SyntenyFinder: A Synteny Blocks Generation and Genome Comparison Tool

Intern: Ilya Minkin
Advisor: Son Pham

Algorithm for finding synteny blocks from nucleotide sequences is presented. The algorithm is able to reconstruct synteny blocks previously found by ad hoc methods.

1. Introduction

Recent growth of number of sequenced genomes arises question about their evolution relationship. In order to perform rearrangement analysis, the genomes must be decomposed into conservative segments, called synteny blocks. Currently existing tools for solving this problem, like DRIMM-Synteny [1], require the genomes to be presented as sequences of enumerated local alignments, or *anchors*. Usually, anchors represent homologous genes. At this moment, there are no general purpose tools that can find synteny blocks from the genomes represented as unannotated nucleotide sequences.

De Bruijn graphs are extensively used in bioinformatics for genome assembly [2, 3]. In this work we address problem of finding synteny blocks from nucleotide sequences. We propose new algorithm for this task based on colored De Bruijn graph.

2. Problem description

Suppose that we are given a set $S = \{s_1, s_2, \ldots, s_n\}$ of chromosomes, and each chromosome is represented as a string over alphabet $\{A, C, G, T\}$. The task of finding synteny blocks is to find a set of so called conserved regions $C = \{C_1, C_2, \ldots, C_n\}$, where each conserved region C_i is a set of substrings of chromosomes from S. Such regions are supposed to be as long as possible. All substrings forming a conserved region C_i must be "similar" to each other according to some criterion of similarity. At this point there

is no generally accepted formal criterion of similarity exist, so the problem of finding synteny blocks is ill-defined.

In this work we introduce new criterion of similarity based on De Bruijn graphs and graph simplifications.

3. Algorithm description

3.1. Gener

Informally speaking, k-dimensional de Bruijn graph is a graph with vertices representing all possible strings of length k (called k-mers). In this graph two vertices u and v are connected by a oriented edge from u to v if exists $(k+1)$-mer w such that u is the prefix of w and v is the suffix of w.

4. Experimental results

Results

5. Conclusion

Conclusion

References

[1] Son K. Pham, Pavel A. Pevzner. DRIMM-Synteny: decomposing genomes into evolutionary conserved segments. Bioinformatics (2010) 26 (20): 2509-2516.

[2] Pavel A. Pevzner, Haixu Tang, Michael S. Waterman. An Eulerian path approach to DNA fragment assembly. Proc. Natl. Acad. Sci. USA. 2001 Aug 14; 98(17): 9748-53.

[3] Zamin Iqbal, Mario Caccamo, Isaac Turner, Paul Flicek, McVean. De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat Genet. 2012 Jan 8;44(2):226-32. doi: 10.1038/ng.1028.

[4] Manolis Kellis, Bruce W. Birren, Eric S. Lander. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature 2004 Apr 8;428 (6983): 617-24.