Suffix Trees (and Relatives) Come of Age in Bioinformatics

Dan Gusfield
Computer Science Department, University of California Davis

Abstract

The book *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology* [1] contains about 125 pages devoted to suffix trees, suffix arrays, and their applications in computational biology. A related data structure, the DAWG is discussed via exercises. The book contains a wide range of applications of suffix trees, and while most have a biological "motivation", at the time I wrote the book (between 1994-1997) I only knew of a few publications where suffix trees and suffix arrays were actually applied to real problems in biology. So much space was devoted to suffix trees because I believed then that they had great potential applications in computational biology, allowing one to efficiently solve a very wide range of complex sequence analysis tasks. Often, the use of suffix trees, or close relatives, allowed a speedup from an exponential number of operations, down to a linear number. Even a reduction from a quadratic number of operations to linear can have dramatic consequences. No other single data structure allowed such impressive speedups, and on such a wide range of problems.

At the time the book was written, several things limited the wider application of suffix trees: large memory requirements; limited locality of reference; the conceptual difficulty of the algorithms; and lack of available code; lack of general exposure in the bioinformatics community (and even the computer science community) to suffix trees.

Much has changed since 1997. Suffix trees and close relatives are now widely taught in graduate level courses on computer algorithms and on bioinformatics; there are several good expositions (I think) on suffix tree algorithms and uses; the space requirements have been substantially reduced; machines memories have greatly increased; additional variants of suffix trees have been introduced that address some of their deficiencies; and suffix tree code is publicly available. As a result, and to some extent as a cause, there are now many more applications in bioinformatics of suffix trees and related data structures.

Since the publication of the book, I have been aware of the wider uses of suffix trees in bioinformatics, but I had not systematically followed the field. This talk (and the paper I am writing based on it) is an attempt to catch up. I will try to survey here several (about twenty in the paper) recent papers that use suffix trees in bioinformatics, and papers that helped enable those applications. I also mention a few results on suffix trees that are presently only motivated by computational biology, but may later have significant application. However, the survey will not be exhaustive, and I apologize to those people whose work I have (unintentionally) overlooked.

I view the recent results as falling into the following overlapping categories: new fundamental algorithms; implementation improvements (mainly space); important new variations on suffix trees and arrays (for example virtual suffix arrays); new algorithms for tandem repeats; algorithms for approximate repeats; algorithms for fast lookups in databases; algorithms to study and display substring frequencies; motif and pattern discovery algorithms; hybrid dynamic programming and pairwise alignment methods; oligo construction and microarray design algorithms; sequence assembly and resequencing; (multiple) whole genome comparisons; miscellaneous applications; less used, related data structures - affix trees, DAWG; unsolved problems.

Reference

[1] Gusfield, D, *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*, Cambridge University Press, Cambridge, 1997

Proceedings of the IEEE Computer Society Bioinformatics Conference (CSB'02)
0-7695-1653-X/02 $17.00 © 2002 **IEEE**