BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 16 2009, pages 2071–2073
doi:10.1093/bioinformatics/btp356

Genome analysis

Reordering contigs of draft genomes using the Mauve Aligner

Anna I. Rissman[1,*], Bob Mau[1], Bryan S. Biehl[1], Aaron E. Darling[2], Jeremy D. Glasner[1] and Nicole T. Perna[1]

[1]Genome Evolution Laboratory, University of Wisconsin-Madison, 425G Henry Mall Suite 4400V, Madison, WI 53706 and [2]Genome Center, University of California-Davis, 451 Health Sciences Dr, Davis, CA 95616

Received on January 7, 2009; revised on May 12, 2009; accepted on June 3, 2009

Advance Access publication June 10, 2009

Associate Editor: Dmitrij Frishman

ABSTRACT

Summary: Mauve Contig Mover provides a new method for proposing the relative order of contigs that make up a draft genome based on comparison to a complete or draft reference genome. A novel application of the Mauve aligner and viewer provides an automated reordering algorithm coupled with a powerful drill-down display allowing detailed exploration of results.

Availability: The software is available for download at http://gel.ahabs.wisc.edu /mauve.

Contact: rissman@wisc.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online and http://gel.ahabs.wisc.edu

1 INTRODUCTION

New high-throughput technologies have greatly reduced the cost of genome sequencing, leading to an abundance of draft-quality genome sequences that may be composed of hundreds or thousands of contigs. Ordering and orienting these contigs into larger units (scaffolds or supercontigs) facilitates genome closure and comparative analyses. Contigs can be ordered based on additional data, such as the presence of discontinuous portions of the same sequencing template (clone or fragment) in two contigs, but this type of information is not available for all projects. However, even without additional data, contig order can be predicted by comparison with a reference genome that is expected to have conserved genome organization.

We present a new method for comparative contig ordering based on iterative genome alignment using Mauve. The reference used may be draft quality itself, or may have divergent genetic content. The Mauve aligner has been used extensively for microbial genome comparisons because it effectively identifies and aligns homologous regions even if genomes have undergone rearrangements, large insertions or deletions, and substantial sequence divergence. Mauve Contig Mover (MCM) provides advantages over methods that rely on matches in limited regions near the ends of contigs, require anchors at both ends of contigs, force users to exclude lineage-specific sequences at contig boundaries, or are unable to resolve which, if any, copies of repeated sequences are consistent with more extensive collinearity (Darling *et al.*, 2004; Richter *et al.*, 2007; van Hijum *et al.*, 2005). An interactive full-genome alignment display shows the relative order of the contigs as well as potential gaps in sequence coverage and regions of possible rearrangement or misassembly. After reordering, Mauve is a useful platform for further detailed comparative sequence analysis that is often the motivation for the sequencing effort itself.

*To whom correspondence should be addressed.

2 METHODS

The Mauve aligner filters and sorts internally identified matches into locally collinear blocks (LCBs). Each LCB represents a region of homologous sequence without rearrangement among the input genomes. Each LCB must be separated from the next by rearrangement in at least one genome (Darling *et al.,* 2004). Contig boundaries (edges) represent potentially artificial LCB edges. Therefore, finding the contig order that minimizes the number of LCBs caused by contig edges is equivalent to finding a likely contig order.

Using the Mauve alignment LCBs, the reordering process occurs in three steps: placing contigs with no apparent conflict in ordering information, placing contigs with conflicting information into intermediary anchor positions, and finally matching LCB ends that extend to contig boundaries. Each step occurs in at most $O(n^2)$ time, where n is the number of LCBs, plus the time required for alignment (Darling *et al.*, 2004). Mauve assumes contigs are in the correct order when filtering matches, so as the order is optimized, alignment results change. Therefore, results are refined through iterative alignment until no further ordering is possible.

MCM outputs a series of Mauve alignments, each representing an iteration of the reordering. In addition to the standard Mauve output, the reorder process produces a FastA file containing the new order and orientation, as well as a list of ordered contigs including name and coordinate location. The standard Mauve visualization can be applied in novel ways to analyze contig order. For example, we have used it to identify potential misassemblies in contigs, and to evaluate the presence or absence of genes split by contig boundaries or by rearrangements. If FastAs representing the order produced by other programs are created, Mauve can also be used to compare results, as in Supplementary Figure 1. Furthermore, annotations from GenBank format input can be viewed, even once reordered.

3 RESULTS

We have used MCM to order contigs for a variety of different bacterial genome projects based on comparison to the single best reference sequence available, and show some of our results in Table 1. These projects include draft genomes assembled from Sanger sequencing as well as short read generating technologies developed by 454 and Illumina, with the most fragmented example involving a 5 Mb genome with more than 1000 contigs. The draft and reference genome combinations selected include comparisons of genomes from different species, the same species, different strains

Table 1. Summary of results of Mauve Contig Mover reorders

Draft genome	Reference genome	Number in draft		Number of contigs/% bp ordered					
		Contigs	bp	Mauve		Projector		OSLay	
P. brasiliensis Pbr1692 (Glasner *et al.,* 2008b)	*P. atroseptica* SCRI1043 (Toth *et al.,* 2004)	1370	4918574	121	95.9	89	90.7	112	93.5
P. caratovorum WPP14 (Glasner *et al.,* 2008b)	*P. atroseptica* SCRI1043 (Toth *et al.,* 2004)	741	4823187	222	96.4	176	90.7	198	93.8
Yersinia pestis FV-1 (Touchman *et al.,* 2007)	*Yersinia pestis* CO92 (Parkhill *et al*., 2001)	400	4472646	355	94.1	353	93.4	Did not finish	
Escherichia coli EC4501 (Glasner *et al.*, 2008a)	*Escherichia coli* Sakai (Hayashi *et al.,* 2001)	250	5677181	140	93.4	107	93.3	177	90.9
E. coli MG1655 mutant* (Glasner *et al.*, 2008a)	*Escherichia coli* MG1655 (Blattner *et al.,* 1997)	1663	4554569	1068	98.8	725	96.4	784	94.6
Erwinia chrysanthemi 3937 v3	*E. chrysanthemi* 3937 v6b (Glasner *et al.*, 2008a)	767	5119283	228	95.4	212	95.4	267	90.7

Data includes the draft sequence and reference sequence used to perform the order, the number of contigs and base pairs contained in the draft, and the number of contigs and percentage of base pairs ordered by Mauve, Projector (van Hijum *et al.,* 2005), and OSLay (Richter *et al.*, 2007). *Pectobacterium* is abbreviated *P*. in table. All drafts were sequenced using 454 technology, except (*), which used Solexa. While we included numbers from OSLay reorders, the structure suggested by the OSLay reorder differs significantly from that of Mauve and Projector, as can be seen in Supplementary Figure 1. Table 2 and Supplementary Table 1 also summarize correctly ordered quantities based on artificially cut genomes.

Table 2. Overview of percents ordered and correctly ordered

Artificial draft	Reference	bp	Percentage ordered of total bp			Percentage correct of total bp		
			Mauve/Projector/OSLay					
P. atroseptica SCRI1043 (Toth *et al.,* 2004)	*P. atroseptica* SCRI1043 (Toth *et al.,* 2004)	5064019	99.4	98.8	97.7	99.4	98.7	95.1
Escherichia coli EDL933 (Perna *et al.,* 2001)	*Escherichia coli* MG1655 (Blattner *et al.,* 1997)	5528133	95.0	91.7	85.9	94.1	82.6	78.4
Yersinia pestis KIM (Deng *et al.,* 2002)	*Yersinia pestis* CO92 (Parkhill *et al*., 2001)	4781603	96.5	96.8	92.7	90.4	61.7	66.8
Overall average			96.8	95.1	91.9	94.2	80.7	78.9

Each row is an average of the orders of the same sequences listed below in Supplementary Table 2. The draft, in each case, was artificially cut into pieces using in-house software. The pieces were ordered using Mauve, Projector and OSLay, and the results compared to the correct order. A piece (contig) was considered out of order if it was out of position relative to the closest correctly ordered contig on either side. The table shows the total number of base pairs, the percentage ordered using each algorithm, and the percent of the total base pairs that were correctly ordered. Draft sequences are prone to errors and omissions that have not been modeled in the artificially partitioned 'drafts' used. Therefore, these figures are meant to bound the number of ordered base pairs. Each row represents different genomes with different divergence, giving an idea of these percentages over a range of data.

and different assemblies of the same genome. Many of the draft genomes available through the Enteropathogen Resource Integration Center (Glasner *et al.*, 2008a) and the ASAP database (Glasner *et al.*, 2006) have been ordered using MCM. Examples are available as Supplementary Material on our web site. Supplementary Figure 1 shows a *Yersinia pestis* strain FV1 draft genome (Touchman *et al.*, 2007) reordered and aligned to the complete *Y. pestis* CO92 reference genome. MCM was able to order 356 out of 400 contigs (4211103 out of 4472646 bp) reducing the alignment from 359 to 11 LCBs. Supplementary Figure 1 also shows the utility of the Mauve Viewer for comparing different suggested contig orders.

We urge caution in interpretation of contig order predicted using MCM or any other algorithm. Many true bacterial genome rearrangements occur at repetitive sequences, which pose challenges for both genome assembly and alignment. Ordering contigs based on comparative analyses can mask true rearrangements anchored in repeats at these contig breaks. Annotations of complete and partial repeats can be viewed in the Mauve alignment display, providing a means of identifying such regions. Conversely, misassemblies can appear as false rearrangements. Mauve clearly displays these regions, and PCR primers may be designed that allows verification of the rearrangement or proof of the misassembly. The Mauve Viewer also allows exploration of alternate positions for contigs with multiple LCBs. A comparison of the number of LCBs between reference and draft to the number between other genomes expected to show similar levels of rearrangement can provide an estimate of this effect. Table 2 summarizes reorders without modeling these effects, showing accuracy between 90.4% and 99.4%. Generally, closer alignments will provide more accurate reorders covering more of the draft. Because MCM maximizes collinearity among genome sequences under comparison it produces alignments that are easily visualized and provides an excellent platform for analysis and finishing of draft genomes.

ACKNOWLEDGEMENTS

The authors thank Eric Cabot and Michael Cox (UW-Madison), John Battista (LSU), and the UW Biotechnology Center for sequence data and analyses of radiation resistant *E.coli* strains (NIH grant #GM067085).

Funding: NIH-NIGMS Award #GM62994 and NSF Award #0412599 (to N.T.P.); Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN266200400040C for the Enteropathogen Resource Integration Center.

Conflict of Interest: none declared.

REFERENCES

Blattner,F.R. *et al.* (1997) The complete genome sequence of Escherichia coli K-12. *Science*, **277**, 1453–1474.

Darling,A.C. *et al.* (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.*, **14**, 1394–1403.

Deng,W. *et al*. (2002) Genome sequence of Yersinia pestis KIM. *J. Bacteriol.*, **184**, 4601–4611.

Glasner,J.D. *et al*. (2006) ASAP: a resource for annotating, curating, comparing, and disseminating genomic data. *Nucleic Acids Res.*, **34**, D41–D45.

Glasner,J.D. *et al.* (2008a) Enteropathogen Resource Integration Center (ERIC): bioinformatics support for research on biodefense-relevant enterobacteria. *Nucleic Acids Res.*, **36**, D519–D523.

Glasner,J.D. *et al.* (2008b) Niche-specificity and the variable fraction of the Pectobacterium Pan-Genome. *Mol. Plant-Microbe Interact.*, **21**, 1549–1560.

Hayashi,T. *et al.* (2001) Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res.*, **8**, 11–22.

Parkhill,J. *et al*. (2001) Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature*, **413**, 523–527.

Perna,N.T. *et al*. (2001) Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. *Nature*, **409**, 529–533.

Richter,D.C. *et al*. (2007) OSLay: optimal syntenic layout of unfinished assemblies. *Bioinformatics*, **23**, 1573–1579.

Toth,I.K. *et al.* (2004) Genome sequence of the enterobacterial phytopathogen *Erwinia carotovora* subsp. atroseptica and characterization of virulence factors. *Proc. Natl Acad. Sci. USA*, **101**, 11105–11110.

Touchman,J.W. *et al*. (2007) A North American *Yersinia pestis* draft genome sequence: SNPs and phylogenetic analysis. *PLoS ONE*, **2**, e220.

van Hijum,S.A. *et al*. (2005) Projector 2: contig mapping for efficient gap-closure of prokaryotic genome sequence assemblies. *Nucleic Acids Res.*, **33**, W560–W566.