Poster C-37

Genome alignment with lineage-specific rearrangement and gene flux rates

Authors:
Aaron E. Darling (*Dept. of Computer Science, Univ. of Wisconsin-Madison*)
Bob Mau (*Dept. of Animal Health and Biomedical Sciences, Univ. of Wisconsin-Madison*)
Nicole T. Perna (*Dept. of Animal Health and Biomedical Sciences and Genome Center, Univ. of Wisconsin-Madison*)

Short Abstract: We assess the relative rates of genome rearrangement, gene acquisition, gene loss, and nucleotide substitution among finished genome sequences from the family Enterobacteriacae. We find significant evidence for lineage specific mutation rates, and develop a new multiple-genome alignment method to account for the observed heterotachy.

Long Abstract:

Genomes evolve via large-scale and local mutational processes that include nucleotide substitution and indels in addition to rearrangement and gene flux: lateral acquisition, gene duplication, and loss. The relative contribution made by each mutation type to adaptive evolution of microbial genomes has yet to be described in a unified framework. We apply existing genome comparison methods [1] to elucidate the relative rates of various mutational processes in a group of sequenced Enterobacteriacae. Pairwise comparisons indicate the Yersinia lineage appears to have a high relative rate of genomic rearrangement [2], while E. coli and Salmonella are recalcitrant to rearrangement, but undergo substantial gene flux (lateral transfer and gene loss). The Yersinia and Shigella are known to be colonized with mobile genetic elements, which induce a high rate of pseudogenization and appear to mediate genomic rearrangement as a byproduct.

Given strong evidence for lineage-specific mutation rates, we develop a new method for multiple genome alignment that accounts for heterotachy when scoring candidate alignments. At a high level, the algorithm consists of four steps: (1) generation of high-scoring local-multiple-alignments, (2) estimation of pairwise breakpoint and genome content distances, (3) progressive multiple genome alignment, and (4) iterative refinement. We now discuss each step in turn.

High-scoring local alignments are identified using an extension of a previously published seed-and-extend hashing method [1] that now uses approximate seed matches [2,3]. The identified local alignments serve as candidate anchors for subsequent pairwise and progressive multiple alignment steps.

Pairwise breakpoint distances are estimated by generating Locally Collinear Blocks (LCBs) from pairwise local alignments. We apply the previously described greedy breakpoint elimination algorithm [1] to identify significant LCBs among each pair of genomes. The weight criterion for LCBs has been modified to down-weight regions containing repetitive sequence elements. In order to generate only high-confidence LCBs, the stopping criteria for breakpoint elimination was modified such that at least 15% of regions covered by local-alignments are discarded. For any pairwise comparison, the resulting number of LCBs is used as an estimate of the pairwise breakpoint distance. We do not compute a full multiple alignment between genome pairs.

Progressive multiple alignment proceeds by aligning progressively distant taxa according to a phylogenetic guide tree. We compute a guide tree using Neighbor-Joining on the pairwise genome content distance matrix. Pairwise sequence alignment is done similarly to above, while profile-to-sequence and profile-to-profile alignment uses novel anchoring and scoring methods. In computing profile alignments, local alignments among extant sequences are projected into alignment column coordinates and any resulting inconsistent local-multiple-alignments are resolved to be consistent—a given nucleotide may be part of at most one local-multiple alignment and may not be aligned to another nucleotide in the same genome. We then use sum-of-pairs breakpoint elimination to filter spurious local alignments. Each pair of genomes has a breakpoint penalty weighted by its previously calculated breakpoint distance. Breakpoints are calculated at each internal node of the guide tree using average breakpoint distance of descendant nodes as a weight. The sum-of-pairs breakpoint elimination results in an candidate set of alignment anchors grouped into Locally Collinear Blocks among two or more sequences. The initial anchor set may sparsely cover the genomes, so a recursive search for additional anchors between existing anchors is repeatedly performed until no additional anchors are found. Once a final set of alignment anchors has been selected, we apply the MUSCLE [5] alignment algorithm in profile-to-profile mode to align the remaining regions between anchors.

The progressive alignment method aligns sequences below each node of the guide tree until it reaches the tree root at which point all sequences have been aligned. The final step of our new algorithm applies MUSCLE to iteratively refine the alignment in 10Kbp windows. Iterative refinement corrects alignment errors made early during the alignment process such as misplacement of gaps.

We evaluate the quality of genome alignments computed using the new method on simulated data sets and draw comparison to previous methods for genome comparison [6,7]. In general, the new method scales more gracefully than the original Mauve algorithm and accommodates large data sets with many more taxa. In particular, the new method identifies and aligns regions conserved among subsets of the genomes under study—an important improvement over the original Mauve alignment algorithm.

[1] Darling ACE, Mau B, Blattner FR, Perna NT.(2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Research 14(7):1394-1403.

[2] Deng W, Burland V, Plunkett G 3rd, Boutin A, Mayhew GF, Liss P, Perna NT, Rose DJ, Mau B, Zhou S, Schwartz DC, Fetherston JD, Lindler LE, Brubaker RR, Plano GV, Straley SC, McDonough KA, Nilles ML, Matson JS, Blattner FR, Perry RD. (2002) Genome sequence of Yersinia pestis KIM. J Bacteriol. 184(16):4601-11.

[3] Choi KP, Zeng FF, Zhang LX. (2004) Good Spaced Seeds for Homology Search, Bioinformatics 20:1053-59

[4] Zhang LX (2004) Personal communication

[5] Edgar RC. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 5:113.

[6] Brudno M, Malde S, Poliakov A, Do CB, Couronne O, Dubchak I, Batzoglou S. (2003) Glocal alignment: finding rearrangements during alignment. Bioinformatics. 19 Suppl 1:i54-62.

[7] Bray N, Pachter L. (2004) MAVID: constrained ancestral alignment of multiple sequences. Genome Research 14(4):693-9.