New method for improved sensitivity and reliability of anchor based genome alignment

Raluca Uricaru[1], Célia Michotey[2], Laurent Noé[3], Hélene Chiapello[2], Eric Rivals[1]

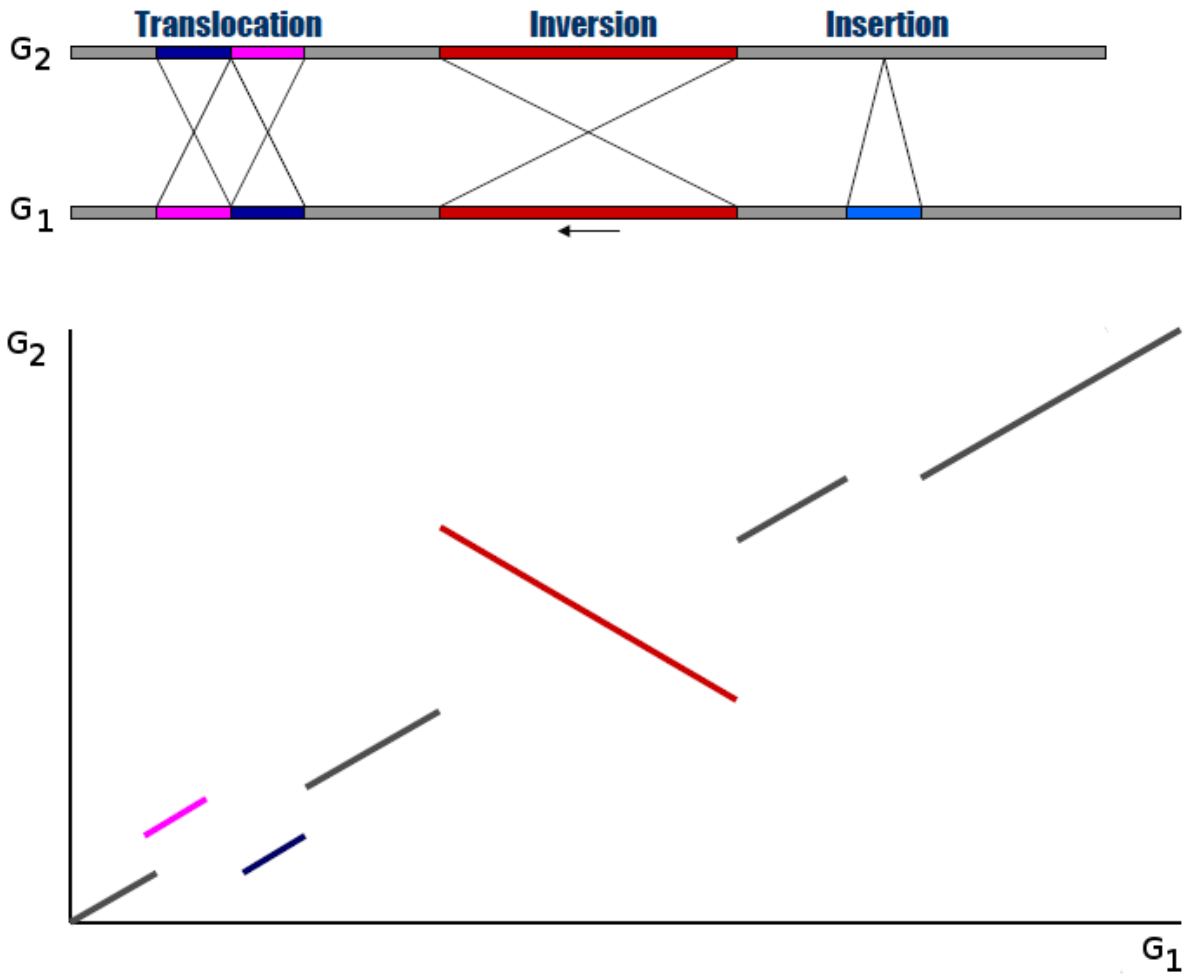
LIRMM (CNRS), **MIG** (INRA), **LIFL** (INRIA)

2 février 2010

Summary

Summary

Pairwise Global Genome Alignment



[S. Kurtz et al. – MUMMER]

Global Genome Alignment Motivations

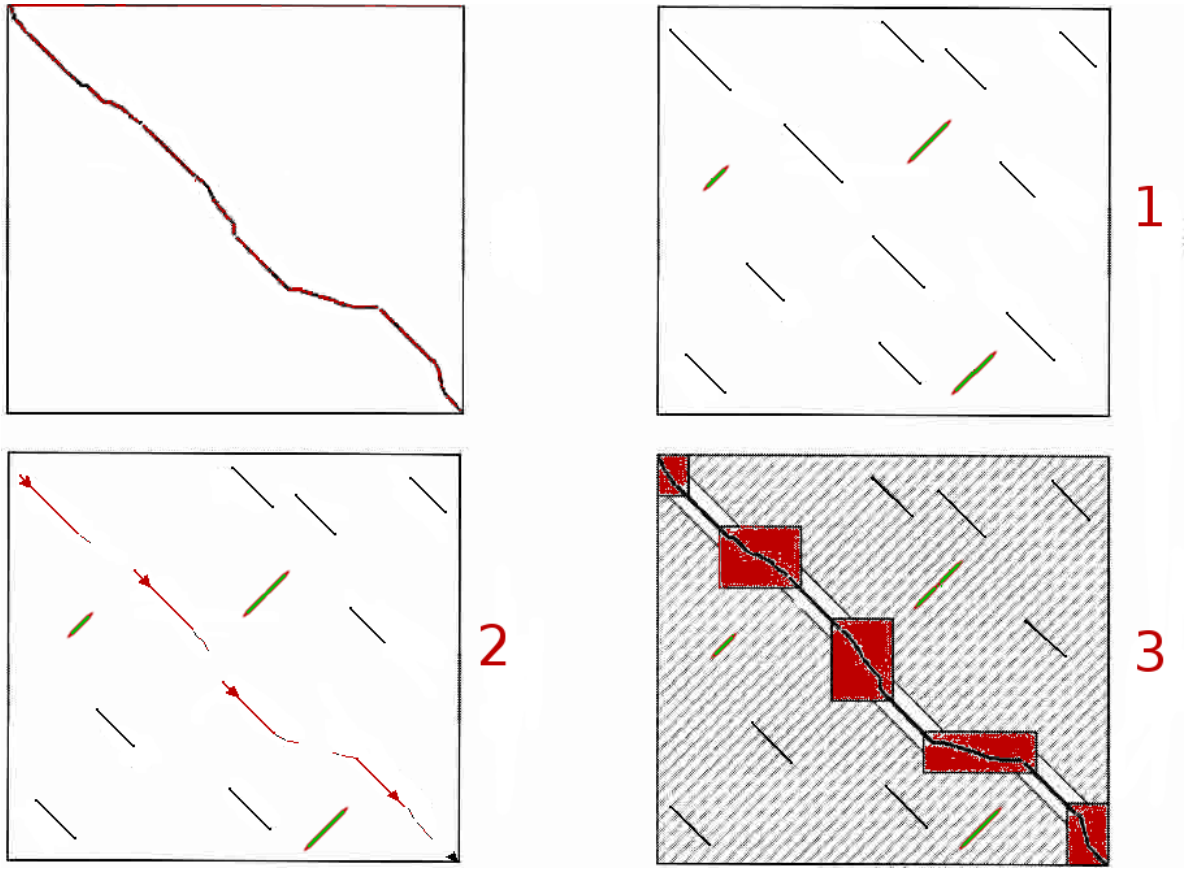
Fundamental task in computational biology :

- identify conserved parts of genomes (*backbone segments*)
  e.g. biological components common between species
- identify differences in sequences (*variable segments*)
  e.g. pathogenic islands
- study of evolutionary relatedness among various groups of organisms
  phylogenetic trees

Summary

Anchor-based, 4-phase Method

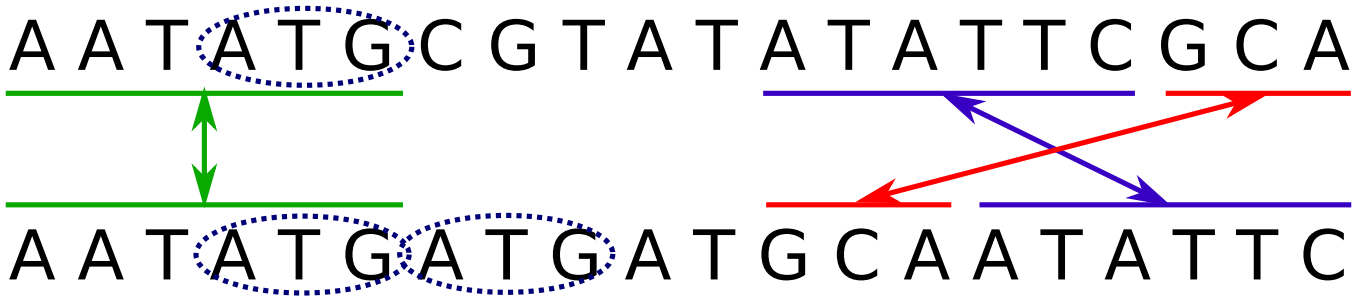
commonly used heuristic – overcomes resource limitations



[M. Brudno et al.]

P1 Computation of Fragments : similarity regions

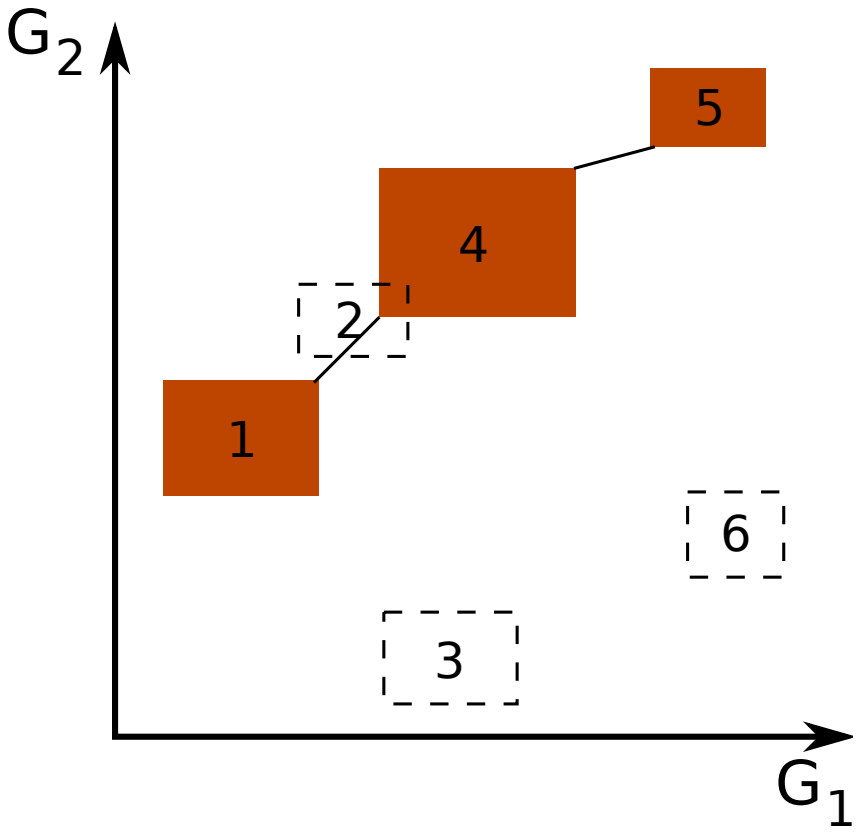
- MUMs (or MRMs, or MEMs, . . .)
  MAXIMAL UNIQUE MATCH = maximal exact match, unique in both sequences



- Problem : keep good MUMs only (threshold for match length)
  usually, MUMs *lengths do not reach 1kb*

P2 Chaining Phase : anchors selection

best scoring chain of collinear, non-overlapping fragments



P3 + P4

P3 Recursive anchoring : **P1** + **P2**

- on each pair of yet unaligned facing regions between the anchors
- use adapted parameters

P4 “Last chance alignment”

- for each pair of short, unaligned facing regions in the end

Anchor-based Genome Alignment Tools

- **MGA** (M. Höhl et al., 2002) – *does not deal with complex rearrangements*

P1	computes MEMs – VMATCH software (S. Kurtz et al., 2004)
P2	selects a collinear set of anchors – Chainer software (M. I. Abouelhoda et al., 2005)
P3+P4	recursive anchoring + "last chance alignment"

Anchor-based Genome Alignment Tools (II)

- **MAUVE** (A. Darling et al, 2004.) – *handles rearrangements*

  P1 searches for approximate matches

  P2 selects a subset of non-overlapping anchors
  relaxes the property of collinearity

  P3+P4 recursive anchoring + “last chance alignment”

Summary

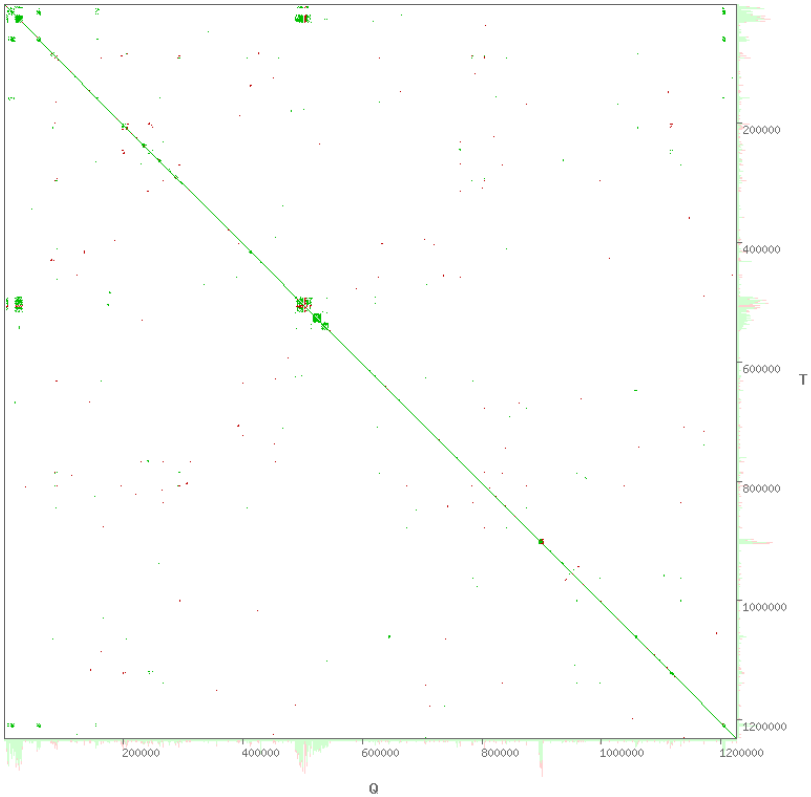
Initial Experiments

- **dataset** :
  236 couples of intra-species bacterial genomes (42 species)
  94 collinear pairs, 142 rearranged pairs
- **tools** :
  2 state-of-art tools : MGA, MAUVE
- **comparison measures** :
  global comparison measures (*coverage%*[1], *id%*[2])
- **remarks** :
  - extreme coverage difference in the same bacteria species
    Prochlorococcus marinus : [0, 78]% **MGA**, [6, 96]% **MAUVE**
  - extreme coverage difference between species
    Staphylococcus aureus : >90%, *Synechococcus sp* : <10%

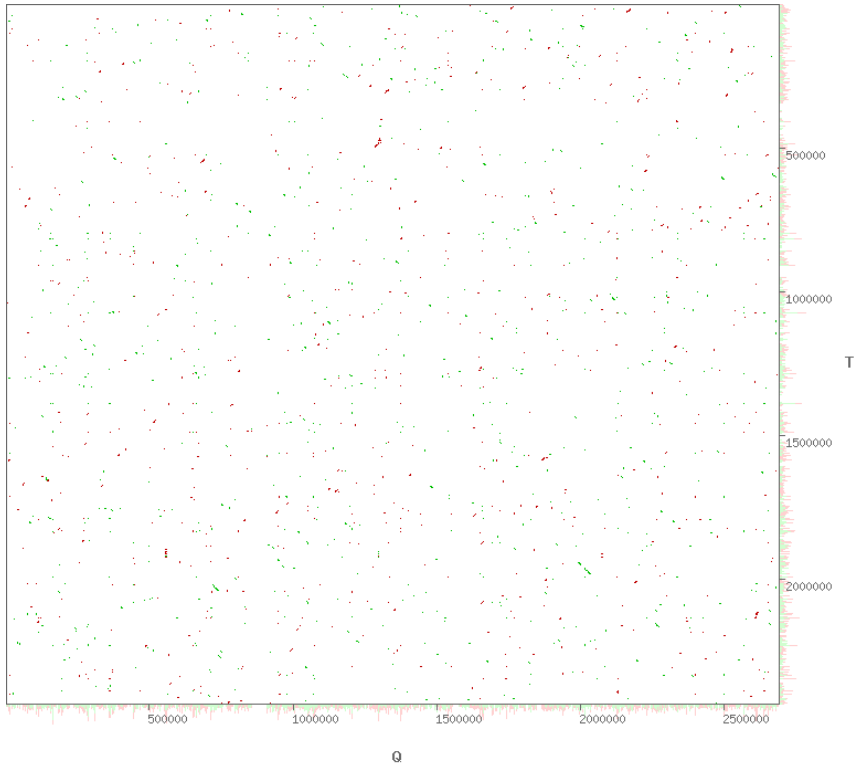
[1] ratio of the total length of aligned segments over the genome size

[2] new definition : the ratio of identical bases in aligned segments over the genome size

Examples of Different Levels of Similarity

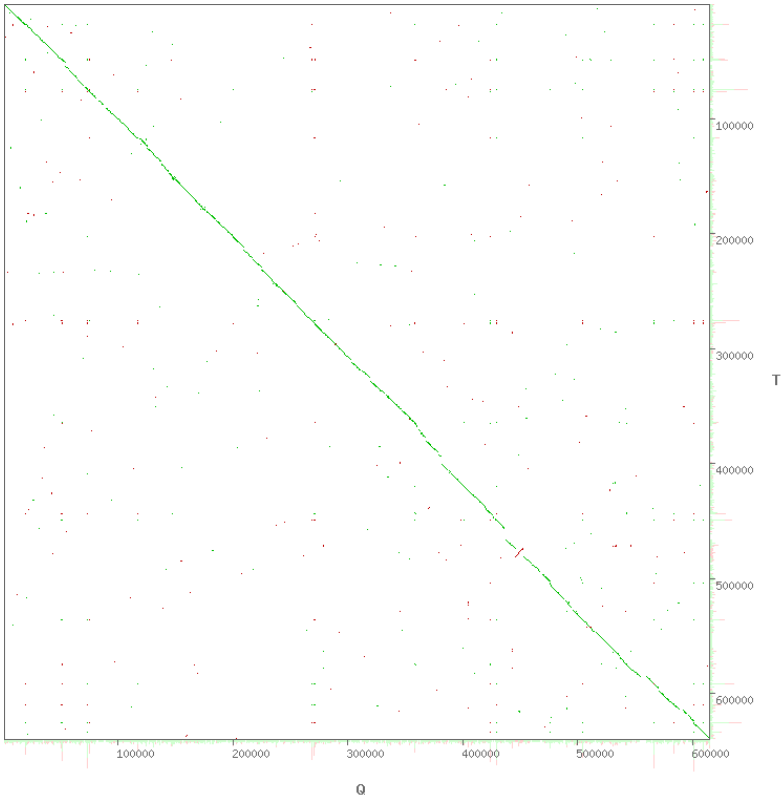


Chlamydia pneumoniae

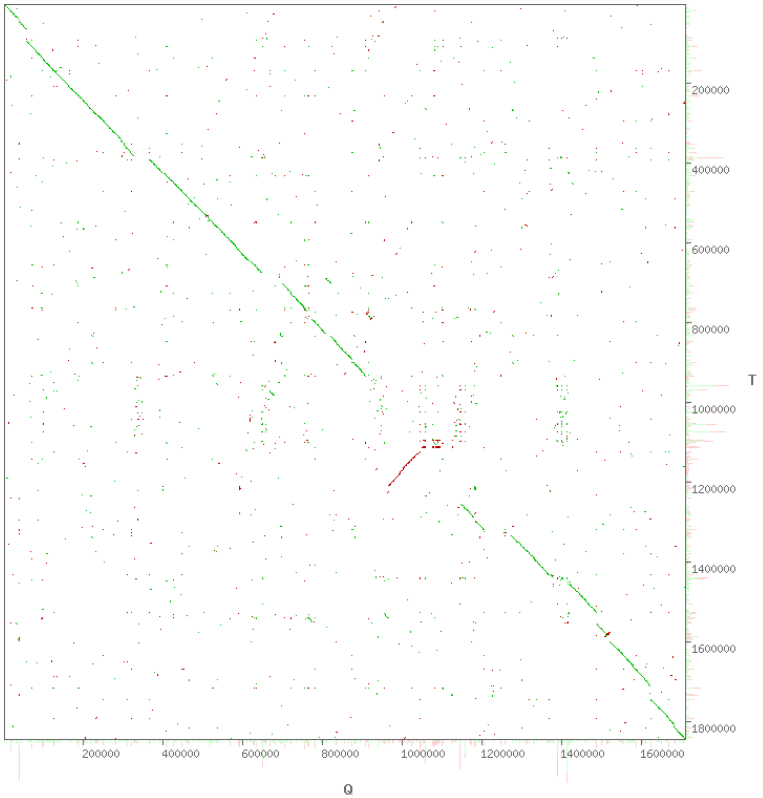


Synechococcus sp.

Divergent Cases



Buchnera aphidicola



Prochloroccocus marinus

Divergent Cases (II)

Species	Collinear	Cov%		Id%		Id% of coverage	
		MGA	MV	MGA	MV	MGA	MV
Buchnera aphidicola							
AE016826	×	37	99	24	62	64	62
BA000003		34	97	23	59	67	60
Prochl. marinus							
CP000111		4	84	3	47	75	56
CP000095		4	84	2	44	50	52

Our Proposal

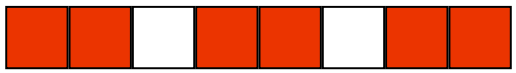
- **Goal** : work on the **anchor-based approach**
  - improve sensitivity and reliability
  - especially for divergent cases
- **Our Proposal** : focus on phase **P1** (**fragments computation**)
  - original idea – **use local similarities (LS) as fragments**, instead of short, unique/approximate matches
  - LS from seed-and-extend methods, like BLAST – HSPs and YASS – SPACED-SEEDS (L. Noé et al., 2005)
  - 1st hypothesis – **LS improve sensitivity**

Our Proposal (II)

Spaced-seeds

patterns that define matches for which we ignore several positions

e.g. *we ignore the* 3*rd and the* 6*th position out of* 8



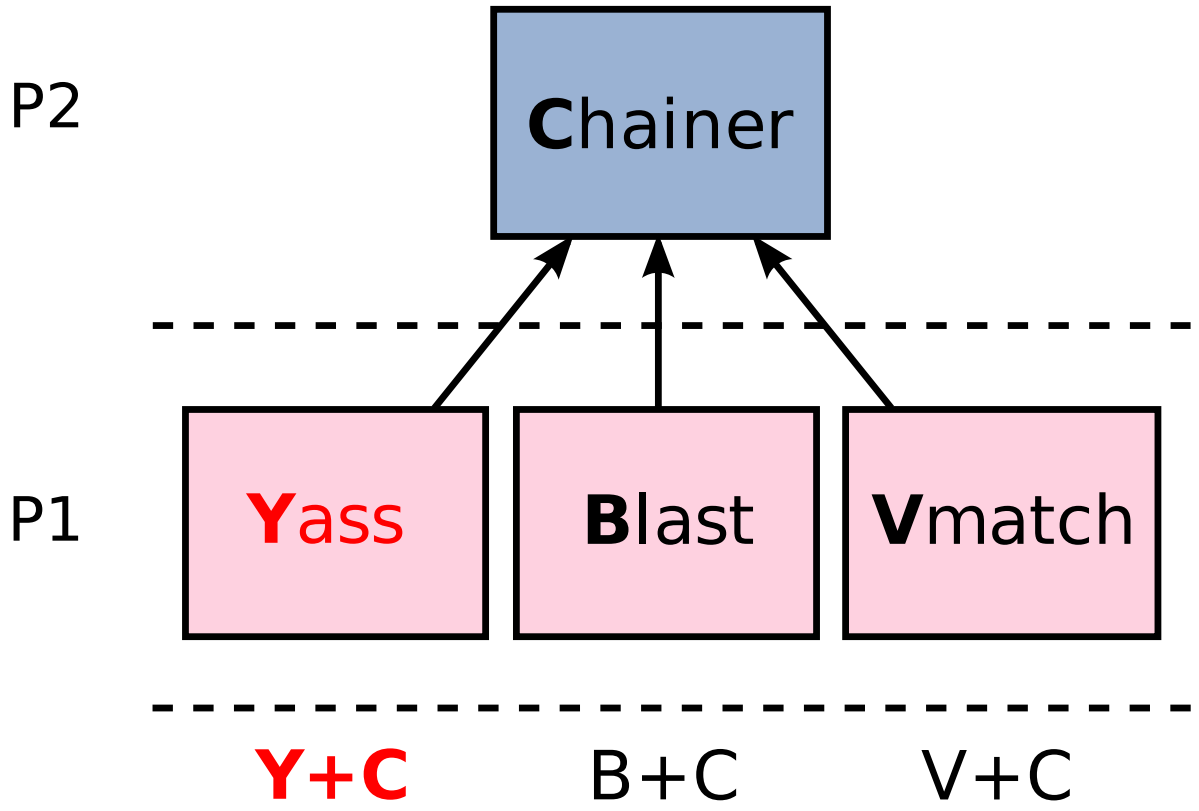
a c a t t g a c
a c c t t c a c

- increased sensitivity of spaced-seeds compared to contiguous-seeds (L. Noé et al., 2005)
- 2nd hypothesis – **for divergent cases, spaced-seed LS significantly improve the sensitivity of the anchor strategy**

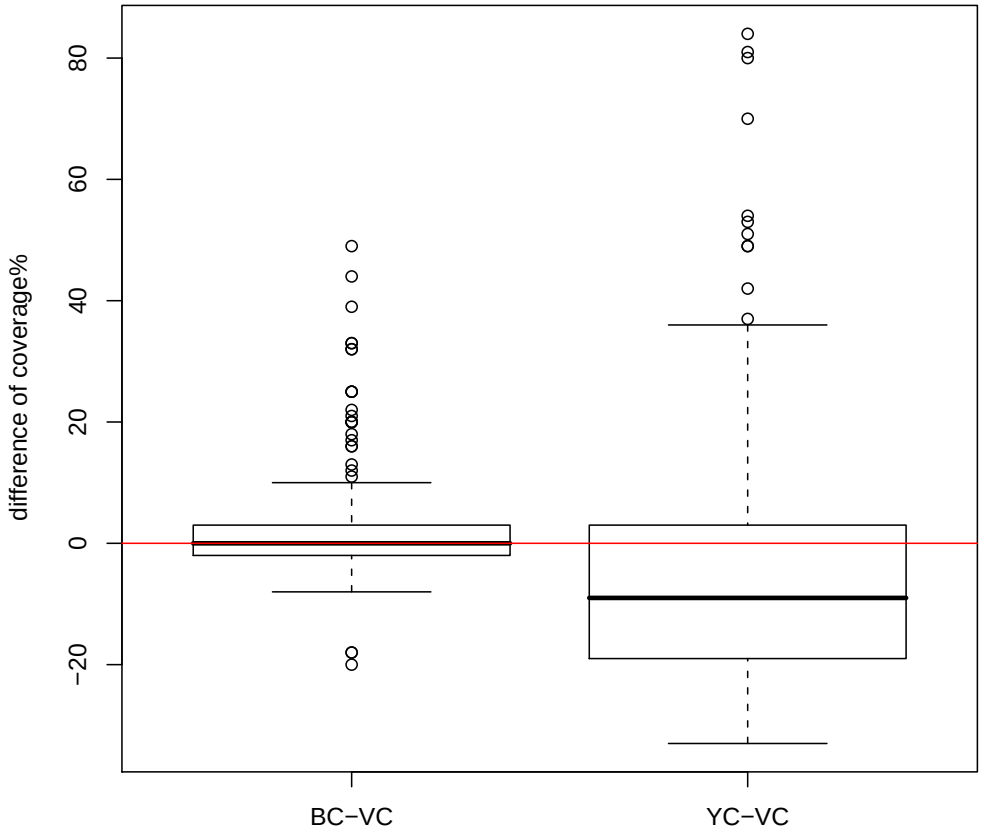
Test Prototype

- we propose 2-phase prototypes based on **P1** + **P2** of the anchor strategy
- **P1** – test the impact of 3 methods for computing fragments :
  - one MEM method : VMATCH (V)
  - two LS seed-and-extend methods : BLAST 2 (B) and YASS (Y)
- **P2** – the classical chaining method : Chainer (C)
- $\longrightarrow$ 3 prototypes : V+C, B+C, Y+C
- same dataset, same comparison measures

Test Prototype (II)



Preliminary Results



Box-and-whisker plots for coverage differences between :

B+C/Y+C and V+C

Conclusion

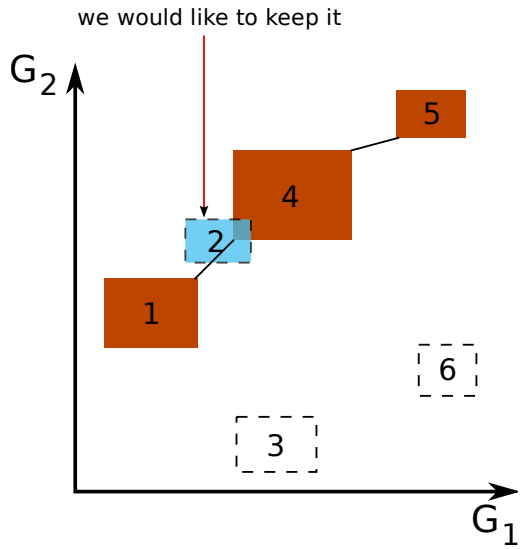
- BC does better than VC in half of the cases
- unexpected : YC is mostly dominated by VC

Remarks on the Preliminary Results

Classical chaining proves to be unsuited to LS, especially spaced-seeds LS :

- extreme lengths, border effects (tandem repetitions, extension phase), YASS clustering method → **overlaps**
- remember *chaining definition* [a] : **overlaps are forbidden**

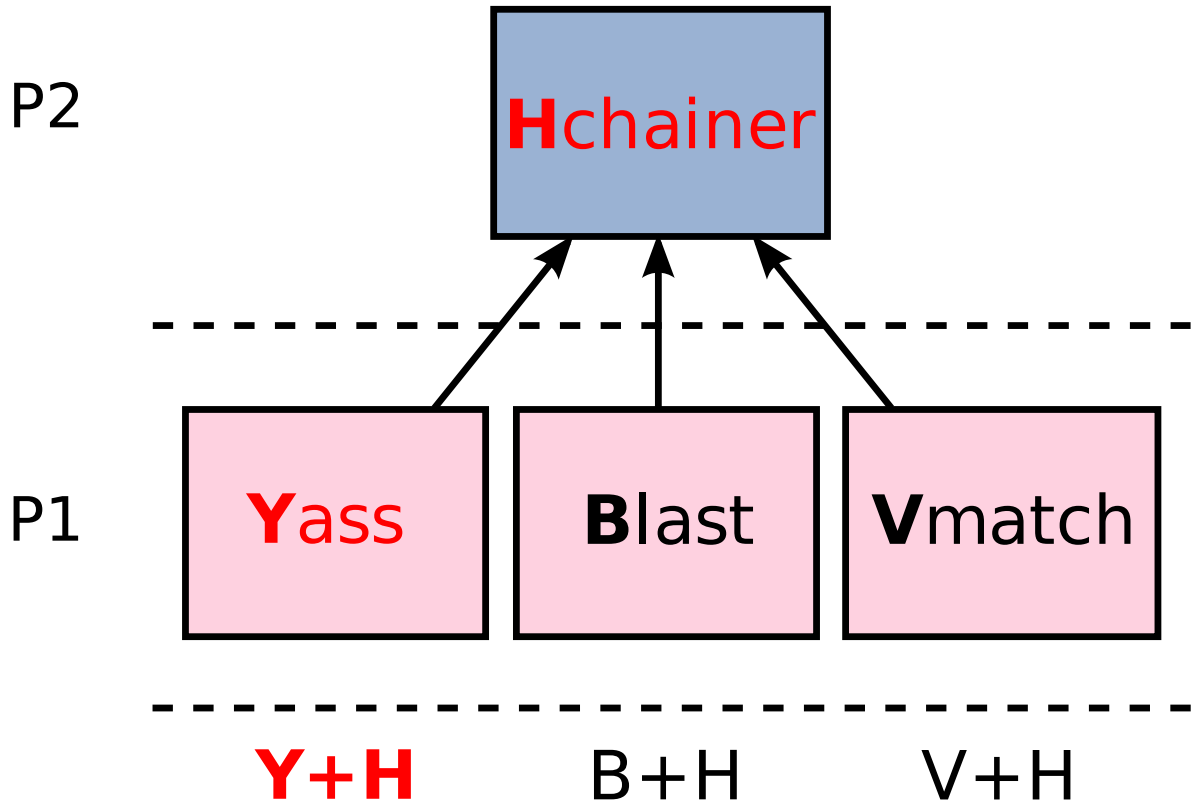
[a] computation of the chain of collinear, **non-overlapping** fragments . . .



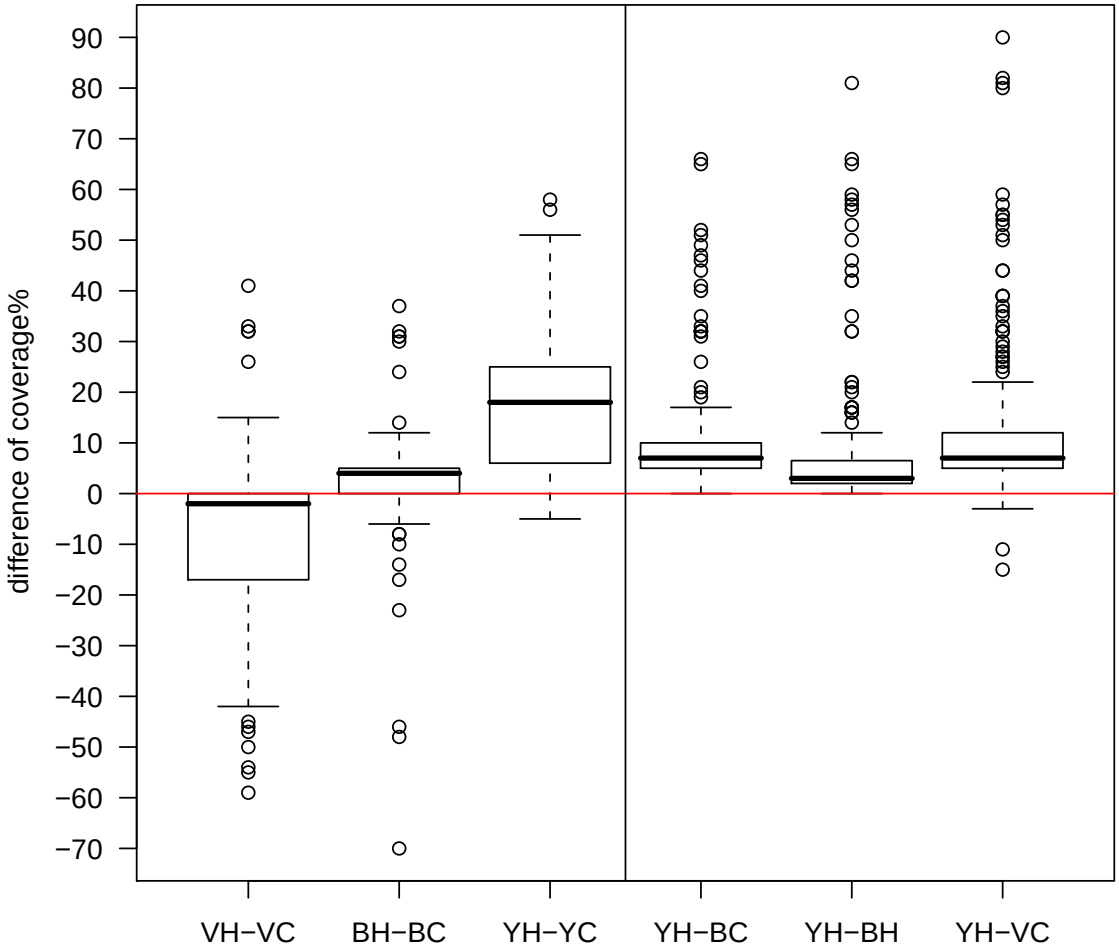
Improvements and Final Test Prototype

- we need a chaining algorithm that deals with overlaps ;
- we propose a novel collinear chaining algorithm that accepts limited overlaps :
  heuristic : **hierarchical greedy method** (H)
- we performed tests for V, B, Y with the novel chaining method, H
  $\longrightarrow$ 6 protypes : V+H, B+H, Y+H

Final Test Prototype (II)



Results for the 6 Prototypes



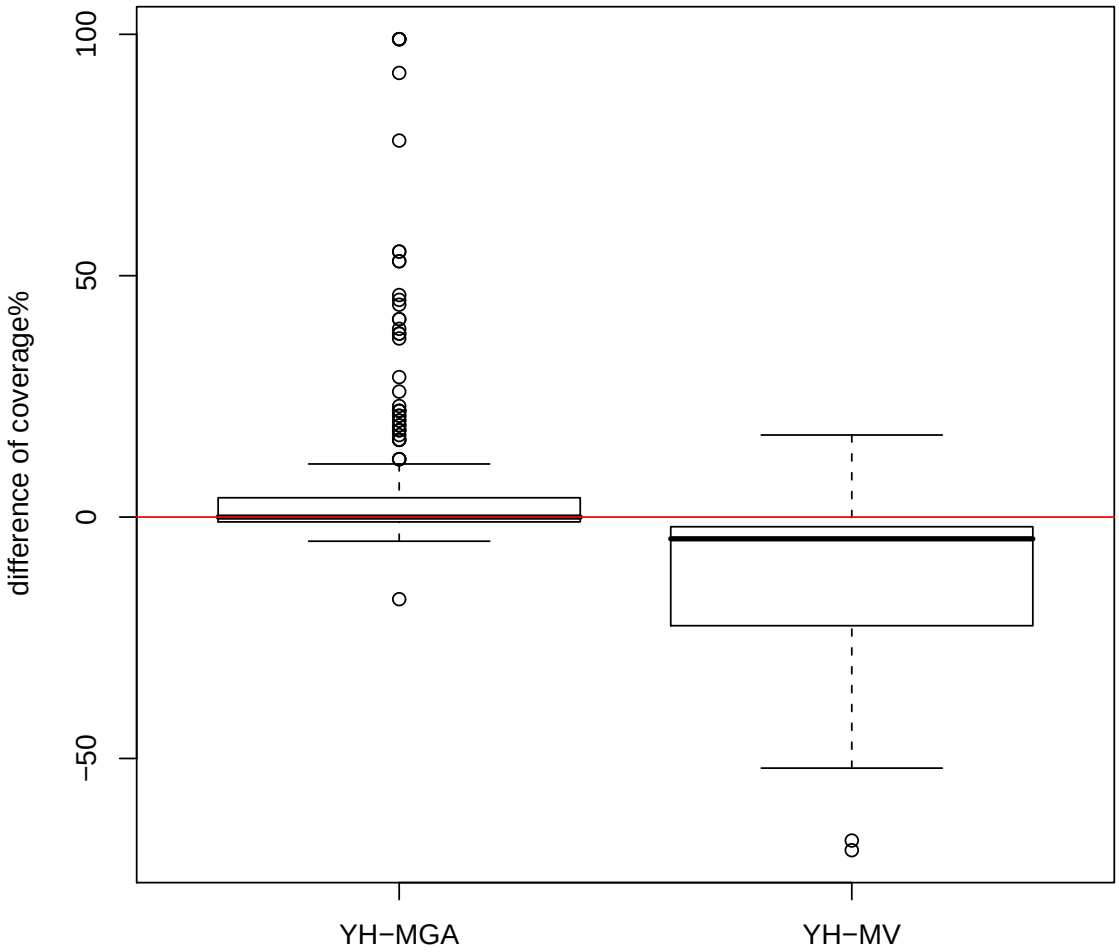
Box-and-whisker plots for coverage differences between the 6 prototypes.

Y+C, Y+H, B+C, B+H, V+C, V+H

1. for all computing fragments methods, M, compute MH-MC
2. **YH compared to the other 3 best combinations**

Conclusion YH obtains higher coverage in the vast majority of cases over all other combinations.

YH versus MGA and MAUVE



Box-and-whisker plots for coverage differences between YH and MGA/MAUVE

Conclusion

1. YH - MGA :
   YH mostly improves or eq. MGA
   $[-17, 99]$%, avg 7.2%
   140 cases ≥ 0%,
   10 cases < -2%
2. YH - MAUVE
   MAUVE improves on YH
   in average, MAUVE covers 13% more nucleotides than YH

Remarks YH vs MAUVE

- the "excellent" results of MAUVE are, as expected, mostly due to its capacity to handle rearrangements
- however, MAUVE high coverages sometimes hide unreliable alignments
  remember the table with divergent cases

  Assumption :
  segments aligned in P4[a] by MAUVE do not necessarily share sequence similarity and are often unreliably aligned

[a] *"last chance alignment" with an exact alignment method*

Summary

Preliminary GRAPE Expertise on *P. marinus* Couple for MAUVE and YH

GRAPE(Lunter et al., 2007)

- for the 2926 MAUVE aligned segments (having id% < 100%)
  - in avg, they have 30% of unalignable positions
  - 21% of them have $>$ 50% of unalignable positions

total segm. length	length of segm. with $>$ 50% unalign. pos.
$\approx$ 2 Mb	$\approx$ 1 Mb

- for the 488 YH fragments
  - in avg, fragments have 6% of unalignable positions
  - 1 fragment has $>$ 50% of unalignable positions

GRAPE Expertise on the Complete Dataset

- tools : YH, MAUVE
- YH : 0.22% of the seq total length, filtered in average (max 3%) ;
- MAUVE : 4.1% of the seq total length, filtered in average (max 55%) ;

Biological Validation Protocol using Orthologous Genes

- pairs of orthologs and their genomic positions are taken from OMA database ;
- the set of genome pairs having the same accession numbers in our set and in OMA : 161 pairs, 34 species ;
- we computed the **nb of orthologs completely included in backbone segments, VS, or overlapping both** ;
- MAUVE and YH coverages are considered as being equal if they have $\leq$ 1% difference on both genomes ;

Biological Validation Results using Orthologous Genes

- from 33 **pairs having equal cov**, YH completely includes in the backbone **more orthologs** than MAUVE in 32 pairs ;

  e.g. *S. pyogenes* pair : MAUVE covers additional 0.8% of the genomes, but completely aligns 5% less orthologs than YH

- for **rearranged cases**, MAUVE achieves better cov and thus better ortholog cov ;

- for **divergent cases**, YH usually surpasses MAUVE ;

  e.g. *P. marinus* pair : YH covers additional 7% of the genomes, but completely aligns 53% more orthologs than MAUVE

Lactococcus Lactis (IL1403 vs SK11) I

Mauve Alignment

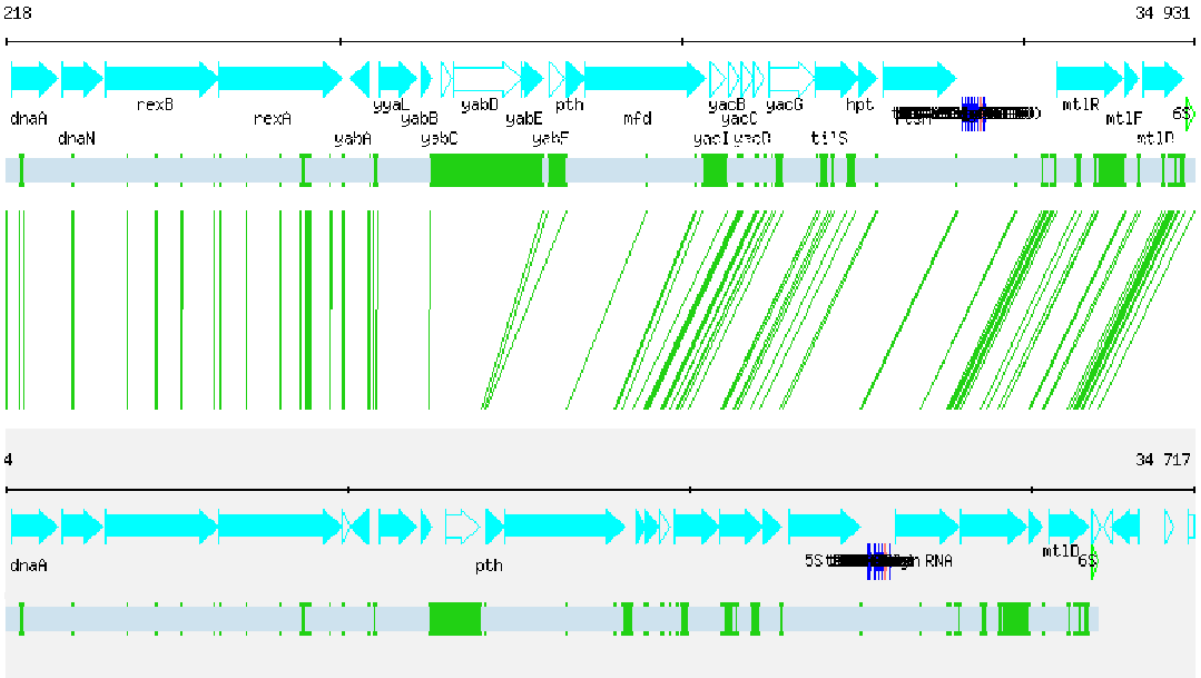


FIG.: Mauve alignment segmentation of the first 34 kbp collinear block of the genomes. Genes are displayed as arrows, and variable segments as green boxes. Mosaic server image.

Lactococcus Lactis (IL1403 vs SK11) II

YH Alignment

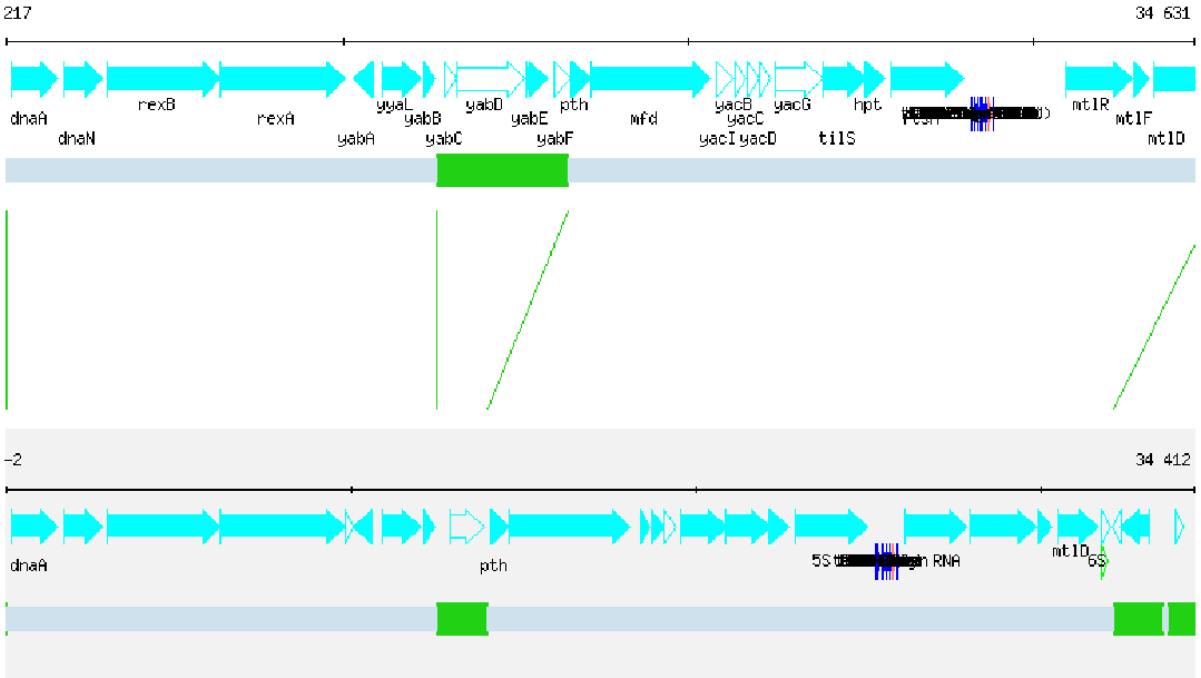


FIG.: YH alignment segmentation of the first 34 kbp collinear block of the genomes. Genes are displayed as arrows, and variable segments as green boxes. Mosaic server image

Summary

Discussion

Spaced-seed LS with adapted chaining . . .

- **improve the sensitivity of classical genome alignment methods**
- **simplify genome alignment visualisation**
  a lot less fragmented main diagonal in the dotplot
- **avoid recursion phase**
  tests show that recursion phase does not significantly improve the alignment in this case

Discussion (II)

Also, spaced-seed LS with adapted chaining . . .

- **produce more accurate alignment bounds than both MGA and MAUVE**
  tests using orthologous genes from OMA *database*
- **produce more reliable alignments (without recursion)**
  LS are selected on their E-value

Work in Advancement

- **exact method** for chaining with proportional overlaps ;
- **extended nb of tools** : PROGRESSIVEMAUVE, LAGAN ;
- **extended dataset** : 694 bacteria couples ;

Support :

CocoGEN
http://www.lirmm.fr/~uricaru/CoCoGEN

Thank you for your attention !

Questions ?