

# Improved sensitivity and reliability of anchor based genome alignment

---

Raluca Uricaru<sup>1</sup>, Célia Michotey<sup>2</sup>, Laurent Noé<sup>3</sup>, H elene Chiapello<sup>2</sup>,  
Eric Rivals<sup>1</sup>

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

---

9 juin 2009



# Summary

- 1 Introduction
- 2 Methods
- 3 Contributions
- 4 Conclusions

# About me

- 3<sup>rd</sup> year PHD student in computer science  
**M**éthodes et **A**lgorithmes pour la **B**ioinformatique (**MAB**) team,  
LIRMM, Montpellier
- supervisors : ERIC RIVALS & HENRI VIAL
- main research field : *bioinformatics*
- PHD subject : *Algorithms for the comparison of complete genomes*
- for further details on my research interests see :  
**[www.lirmm.fr/~uricaru/](http://www.lirmm.fr/~uricaru/)**

# Summary

- 1 Introduction
- 2 Methods
- 3 Contributions
- 4 Conclusions

# Two Kinds of Sequence Alignment

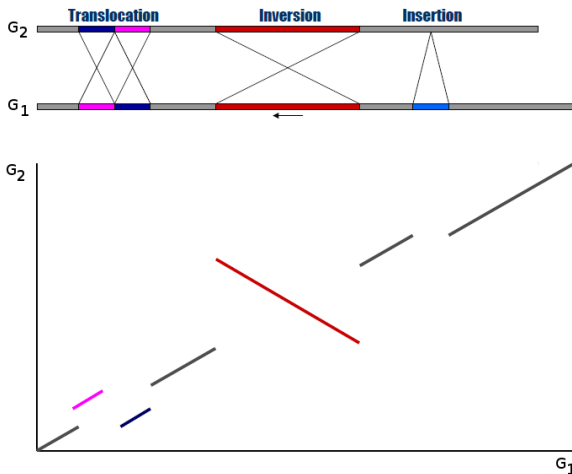
## Local alignment

*finds and extracts **sets of regions**, one region from each of the given sequences, which exhibit **high similarity** (e.g. **Blast**)*

## Global alignment of complete genomes

*computes the best scoring solution for **unique mapping** of similar regions between sequences*

# Pairwise Global Genome Alignment



[S. Kurtz et al. – MUMMER]

Improved sensitivity and reliability of anchor based genome alignment

# 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

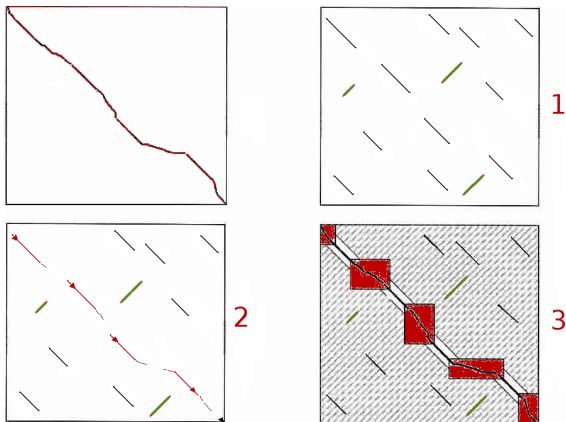
# Summary

- 1 Introduction
- 2 Methods**
- 3 Contributions
- 4 Conclusions



# 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, ...)

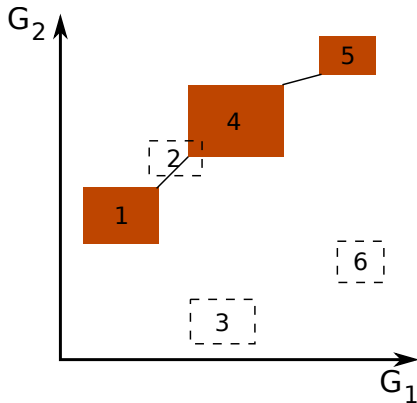
**MAXIMAL UNIQUE MATCH** = maximal exact match, unique in both sequences



- 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

- 1 Introduction
- 2 Methods
- 3 Contributions**
- 4 Conclusions

# 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*%<sup>1</sup>, *id*%<sup>2</sup>)

- **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%

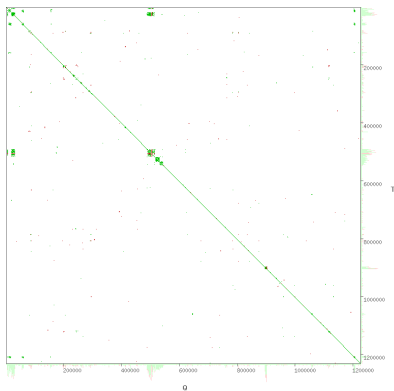
---

<sup>1</sup>ratio of the total length of aligned segments over the genome size

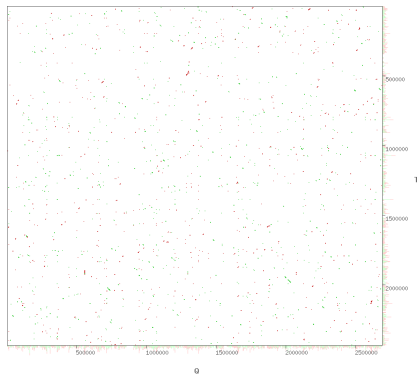
<sup>2</sup>**new definition** : the ratio of identical bases in aligned segments over the genome size



# Examples of Different Levels of Similarity

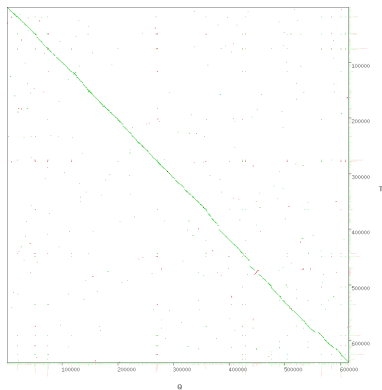


*Chlamydia pneumoniae*

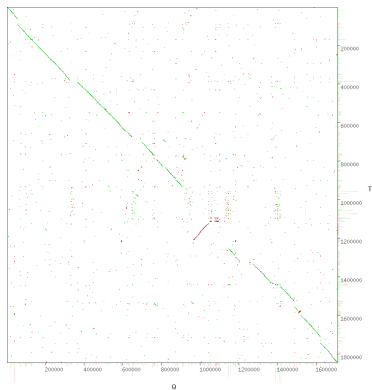


*Synechococcus sp.*

# Divergent Cases



*Buchnera aphidicola*



*Prochlorococcus marinus*

## Divergent Cases (II)

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

These types of cases need to be improved ...

# 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 3rd and the 6th position out of 8*



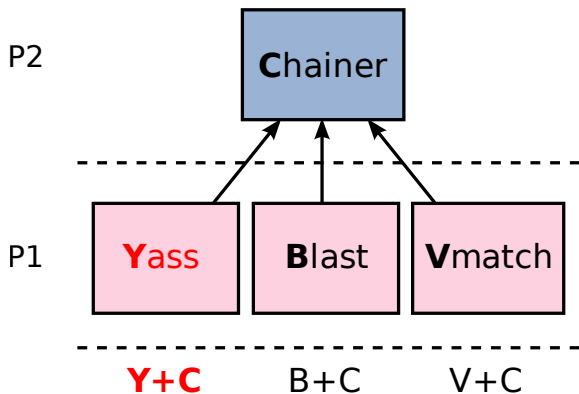
a c a t t g a c  
a c c t t c a c

- 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

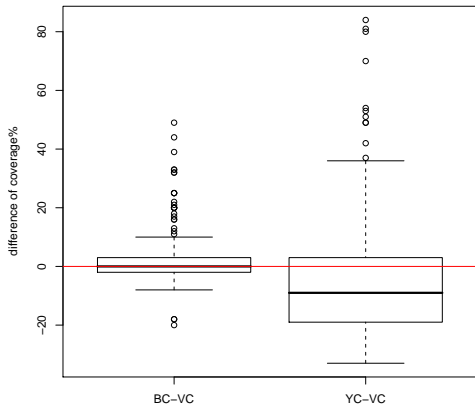
# Test Prototype

- we propose a 2-phase method 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)
- → **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

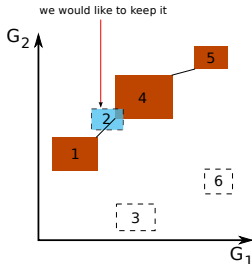


## Remarks on the Preliminary Results

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*<sup>a</sup> : **overlaps are forbidden**

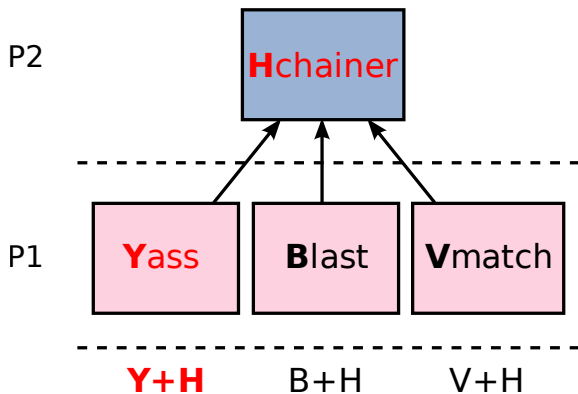
<sup>a</sup>computation of the chain of collinear, **non-overlapping** fragments ...



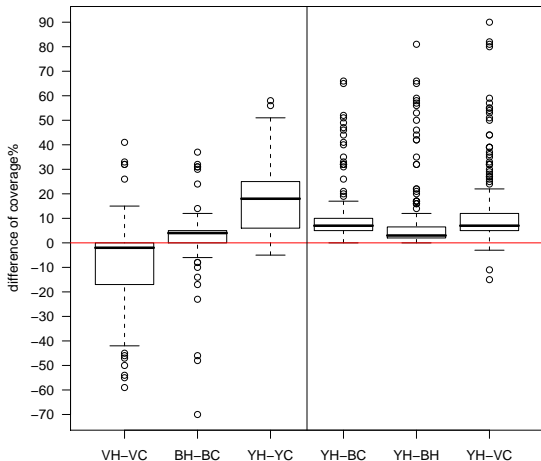
# Improvements and Final Test Prototype

- 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  
→ **6 prototypes** : V+H, B+H, Y+H

## Final Test Prototype (II)



# Results for the 6 Prototypes



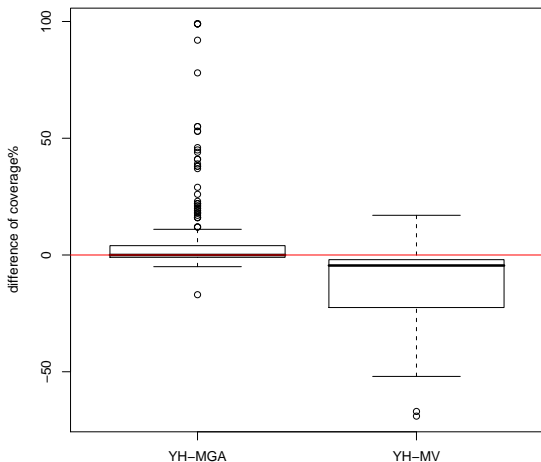
Box-and-whisker plots for coverage differences between the **6 prototypes**.

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  $\geq 0\%$ ,  
 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<sup>a</sup> by MAUVE do not necessarily share sequence similarity and are often unreliably aligned**

---

<sup>a</sup>*“last chance alignment” with an exact alignment method*

## Verify Assumption : Studies on Reliability

Strains	Collinear	Cov%		Id%		Cov% Mv-YH	Id% Mv-YH	Id% addit. cov Mv
		YH	Mv	YH	Mv			
<b>Buchnera aphidicola</b>								
AE016826	X	83	99	56	62	16	6	37
BA000003		78	97	53	59	19	6	31
<b>Prochl. marinus</b>								
CP000111		59	84	39	47	25	8	32
CP000095		55	84	36	44	29	8	27

- MAUVE id% of coverage : Buchn. aph. 62%, Prochl. mar. 54%
- YH id% of coverage : Buchn. aph. 68%, Prochl. mar. 66%

# Further Studies on Reliability

## Prochlorococcus marinus case

Use GRAPE(Lunter et al., 2007) to estimate the quality of MAUVE **P4** segments and YH fragments.

- for the 2926 MAUVE **P4** segments
  - in avg, they have **30%** of unalignable positions
  - **21%** of them have **> 50%** of unalignable positions

segment length	length of segments with <b>&gt; 50% unalign. pos.</b>
≈ 2 Mb	≈ 1 Mb

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



# Summary

- 1 Introduction
- 2 Methods
- 3 Contributions
- 4 Conclusions**

# 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*

## Future work

- an exact algorithm for chaining with limited overlaps (in work)
- extend our method to **multiple genome alignment**
- relax the constraint of collinearity in order to deal with complex rearrangements : translocations, transpositions  
**NP-complete problem**
- extend experiments to **complex genomes**

Support :

CocoGEN

<http://www.lirmm.fr/~rivals/>

PlasmoExplore

<http://www.lirmm.fr/~brehelin/PlasmoExplore/>

Thank you for your attention !

Questions ?