Reconstruction of Ancestral Gene Order after Segmental Duplication and Gene Loss

Jun Huan[1], Jan Prins[1], Wei Wang[1], Todd Vision[2]
Departments of [1] Computer Science and [2]Biology
University of North Carolina at Chapel Hill
{huan, prins, weiwang}@cs.unc.edu, tjv@bio.unc.edu

Abstract

As gene order evolves through a variety of chromosomal rearrangements, conserved segments provide important insight into evolutionary relationships and functional roles of genes. However, gene loss within otherwise conserved segments, as typically occurs following large-scale genome duplication, has received limited algorithmic study. This has been a major impediment to comparative genomics in certain taxa, such as plants and fish.

We propose a heuristic algorithm for the inference of ancestral gene order in a set of related genomes that have undergone large-scale duplication and gene loss. First, approximately conserved (i.e. homologous) segments are identified using pairwise local genome alignment. Second, homologous segments are iteratively clustered under the control of two parameters, (1) the minimal required number of shared genes between two clusters and (2) the maximal allowed number of rearrangement breakpoints along the lineage leading to each descendant segment. Finally, we compute an estimated ancestral gene order for each cluster that is optimal in some sense.

We evaluate the performance of this algorithm on simulated data that models a genome evolving by large-scale duplication, duplicate gene loss, transposition, translocation, and inversion. The results suggest that long segments of ancestral gene order may be reconstructed following moderate levels of rearrangement with only minor loss of accuracy.

1. Introduction

In previous work, we provided a mathematical framework and algorithm for identifying homologous segments in a pairwise genome alignment [3]. However, application of this approach is difficult when genomes are distantly related, in large part due to the occurrence of lineage-specific genomic duplication (*i.e. polyploidy*) and gene loss (*i.e. diploidization*) events. Such events are now known to have occurred regularly in eukaryotic genome evolution [8].

The loss of one or the other copy of a large fraction of duplicated gene pairs following segmental or global duplication serves to obscure the presence of many segmental homologs [4]. Such highly diverged homologs, which lack a sufficient number or density of shared genes and therefore are hard to identify by local genome alignment, have been referred to as *ghosts* [6]. A number of strategies for the identification of ghosts have been proposed. One is to identify homologs from multiple, rather than pairwise, genome alignment [6, 7]. Another is to incorporate reconstructed ancestral genomes directly into genome alignment algorithms [1, 2]. However, methods for inferring ancestral gene order have not been well studied and little is known about the accuracy of the reconstructions that can be obtained.

In the present work, we study the problem of reconstructing ancestral marker (*i.e.* gene) order in the presence of global duplication, marker loss and other rearrangement events that permute marker order. We have developed an algorithm for solving this problem called eAssembler (for *e*volutionary *Assembler*). Our approach differs from previous work in this area [1, 2] by taking advantage of the overlap among different pairs of segmental homologs, thus reconstructing ancestral segments that contain more distinct markers than any single pair. We assess the performance of the algorithm with simulation experiments.

2. eAssembler Overview

The eAssembler algorithm is designed to infer the ancestral order of markers in a set of descendant genome blocks. In the first step, a clustering algorithm is used to determine which segments are to be assembled and in what order. Initially, the algorithm places each pair of homologous segments in a separate cluster. The algorithm iteratively joins two existing clusters P and Q that satisfy two conditions, governed by parameters t and k. First, P and Q must share at least k markers. Second, there must exist a median m (a permutation of all markers in either P or Q) such that the distance between every segment in the two clusters and m is

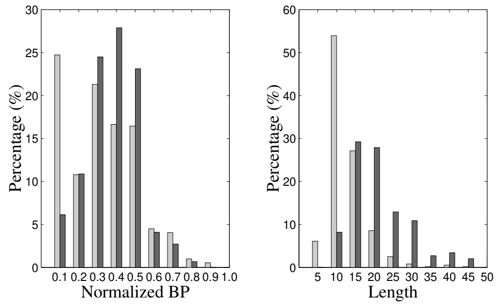

Figure 1. Comparison of output from pairwise genome alignment using FISH (light gray) and after further analysis by eAssembler (dark gray). Left: distribution of normalized BP distances, in unit of BP/marker. Right: length distribution of segments/contigs, in number of markers. Mean normalized BP distance: 0.260 (FISH) and 0.316 (eAssembler) Mean length of segment: 10.6 and contig: 19.9.

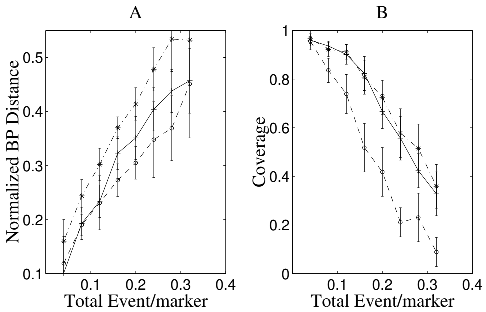

Figure 2. Normalized BP distance and coverage under varying numbers of rearrangement events, using representative parameters. The three lines show results for different fixed proportions of deletion : inversion : transposition : translocation. 12:6:1:1(solid), 11:3:3:3 (dashed), 9:9:1:1 (dot-dashed) Ten genomes were simulated and assembled for each point. Vertical bars show one standard deviation.

no greater than t. We use the *induced breakpoint distance* to measure the dissimilarity between two marker strings with possibly unequal contents [5]. If there are multiple such joins available among the current clusters, a join with the maximal number of shared markers is chosen. The clustering algorithm stops when no further joins are possible. An important consequence of this clustering procedure is that a join may occur even in cases where pairs of individual segments lack sufficient overlap to be joined by themselves. This contrasts with previously published approaches in which clusters can consist of only two segments [1, 6]. In the second step, the algorithm computes the *optimal median* for each cluster. This is defined as one of the (potentially large number of) medians that satisfies the distance constraint and also has minimal sum of distances. The optimal median is taken to be the estimate of the ancestral marker string. By analogy with sequence assembly, we refer to the optimal median of each cluster as a *contig*.

3. Experimental Study

We simulated genomes evolving over time and analyzed the final products using eAssembler in order to assess the quality of the reconstructions obtained. Our model of evolution in a multiple-chromosome genome combines global duplication and single-marker deletion with three types of operations acting on gene order alone (inversion, transposition, and reciprocal translocation).

We used two assessments, coverage and normalized breakpoint distance, to measure the quality of the reconstructions. We calculate coverage as the ratio of the number of distinct markers in all contigs to the number of markers in the original genome. For a contig, the normalized breakpoint distance (hereafter referred to simply as the *BP distance*) is defined as the ratio of its induced breakpoint distance to its length in markers.

We compared the quality of eAssembler contigs to those produced by pairwise genome alignment (using FISH [3]). The result is presented in Figure 1. The sensitivity of eAssembler to the total number of rearrangements is presented in Figure 2.

4 Discussion

The simulation results show that long contigs with only minor rearrangements from the ancestral order can be obtained using the eAssmbler algorithm. Such contigs can be used to substantially improve the detection sensitivity of local genome alignment algorithms [2].

References

[1] G. Blanc, K. Hokamp, and K. H. Wolfe. *Genome Res*, 13:137–44, 2003.
[2] J. E. Bowers, B. A. Chapman, J. Rong, and A. H. Paterson. *Nature*, 422:433–438, 2003.
[3] P. P. Calabrese, S. Chakravarty, and T. J. Vision. *Bioinformatics*, 19:i74–i80, 2003.
[4] H.-M. Ku, T. Vision, J. Liu, and S. D. Tanksley. *Proc Natl Acad Sci USA*, 97:9121–9126, 2000.
[5] D. Sankoff, D. Bryant, M. Deneault, B. F. Lang, and G. Burger. *J Comput Biol*, 7:521–536, 2000.
[6] C. Simillion, K. Vandepoele, M. C. V. Montagu, M. Zabeau, and Y. V. de Peer. *Proc Natl Acad Sci USA*, 99:13627–32, 2002.
[7] K. Vandepoele, C. Simillion, and Y. V. de Peer. *Trends in Genetics*, 18:606–608, 2002.
[8] K. H. Wolfe. *Nat Rev Genet*, 2(5):333–41, 2001.