# CIGAL: Common Intervals Global ALigner

**Guillaume Blin**[1], **Annie Chateau**[2], **Cedric Chauve**[2], **Yannick Gingras**[2]

(1) Institut Gaspard Monge, France (2) CGL, Université du Québec à Montréal, Canada

## INTRODUCTION

In most of existing global aligner, an anchor-based strategy is used: the alignment is guided by exact or non-exact local alignments, then is completed by different methods to produce a global alignment [5]. We recycle this idea to align gene orders. Our data consist in two genomes represented by two sequences of signed identifiers. Those identifiers can be genes, gene families, or any other kind of genomic markers.

We define exact anchors with duplications by considering the generalized common intervals between two genomes [3]. The problem of finding a maximal cover with a minimal number of common intervals involved is **NP-complete**, which we prove here. Thus we use a heuristic approach to compute a covering of the genomes by our anchor. We complete the alignment by releasing conditions on the common intervals in an iterative process.

The alignment can be used to compute breakpoints, conserved intervals or common intervals distances between species, in a comparative genomic purpose like phylogenomic reconstruction, or visualization of repeats [1].

## COMMON INTERVALS WITH DUPLICATIONS

Given an alphabet $\Sigma$, we define a *character set* as a set of elements of $\Sigma$.
Given one genome $A$, the alphabet $\Sigma$ of its gene content, and a character set $S$, we define a *CS-location* of $S$ in $A$ as an occurrence of a word on $S$ in $A$. The location is *maximal* if this word cannot be extended on the right or on the left.
A CS-location represents a contiguous region in $A$ with gene content exactly $S$.
A CS-factor between two or more genomes is a character set which has at least one CS-location in each genome.

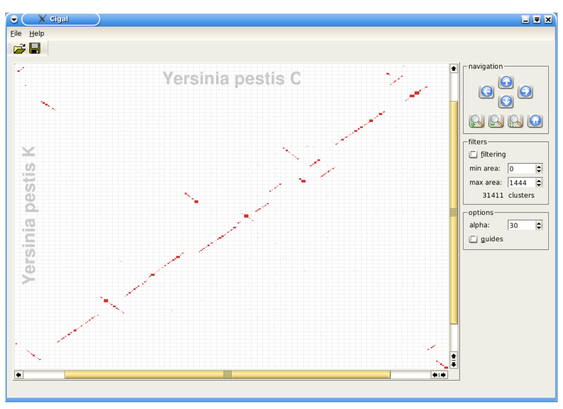
**Exemple:** $S = \{7, 8, 9, 10, 11\}$

```
Genome A:1 2 3 12 11  8  7 9 9 10  4 3 1 1  5
Genome B:6 1 4  8  9 10 11 7  2 13 13 7 8 9 10 11
```

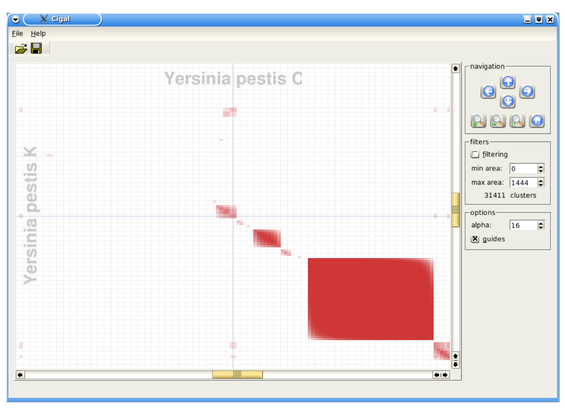
In the following, a *common interval* will refer to the maximal CS-location of a character set in a set of genomes.

## COMPUTATION OF THE COMMON INTERVALS

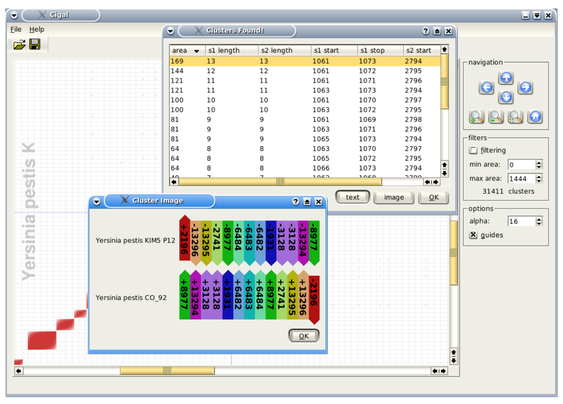
Schmidt and Stoye provide in [6] a quadratic algorithm to compute all the common intervals between two genomes. This algorithm is implemented in the Gecko software [3, 6]. We completed the implementation of this algorithm with a visualizer, **Cigal**, that displays a genome per axis, and the common intervals as a set of boxes.



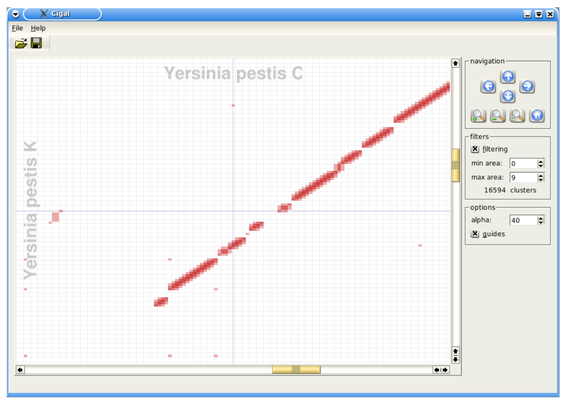

(a) General view of the common intervals between two strands of *Yersina Pestis*




(b) Zoom on a region




(c) Detail of a box with the genes identificator

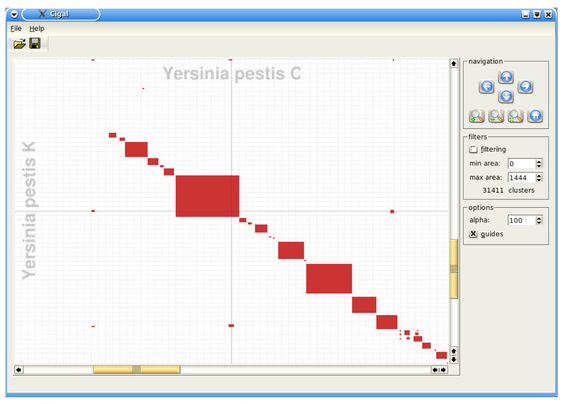



(d) Filtering of big boxes (max area = 9)

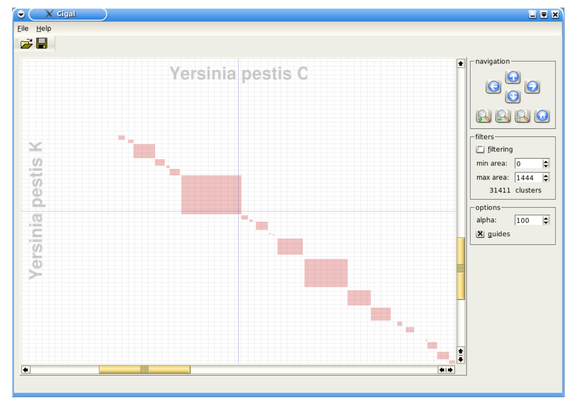
## THE **Minimum Box Full Covering** PROBLEM IS NP-COMPLETE

The **Minimum Box Full Covering** problem is the following problem : Given a set of boxes $B = \{B_1, B_2, \ldots, B_n\}$ and a positive integer $s'$, the problem asks to find a subset $B' \subseteq B$ of cardinality lower than or equal to $s'$, such that (1) given any pair $(B_i, B_j)$ of boxes of $B'$, $B_i$ and $B_j$ are compatible and (2) given any box $B_m \in B$ such that $B_m \notin B'$, $\exists B_i \in B'$ such that $B_i$ and $B_m$ are not compatible ($B'$ is said to be *maximal*).

We provide a polynomial-time reduction from the Minimum Common String Partition problem which has been proved to be NP-complete in [2]. Thus we use a greedy heuristic, which gives good approximation. We sort the boxes by area and iteratively eliminate the incompatible boxes.



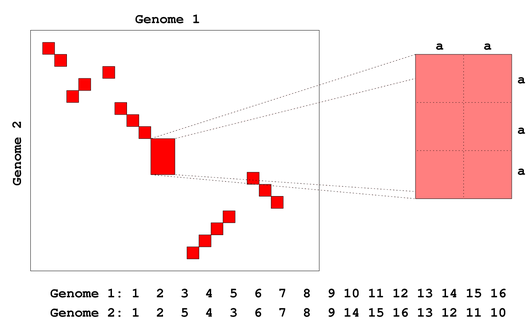

(a) The big box in the middle will be chosen instead of the little ones sharing the same rows or columns

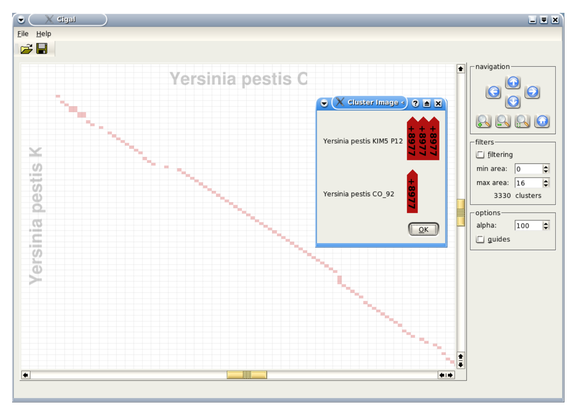



(b) Same region after the covering

## EXTRACTING A PERMUTATION

Once a box covering is computed, we assign orthologs in each box, except when there is a duplication. We obtain a sequence of boxes of size 1 and some boxes of size $\geq 1$ but containing just one gene family. The output permutation is the permutation of the boxes order in one genome and the other.

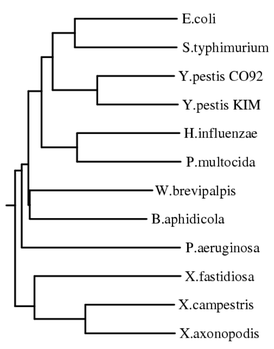


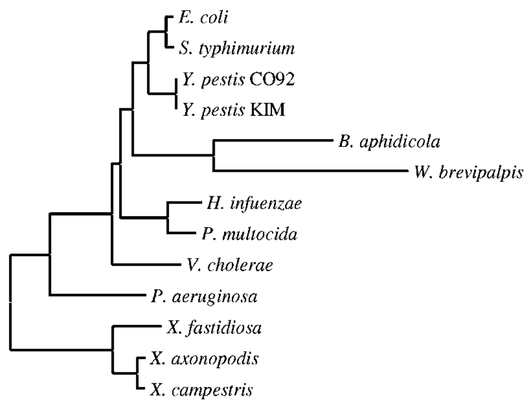

Using the permutation, it is possible to compute some interesting distances using the breakpoints, the common intervals or the conserved intervals between the output permutation and the identity.

## PHYLOGENETIC ISSUES

We computed a distance matrix for the 12 $\gamma-$proteobactia studied in [4] and [1]. The phylogeny was obtained by the `fitch` command available in the PHYLIP package.




(a) The tree obtained with Cigal




(b) Reference tree presented in [4]

## CONCLUSION

This software tool can be used to compute distances in a phylogenomic purpose, but can also be useful in orthologs assignment or complete genomes alignment.
Future works will include algorithms optimization, more options and the extension to multiple genomes alignment.

## References

[1] G. Blin, C. Chauve, G. Fertin. Genes order and phylogenetic reconstruction: application to gamma-Proteobacteria. Comparative Genomics 2005, LNBI 3678:11–20, 2005.

[2] A. Goldstein, P. Kolman and J. Zheng. Minimum Common String Partition Problem: Hardness and Approximations. ISAAC 2004: 484–495, 2004.

[3] S. Heber and J. Stoye. Finding all common intervals of $k$ permutations. In the 12th Annual Symposium in Combinatorial Pattern Matching, CPM 2001, LNCS 2089:207–218, 2001.

[4] E. Lerat, V. Daubin and N.A. Moran. From Gene Trees to Organismal Phylogeny in Prokaryotes:The Case of the $\gamma$-Proteobacteria PLoS Biol 1(1): 101–109, 2003.

[5] C. Notredame. Recent progress in multiple sequence alignment: a survey. Pharmacogenomics. 3(1):131–44, 2002.

[6] T. Schmidt and J. Stoye. Quadratic Time Algorithms for Finding Common Intervals in Two and More Sequences. CPM 2004, 347–358, 2004.

**Address for correspondence: Départment d'Informatique, UQÀM, CP 8888, succ. centre-ville, Montréal (QC), H3C 3P8**