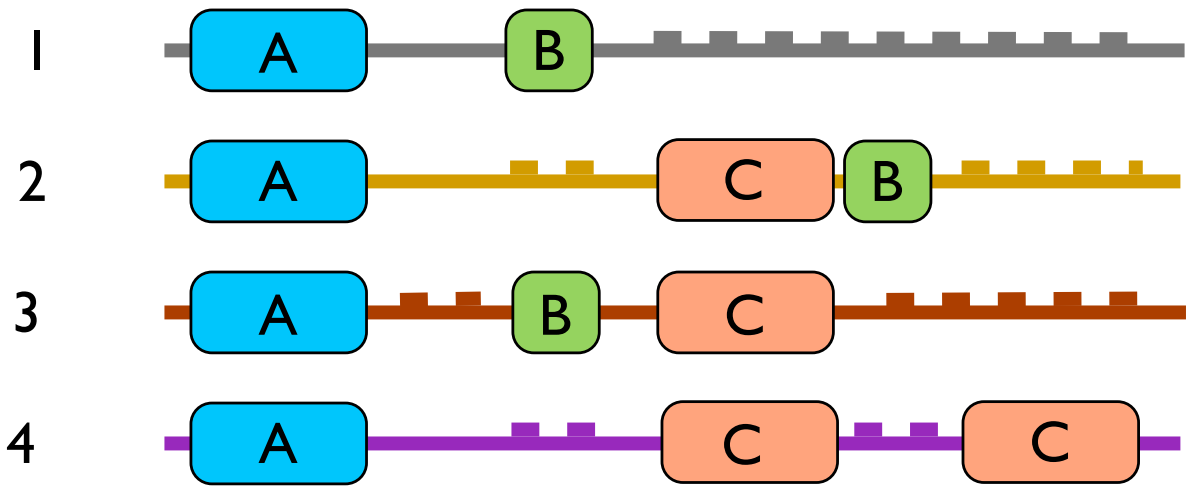


- Diverged by rearrangements of modular units, e.g. domains
- Multidomain proteins (MDPs) difficult to align collinearly

Multidomain protein toy example

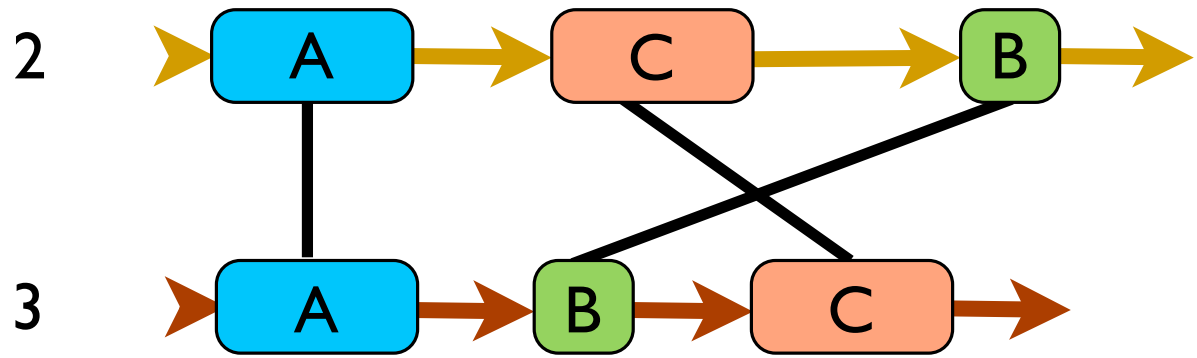


Collinear alignment



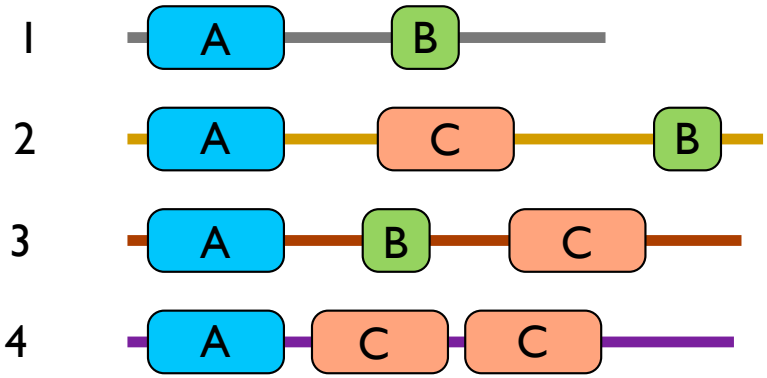
- It's not possible to align all similar domains without reordering

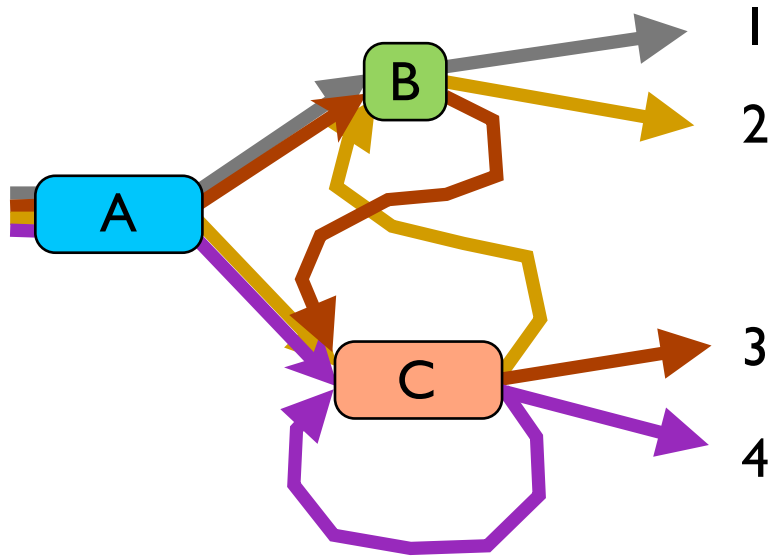
Graph representation of alignments



- Arcs: input sequences
- Edges: matches
- Some edges may be inconsistent: mixed cycles

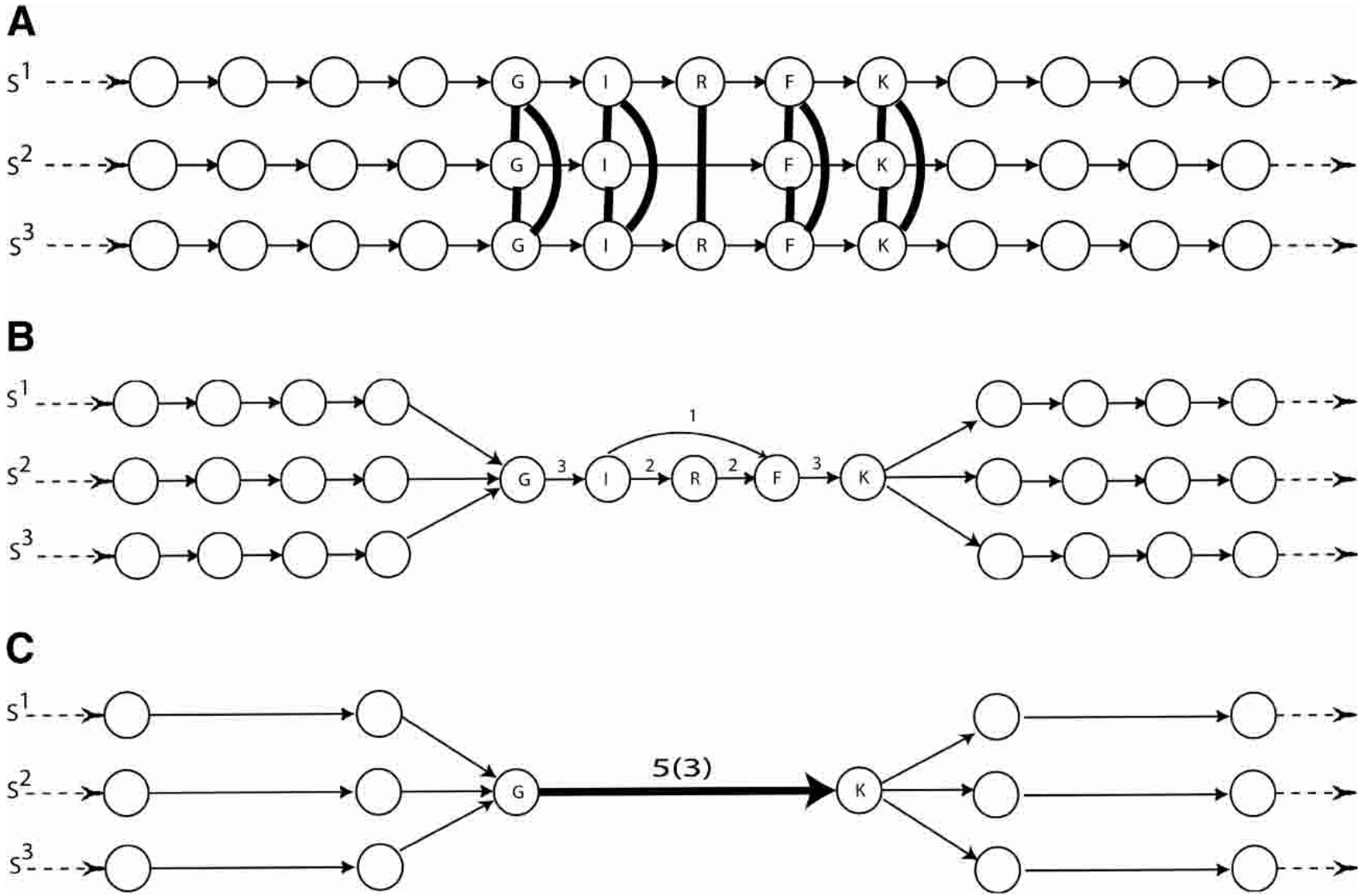
Non-collinear alignment



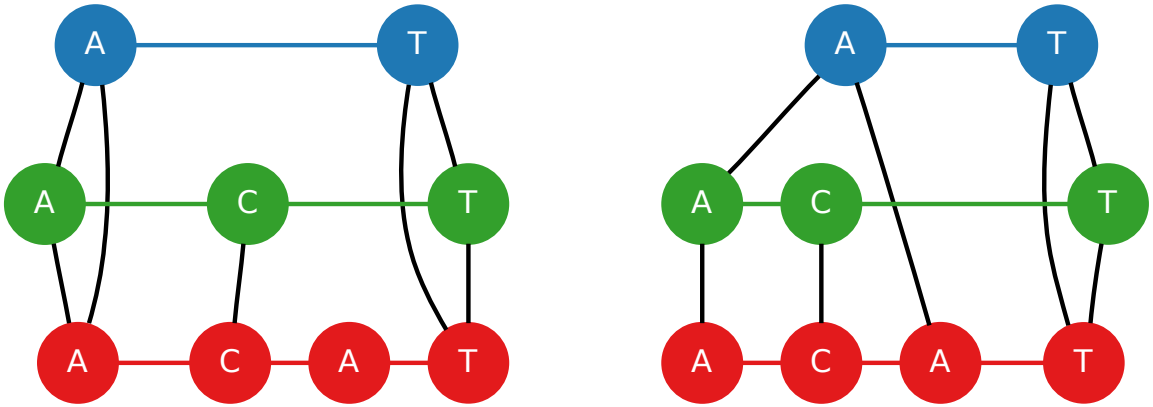


- Allow *large* cycles of *similar* segments

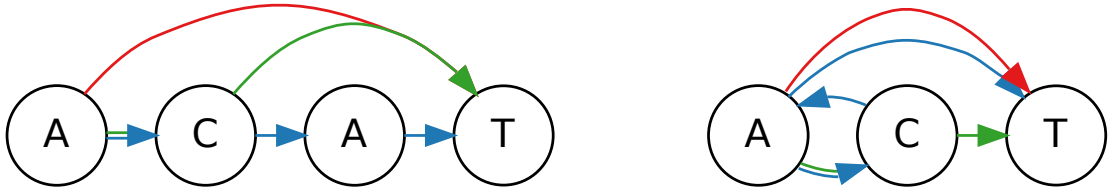
Construction of the A-Bruijn Graph



Whirls and inconsistencies



A-graph



A-Bruijn graph

Evaluating ABA

J. Wiedenhoeft

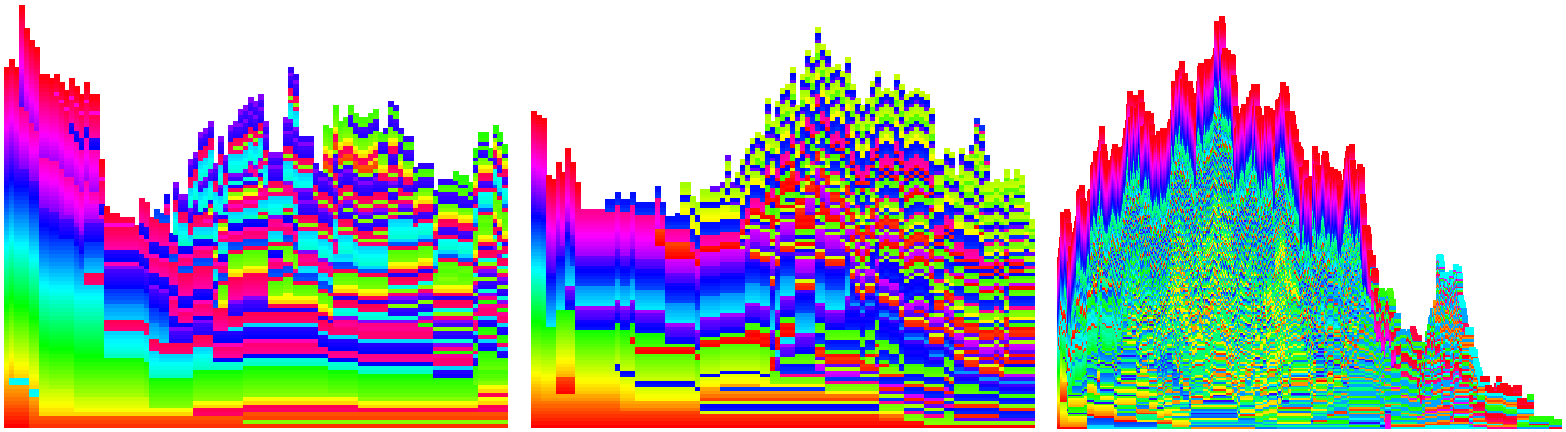
- simulate sequence evolution using PAM (*point accepted mutation*)
- two models of sequence evolution
  - geometric duplication/deletion model
  - rearrangement according to fragility model
- true homology can be tracked to provide a gold standard

PAM sequence evolution

- amino acid substitution modeled as a Markov process
- PAM = transition matrix
- using ABA's BLAST subroutine with PAM30 provides a null model of character homology

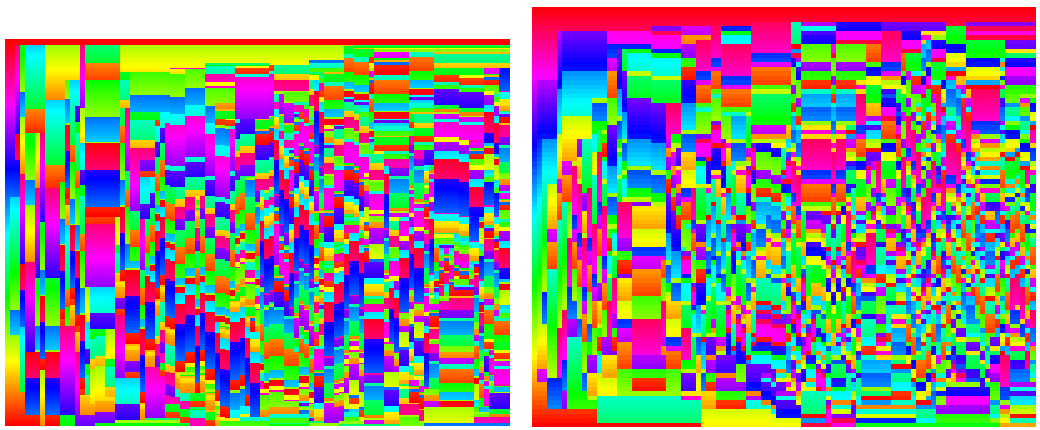
Geometric duplication/deletion model

- pick position by uniform distribution
- determine deletion or duplication by binomial distribution
- determine direction by binomial distribution
- determine length by geometric distribution



Fragility model

- models only translocations
- successful translocation *increases* the chance of a segment being translocated again $\Rightarrow$ models conservation of substructures
- boundaries weighted by length of substructure
- borders of substructures are preferred as insertion spots $\Rightarrow$ prevents disruption of other substructures

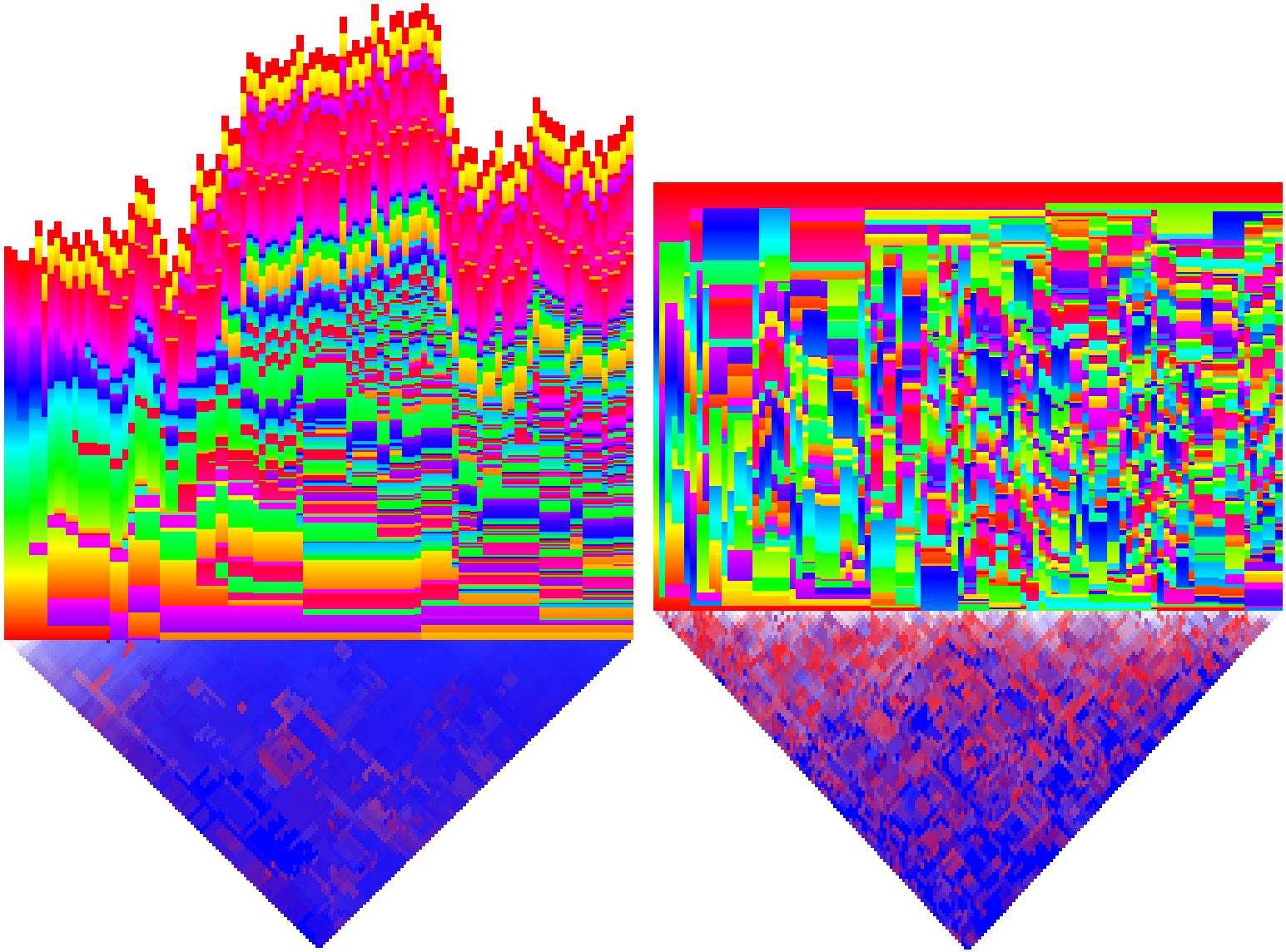


Score

- true negatives are vast due to the low number of paralogs and the alignment bias (BLAST)
- hence precision and accuracy are not suitable measures

$$\frac{\mathrm{FP} + \mathrm{FN}}{\mathrm{FP} + \mathrm{FN} + \mathrm{TP}}$$

Results



Analysing Multidomain Proteins with ABA

M. Homilius

- Noncolinear alignment applied on multidomain proteins (MDPs).
- Histone Deacetylation / Chromatin Remodeling Complexes.

HATs / CRCs

Regulation of gene expression.

Function of chromatin-remodeling complexes

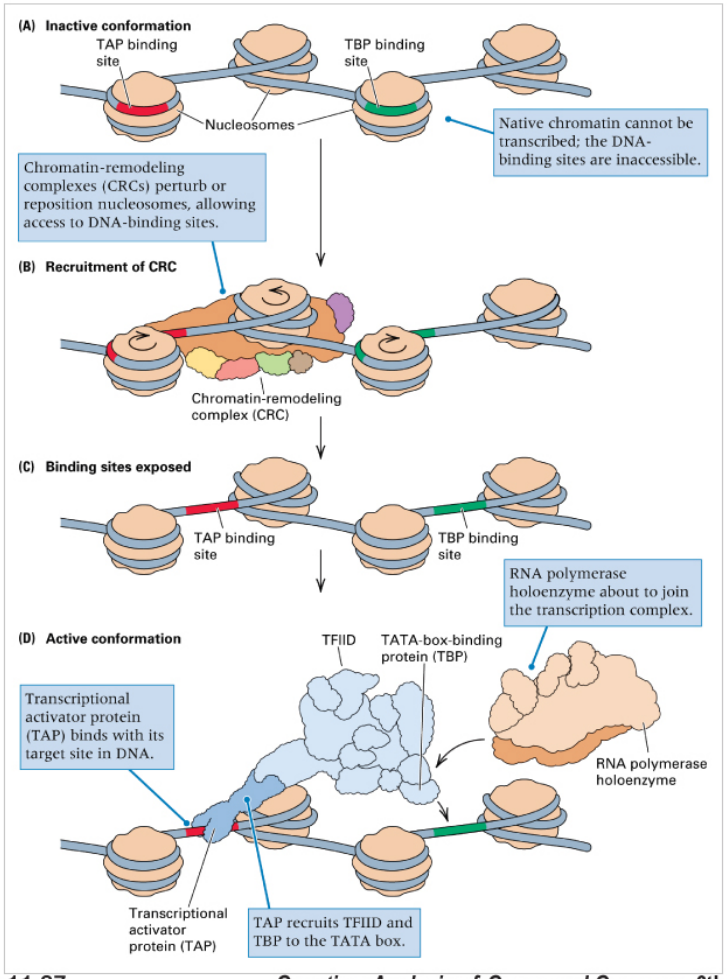


Figure 11.27 *Genetics: Analysis of Genes and Genomes*, 6th Edition
Hartl, Jones

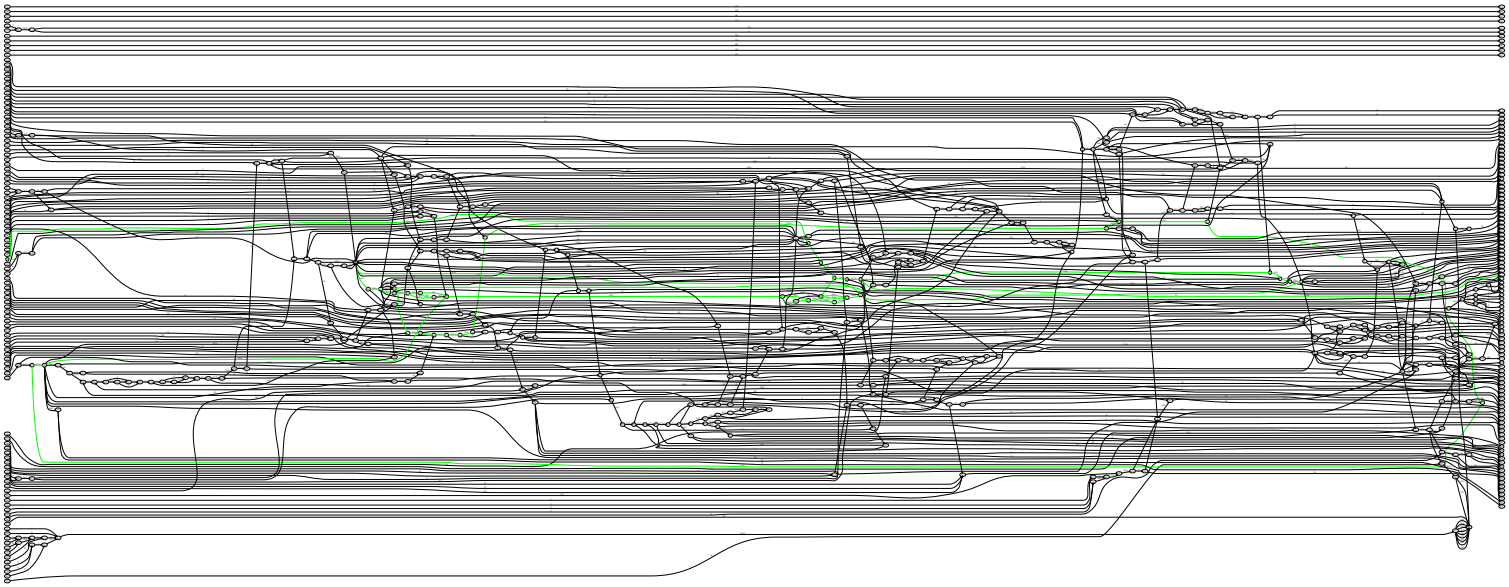
Dataset

- 262 proteins found in literature and manually annotated.
- Thanks to Sebastian, Ivan and Christoph!
- From *S. cervisiae*, *S. pombe*, *D. melanogaster* and *H. Sapiens*

Questions

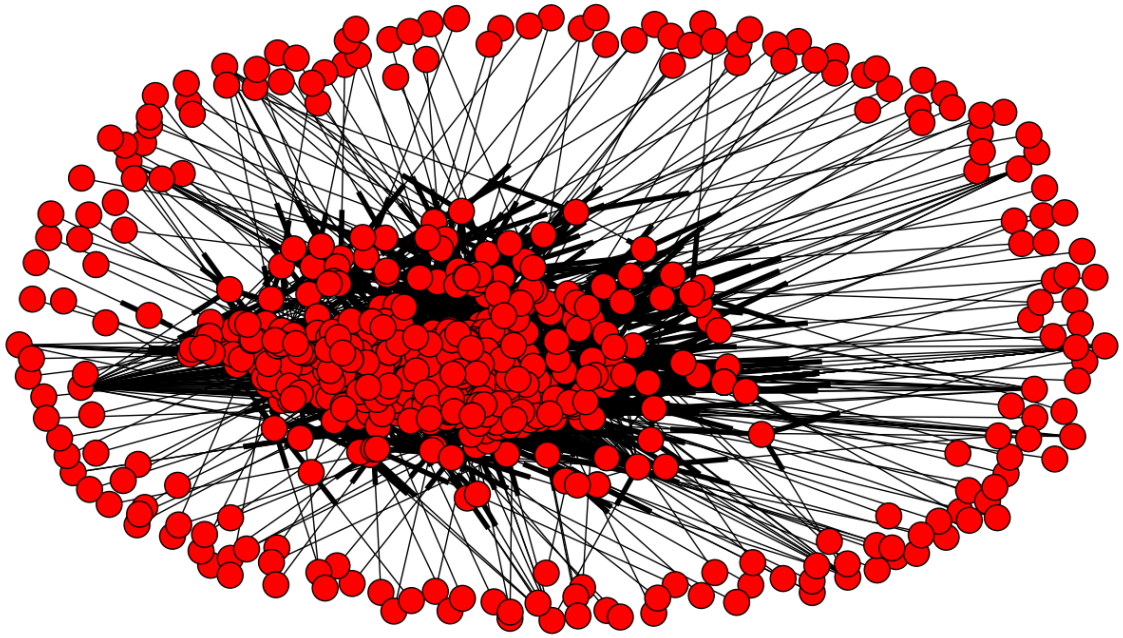
- Can ABA recognize domain-like structures?
- Do domains move around in the complexes?
- What structures occur often?

Output of ABA



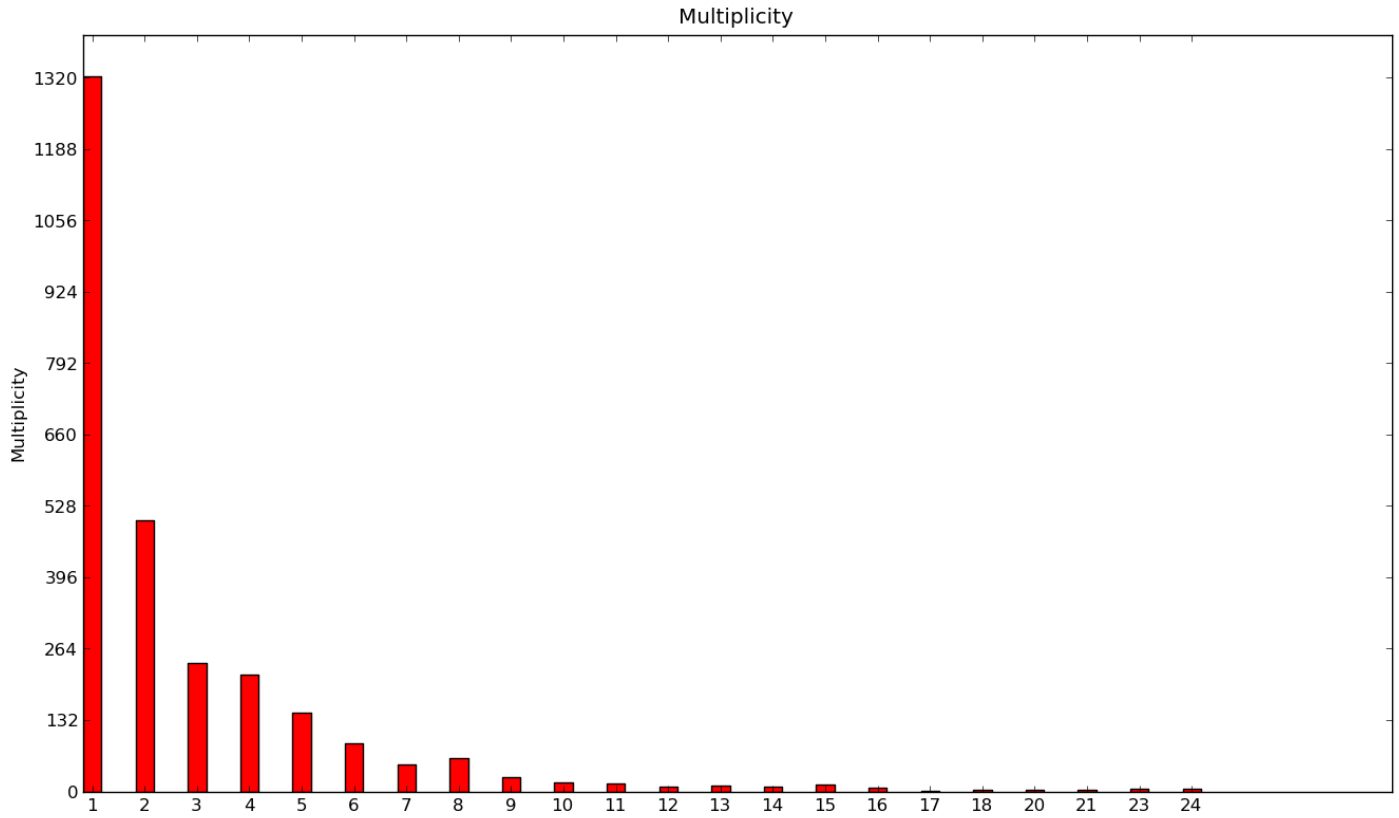
- Applied to only 2 species.
- Rendering takes a long time.
- Hard to interprete (manually).

Parsing output of ABA



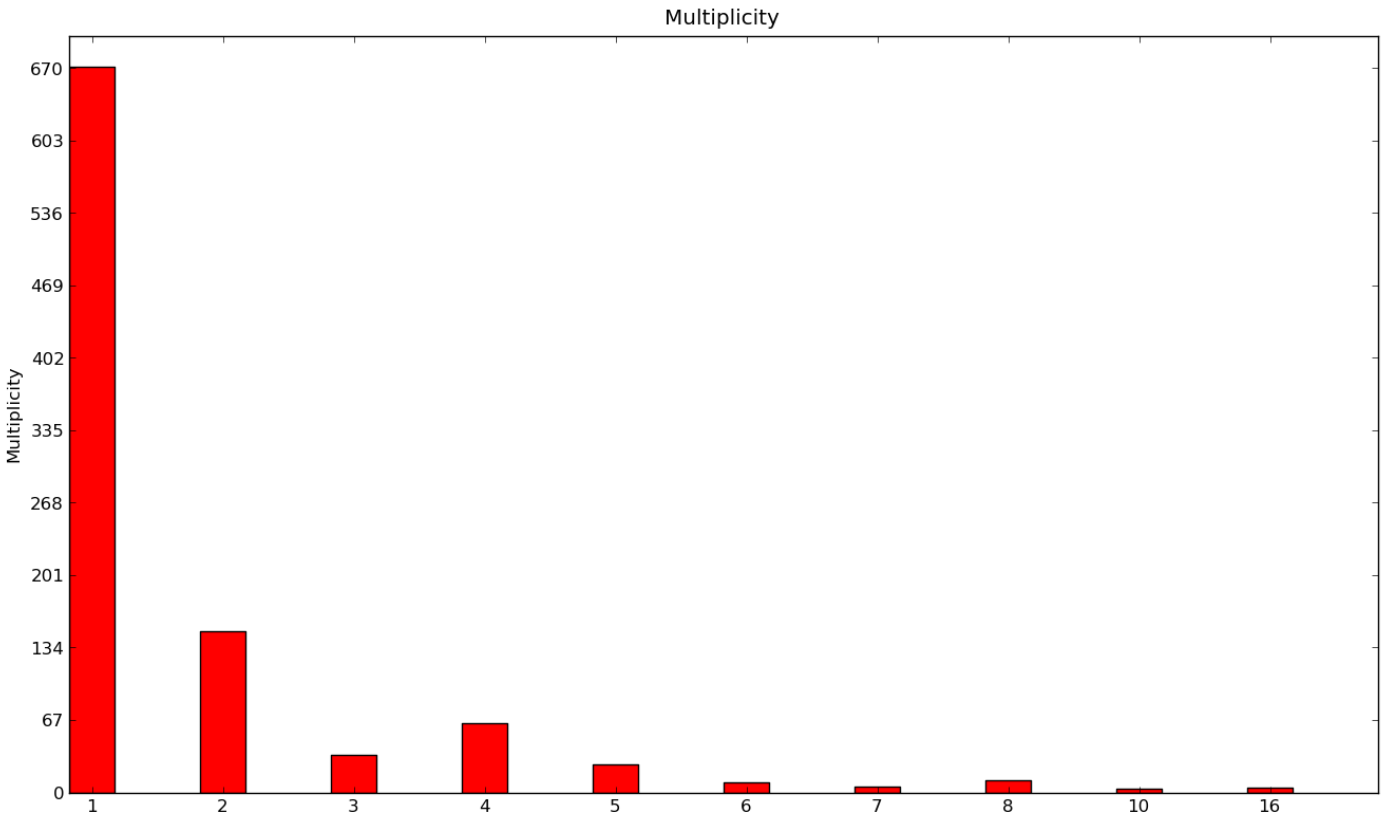
- Applied to 4 species.
- Reconstructed A-Bruijn Graph from ABA-Output.

Distribution of edge multiplicity



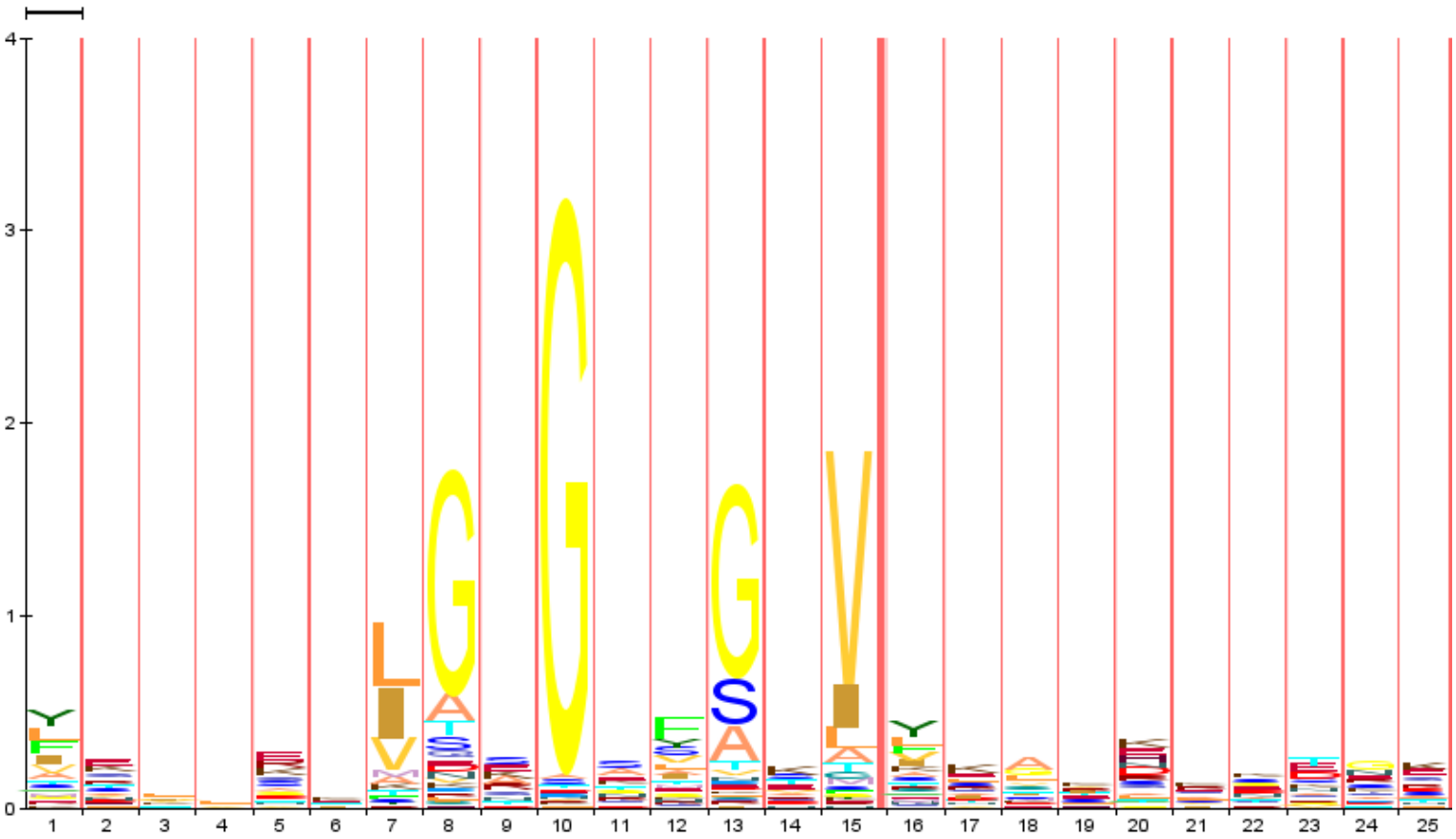
- High-weight edges point out to conserved and repeated elements.
- Within and across proteins.
- (Girth parameter did not seem to work.)

Distribution of edge multiplicity (filtered)



- Filtered distribution of the multiplicity of edges (length > 40).

Comparison with PFAM-Annotation

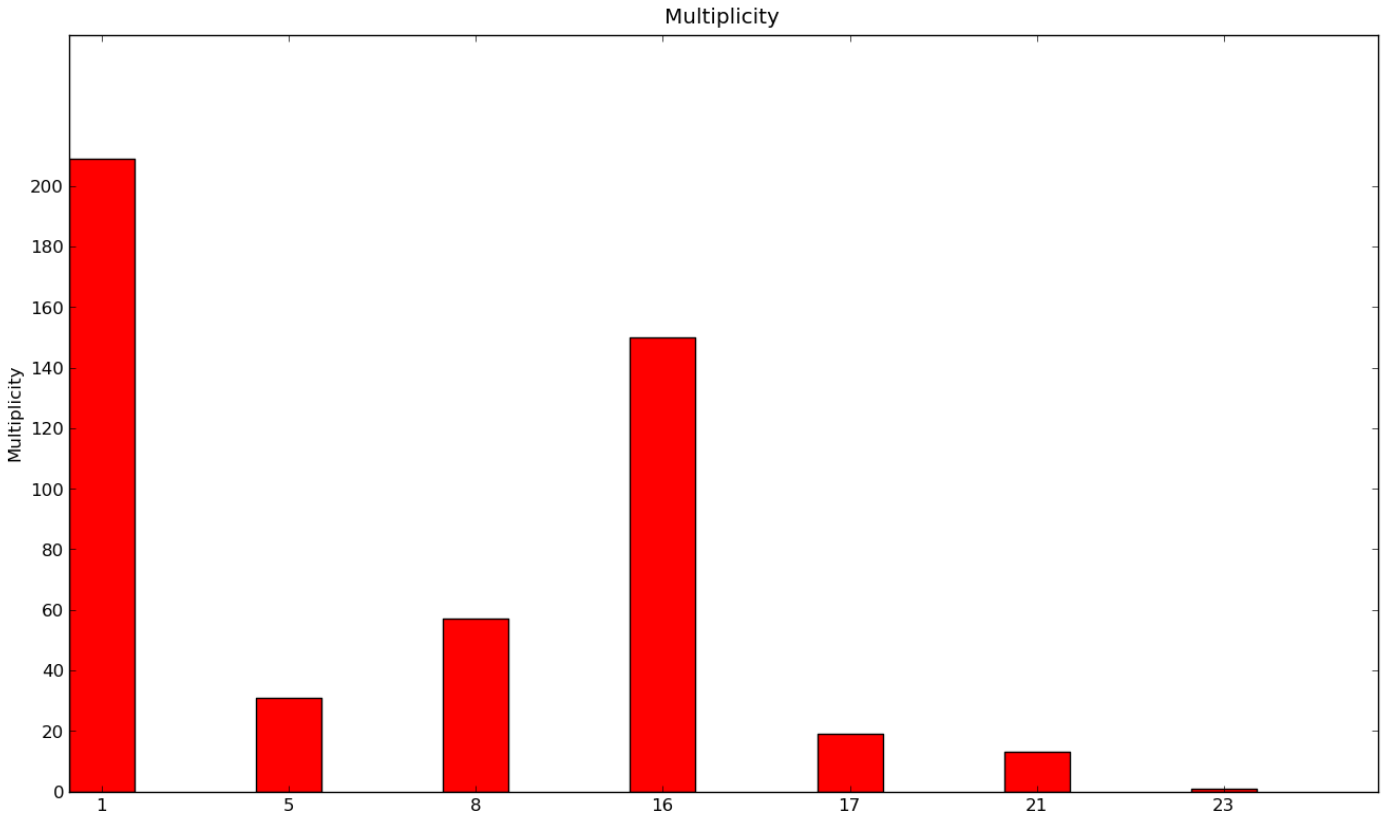


- Hidden markov models learned from multiple sequence alignments.

Comparison with PFAM-Annotation

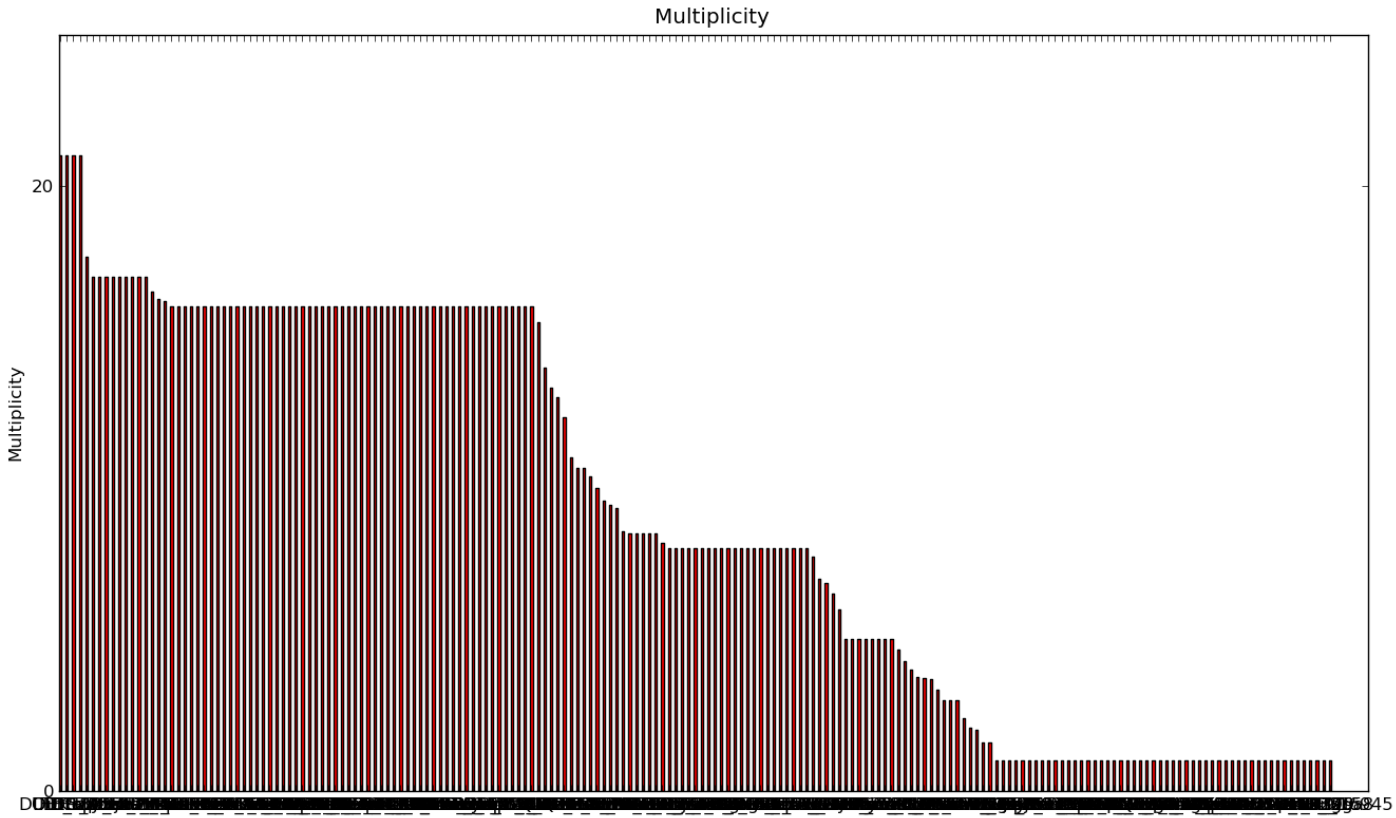
- Annotated all proteins with PFAM/HMMER.
- Detected 561 domains (not unique).

Distribution of edges with domains



- $\approx$ 210 edges of multiplicity 1.
- $\approx$ 150 edges of multiplicity 16.

Repeated domains



- Domains seem to share edges in ABA-graph.

Repeated domains

Domain	Average Multiplicity
DUF1679	21.0
Elf1	21.0
DUF1825	21.0
Fib_alpha	21.0
ZZ	17.7
Otopetrin	17.0
CDK5_activator	17.0
...	...
RFX_DNA_binding	1.0
zfC5HC2	1.0
DUF1542	1.0
Rep_N	1.0
DUF3619	1.0
TIP49	1.0
HTH_Mga	1.0

Whats next?

- Do ABA-edges correlate with found domains?
- Apply real null model. Significance tests.
- Can ABA be used to complement the domains found with HMMER?

Non-Collinear Alignment: Reannotation of genomes.

Carosonella ruddii: an interesting thing

- unclassified γ-proteobacteria. (Like e.g. *E.Coli*)
- Sequenced 2006.

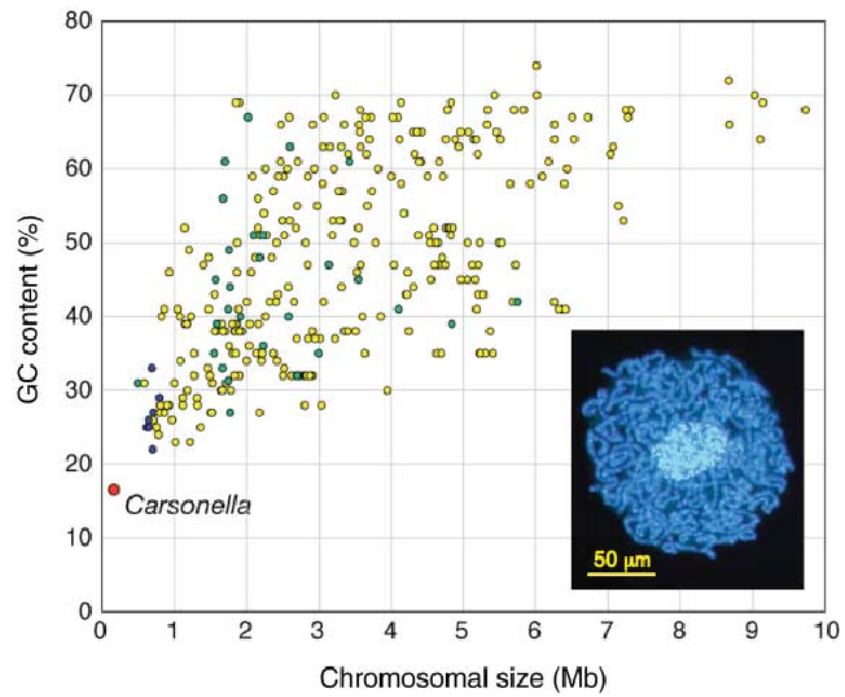
Carosonella ruddii

what is it?

- Smallest bacterial genome known.→ 160 Mb (!). *E.Coli* has 4,5 Gb

Smallest genome before Carsonella

- 362 protein-coding genes in Buchnera aphidicola BCc



Carosonella ruddii

what is it?

- CG-Content: Very low (16%). *E.coli*: (50%)

GC-Content

GC Content is defined as: GC-content (or guanine-cytosine content), in molecular biology, is the percentage of specific bases on a DNA molecule which are either guanine or cytosine.

Carosonella ruddii

what is it?

- CG-Content: Very low (16%). *E.coli*: (50%)
- First annotation: 213 genes. *E.coli*: 4400 genes

Minimal set of genes for life

- : Moya A. et al. proposed 2003 that the minimal gene set for a endosymbiotic life is close to 313.

Interesting question

- DNA replication and repair system is strongly degraded.

Interesting question

- DNA replication and repair system is strongly degraded.
- Transcriptioin machinery is reduced to core subunits of RNA Polymerase (no promotor-recognition)

Interesting question

- DNA replication and repair system is strongly degraded.
- Transcriptioin machinery is reduced to core subunits of RNA Polymerase (no promotor-recognition)
- Translation machinery is highly reduced. (three essential rRNAs are present)

Interesting question

- DNA replication and repair system is strongly degraded.
- Transcriptioin machinery is reduced to core subunits of RNA Polymerase (no promotor-recognition)
- Translation machinery is highly reduced. (three essential rRNAs are present)
- No Shine-Dalgarno sequence present (the way it is defiend)

16S rRNA and Shine-Dalgarno Sequence

- Shine-Dalgarno (SD) is a regulatory sequence strongly involved in translation of bacterial poly-cystronic mRNAs.

Interesting question

Is Carsonella ruddii a living cell?

- 9 aminoa-cyl-tRNA synthetases and 15 out of 50 essential ribosomal protein are **missing** or degraded.

Interesting question

Is Carsonella ruddii a living cell?

- 9 aminoa-cyl-tRNA synthetases and 15 out of 50 essential ribosomal protein are **missing** or degraded.

Two different theories

- C.ruddii is a bacteria which undergoes the change to endosymbiont.
- C.ruddii is an former primary endosymbiont, is being driven towards its extinction and replacement by a new symbiont.

Current Annotation

What has been done until now

- 2006: First annotaion (213 genes)
- 2007: Second annotation
- Both teams used well known Gene-prediction algorithms + **collinear** alignment

Current Annotation

What has been done until now

- 2006: First annotaion (213 genes)
- 2007: Second annotation
- Both teams used well known Gene-prediction algorithms + **collinear** alignment

Current Annotation

What has been done until now

- 2006: First annotaion (213 genes)
- 2007: Second annotation
- Both teams used well known Gene-prediction algorithms + **collinear** alignment
- Problem: Over-annotation of function of genes. Many genes that are believed to be orthologous are much shorter and therefore deffer in their function.

My goal

use an non-collinear alignment algorithm to reannotate the whole genome of C.ruddii

Reannotation

Algorithms

- SuperMap + S-LAGAN
- A-Bruijn Alignment (ABA)

S-LAGAN

Species used

- Carsonella Ruddii PV (160 kb genome, 213 genes)
- Buchnera aphidicola BCc (Cc) (+ a plasmid) : 450 kb. (397 genes)
- Candidatus Blochmannia floridanus: 705 kb. (631 genes).
- Wigglesworthia glossinidia (+ a plasmid): 698 kb. (651 genes)
- Baumannia cicadellinicola str. Hc: 686 kb (651 genes)

S-LAGAN

plus Supermap

- A guiding tree (evolutionary tree) was build out of 16S-rRNAs of the species.

S-LAGAN

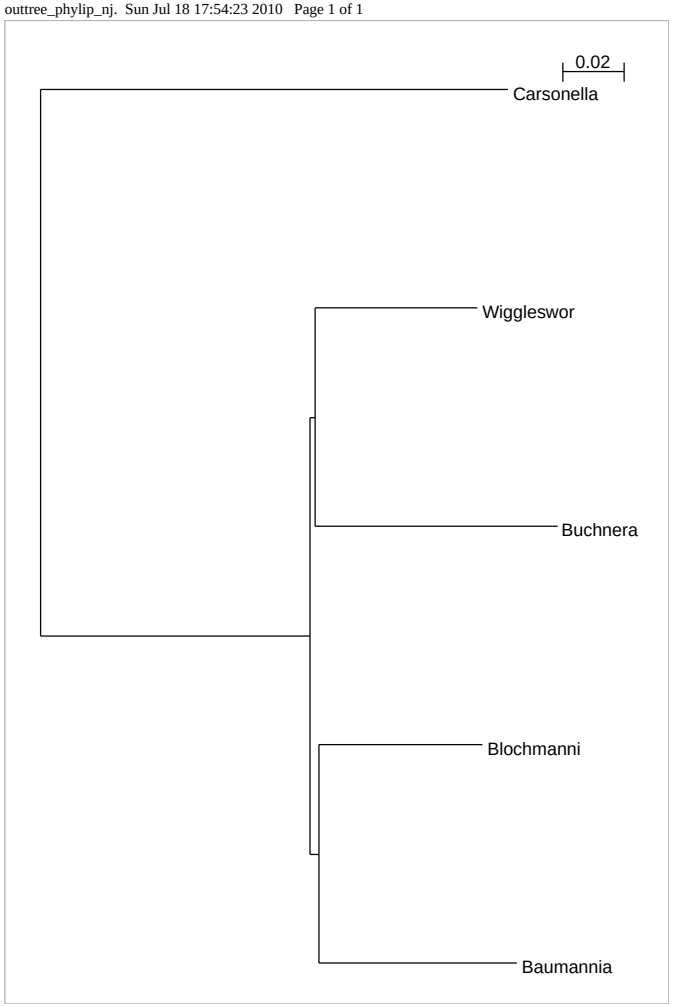
plus Supermap

- A guiding tree (evolutionary tree) was build out of 16S-rRNAs of the species.
- Neighbor joining tree
- Maximum likelyhood tree

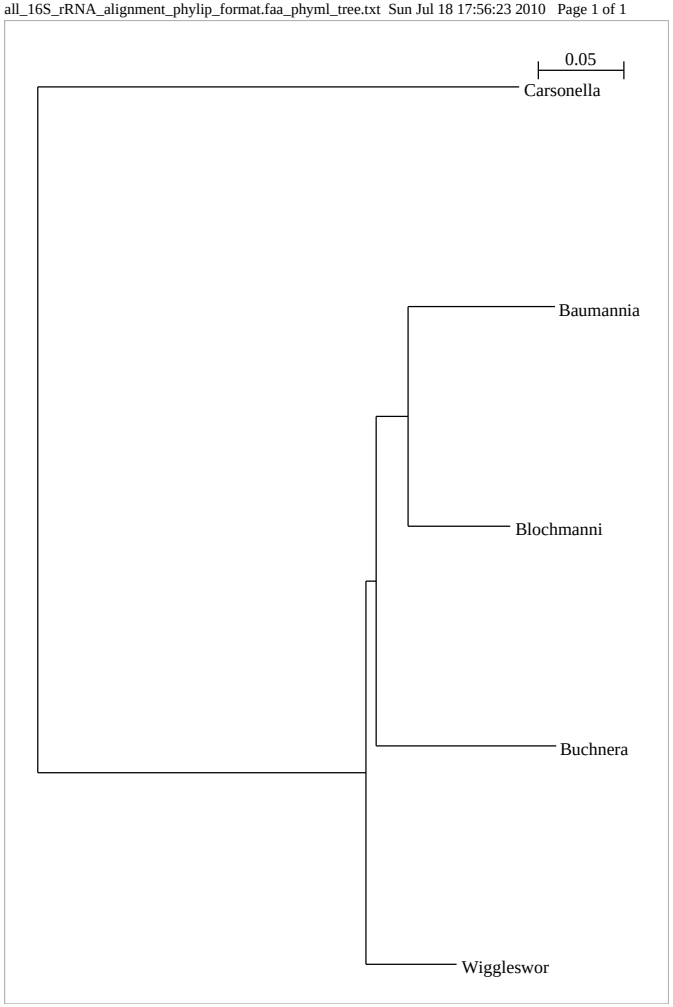
Trees

of 16S-rRNA sequence

outtree_phylip_nj. Sun Jul 18 17:54:23 2010 Page 1 of 1

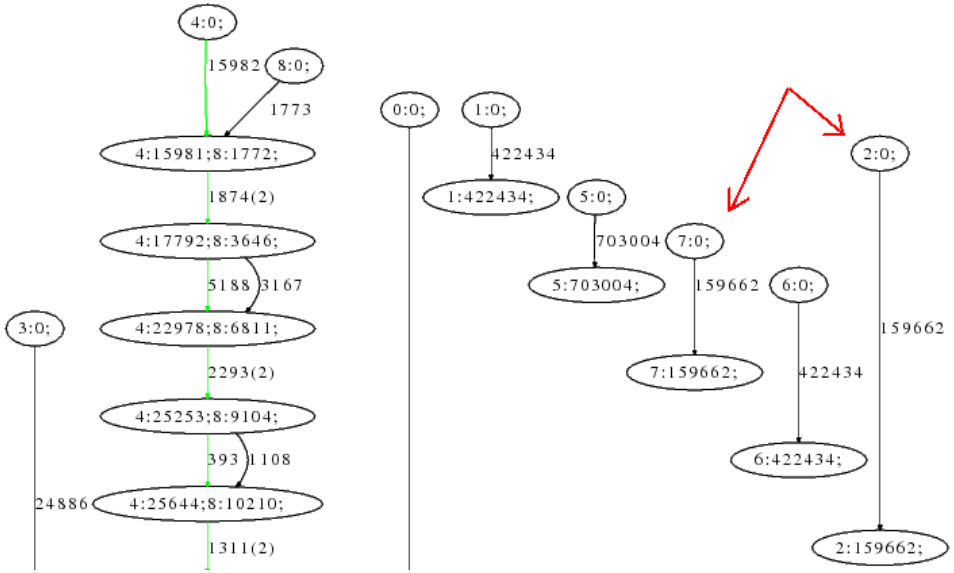


all_16S_rRNA_alignment_phylip_format.faa_phyml_tree.txt Sun Jul 18 17:56:23 2010 Page 1 of 1



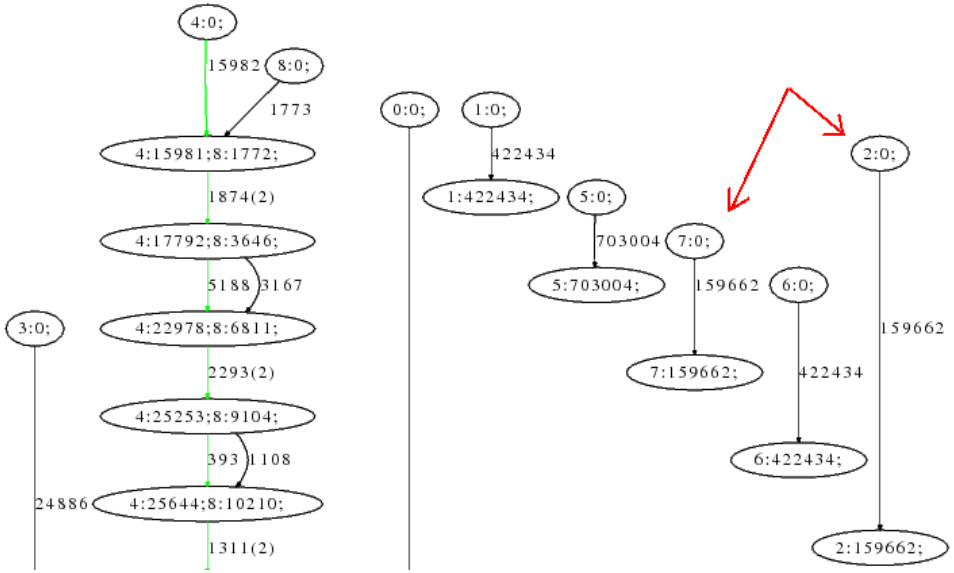
ABA

Using “my” 5 Species



ABA

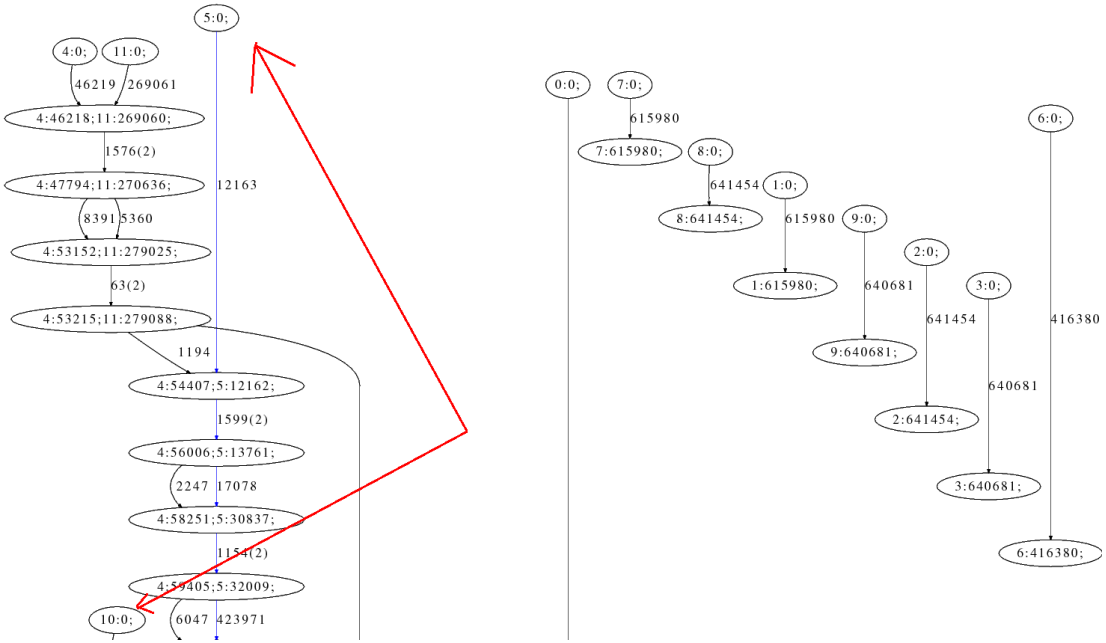
Using “my” 5 Species



- Species
  - 0 and 5: Wigglesworthia
  - 1 and 6: Buchnera aphidicola
  - 2 and 7: Carsonella Ruddii
  - 3 and 8: Blochmannia
  - 4 and 9: Baumannia

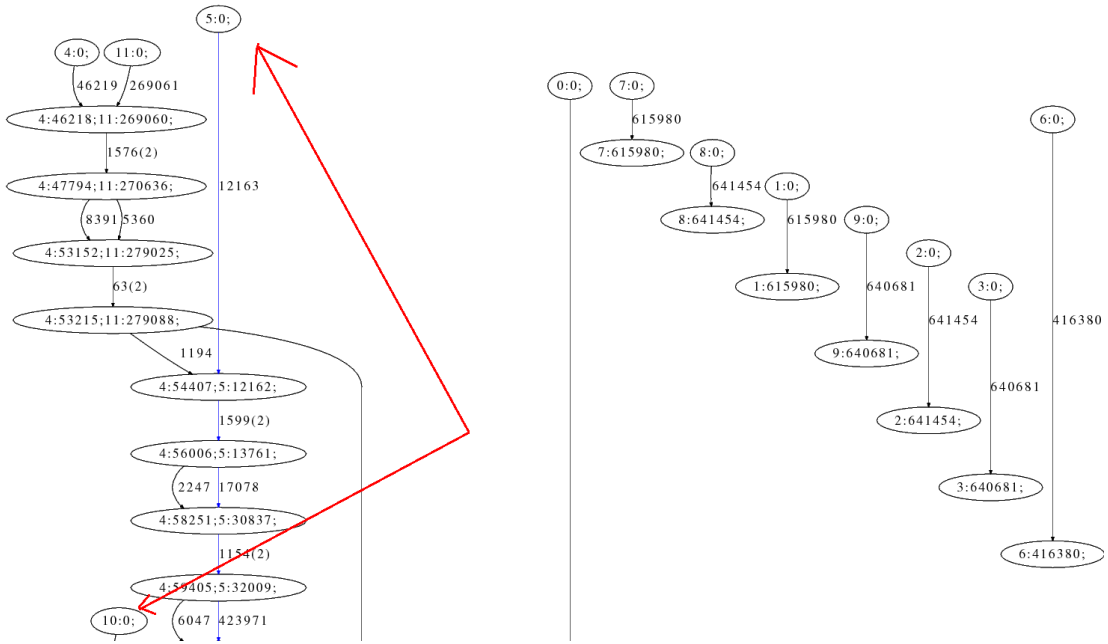
ABA

Using “Moya’s” Species



ABA

Using "Moya's" Species



- Species
  - 0 and 6: Buchnera aphidicola str. Cc
  - 1 and 7: Buchnera aphidicola str. Bp
  - 2 and 8: Buchnera aphidicola str. Sg
  - 3 and 9: Buchnera aphidicola str. APS
  - 4 and 10: Carsonella ruddii
  - 5 and 11: E.Coli

ABA

Only using Carsonella and E.Coli

2 Species (Carsonella and E.Coli) produce the same alignment as 6 Species from Moya paper

ABA

Gene prediction

7 genes of 213 were cut by the prediction in **C.ruddii**. 22 genes of 4494 were cut by the prediction in **E.Coli**.

ABA

Gene prediction

7 genes of 213 were cut by the prediction in **C.ruddii**. 22 genes of 4494 were cut by the prediction in **E.Coli**.

Example

region 0 - 46219 : 56 genes
region 46219 - 47795 : 0 genes
region 47795 - 53155 : 10 genes
region 53155 - 53218 : 0 genes
region 53218 - 54412 : 4 genes
region 54412 - 56011 : 0 genes
region 56011 - 58258 : 4 genes
region 58258 - 59412 : 0 genes
region 59412 - 65459 : 8 genes
region 65459 - 67041 : 1 genes
region 67041 - 70177 : 4 genes

ABA

Gene prediction

7 genes of 213 were cut by the prediction in **C.ruddii**. 22 genes of 4494 were cut by the prediction in **E.Coli**.

ABA

Gene prediction

7 genes of 213 were cut by the prediction in **C.ruddii**. 22 genes of 4494 were cut by the prediction in **E.Coli**.

Example

region 0 - 46219 : 56 genes
region 46219 - 47795 : 0 genes
region 47795 - 53155 : 10 genes
region 53155 - 53218 : 0 genes
region 53218 - 54412 : 4 genes
region 54412 - 56011 : 0 genes
region 56011 - 58258 : 4 genes
region 58258 - 59412 : 0 genes
region 59412 - 65459 : 8 genes
region 65459 - 67041 : 1 genes
region 67041 - 70177 : 4 genes

A possible future

- There are still at least 29 genes with no assigned function.
- Insightes into the possibility to create symbiotic life.

Project: Reimplementation of S-LAGAN Using SeqAn

F. Heeger, S. Specovius

1. Introduction to S-LAGAN
2. Implementation and Problems
3. Results

S-LAGAN

Shuffle-Limited Area Global Alignment of Nucleotides

- S-LAGAN computes **glocal** alignments of 2 sequences

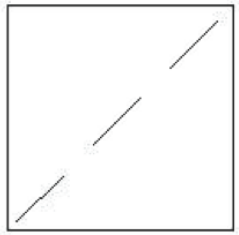
  $\rightarrow$ Set of local alignments which cover the whole sequence

- S-LAGAN is able to handle rearrangements

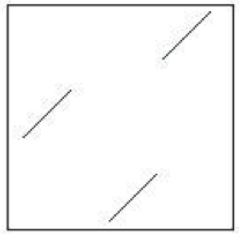
S-LAGAN

Rearrangements

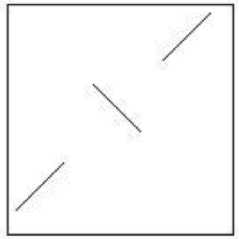
- No rearrangements



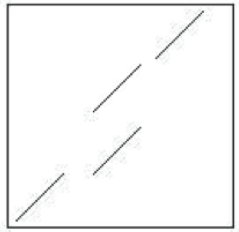
- Translocation



- Inversion



- Duplication



S-LAGAN

Overview

1. Computation of local alignments
2. Chaining
3. Realignment of consistent subchains

S-LAGAN

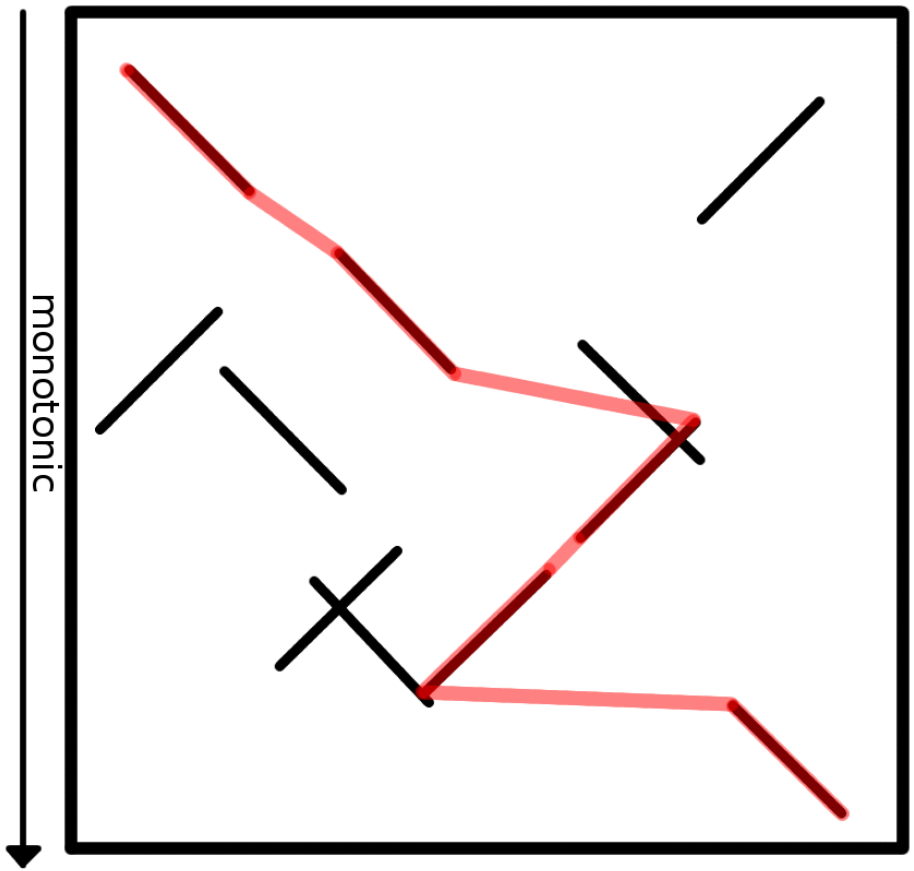
1. Computation of local alignments

- S-LAGAN uses CHAOS for this step
- Applies CHAOS twice
  - $\rightarrow$ Sequence 1 with sequence 2
  - $\rightarrow$ Sequence 1 with reverse complement of sequence 2

S-LAGAN

2. Chaining

1-monotonic



S-LAGAN

3. Realignment of consistent subchains

- Consistent (co-linear) subchains are globally aligned
- S-LAGAN uses LAGAN for this step

Implementation and Problems

Goal

- Implementation in SeqAn
- Extract Chaos from SeqAn implementation of LAGAN
- Implement 1-monotonic chaining
- Use existing SeqAn implementation of LAGAN

Implementation and Problems

Local Alignments

- Find seeds with q-gram index
- Merge overlapping seeds
- Chain seeds with Chaos algorithm

  $\rightarrow$ Segmentation Fault on certain data

  $\rightarrow$ Only gap-free local matches

Implementation and Problems

Chaining

- Graph with nodes representing local matches
- Edges to all matches, which can be chained 1-monotonic
  $\rightarrow$ Heaviest path (Bellman-Ford Algorithm)
- $\mathcal{O}(n^3)$

Implementation and Problems

Realign Consistent Subchains

- Find consistent subchains
- Align them with global alignment algorithm
- LAGAN runs into an endless loop on certain data

  $\rightarrow$ Use Needleman-Wunsch Algorithm

Results

Our implementation...

- is very slow
- can be used on small data, like virus genomes ($\sim$ 5000 bp)
- finds manually inserted rearrangements

Introduction OSL

Motivation

Assume there are two assemblies obtained from different assemblers:

Introduction OSL

Whole **G**enome **S**hotgun Approach (WGS)

Aim: Assemble a genome sequence from given reads.

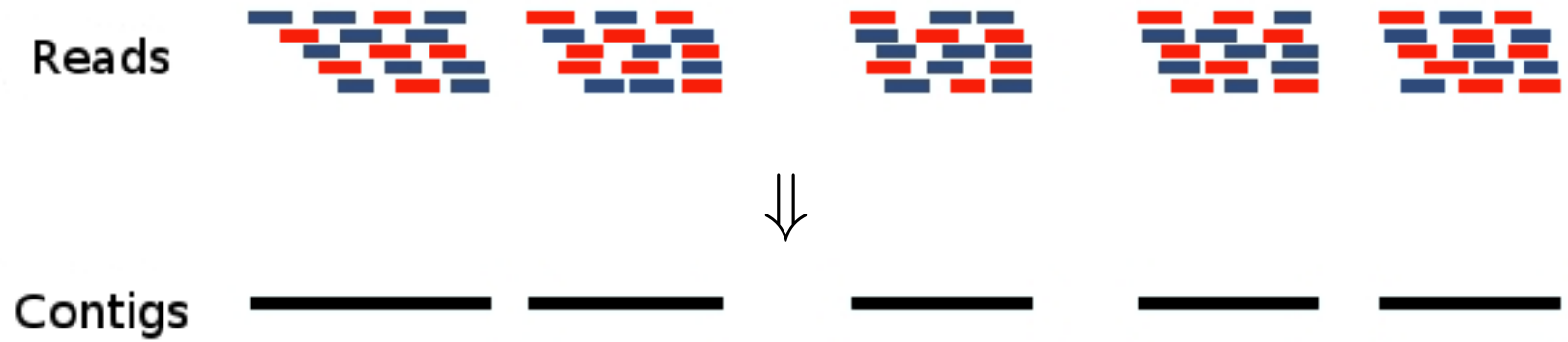
- **Reads**
  - → Collection of short sequences
  - → Obtained from an automated sequencer
  - → Orientation is not known

Introduction OSL

Whole **G**enome **S**hotgun Approach (WGS)

Assemble overlapping reads together to obtain contigs.

- **Contigs**
  $\rightarrow$ Large, contiguous fragments of assembled reads



Problem

- Order and orientation of contigs is unknown

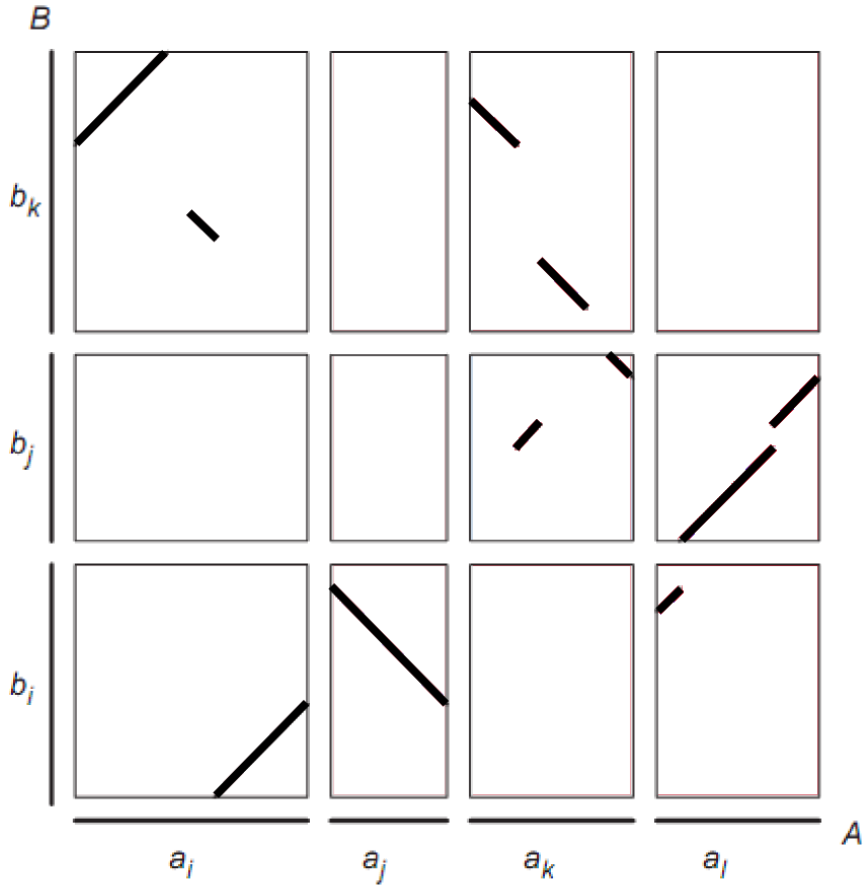
↓

Search for a good assembly layout !

Optimal **S**yntenic **L**ayout of unfinished assemblies

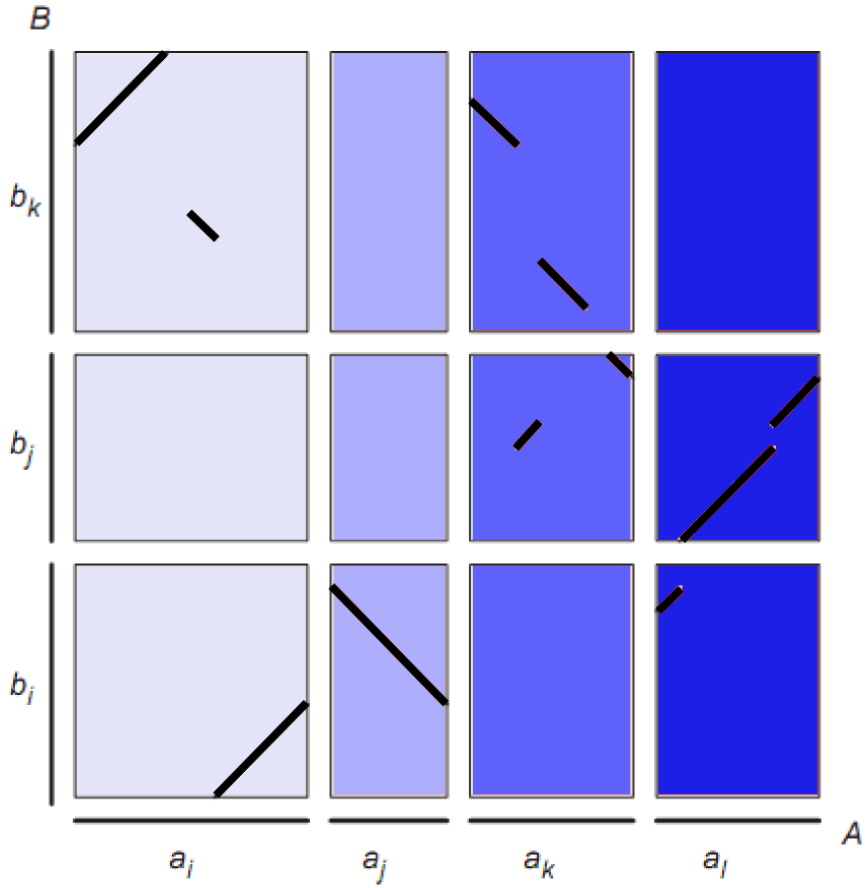
OSL Idea

- Maximize no. of extended local diagonals



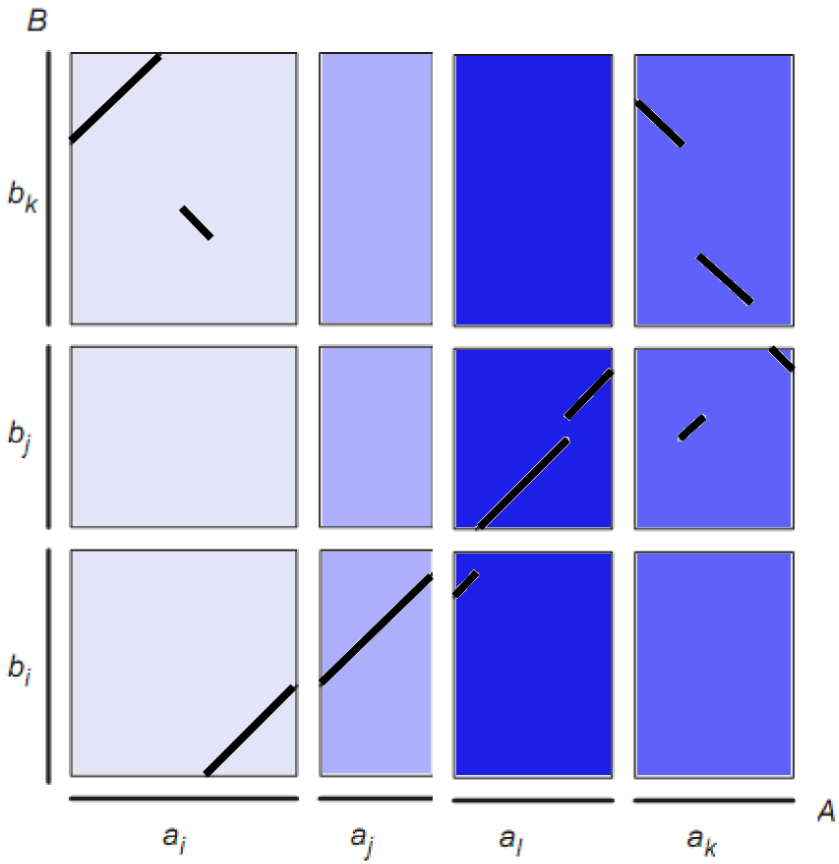
OSL Idea

- Maximize no. of extended local diagonals
- permute and flip contigs of assembly A



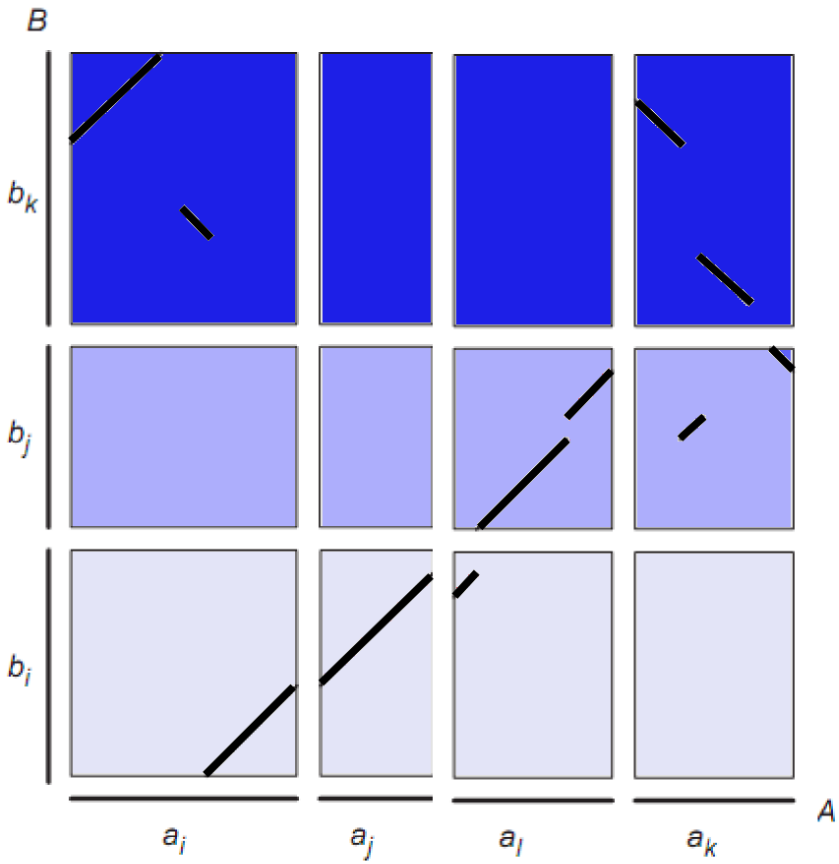
OSL Idea

- Maximize no. of extended local diagonals
- permute and flip contigs of assembly A



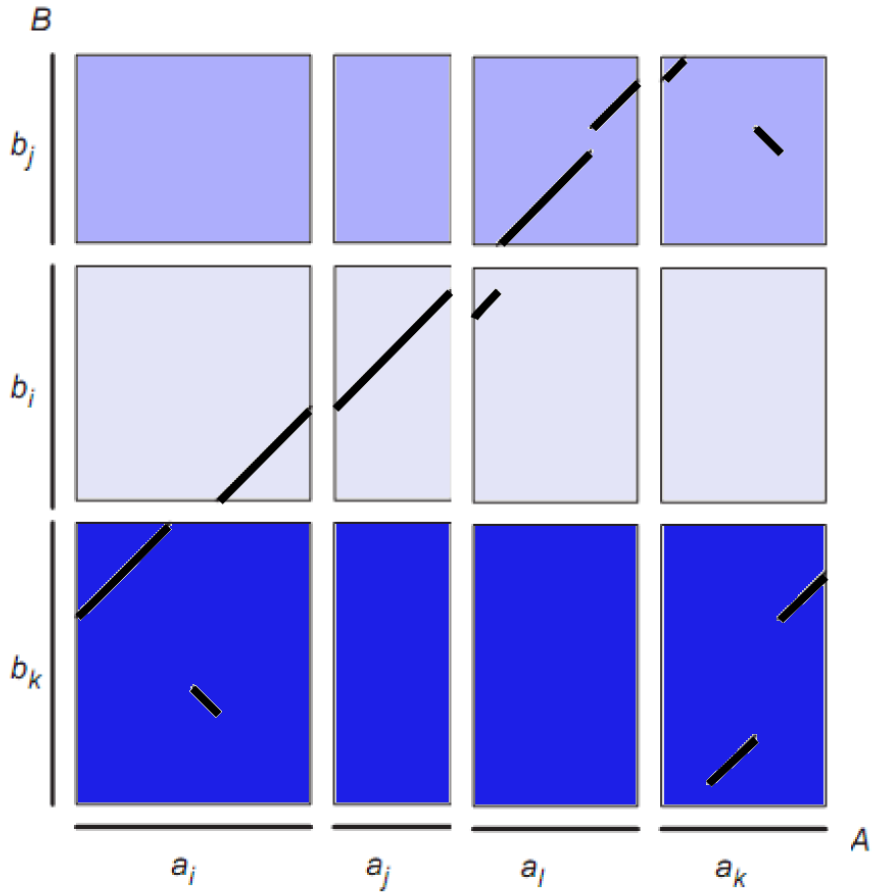
OSL Idea

- Maximize no. of extended local diagonals
- permute and flip contigs of assembly A
- switch roles of A and B



OSL Idea

- Maximize no. of extended local diagonals
- permute and flip contigs of assembly A
- switch roles of A and B



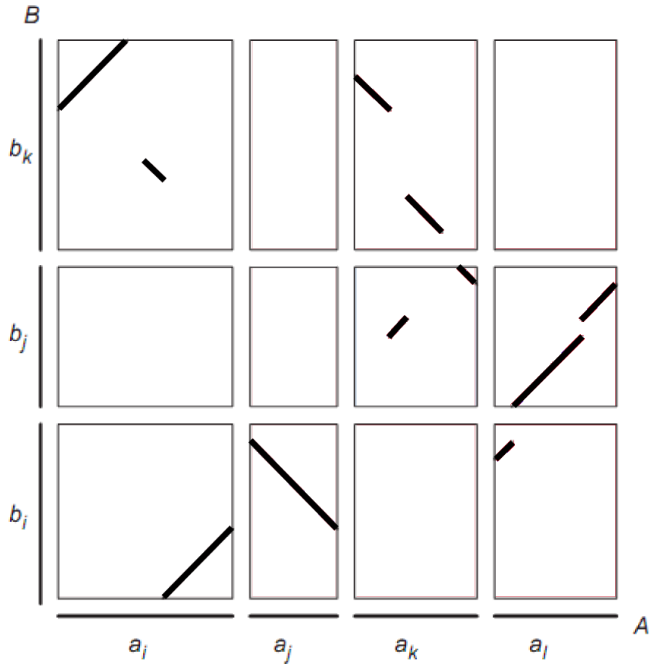
Independency in constructing the layouts of A and B !

The OSL Problem

Basics

Assemblies

$A = (a_1, \ldots, a_p)$
$B = (b_1, \ldots, b_q)$



The OSL Problem

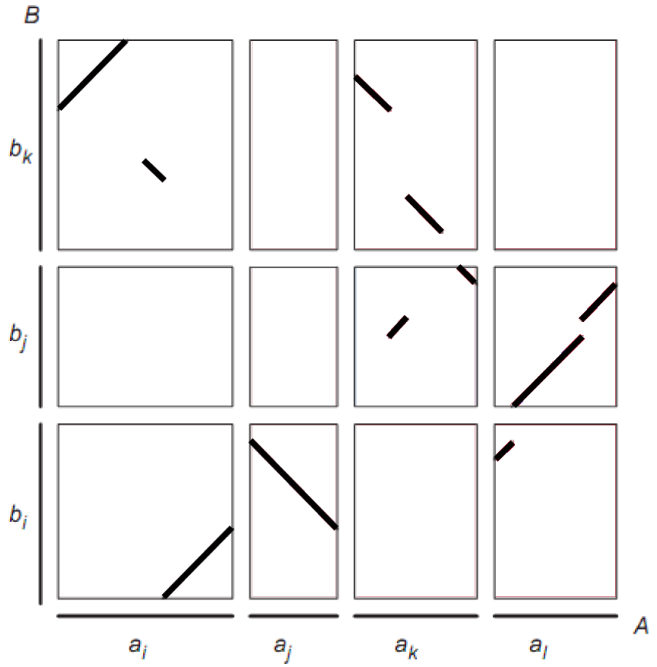
Basics

Assemblies

$A = (a_1, \ldots, a_p)$
$B = (b_1, \ldots, b_q)$

Set of Matches

$M = (m_1, \ldots, m_r)$

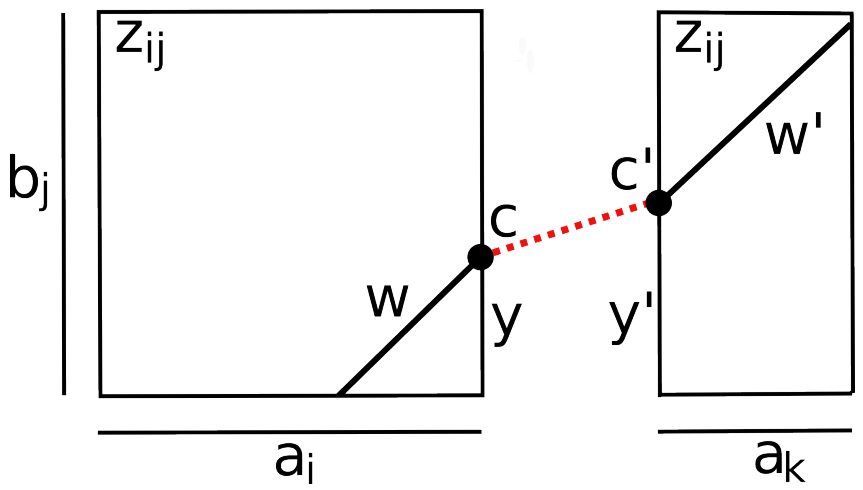


The OSL Problem

Layout

Local diagonal extension

c and c' form a *local diagonal extension* iff $y \sim y'$ and $o = o'$



The OSL Problem

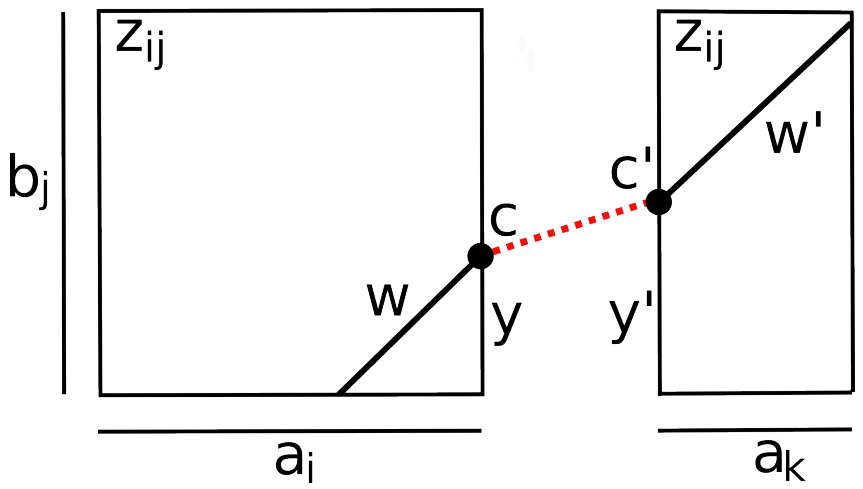
Layout

Local diagonal extension

c and c' form a *local diagonal extension* iff $y \sim y'$ and $o = o'$

Weight of extension

$w + w' - \mid y - y' \mid$



Project: Assembly Comparison

Goal

1. Assemble a set of reads with two different Assemblers
2. Compare the results using Layout Software

   $\rightarrow$ OSLay

Project: Assembly Comparison

1. Assemble a set of reads with two different Assemblers

   - Reads of Chromosom 21
   - Assembler: Mira and Celera (WGS)

Project: Assembly Comparison

1. Assemble a set of reads with two different Assemblers

   - Reads of Chromosom 21
   - Assembler: Mira and Celera (WGS)

Project: Assembly Comparison

Problems:

- WGS Assembler doesn't work with given reads

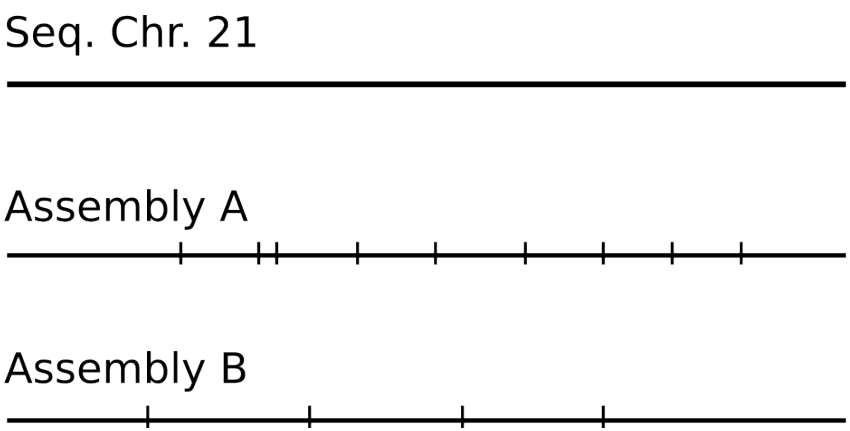
↓

Plan B:

- Take given sequence of chr. 21
- Create artificial contigs

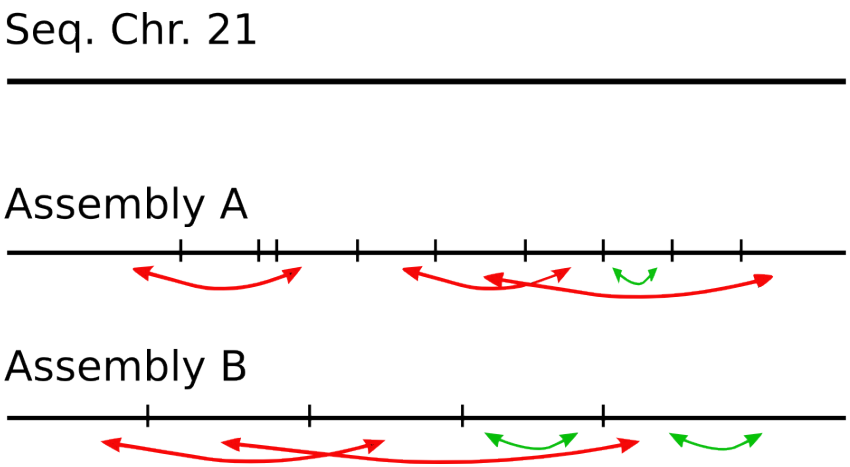
Project: Assembly Comparison

Create artificial contigs:



Project: Assembly Comparison

Create artificial contigs:



Project: Assembly Comparison

BLAST

Assemblies are from the same sequence

↓

Megablast

Project: Assembly Comparison

OSLay

OSLay is the implementation of the OSL algorithm.

Input:

- target assembly
- reference assembly
- matches (e.g. BLAST)

Output:

- original layout
- new layout

Project: Assembly Comparison

OSLay

Problem:

- Input too large for OSLay
- Chr. 21 $\sim$ 34 MB

$\downarrow$

Plan B:

- segment of 210 KB

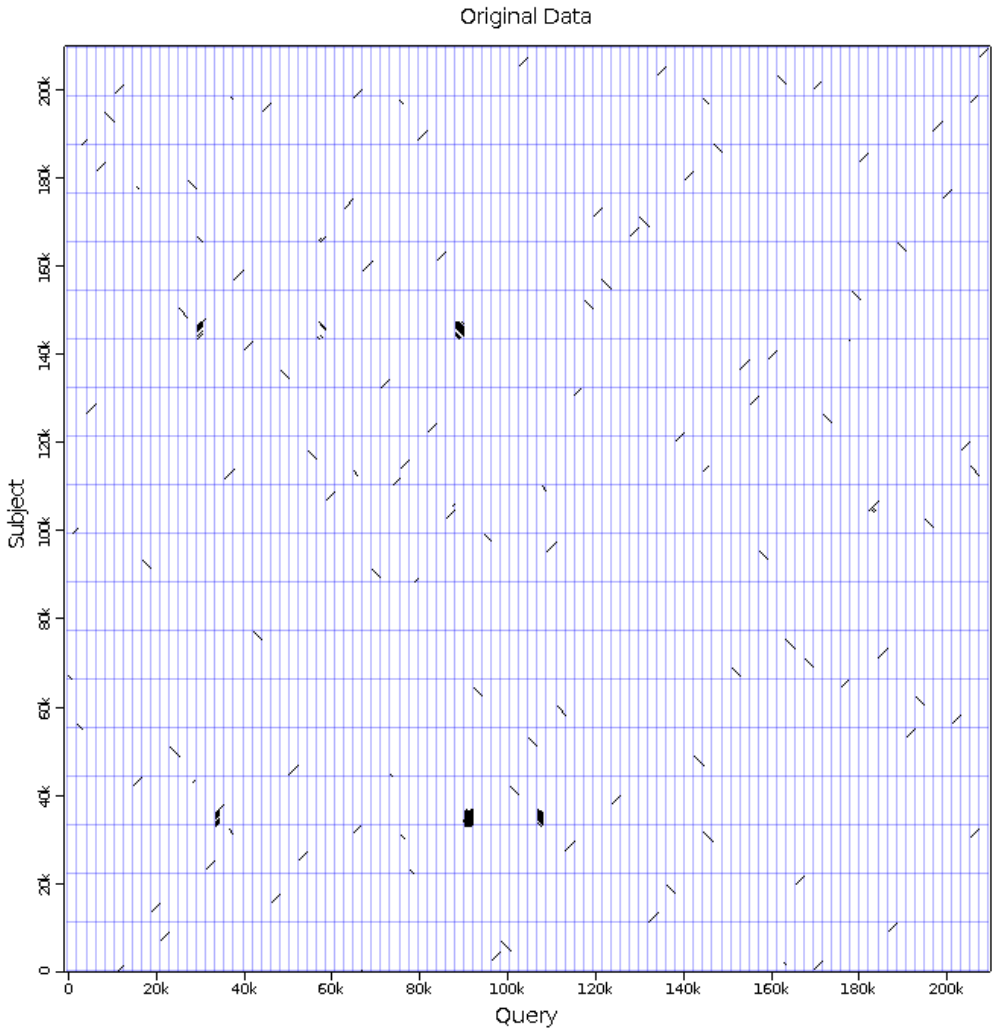
Project: Assembly Comparison

OSLay

- Assembly A: sequence divided by 100
- Assembly B: sequence divided by 19

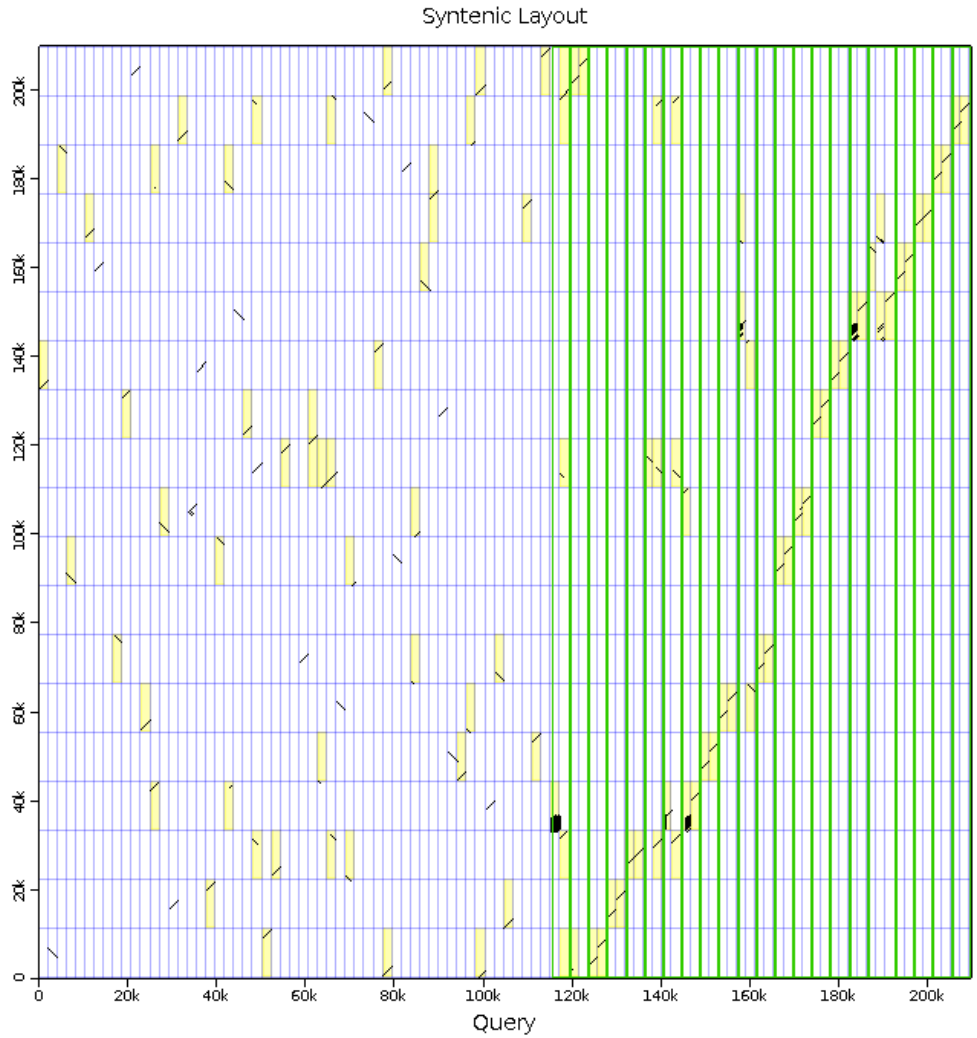
Project: Assembly Comparison

OSLay



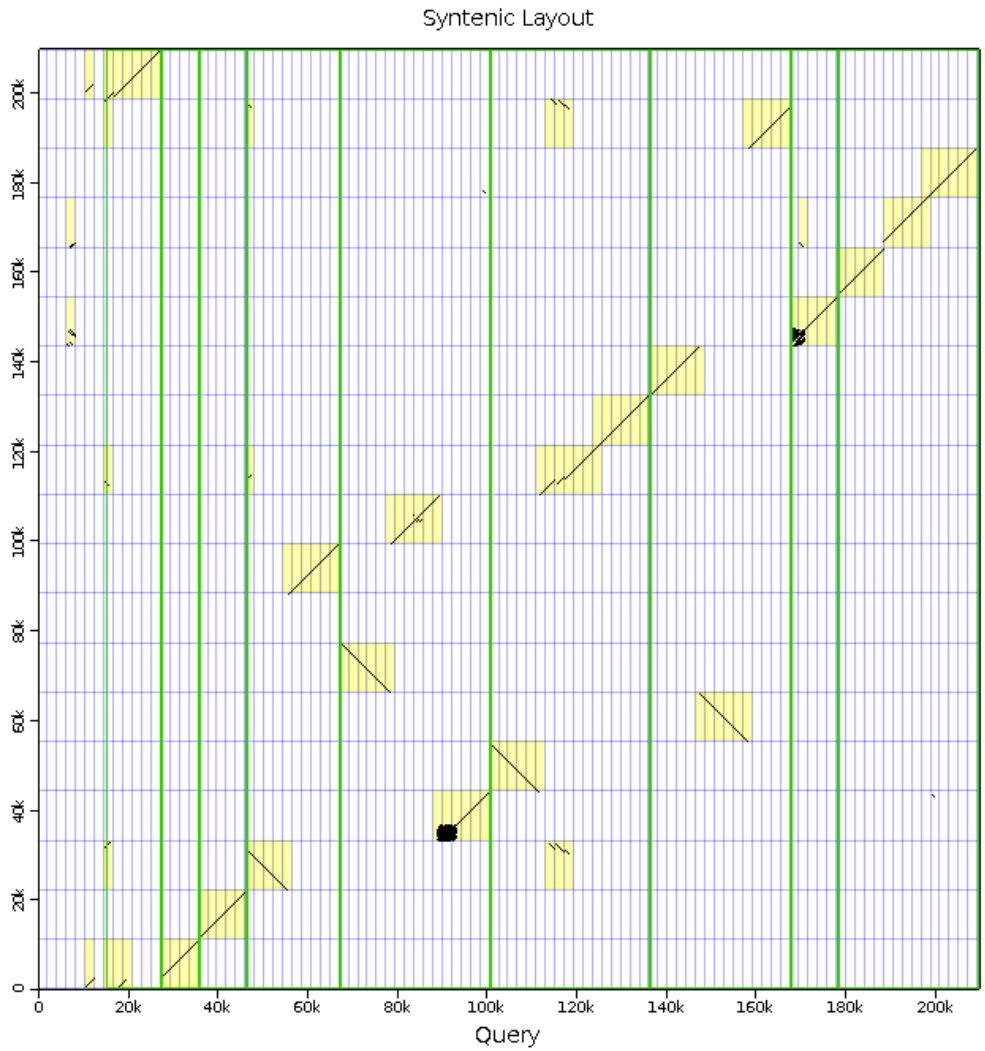
Project: Assembly Comparison

OSLay



Project: Assembly Comparison

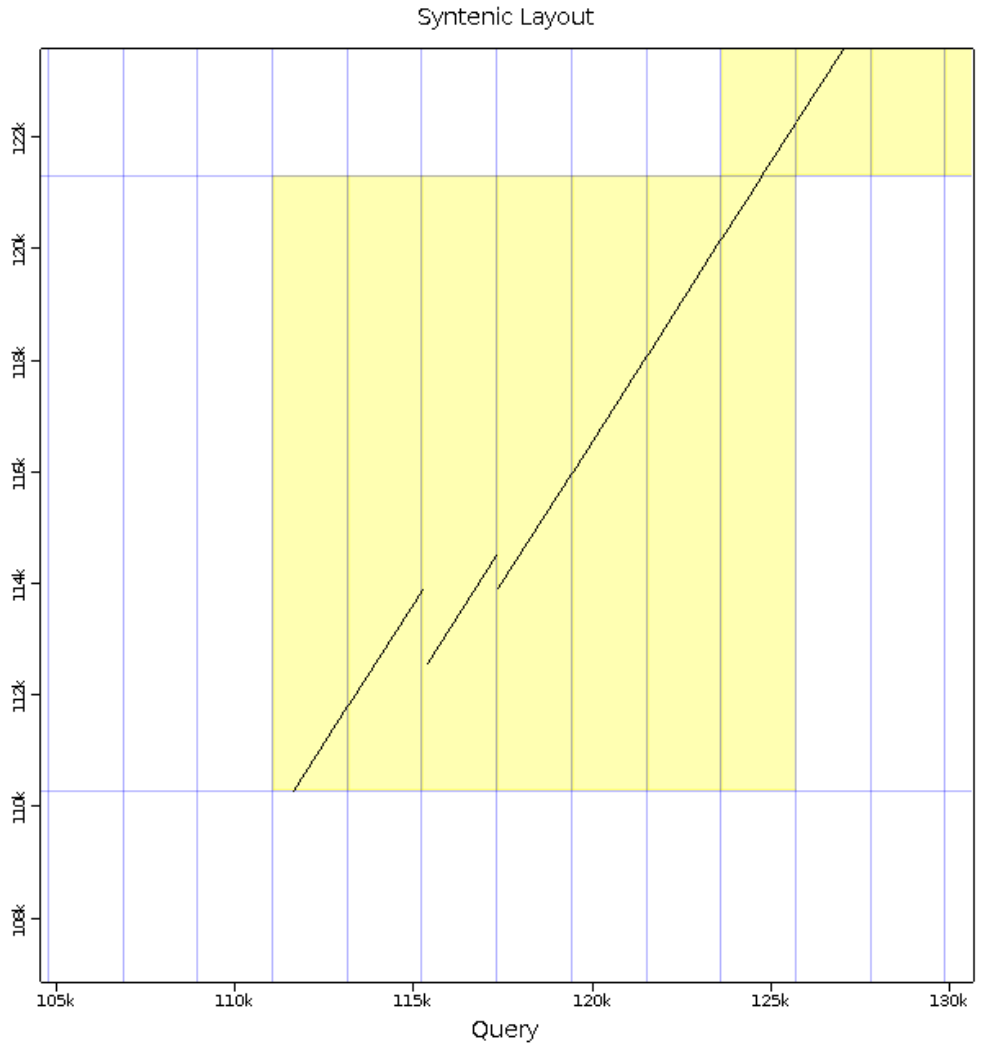
OSLay



Project: Assembly Comparison

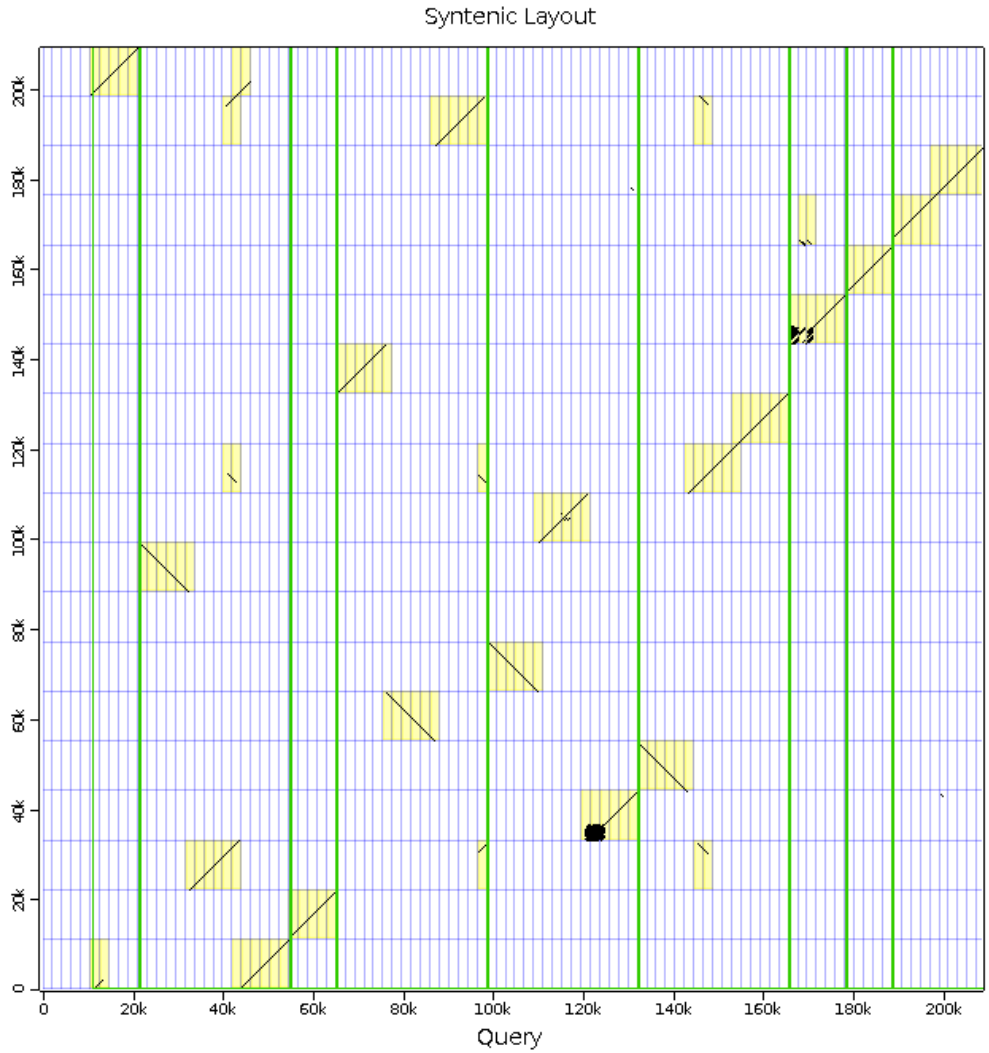
OSLay

False connections:



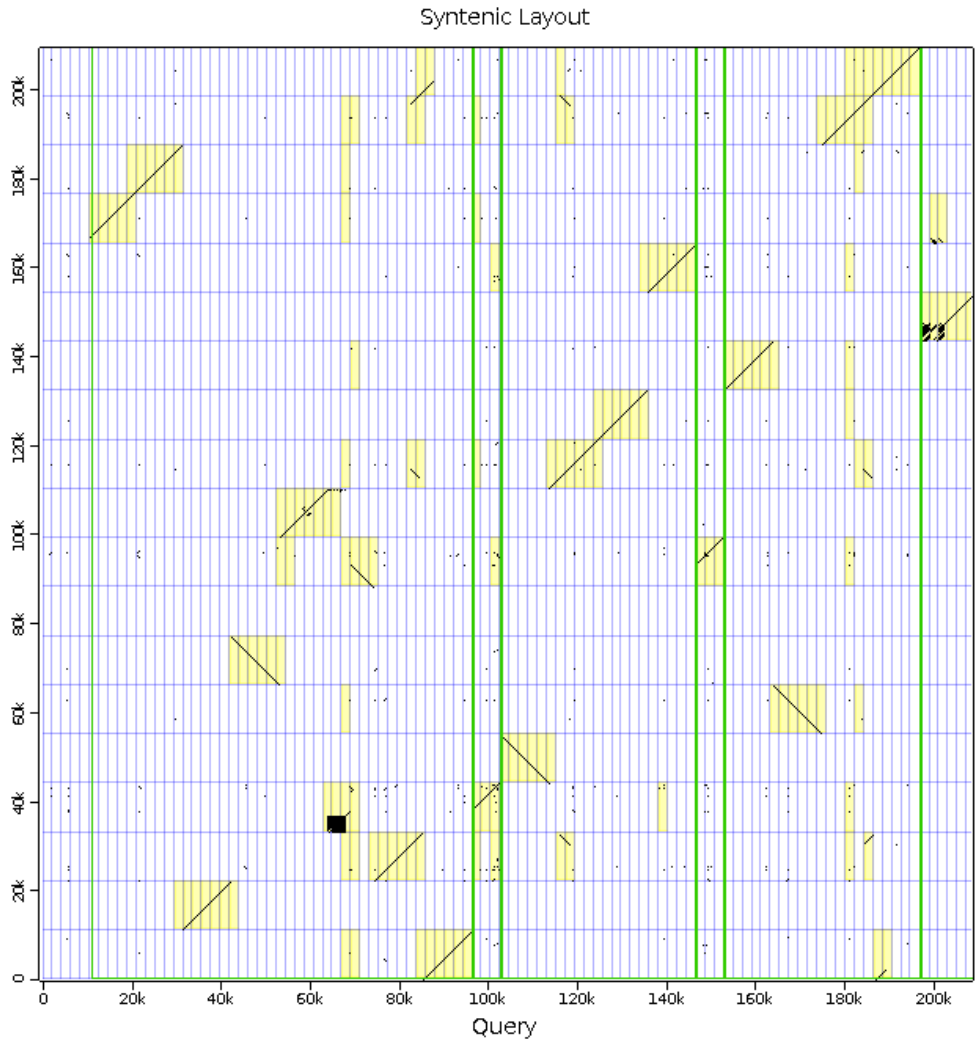
Project: Assembly Comparison

OSLay



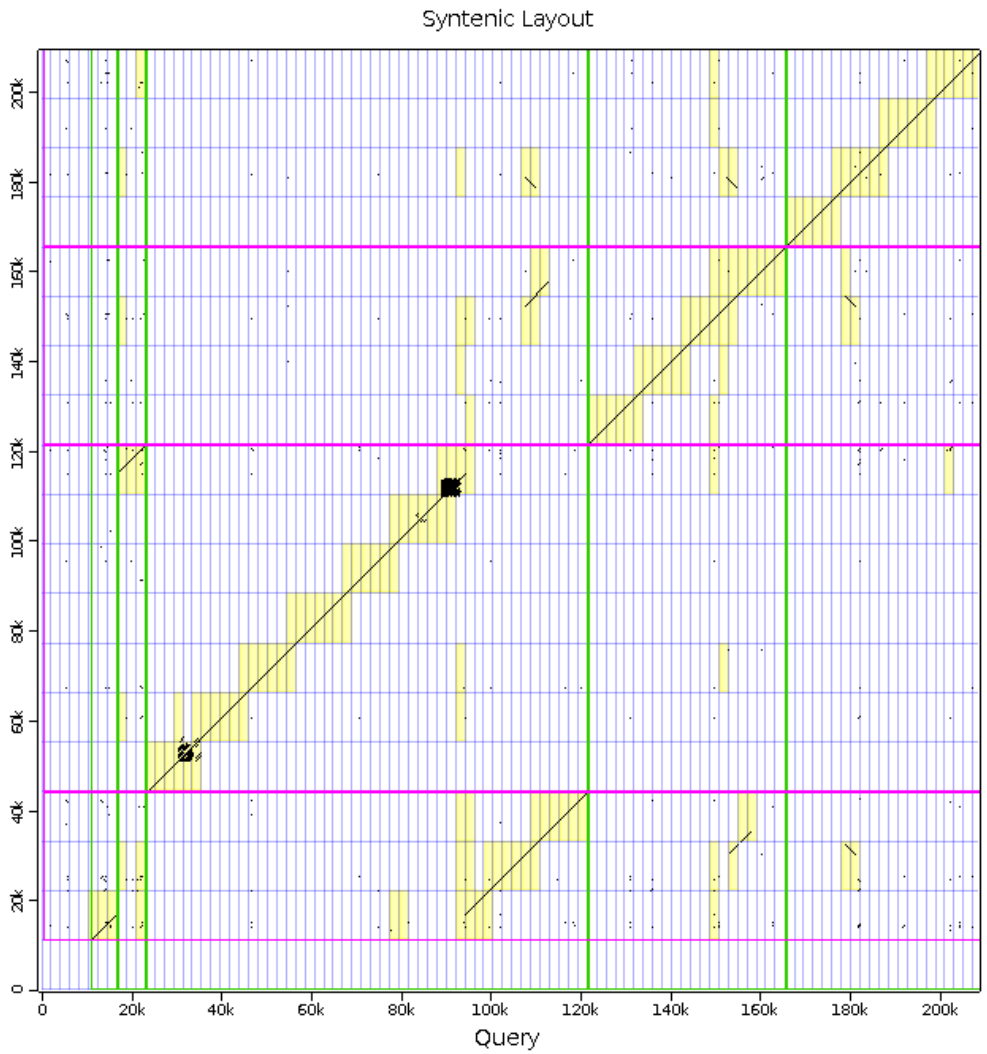
Project: Assembly Comparison

OSLay



Project: Assembly Comparison

OSLay



Project: Assembly Comparison

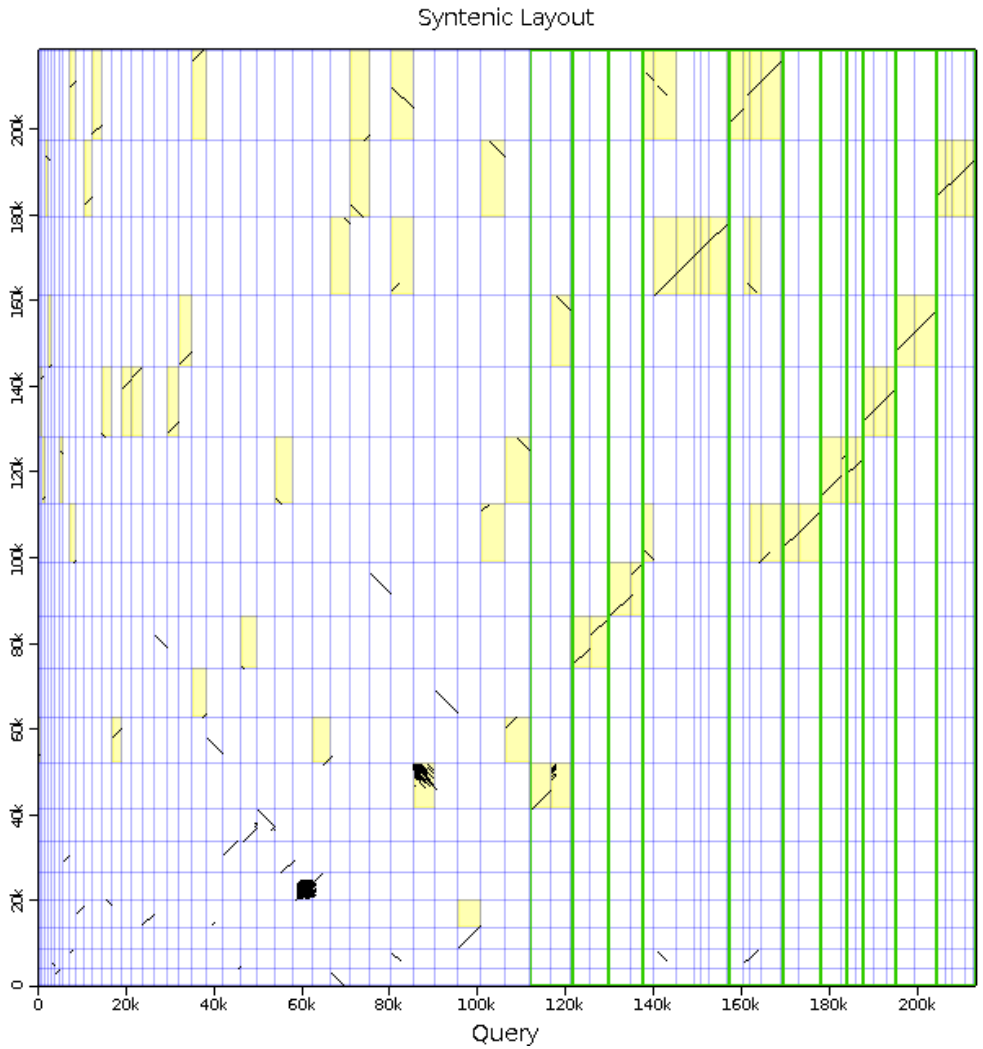
OSLay

Create contigs with random length:

- Assembly A: lengths between 500 and 5000 bp ($\sim$ 100 contigs)
- Assembly B: lengths between 1000 and 200000 bp ($\sim$ 20 contigs)

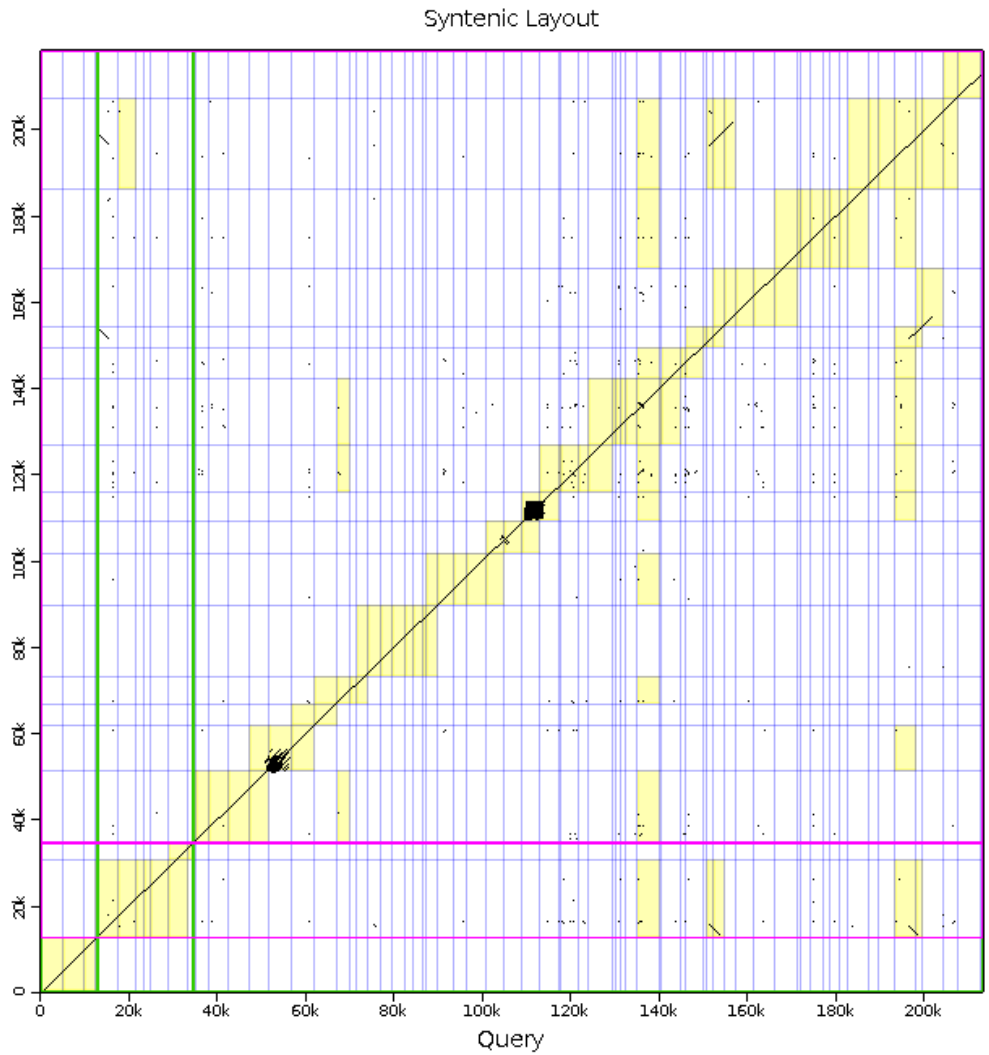
Project: Assembly Comparison

OSLay



Project: Assembly Comparison

OSLay



Project: Assembly Comparison

OSLay

Discussion

- Works only with similar sequences
- But: Contig borders of Assemblies should be different
- Just for small genomes

References

Brudno, M., Do, C. B., Cooper, G. M., Kim, M. F., Davydov, E., Comparative, N., Program, S., Green, E. D., Sidow, A., and Batzoglou, S. (2003a).
LAGAN and Multi-LAGAN : Efficient Tools for Large-Scale Multiple Alignment of Genomic DNA Outline of Algorithms.
Genome Research, (Taylor 1988):721–731.

Brudno, M., Malde, S., Poliakov, A., Do, C., Couronne, O., Dubchak, I., and Batzoglou, S. (2003b).
Glocal alignment: Finding rearrangements during alignment.
Bioinformatics, 19(Suppl 1):i54.

Parker, D. S. and Lee, C. J. (2003).
Multiple Partial Order Alignment as a Graph Problem.
Science (New York, N.Y.).

Pevzner, P. A., Tang, H., and Tesler, G. (2004).
De novo repeat classification and fragment assembly.
Genome Research, 14(9):1786–96.

Raphael, B., Zhi, D., Tang, H., and Pevzner, P. (2004).
A novel method for multiple alignment of sequences with repeated and shuffled elements.
Genome research, 14(11):2336–46.