

Animating Suffix Tree Algorithms

Alice Paul

June 7th, 2011

Outline

Genome Sequencing

- Motivation

- Genome Sequencing Process

- Use of Suffix Trees

Suffix Trees

- Definition

- Algorithms

- Other Applications

Algorithm Animation

- Intro to Gato

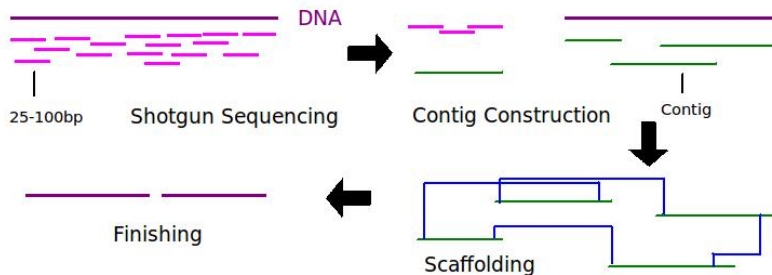
- Uses

References

Meet Tom



Genome Sequencing

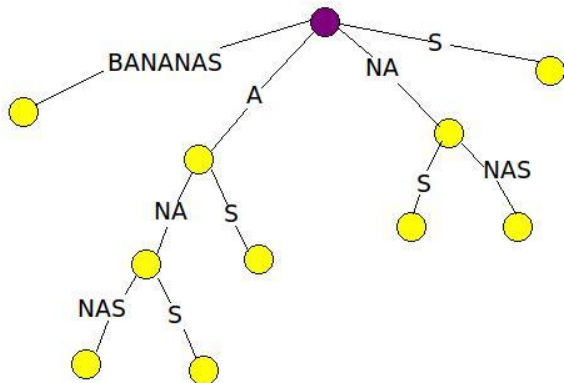


What is important in all of this?

- ▶ Exact pattern matching
- ▶ Repeat patterns
- ▶ Allowing for some mismatches (insertions, deletions, substitutions)

Suffix trees allow us to answer many different questions about patterns efficiently.

What is a suffix tree?



Note: Every substring is a prefix of a suffix in the tree. This allows us to look up all patterns, not just suffixes.

Definition

Definition

A **suffix tree** for a m -character string S is a rooted directed tree with m leaves labeled 1 to m that satisfies the following conditions:

- ▶ Each internal node, other than the root, must have at least two children.
- ▶ Each edge is labeled with a non-empty string.
- ▶ No two edges out of a node can start with the same character.
- ▶ **For any leaf i the concatenation of edges from the root to i exactly spells out $S[i \dots m]$.**

Algorithm of the Year 1973!

Three main algorithms:

- ▶ Weiner's Algorithm 1973
- ▶ McCreight's Algorithm 1976
- ▶ Ukkonen's Algorithm 1995

More Suffix Tree Applications in Bioinformatics

- ▶ Generalized Suffix Trees
- ▶ Longest common substrings
- ▶ Finding complemented palindromic sequences as possible restriction enzyme sites
- ▶ Identifying frequently recurring substrings (Tandem repeats)

[illegible]

Uses of Algorithm Animation in Gato

- ▶ Allows the user to trace the steps of the algorithm
- ▶ Graph algorithms might look daunting on paper, but can be easy to visualize
- ▶ Develop my own understanding of string algorithms in general and uses in bioinformatics

References:

- ▶ Alkan, Can, Bradley P. Coe, and Evan E. Eichler. "Genome Structural Variation Discovery and Genotyping." *Nature Reviews Genetics* 12 (2011): 363-76. Print.
- ▶ Cirulli, Elizabeth T., and David B. Goldstein. "Uncovering the Roles of Rare Variants in Common Disease through Whole-genome Sequencing." *Nature Reviews Genetics* 11 (2010): 415-25. *Nature Reviews Genetics*. Web. 4 June 2011.
<http://www.nature.com/nrg/journal/v11/n6/full/nrg2779.html>.
- ▶ "Genome Sequence Assembly Primer." *UMD Center for Bioinformatics and Computational Biology*. University of Maryland. Web. 04 June 2011.
http://www.cbcb.umd.edu/research/assembly_primer.shtml.
- ▶ Gibson, Jerry D. *The Mobile Communications Handbook*. Boca Raton: CRC, 1999. Print.
- ▶ Gusfield, Dan. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge [England: Cambridge UP, 1997. Print.
- ▶ Schliep, Alexander. "CATBox: An Interactive Course in Combinatorial Optimization." *Schliep.org*. Web. 04 June 2011.
<http://schliep.org/CATBox>.