

Improved sensitivity and reliability of anchor based genome alignment

Raluca Uricaru¹, Célia Michotey², Laurent Noé³, Hélène Chiapello², Eric Rivals¹

¹ LIRMM, CNRS and Université de Montpellier 2
161, rue Ada, 34392 Montpellier cedex 5, France
{uricaru, rivals}@lirmm.fr

² INRA UR1077, Unité Mathématique, Informatique & Génome,
Domaine de Vilvert, 78352, Jouy-en-Josas, France
{helene.chiapello, celia.michotey}@jouy.inra.fr

³ LIFL - INRIA Université de Lille I, Villeneuve d'Ascq, France
noe@lifl.fr

Abstract: *Whole genome alignment is a challenging problem in computational comparative genomics. It is essential for the functional annotation of genomes, the understanding of their evolution, and for phylogenomics. Many global alignment programs are heuristic variations on the anchor based strategy, which relies on the initial detection of similarities and their selection in an ordered chain. Considering that alignment tools fail to align some pairs of bacterial strains, we investigate whether this is intrinsically due to the strategy or to a lack of sensitivity of the similarity detection method. For this, we implement and compare 6 programs based on three different detection methods (from exact matches to local alignments) on a large benchmark set. Our results suggest that the sensitivity of well known methods, like MGA or Mauve, can be greatly improved in the case of divergent genomes if one exploits spaced seeds at the detection phase. In other cases, such methods yield alignments that cover nearly the whole genome. Then, we focus on global reliability of alignments: should an aligned pair of segments be included in the global genome alignment? We investigate this reliability according to both the segment "alignability" and to inclusion of orthologs. Again, we provide evidence that for both close and divergent genomes, one of our programs, YH, achieves alignments with sometimes a lower coverage, but a higher inclusion of orthologs. It opens the way to the first reliable alignments for some highly divergent species like *Buchnera aphidicola* or *Prochlorococcus marinus*.*

Keywords: Global genome alignment, anchor based strategy, spaced seeds

1 Introduction

Whole genome comparisons offer a unique opportunity to investigate globally the mechanisms of evolution in closely related species, are a key to the inference of functional elements in both coding and non coding regions, and serve as a basis in phylogenomics [1,2]. In particular, the conserved parts of genomes, forming the so called *backbone segments*, indicate the biological components common to several species or strains, while differences in sequences, *variable segments*, are likely responsible for what distinguishes them (*e.g.*, pathogenic islands). Genome alignment can deliver both at once.

Due to the genome sizes and to the task complexity, genome alignment tools implement heuristic algorithms. The most used scheme is the **anchor based strategy** (*e.g.*, [3,4,5]), which operates in four phases. It starts by detecting an initial set of pairwise similarity regions (phase 1) and, through a *chaining* phase, selects a non-overlapping maximum-weighted subset of those similarities (phase 2), called *anchors*. Phases 1 and 2 are recursively applied to each pair of yet unaligned regions (phase 3). The last phase consists in systematically applying classical heuristic alignment tools (*e.g.*, ClustalW) to all short region pairs still left unaligned.

The Mosaic database stores the alignments of backbone segments for every pair of strains of the same bacterial species. The backbones are obtained by first aligning the genomes with either MGA or Mauve (two anchor based tools), then by post-processing the alignment to remove segment pairs whose percentage of identity falls below 76%. This post-processing, although based on an arbitrary threshold, is still applied to avoid unreliable alignments. Moreover, some pairs of strains are absent from the database because the backbones covered less than 50% of the genome. It is unanswered whether cases of small coverage are due to a lack of sensitivity of the methods or to an intrinsic limitation of the strategy.

Here, we investigate this issue by implementing and comparing six methods that combine three similarity detection methods and two chaining algorithms, and by comparing the results on a large benchmark made of all pairwise intra-species bacterial genomes. As they simulate the first 2 phases of the strategy, those methods can also be compared to MGA and Mauve results. It turns out that one of the programs that exploits spaced seeds to search for similarity regions, YH, allows to align divergent collinear genomes, for which MGA and Mauve failed to produce reliable alignments. Moreover, when comparing the proportion of orthologous genes included in the alignments, YH seems to overcome some reliability problems encountered by other methods, including on well-known cases like *E. coli*.

In the sequel, Section 2 presents our programs, the benchmark data, and establishes a protocol for the evaluation of global alignment. In Section 3, we evaluate the performance of those methods from both computational and biological view-points, while we discuss the results in Section 4.

2 Methods

Genome Alignment Programs MGA and Mauve are two archetypal anchor based alignment tools, are widely used, documented and proved to be more accurate than MUMer and SLagan [5,6]. They differ by two aspects: in phase 1, MGA searches for similarity regions that are *maximal exact matches* (MEMs) with the program Vmatch [3], while Mauve finds *approximate matches* using a special type of spaced seeds [4]. In phase 2, MGA executes Chainer [7], a program that selects the highest scoring non-overlapping set of collinear matches (*consistent chain* [5]); Mauve uses a greedy breakpoint elimination algorithm [8] that generates an approximate solution to the maximum-weighted non collinear anchoring problem. Hence, MGA treats collinear genome pairs, while Mauve handles rearrangements.

Our 6 programs combine one of 3 similarity detection methods (Vmatch, Blast v2, Yass) and one of 2 chaining algorithms (Chainer and Hierarchical chainer). Contrarily to Vmatch, Blast v2 and Yass find similarities that are local alignments with either contiguous or spaced seeds [9]. We named our 2×3 combinations by the initials of the methods they combine: VC, BC and YC with Chainer, VH, BH and YH with Hierarchical Chainer. By comparing those, we can measure the impact of each element on the final alignment.

The *Hierarchical chainer* implements a greedy chaining that allows for limited overlaps, in place inversions, and privileges region pairs with stronger similarities [10]. Similarities are ordered by decreasing *numbers of identities* (nid) and processed in groups according to several intervals of nid, starting with the largest ones. In each group, we consider first similarities located on the dotplot main diagonal (*i.e.*, shift of 0) and continue with increasing shifts. collinear similarities are being chained from the left end on the reference sequence, in a greedy manner.

Genome Sequences and Comparisons We considered all (236) pairs of bacterial strains of the same species whose complete genomes are available in GenomeReviews database as of mid-2008 [11]. The Mauve and MGA alignments, and the backbone segments positions for each pair were obtained from the Mosaic database [12], except for 37 pairs corresponding to five divergent species (*Buchnera aphidicola*, *Prochlorococcus marinus*, *Pseudomonas fluorescens*, *Rhodospseudomonas palustris* and *Synechococcus sp*) that were recomputed with the same protocol, since excluded from the database due to poor backbone coverage. For all pairs, we compute the alignments with our 6 programs. For 13 pairs, Vmatch yields erroneous results (detects non existing MEMs), which were excluded from further analysis. The backbone according to YH is the intersection of the set of anchors on each genome.

Criteria for Evaluation To compare alignments, all genome aligner publications use global quantitative criteria like the *percentage of identity* (%id) and the *coverage*, however not necessarily with the same definition. The usual definition of the %id, percentage of identical base pairs over the total alignment length (as in Mosaic), makes it incomparable between alignments. We define the *coverage* as the total length of aligned segments, and the %id as the ratio of identical bases in aligned segments (of the coverage) *over the genome length*.

To measure the reliability of the genome alignments, we compare their intersection with the sets of orthologous genes as defined in the OMA database [13]. We retrieved from OMA the list of orthologous genes and their positions for 12 pairs. For each, we compute the number and % of the genomic sequence of orthologous genes included in the backbone.

3 Results

The six programs were applied on every pair of intra-species bacterial genomes (see Section 2). The results were compared to those obtained using MGA, Mauve (collected from Mosaic) with respect to the criteria defined above. Result tables and additional information can be found at the following location: http://www.lirmm.fr/~uricar/Appendix_JOBIM09.html (Appendix).

Present Achievements

The first striking result lies in the difference of coverage between different species obtained by MGA and Mauve. For some species all pairwise alignments cover more than 90% of the genome (*e.g.*, on *Streptococcus thermophilus*), while in others the coverage is below 10% (*e.g.*, *Synechococcus sp*). One also observes, but more rarely, species for which the coverage of both methods varies greatly among pairs of strains (for *P. marinus*, the coverage of MGA varies in [0, 78]% and that of Mauve in [6, 96]%).

It is clear that these programs succeed in aligning some genome pairs and fail in others, which could be due either to a high level of divergence that makes the sequences unalignable (see Appendix) or to a methodological failure in detecting similarity regions or in chaining.

Local Similarities (LS) vs MEMs

The similarity detection phase is mainly responsible for the sensitivity of an anchor based method. Indeed, the chaining phase only discards potential anchors, however it may be unadapted to the type of similarities used. Here, to assess the impact on sensitivity, we compare the genome coverage obtained after the phases 1 and 2 of three different similarity detection methods combined with two chaining algorithms on a large panel of genome comparisons. Similarities are either short exact matches (MEMs), BLAST local alignments or local alignments based on spaced seeds (YASS). Figure 1 shows the difference of genome coverage between pairs of methods as box plots.

The left plot, which compares the effect of MEMs versus spaced seeds LS combined with Chainer, demonstrates that a classical chaining algorithm is unadapted to LS. This is due to overlaps between long local similarities, which are prohibited in the chain and cause Chainer to discard large alignment regions, especially for highly similar genomes. We thus designed a new chaining method allowing for overlaps, called Hierarchical chaining (see Section 2).

The central part of Figure 1 shows which chaining method suits a given type of similarities. Chainer does well with MEMs (VH-VC), Hierarchical performs in average better than Chainer with BLAST LS (BH-BC), and always surpasses it with spaced seeds LS (YH-YC). The right part compares the combination YH with the other best combinations. It clearly shows that YH surpasses all other methods in coverage and can even achieve important differences. Let us take the case of *P. marinus* strains *CP000111_GR* vs *CP000095_GR* as a running example of divergent strains: YH obtains 57% coverage, while BH covers 3% of the genome and VC not even 1%.

Our comparisons provide clear evidence that using spaced seeds in the similarity detection phase improves the coverage, and therefore the global sensitivity of the anchor strategy. Thus, in some cases, alignment failure was due, indeed, to a lack of sensitivity of the anchor detection phase.

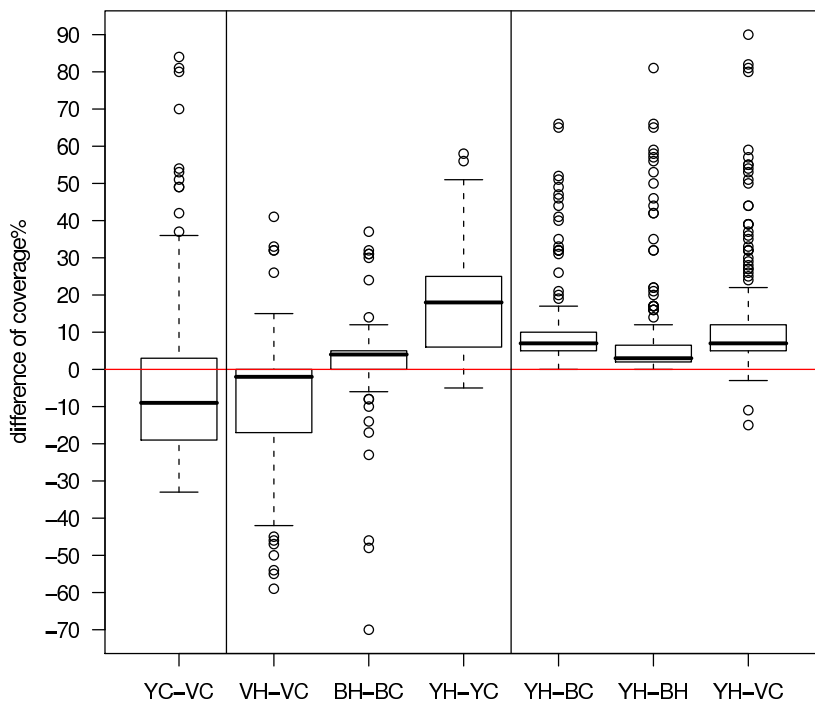


Figure 1: Boxplots of genome coverage differences between methods over 236 genome pairs. *E.g.*, YH-VC means, for each pair, the genome coverage of YH method minus that of VC in %. In a boxplot, circles are outliers. Left part: YC-VC coverage. Central part: comparison of each similarity detection method combined either with Chainer or our Hierarchical chaining. The 4th boxplot plus the right part: YH compared to other combinations; YH obtains larger coverage in the vast majority of cases over all four other combinations.

YH vs MGA/Mauve

Although MGA and Mauve execute two additional steps compared to our programs, the comparison of their results allows to see which proportion of the genome can be aligned solely with the chain of anchors and how it contributes to the percentage of identities.

On the 236 genome pairs, the coverage difference YH-MGA varies from -17% to 99% , with an average of 7.2% , and is zero or positive for 140 pairs and $< -2\%$ for only 10 pairs. The difference in %id varies from -16% to 99% with an average of 7% , and is zero or positive for 183 pairs and $< -2\%$ for only 2 pairs. Hence, the hierarchical program either achieves a result similar to MGA or improves on it in both aspects: coverage and %id.

In average, over the 236 genome pairs, the alignments of Mauve cover 13% more nucleotides than that of YH (variation within $[-14, 69]\%$) and have nearly 10% more identities. This is mostly due to its ability to handle rearrangements. However, as detected in Mosaic, a high coverage sometimes hides unreliable alignments. The segments that are aligned with ClustalW in the fourth phase (these are termed *aligned gaps* by Mauve) do not necessarily share sequence similarity and are often unreliably aligned. We investigated whether unreliable segments have a high impact on the coverage especially for highly divergent strains.

Our running example with a pair of *P. marinus* strains is in fact a typical situation for the divergent bacterial cases. In this case, Mauve covers in average with 27% more than YH (84% , corresponding to 1491kb vs 57% corresponding to 1012kb), while the difference in identity percentage is only of 8% in its favour. As both the coverage and the %id are ratios over the genome length, we can say that it covers 27 additional % (479kb) of the genome with only 8% more identities (142kb). It suggests that some pairs of aligned segments could well be false positives (*i.e.*, should not be part of the alignment). Indeed, by plotting the cumulative coverage with segments below a given threshold of %id, we found that Mauve covers 22 , resp. 30% , with segments whose %id lies ≤ 50 , resp. $\leq 55\%$.

Finally, we looked at the reliability and the accuracy of our backbones with a biological view-point. For this we compared the percentage of nucleotides from orthologous genes and the number of such genes included in YH, MGA, or Mauve backbones. In our *P. marinus* example, 60% of the nucleotides are part of YH backbone, compared to only 7% for Mauve and 3% for MGA. Even for the well studied *E. coli* comparison (K12 vs Sakai), where all three tools report a coverage 80% on Sakai genome, YH completely includes in its backbone 8% more orthologous genes than the other tools do. Even if it can be improved, this suggests that YH gives accurate and reliable backbones, with more precise segment bounds.

4 Discussion

In this work, we conducted one of the first evaluations of anchor-based genome alignment methods on a large set of intra-species bacterial genome alignments. For this, we propose a protocol and implement several programs performing only the first two steps of the anchor based strategy.

First, it appears that even for short, closely related, and sometimes collinear genomes, pairwise alignment is incompletely solved by nowadays programs. Second, the anchor chain they compute can be improved by using local alignments instead of shorter exact or approximate matches as similarities, provided that the chaining algorithm authorises overlaps between adjacent anchors. This improvement measured in terms of genome coverage and of %id is more pronounced if local alignments are detected with highly sensitive spaced seeds

[9]. Third, even if Mauve often achieves higher coverage than our method YH, the reliability of some of its regions aligned in the fourth phase is questionable, and their %id argues in favour of discarding them from the output (see the *P.marinus* example).

With its publication, Mauve opened the way to a better handling of rearrangements; nonetheless our results suggest the similarity detection could be improved, and thereby the global reliability of the complete alignment. The comparison of the coverage of known orthologs between MGA, Mauve, and YH corroborates these findings. Interestingly, our program YH performs drastic improvements where both MGA and Mauve fail: on species with highly divergent strains like *B. aphidicola*, *P.marinus*.

Besides this gain in coverage and percent identities over MGA or sometimes Mauve, YH runs faster (a maximum running time of 102 s. and an average of 10 s.) and brings qualitative ameliorations. Its chain contains 150 anchors in average vs several thousands for MGA and Mauve, making it simpler to visualise and to grasp. Moreover, all local alignments it includes have an associated E-value that lies above a given threshold, ensuring they are statistically significant, which is not the case in MGA or Mauve alignments. Altogether, YH could be useful to automatically determine the backbone (a goal of Mosaic) without further post-processing based on an arbitrary threshold.

Acknowledgements: RU benefits from a PhD fellowship from the French Ministry of Research. This work is supported by the ANR project CoCoGen (BLAN07-1_185484).

References

- [1] Bigot, S., Saleh, O., Lesterlin, C., Pages, C., Karoui, M.E., Dennis, C., Grigoriev, M., Allemand, J.F., Barre, F.X., Cornet, F.: KOPS: DNA motifs that control *E. coli* chromosome segregation by orienting the FtsK translocase. *EMBO J.* **24** (2005) 3770–3780
- [2] Delsuc, F., Brinkmann, H., Philippe, H.: Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics* **6** (2005) 361–375
- [3] Hohl, M., Kurtz, S., Ohlebusch, E.: Efficient multiple genome alignment. *Bioinformatics* **18**(S1) (2002) S312–320
- [4] Darling, A.C., Mau, B., Blattner, F.R., Nicole T. Perna: Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. *Genome Res.* **14**(7) (2004) 1394–1403
- [5] Brudno, M., Malde, S., Poliakov, A., Do, C.B., Couronne, O., Dubchak, I., Batzoglou, S.: Glocal alignment: finding rearrangements during alignment. *Bioinformatics* **19**(S1) (2003) i54–62
- [6] Kurtz, S., Phillippy, A., Delcher, A., Smoot, M., Shumway, M., Antonescu, C., Salzberg, S.: Versatile and open software for comparing large genomes. *Genome Biology* **5**(2) (2004) R12
- [7] Abouelhoda, M.I., Ohlebusch, E.: Chaining algorithms for multiple genome comparison. *J. of Discrete Algorithms* **3** (2005) 321–341
- [8] Blanchette, M., Bourque, G., Sankoff, D.: Breakpoint phylogenies. In Miyano, S., Takagi, T., eds.: *Genome Informatics*. (1997) 25–34
- [9] Noe, L., Kucherov, G.: YASS: enhancing the sensitivity of DNA similarity search. *Nucl. Acids Res.* **33**(S2) (2005) W540–543
- [10] Roytberg, M.A., Ogurtsov, A.Y., Shabalina, S.A., Kondrashov, A.S.: A hierarchical approach to aligning collinear regions of genomes. *Bioinformatics* **18**(12) (2002) 1673–1680
- [11] Sterk, P., Kersey, P.J., Apweiler, R.: Genome Reviews: Standardizing Content and Representation of Information about Complete Genomes. *OMICS: A Journal of Integrative Biology* **10**(2) (2006) 114–118
- [12] Chiapello, H., Bourgait, I., Sourivong, F., Heuclin, G., Gendrait-Jacquemard, A., Petit, M.A., El Karoui, M.: Systematic determination of the mosaic structure of bacterial genomes: species backbone versus strain-specific loops. *BMC Bioinformatics* **6**(1) (2005) 171
- [13] Schneider, A., Dessimoz, C., Gonnet, G.H.: OMA Browser Exploring orthologous relations across 352 complete genomes. *Bioinformatics* **23**(16) (2007) 2180–2182