

New Graduate Course

CS 591 -- B1
TR 3:30-5:00
Fall 2004

Instructor: Gary Benson

Course Title: Pattern Matching and Pattern Detection Algorithms with Applications in Biological Sequence Analysis

Credits: 4

Prerequisites

Solid background in algorithms (CS 330 or equivalent) and permission of instructor.

Overview

Pattern matching and pattern detection problems involve finding occurrences of patterns in data. In pattern matching problems, the pattern is given; in pattern detection only a general pattern model is available. A simple and well studied example of the former is finding all occurrences of a word in a text. An example of the latter is finding a common motif in a collection of biological sequences. Pattern matching and pattern detection algorithms are developed to efficiently solve these problems. Modern experimental procedures in biology generate enormous quantities of sequence data, *i.e.* DNA, RNA and protein sequences. The availability of these sequences, has spurred the design of algorithms specifically adapted for their analysis. This graduate level course will examine a variety of algorithmic techniques for pattern matching and pattern detection with applications geared to biological sequence analysis. Readings will come from a mixture of journal articles and textbooks. Algorithms will be discussed in the context of correctness, time and space complexity, sensitivity/specificity trade-offs, and the generality of their approach. They will be placed in an historical context to show how key ideas have led from one level of sophistication and performance to the next.

Topics

Exact Pattern Matching – Pattern to Text

- ? Finite automata and the Knuth-Morris-Pratt Algorithm
- ? Boyer-Moore algorithm
- ? Suffix trees
- ? Dictionary Matching

Approximate Matching – Sequence to Sequence

- ? Dynamic Programming and Alignment
- ? Specialized Alignment Techniques
- ? Tandem Alignment
- ? Composition Alignment

Approximate Matching – Sequence to Library

- ? BLAT

? Sim4

Approximate Matching – Library to Sequence

? RepeatMasker

? Censor

Pattern Detection

? Pattern Enumeration Methods

? Short Word Methods

Potential Readings

R. Baeza-Yates and C. Perleberg. Fast and practical approximate string matching. In *Third Annual Symposium on Combinatorial Pattern Matching*, pages 185-192, 1992.

G. Benson. Composition Alignment. *Proceedings of the Third International Workshop on Algorithms in Bioinformatics (WABI 2003)*, pp. 447-461, 2003.

G. Benson. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, 27:573-580, 1999.

G. Benson. Tandem cyclic global alignment. *Proceedings of the 12th Annual Symposium on Combinatorial Pattern Matching Jerusalem*. July 2001.

C. Burge and S. Karlin. Prediction of complete gene structures in Human genomic DNA. *J. Mol. Biol.* 268:78-94 (1997).

I. Jonassen, J. Collins and D. Higgins. Finding flexible patterns in unaligned protein sequences. *Protein Science*, 4:1587-1595, 1995.

T. Cormen, C. Leiserson and R. Rivest. *Introduction to Algorithms*. MIT Press, 1990.

W. Feller. *An introduction to probability theory and its applications*, volume I. John Wiley & Sons, 3rd edition, 1968.

L. Florea, G. Hartzell, Z. Zhang, G.M. Rubin, and W. Miller. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Research*, 8:967-974, 1998.

Dan Gusfield. *Algorithms on strings, trees, and sequences: computer science and computational biology*. Cambridge University Press, Cambridge, England, 1997

J. Jurka, P. Klonowski, V. Dagman, and P. Pelton. CENSOR - a program for identification and elimination of repetitive elements from DNA sequences. *Computers Chem.*, 20:119-121, 1999.

G. Landau and U. Vishkin. Fast parallel and serial approximate string matching. *Journal of Algorithms*, 10:157-169, 1989.

M.-Y. Leung, B. Blaisdell, C. Burge, and S. Karlin. An efficient algorithm for identifying matches with errors in multiple long molecular sequences. *Journal of Molecular Biology*, 221:1367-1378, 1991.

W. Miller and E. Myers. Approximate matching of regular expressions. *Bulletin of Mathematical Biology*, 51:5-37, 1989.

A. Neuwald and P. Green. Detecting patterns in protein sequences. *Journal of Molecular Biology*, 239:698-712, 1994.

P. Pevzner and M. Waterman. Multiple filtration and approximate pattern matching. *Algorithmica*, 13:135-154, 1995.

I. Rigoutsos and A. Floratos. Motif discovery without alignment or enumeration. In *Proceedings, Second Annual International Conference on Computational Molecular Biology (RECOMB 98)*, pages 221-227, 1998.

S. Salzberg, A. Delcher, S. Kasif, and O. White. Microbial gene identification using interpolated Markov models. *Nucleic Acids Research*, 26(2): 544-548, 1998.

J. Schmidt. All shortest paths in weighted grid graphs and its application to finding all approximate repeats in strings. *SIAM Journal of Computing*, 27:972-999, 1998.

A. Smit and P. Green. RepeatMasker at <ftp.genome.washington.edu/RepeatMasker.html>.

H. Smith, T. Annau, and S. Chandrasegaran. Finding sequence motifs in groups of functionally related proteins. *PNAS*, 87:826-830, 1990.

M. Suyama, T. Nishioka, and J. Oda. Searching for common sequence patterns among distantly related proteins. *Protein Engineering*, 8:1075-1080, 1995.