Analysis of Algorithms on Sequences with Applications: (Multimedia, Security, Bioinformatics)

Professor: W. Szpankowski
E-mail: spa@cs.purdue.edu

Overview. This course is about designing and analyzing algorithms and data structures on strings (e.g., digital trees, pattern matching, Lempel-Ziv schemes) and their applications to multimedia, security and bioinformatics. Following the general precept: *Give a man a fish and you feed him for the day; teach him to fish and you feed him for his lifetime*, we will concentrate on methodology of precise analysis of algorithms. The area of (precise) analysis of algorithms was born on July 27, 1963, when D. E. Knuth wrote his "Notes on Open Addressing" about hashing tables with linear probing. Since 1963 the field has been undergoing substantial changes.

In the first part of the course, we focus on techniques of analysis, but every method is illustrated by a variety of specific problems that arose from algorithms and data structures on strings. Our choice stems from the fact that there has been a resurgence of interest in algorithms on sequences and their applications in computational biology, security, multimedia compression, and information theory.

Approximate Course Outline

1. *Data structures and algorithms on sequences* (digital trees, Lempel-Ziv data compression algorithms, pattern matching algorithms, shortest common superstring, string editing problem).
2. *Probabilistic and analytical models* (probabilistic models on strings, review from probability, review of complex analysis).
3. *The first and second moment methods* (height in tries and suffix trees, microarrays).
4. *Elements of information theory* (lossless and lossy data compression, Lempel-Ziv schems, pattern matching).
5. *Generating functions* (recurrences, pattern occurrences in a text).
6. *Applications*:
   - Redundancy of popular data compression schemes.
   - Finding motifs in DNA.
   - Reliable threshold for intrusion detection.

Books used in the course:

- W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*, Wiley, New York, 2001 (preliminary version available at http://www.cs.purdue.edu/homes/spa/book.html).
- M. Mitzenmacher and E. Upfal, *Probability and Computing : Randomized Algorithms and Probabilistic Analysis*, Cambridge University Press, 2005.

Course Notes: I plan to run this course as a seminar with lectures, problems solving, and student presentations at the end of the class.