

Whole Genome Alignment

BMI/CS 776

www.biostat.wisc.edu/~craven/776.html

Mark Craven

craven@biostat.wisc.edu

February 2002

Announcements

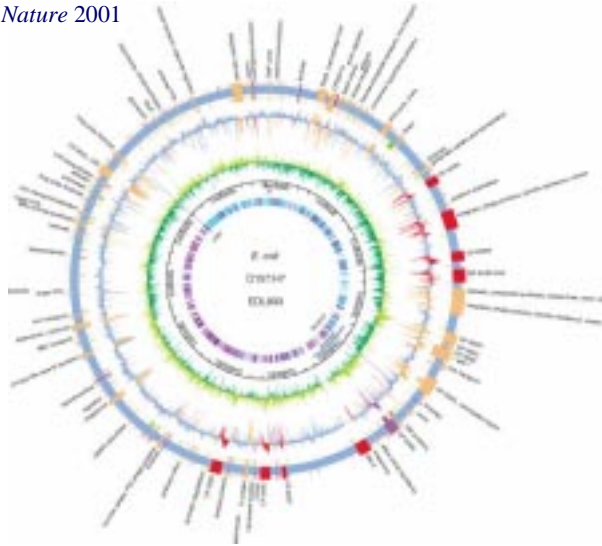
- talk of interest today: *Divergence Time and Evolutionary Rate Estimation with Multilocus Data*
Jeffrey Thorne, North Carolina State University
4:00pm, 1221 Computer Sciences
- guest lectures next week:
 - Prof. Christina Kendzierski on quantitative trait loci (QTL) mapping
 - Prof. Rich Maclin on keyphrase extraction to annotate high-throughput experiments
- reading for the week of 2/25: Chapter 3 of Durbin et al.

Whole Genome Alignment: Task Definition

- Given
 - a pair of genomes (or other very large scale sequences)
 - a method for scoring the similarity of a pair of characters
- Do
 - construct global alignment: identify matches between genomes as well as various non-match features

E. Coli Whole Genome Alignment

Perna et al., *Nature* 2001



Why Not Use Standard DP Methods?

- size of sequences being compared
memory, run-time issues
- features accounted for
standard alignment: point mutations, insertions, deletions
whole genome alignment: also transpositions, differences
in tandem repeats, etc.

The MUMmer System

- Delcher et al., *Nucleic Acids Research*, 1999
- given genomes A and B
 - find all maximal, unique, matching subsequences (MUMs)
 - extract the longest possible set of matches that occur in the same order in both genomes
 - close the gaps
 - output the alignment

Features Identified by MUMmer

- single nucleotide polymorphisms (SNPs)
- regions of divergence > 1 SNP
- large inserts
- repeats
- tandem repeats: two or more adjacent, approximate copies of a DNA pattern

Step 1: MUM Decomposition

- maximal unique match (MUM):
 - occurs exactly once in both genomes A and B
 - not contained in any longer MUM

Genome A: tcgatacGACGATCGCGGCCGTAGATCGAATAACGAGAGAGCATAAacgactta
Genome B: gcattaGACGATCGCGGCCGTAGATCGAATAACGAGAGAGCATAAtccagag



- key insight: a significantly long MUM is certain to be part of the global alignment

Suffix Trees

- the key idea in identifying MUMs is to build a suffix tree for genomes A and B

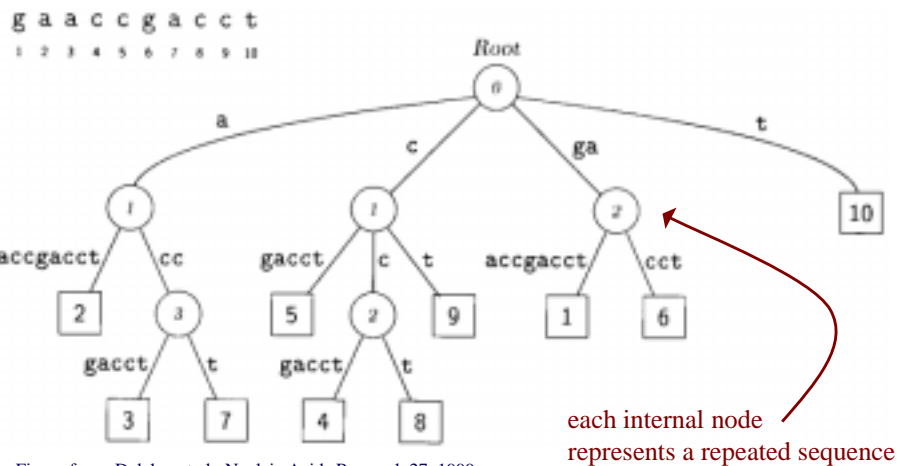


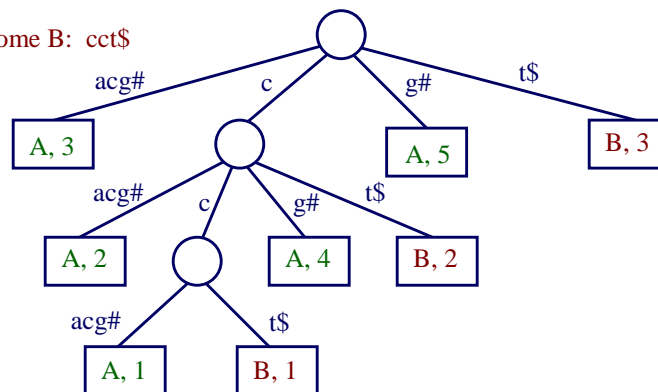
Figure from: Delcher et al. Nucleic Acids Research 27, 1999

MUMs and Suffix Trees

- add suffixes for both genomes A and B to tree
- label each leaf node with genome it represents

Genome A: ccacg#

Genome B: cct\$

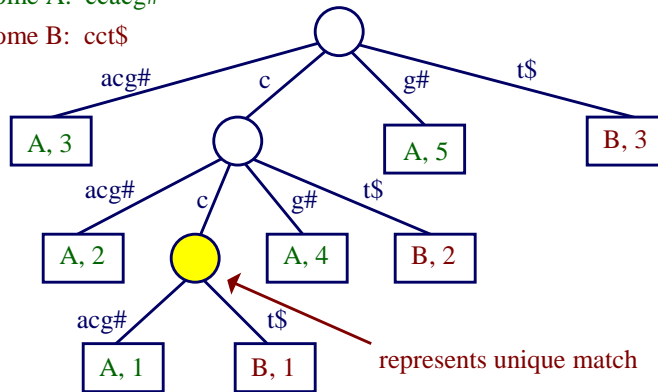


MUMs and Suffix Trees

- a unique match: internal node with 2 children: leaf nodes from different genomes
- but these matches are not necessarily maximal

Genome A: ccacg#

Genome B: cct\$

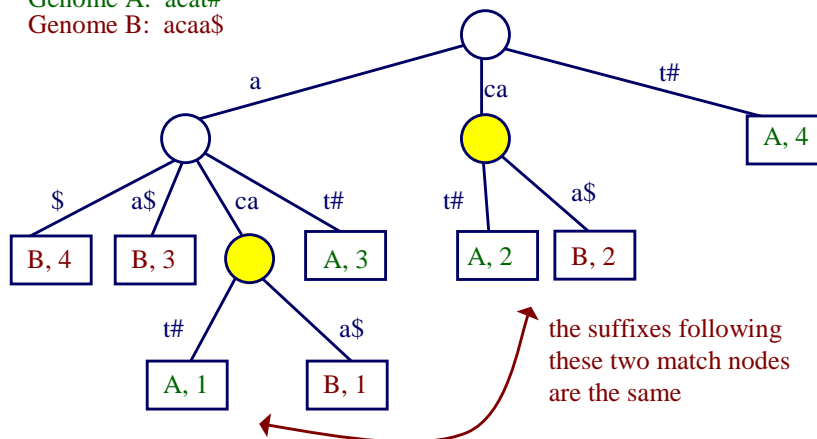


MUMs and Suffix Trees

- to identify maximal matches, can compare suffixes following unique match nodes

Genome A: acat#

Genome B: acaa\$



Suffix Trees

- can build in linear time (in lengths of genomes)
- can identify all MUMs in linear time (one scan of tree)
- space complexity is linear (exactly one leaf and at most one internal node for each base)
- main parameter of system: length of shortest MUM that should be identified (20 - 50bp here)

Step 2: Find Longest Subsequence

- sort MUMs according to position in genome A
- solve variation of Longest Increasing Subsequence (LIS) problem to find sequences in ascending order in both genomes

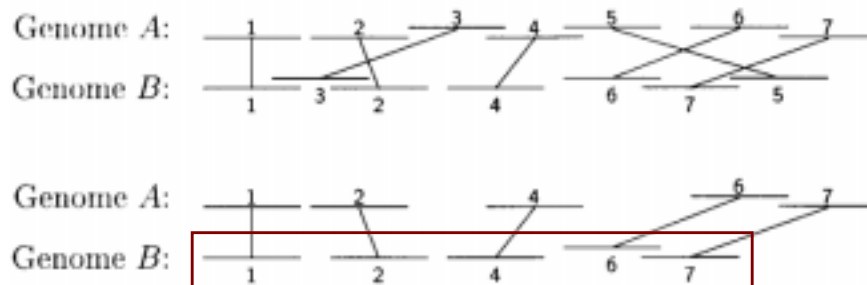


Figure from: Delcher et al. Nucleic Acids Research 27, 1999

Step 3: Close the Gaps

- SNPs:
 - between MUMs: trivial to detect
 - otherwise: handle like repeats
- inserts
 - transpositions (subsequences that were deleted from one location and inserted elsewhere): look for out-of-sequence MUMs
 - simple insertions: trivial to detect

Step 3: Close the Gaps

- polymorphic regions
 - short ones: align them with dynamic programming method
 - long ones: call MUMmer recursively w/ reduced min MUM length
- repeats
 - detected by overlapping MUMs

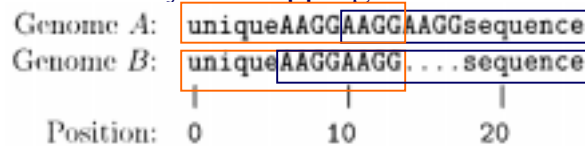


Figure from: Delcher et al. Nucleic Acids Research 27, 1999