SyntenyFinder: A Synteny Blocks Generation and Genome Comparison Tool

Ilya Minkin[1], Nikolay Vyahhi[1], and Son Pham[2]

[1] St. Petersburg Academic University, St. Petersburg, Russia
[2] University of California, San Diego, USA

Introduction

Recent advances in sequencing and genome assembling technologies are resulting in many finished genomes. The comparison of these genomes has been emerging as a powerful tool for genome interpretations. These tasks often require genomes to be decomposed to a collection of synteny blocks – regions of conserved DNA.

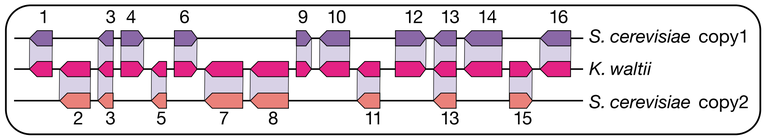


Figure 1: Example of a synteny block in two yeast genomes (Kellis2004), rectangles with arrows depict homologous genes

We propose *SyntenyFinder* – a tool for finding synteny blocks in genomes represented as nucleotide sequences. Our approach is based on de Bruijn graph and can be applied to closely related genomes.

De Bruijn graphs in synteny block mining

Given a string $\boldsymbol{S}$ and a natural number $\boldsymbol{k}$ we construct the de Bruijn graph $\boldsymbol{G(k)}$ as follows:

- For each unique $\boldsymbol{k}$-substring add a vertex to $\boldsymbol{G(k)}$
- For each $\boldsymbol{(k+1)}$-substring $\boldsymbol{w}$ in $\boldsymbol{S}$ add to $\boldsymbol{G(k)}$ an edge that connects $\boldsymbol{k}$-prefix of $\boldsymbol{w}$ with $\boldsymbol{k}$-suffix of $\boldsymbol{w}$
- Label edges with positions of the corresponding $\boldsymbol{(k+1)}$-mers

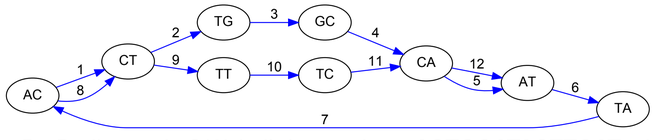


Figure 2: De Bruijn graph $\boldsymbol{G(2)}$ built from string $\boldsymbol{"ACTGCATACTTCAT"}$, bulge indicates a mismatch

In this graph we allow only paths that have consecutive labels on edges. Graph $\boldsymbol{G(k)}$ has following properties:

- Each valid path in the graph spells a substring from $\boldsymbol{S}$
- Non-branching paths in the graph indicate exact repeats in $\boldsymbol{S}$
- Variations in repeats create *bulges* in $\boldsymbol{G(k)}$
- Bulges are formed by two disjoint valid paths with shared ends

We remove bulges with size larger than some predefined constant δ to obtain long non-branching paths. This process is called *simplification*.

Double strandness issue

DNA has two strands and synteny blocks can be located on both. We address this problem by:

- Building graphs for both strands separately
- Labeling edges in these graphs with two different colors
- Merging two graphs and work with the resulting graph

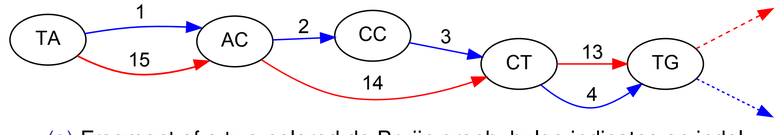
SyntenyFinder algorithm

Given two numbers $\boldsymbol{k}$ and δ and a set $\boldsymbol{S} = \{\boldsymbol{S_1}, \boldsymbol{S_2}, \ldots, \boldsymbol{S_n}\}$ of chromosomes represented as nucleotide strings, our algorithm works as follows:

- Concatenate all chromosomes in $\boldsymbol{S}$ into the supergenome $\hat{\boldsymbol{S}}$
- Construct graph $\boldsymbol{G^+(k)}$ from $\hat{\boldsymbol{S}}$ and color all its edges *blue*
- Construct graph $\boldsymbol{G^-(k)}$ from reverse-complementary of $\hat{\boldsymbol{S}}$ and color all its edges *red*
- Obtain $\boldsymbol{G(k) = G^+(k) \cup G^-(k)}$
- Change $\hat{\boldsymbol{S}}$ so that $\boldsymbol{G(k)}$ doesn't contain bulges having size $< \delta$
- Output non-branching paths as synteny blocks

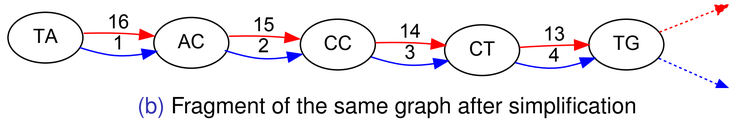
At any moment during simplification there is a one-to-one correspondence between the string and the graph – any changes in $\hat{\boldsymbol{S}}$ are immediately reflected in $\boldsymbol{G(k)}$.
We also keep some information on the edges to be able to reconstruct original coordinates of the found synteny blocks.

5' TACCTG ... TCAGTA 3'
3' ATGGAC ... AGTCAT 5'



(a) Fragment of a two-colored de Bruijn graph, bulge indicates an indel

5' TACCTG ... CAGGTA 3'
3' ATGGAC ... GTCCAT 5'



(b) Fragment of the same graph after simplification

Figure 3: Illustration of the bulge removal

Results

We benchmark SyntenyFinder on two datasets.

- Four strains of the bacteria *Pseudomonas aeruginosa*
- Two different yeast species: *K.waltii* and *S. cerevisiae*

Dataset	Total size	k	δ	Multiplicity	Count	Coverage
Bacteria	27 MBP	1000	5000	4	140	90%
Yeast	23 MBP	1000	20000	3	270	64%

For the yeasts we used local alignments from (Kellis2004) to enrich number of common $\boldsymbol{k}$-mers in conserved regions. Our blocks share 80% of their basepairs with the blocks described in (Kellis2004).

Discussion

SyntenyFinder can be efficiently used for finding synteny blocks in closely related genomes represented as nucleotide sequences. Benchmarks show that with some modifications our method can be applied to more distant genomes. Our near plans include:

- Release version for closely related species
- Extend SyntenyFinder to a wider range of genomes
- Incorporate it into genome rearrangements analysis tools

Acknowledgments

We would like to thank Pavel Pevzner, Steve O'Brien, Alla Lapidus, Matt Schultz and Dinh Diep for their help and contributions to many helpful discussions.
This work was supported by the Government of the Russian Federation (grant 11.G34.31.0018) and the National Institutes of Health (NIH grant 3P41RR024851-02S1).

References

Kellis M, Birren BW, Lander ES
Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae.
Nature, 2004 Apr 8; 428(6983):617-24.

Pham KS, Pevzner PA
DRIMM-Synteny: decomposing genomes into evolutionary conserved segments.
Bioinformatics, 2010; 26(20):2509-2516.