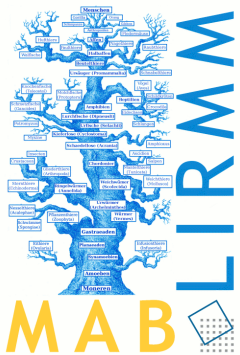
CoCoGen meeting

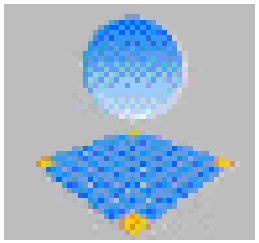
Accuracy of the anchor-based strategy for genome alignment

Raluca Uricaru

LIRMM, CNRS Université de Montpellier 2

3 octobre 2008





Summary

Summary

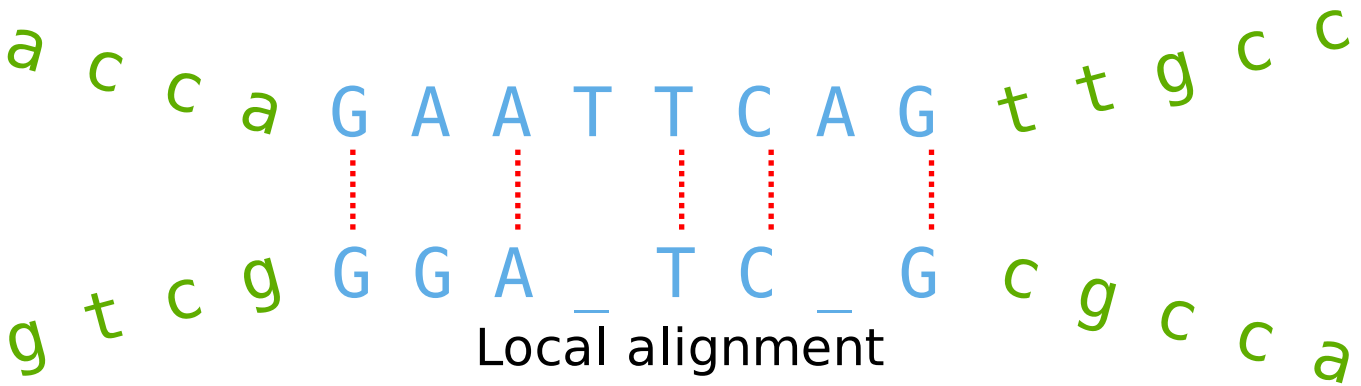
Motivation

Do I really have to explain this to YOU ?

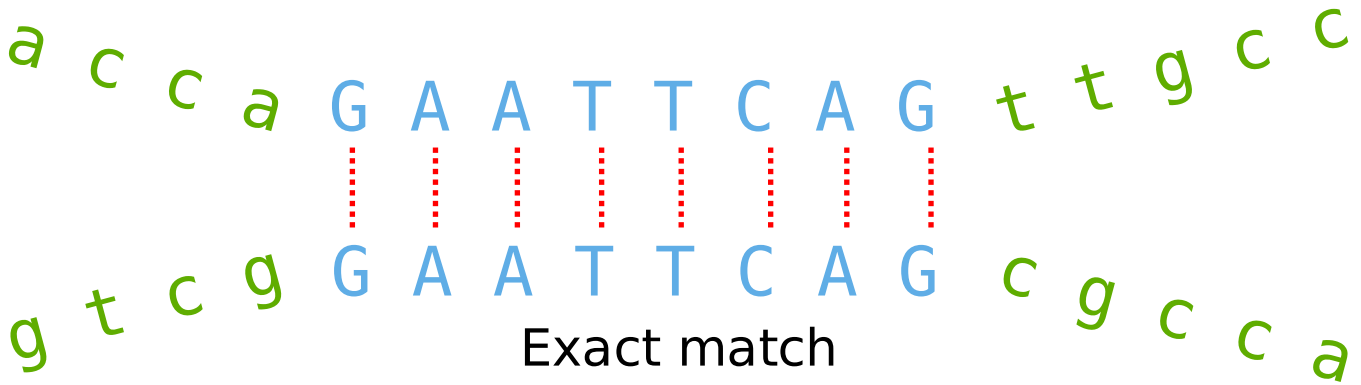
We are working along with the people from the

MOSAIC project
segmentation of bacterial genomes :
backbone & **variable segments**

Local alignments vs exact matches



Local alignment



Exact match

Genome alignment approaches

Local alignment

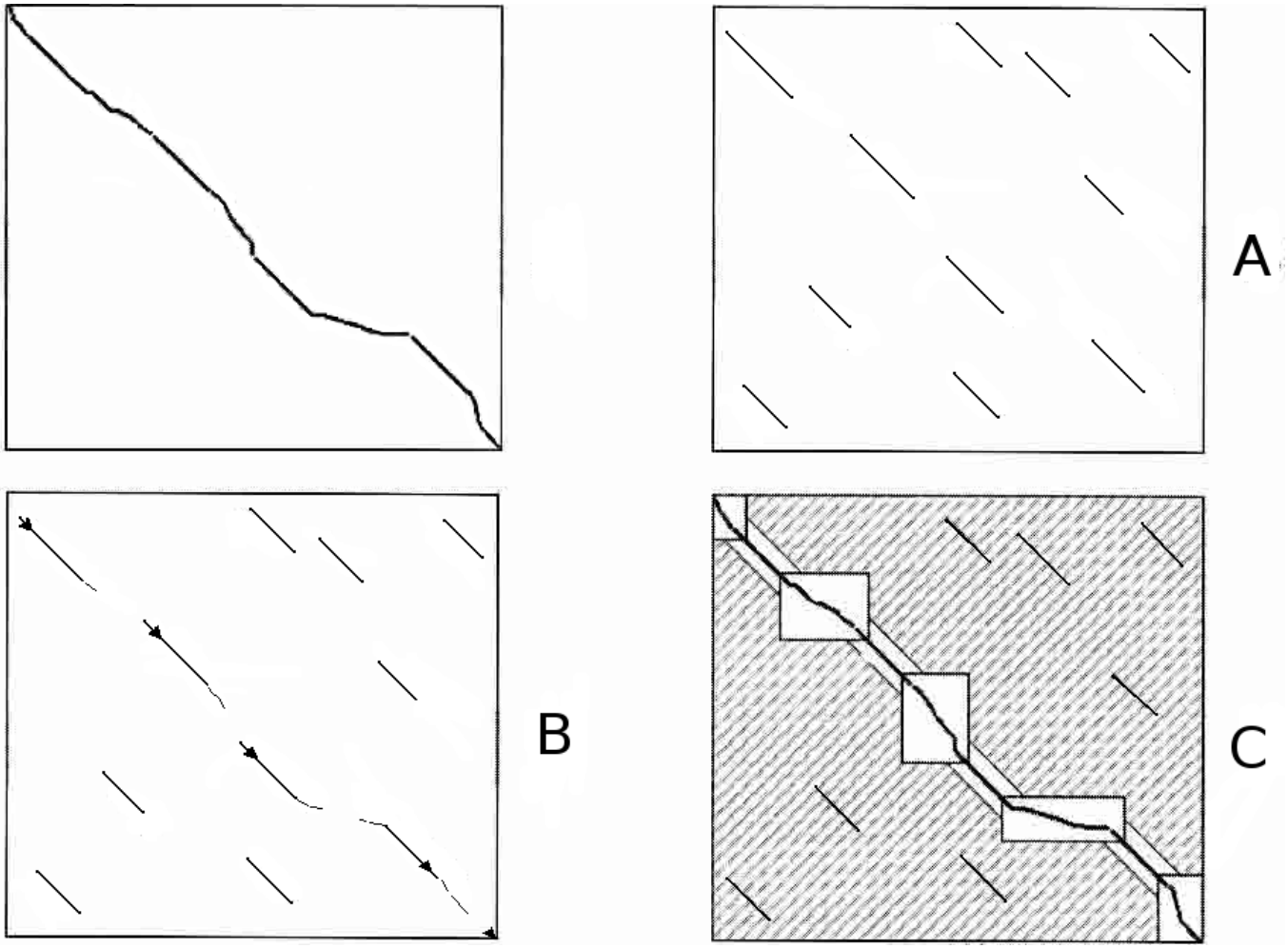
- in case of **genomic rearrangements**
- align common regions instead of the entire genomes
- for big, distant, rearranged genomes (e.g. *eukaryotes*)

Global alignment

- in case of collinear genomes
- technique : anchor-based, 3-phase method
- for closely related, non-rearranged genomes (e.g. *bacteria*)

Summary

Anchor-based, 3-phase method



Anchor-based, 3-phase method (2)

1. Computation of fragments
   pairwise similarity regions

2. Anchors computation (*chaining phase*)
   - *basis of the alignment*
   - *subset of non-overlapping fragments forming an optimal scoring, global chain*

3.
   - Recursive anchoring
     for each pair of yet unaligned regions between the anchors
   - "last chance align." with **ClustalW** [J.D. Thompson et al., 1994]

Global alignment (GA) programs

Present global alignment programs are heuristic variations on the **anchor based strategy**

- **MGA** [M. Hohl et al., 2002] (*does not deal with rearrangements*)
  1. compute **M**aximal **E**xact **M**atches (VMatch [S. Kurtz, 2003])
  2. select the set of anchors (Chainer [M.I. Abouelhoda et al., 2005])
     subset of non-overlapping MEMs forming the maximal scoring **consistent chain**

     consistent chain = increasing on both genomes
  3. recursive anchoring
     "last chance alignment"

Global alignment programs (2)

- **MAUVE** [A.C.E. Darling et al., 2004] (*handles rearrangements*)

  1. search for approximate matches using **spaced seeds**
  2. select a subset of non-overlapping anchors
     without imposing the consistency property

     use a heuristic consisting in finding a minimum partition in **L***ongest* **C***ollinear* **B***locks*
  3. *recursive anchoring*
     “last chance alignment”

Global alignment programs (3)

Non investigated questions on the *anchor based strategy*

Lack of evalution for :

- global accuracy
  benchmark, evaluation criteria & significance measure
- the influence of exact & approximate matches
- the impact of $\neq$ types of matches on chaining

1. Does the *anchor based strategy* produce satisfactory results on non-rearranged genomes ?
2. How can one deal with rearrangements ?

2 initial prototypes

To address the 1st previous question, we designed 2 prototypes for pairwise genome alignment without rearrangements

1. **VMatch + Chainer** (VC)

   simulate the first 2 phases of **MGA**

2. **Yass + Chainer** (YC)

   same outline as **MGA**
   replace **VMatch** by **Yass** [L. Noe et al., 2005]

   Yass = *fast similarity search tool based on spaced seeds*

   use local alignments instead of MEMs

Comments on the 2 prototypes

- frequent overlaps due to the use of local alignments (2^{nd} prot.)
  - overlaps should be taken into account in the chaining
  - proposed solution : "hierarchical chaining method" (3^{rd} prot.)
- not dealing with rearrangements may generate poor results :
  - singular inversions, complementary strands (all 3 prot.)
  - circularity (3^{rd} prot.)
  - other rearrangement events :
    successive inversions, translocations, duplications, . . .

    need chaining without consistency property : NP-complete prob ?

the 3rd prototype

3. **Yass + Hierarchical** (YH)

   replace the first 2 phases of **MGA** :

   - *use local alignments instead of MEMs*
   - *replace Chainer by* **greedy, hierarchical chaining method**

   Hierarchical method

   - *keep stronger simil. regardless of their conflicts with weaker ones*
   - *multiple level chaining heuristic, dealing with overlaps*

Summary

Benchmark

MOSAIC [H. Chiapello et al., 2005]

- public curated database of backbone alignments for intra species bacterial genomes
- 42 species from Genome Reviews [P. Sterk, 2006], except 5

  Buchnera aph., *Prochl. marinus*, *Synechococcus sp*, *Pseudomonas fluor.*, *Rhodopseudomonas pal.*

  excluded due to not satisfactory alignments
- alignments are made with :
  MAUVE (rearranged) & **MGA** (non-rearranged)
- parameters calibrated on *E-coli*
- backbone alignments were post-processed
  keep aligned gaps $\geq$ 76% *id%*

Global evaluation criteria

- complexity of the anchor chain
  number of anchors in the chain
- coverage
  total length of regions common to both sequences
  filtered cov. = cov. without aligned gaps <76% id%
- identity percentage (id%)
  ratio of identical bp in the segm. from the cov., over gen. len.
  filtered id%. = id% without aligned gaps <76% id%

Comparison protocol

- 42 bacterial species (236 intra-species genome pairs)
  - close, moderately divergent, rearranged genomes

  ⋆ *for exceptions, see the 5 species excluded from MOSAIC*
  - 94 collinear pairs, 142 rearranged pairs
  - existence of benchmark (**MOSAIC**)
- compare alignments obtained with **MGA**, **MAUVE**, **VC**, **YC** & **YH**
- 3 evaluation criteria (*see prev. slide*)
- tested several parameters and options, kept the best ones

Present achievements (MOSAIC)

MGA and MAUVE

- species with average good coverages
  Staphylococcus aureus : >90% both **MGA** and **MAUVE**
- coverage differences in different species
  Streptococcus thermophilus : 97% both **MGA** and **MAUVE**
  Synechococcus sp : <10% both **MGA** and **MAUVE**
- coverage differences inside a species
  Prochlorococcus marinus : [0, 78]% **MGA**, [6, 96]% **MAUVE**

Programs succeed in aligning some pairs and fail in others

Causes :

high level of divergence, rearrangements

Local similarities vs MEMs

YC/YH vs VC

YC vs VC

- compute the difference in cov. **YC**-**VC**
- $[-69, 81]$%, positive in 89 (over 236) cases
- unexpected bad results for close genome cases

  Cause : large overlaps of local align.
  Solution : **YH**

YH vs VC

- **YH** improves the cov. over **VC** : 11% in avg
- **YH** improves the cov. over **YC** : 19% in avg
- Example CP000046 vs BA000018 (*Staphylococcus aureus*)
  YC 64%, **VC** 88%, **YH** 99%

Hierarchical method vs MOSAIC

YH vs MGA/MAUVE

YH vs MGA on the 94 collinear pairs

- compute the difference in cov/id% **YH-MGA**
- cov : $[-4, 45]$%, 3% in avg
- positive in 52 cases, <-2% in 3 cases
- id% : $[1, 45]$%, 9% in avg

YH vs MAUVE on the 142 rearranged pairs

- **MAUVE** covers 21% more nucleotides than that of **YH**
- **MAUVE** has 18% more identities than **YH**

Summary

Discussion : evaluation of the anchor-based strategy

- we do not solve rearrangements and we do not do recursion steps
  However, we obtain
- gain in cov, id% and simplicity of chain complexity
  of **YH** over **MGA** and **VC**

  nb of anchors : <150 **YH**, >3000 **MAUVE**, >13000 **MGA**

Discussion : evaluation of the anchor-based strategy (2)

- in terms of filtered coverage and id%, **YH** improves **MGA** . . .
- and **MAUVE**, only in the case of

  divergent genomes (thanks to spaced seeds ?)
  what is happening with the aligned gaps ?

  Example (*Buchnera aphidicola* - 3 genome pairs)
  filtered coverage : **YH** improves **MGA**/**MAUVE** by 69%
  filtered id% : **YH** improves **MGA**/**MAUVE** by 47%

Conclusion

- pairwise alignment is incompletely solved nowadays
- heuristic anchor methods can be improved :
- quantitatively
  - local similarities
  - chaining with overlaps
- qualitatively : E-values for local similarities
  - improved backbone alignment
  - avoid alignment post-processing
- in terms of chain complexity
  - computationally less expensive
  - easy to biologically interpret

Short time perspectives

- recursive anchoring
  - successively search for similarities in non-aligned regions
  - "chained spaced seeds"

    knowing that the spaced seed (set) A is not found on step i, compute the best spaced seed (set) B for step i + 1
- dealing with rearrangements :
  method for non-consistent global alignment allowing overlaps

Short time perspectives (2)

- biological validation

  global, average values are not entirely relevant
- reinforce (test) the gain from using spaced seeds

  compare to BLAST
- use negative filters

  detect unique regions
  - validation method on the non-aligned regions
  - pre-processing step

Bibliography

J.D. Thompson, D.G. Higgins, T.J. Gibson
CLUSTAL W : improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice
Nucleic Acids Research, 1994

L. Noe and G. Kucherov
YASS : enhancing the sensitivity of DNA similarity search
Nucleic Acids Research, 2005

H. Chiapello, I. Bourgait, F. Sourivong, G. Heuclin, A. Gendrault-Jacquemard, M-A Petit and M. El Karoui
Systematic determination of the mosaic structure of bacterial genomes : species backbone versus strain-specific loops
BMC Bioinformatics, 2005

Bibliography (2)

M. Hohl, S. Kurtz and E. Ohlebusch
Efficient multiple genome alignment
Bioinformatics, 2002

A.C.E. Darling, B. Mau, F.R. Blattner and N.T. Perna
Mauve : Multiple Align. of Conserved Gen. Seq. With Rearr.
Genome Research, 2004

M.I. Abouelhoda and E. Ohlebusch
Chaining Algorithms for Multiple Genome Comparison
JDA, 2005

S. Kurtz
VMatch, technical report, 2003

P. Sterk, P.J. Kersey and R. Apweiler
Genome Reviews : Standardizing Content and Representation of Information about Complete Genomes
OMICS, 2006

- Support :

  CocoGEN
  http://www.lirmm.fr/~rivals/

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

- Supplementary material available at :
  www.lirmm.fr/~uricaru/Appendix_RecombCG.html

Thank you for your attention !

Questions ?