

Multiple Alignment of Rearranged Genomes

Aaron E. Darling¹ * Bob Mau² Mark Craven³ Nicole T. Perna²

1. Dept. of Computer Science, 2. Dept. of Animal Health and Biomedical Sciences, 3. Dept. of Biostatistics and Medical Informatics, University of Wisconsin–Madison, Madison, WI USA

Abstract

The nature of large-scale evolutionary processes that shape genomes over time fundamentally differs from the forces governing local evolution within individual genes. Large-scale events such as horizontal transfer, genome rearrangements, gene duplication, and gene loss obscure the notion of orthology and have demanded new models of evolution. Whether or not explicitly designed to do so, multiple genome alignment tools must cope with such large-scale changes in addition to local changes such as nucleotide substitution and indels. Using simulated genomes containing both large and small-scale evolutionary changes, we present an alignment quality comparison of Mauve, a multiple genome aligner that considers large-scale evolutionary events, to alignments generated by other state-of-the-art genome alignment systems. Our results indicate that in the presence of large-scale rearrangement events, Mauve has superior accuracy.

Availability

Mauve and the simple genome evolver (sgEvolver) are available from <http://gel.ahabs.wisc.edu/mauve>

1 Introduction

High-throughput DNA sequencing technology has enabled researchers to rapidly determine the genome sequences of a wide variety of organisms, laying the foundation for comparative genomics. The forces governing genome evolution fundamentally differ from those governing gene evolution. Recombination events such as inversion and rearrangement, horizontal transfer, gene duplication, and gene loss occur frequently on the genome-scale but seldom within genes. Furthermore, the rates and patterns of each event depend on the particular set of genomes being compared. For example, observations of gene duplication and repetitive sequences are much more common among higher eukaryotes than bacteria, while genome rearrangements can be readily observed between both closely-related and divergent organisms of all

types (Tillier and Collins, 2000). These additional evolutionary mechanisms distinguish the genome comparison and alignment task from traditional sequence alignment. Accordingly, several new methods and tools for genome comparison have been developed.

Genome comparison tools typically make a set of assumptions about the nature of evolutionary events that genomes have undergone. When selecting a genome comparison tool, knowledge of the assumptions and performance of each tool is essential to make an informed decision. We have evaluated available multiple genome alignment systems to gauge their robustness to various types of evolutionary events. Because manual curation of a multiple genome alignment on actual genome sequence is too costly a process, there is no “gold standard” alignment to use when assessing the quality of calculated alignments. Instead we designed a model of genome evolution that allows software to simulate evolution, producing a set of evolved sequences and an alignment of nucleotides conserved through the process of genome evolution. Genome alignment tools can thus be evaluated on their ability to reproduce the correct alignment of the simulated genomes.

2 Genome Alignment Methods

Early research into genome alignment focused on scaling traditional $O(n^2)$ pairwise alignment methods to handle much longer genome sequences. Pairwise genome alignment tools such as MUMmer (Delcher et al., 1999), GLASS (Batzoglou et al., 2000), and WABA (Kent and Zahler, 2000) pioneered the use of anchoring to accelerate the alignment process. Anchored alignment typically proceeds in three steps. First, the aligner identifies a set of local alignments in regions of high similarity among the genomes. Next, a subset of the regions identified in the first step are selected as alignment anchors, based on whether the tool believes they are part of the correct alignment. Finally, the alignment anchors are used to restrict the number of possible alignments considered when performing an $O(n^2)$ gapped alignment using dynamic programming. In order to complete a gapped alignment, many tools assume that the genomes are collinear – that no significant inversion or rearrangement events took place since their divergence.

*1656 Linden Dr., Madison, WI 53706 USA, darling@cs.wisc.edu

3 Mauve: Multiple genome alignment with rearrangements

Mauve implements an anchored alignment algorithm designed to address the presence of significant inversions and rearrangements in a set of genomes to be aligned (Darling et al., 2004a). Mauve's alignment algorithm first identifies all unique subsequences that match exactly in two or more of the genomes under study (multi-MUMs) and that are longer than some minimum length. These multi-MUMs serve as potential alignment anchors and are found using the algorithm implemented in GRIL (Darling et al., 2004b). Mauve then calculates a distance matrix using the fraction of nucleotides shared in the multi-MUMs among each pair of genomes as a distance metric. A phylogenetic guide tree is calculated using the distance matrix and Neighbor-Joining. Using only the multi-MUMs existing in all genomes, Mauve performs greedy breakpoint elimination to identify significant regions of collinearity called locally collinear blocks. The identified collinear regions are then the subject of anchored global alignment using Clustal-W (Thompson et al., 1994) and the previously calculated guide tree. To limit running time, Mauve restricts the size of regions aligned by Clustal-W to 10Kbp.

4 Simulating Evolution

We have designed a simple model of genome evolution that attempts to capture the processes governing evolution of bacterial genomes. Given a rooted phylogenetic tree, an ancestral sequence, a "donor" sequence for insertions, and rates for each type of genome evolution, our model specifies the process of evolving the ancestor genome into the leaf genomes. In these experiments, nucleotide substitution sites are unevenly distributed according to the gamma distribution. Indel sites are uniformly distributed and indel sizes are sampled from a Poisson distribution with mean 3bp. Horizontal transfer events are modelled to occur with two distributions, small events with average size 200bp and large events with size uniformly distributed between 10Kbp and 60Kbp. Inversions are modelled with a mean size 50Kbp. Translocations are not explicitly modelled but occur due to overlapping inversion events. Locations for indel, inversion, and horizontal transfer events are sampled uniformly throughout the genome, and all events are simulated to have taken place at a point in time given by a marked Poisson process over the phylogenetic tree. Finally, our model assumes constant genome size so deletion events are sampled with equal frequency to events that introduce new sequence. Calculated alignments are scored against the correct alignments produced by the evolver using the sum-of-pairs scoring scheme also used in BaliBASE (Thompson et al., 1999). This scoring scheme yields an accuracy measure that quantifies the percentage of pairs of nucleotides aligned identically in both the correct and calculated alignments.

5 Results and Conclusions

We performed several experiments using our simulated evolution environment to compare the accuracy of Mauve to Multi-LAGAN (Brudno et al., 2003a), Shuffle-LAGAN (Brudno et al., 2003b), and MAVID (Bray and Pachter, 2004). Our first experiment compares the ability of each multiple genome aligner to align genomes in the presence of increasing nucleotide substitution and indel rates. This experiment demonstrates the accuracy of Multi-LAGAN, a sensitive cross-species comparison tool, when no rearrangements have taken place. The second experiment, evaluates alignment quality in the presence of increasing rates of inversion and horizontal transfer, with low substitution and indel rates. This experiment demonstrates that Mauve clearly excels at aligning genomes with rearrangements.

Acknowledgements

Funding for this research was provided by NIH Grant GM62994-02. In addition, A. Darling was supported in part by NLM Training Grant 1T15LM007359-01.

References

- Batzoglou, S., Pachter, L., Mesirov, J. P., Berger, B., and Lander, E. S. (2000). Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res*, 10(7):950-8.
- Bray, N. and Pachter, L. (2004). MAVID: constrained ancestral alignment of multiple sequences. *Genome Res*, 14(4):693-9.
- Brudno, M., Do, Chuong, B., Cooper, Gregory, M., Kim, Michael, F., Davydov, E., Green, Eric, D., Sidow, A., and Batzoglou, S. (2003a). LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res*, 13(4):721-31.
- Brudno, M., Malde, S., Poliakov, A., Do, Chuong, B., Couronne, O., Dubchak, I., and Batzoglou, S. (2003b). Global alignment: finding rearrangements during alignment. *Bioinformatics*, 19 Suppl 1:154-162.
- Darling, A. C. E., Mau, B., Blattner, F. R., and Perna, N. T. (2004a). Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Research*, 14(7):In press.
- Darling, A. E., Mau, B., Blattner, F. R., and Perna, N. T. (2004b). GRIL: Genome rearrangement and inversion locator. *Bioinformatics*, 20(1):122-124.
- Delcher, A. L., Kasif, S., Fleischmann, R. D., Peterson, J., White, O., and Salzberg, S. L. (1999). Alignment of whole genomes. *Nucleic Acids Res*, 27(11):2369-76.
- Kent, W. J. and Zahler, A. M. (2000). Conservation, regulation, synteny, and introns in a large-scale *C. briggsae*-*C. elegans* genomic alignment. *Genome Res*, 10(8):1115-25.
- Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673-80.
- Thompson, J. D., Plewniak, F., and Poch, O. (1999). A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res*, 27(13):2682-90.
- Tillier, E. R. and Collins, R. A. (2000). Genome rearrangement by replication-directed translocation. *Nat Genet*, 26(2):195-7.