Analytic Pattern Matching with Applications: (DNA, Lempel-Ziv, and Trees)

Professor: W. Szpankowski
E-mail: szpan@purdue.edu

Overview. This course is about analytic and combinatorial methods in computer science with applications to pattern matching, information theory, bioinformatics and analysis of massive data. It is based on yet unpublished book by P. Jacquet and W. Szpankowski by the same title as the course. Our focus is on general pattern matching algorithms (exact, approximate, generalized, subsequence) treated by precise techniques of analytic combinatorics such as generating functions and complex asymptotics. After learning these new tools we apply them to several interesting applications from data compression (e.g., Lempel-Ziv and Tunstall schemes) to bioinformatics (motifs in DNA) to massive data (classification of twitter messages).

Approximate Course Outline

A. Analytic Combinatorics

1. Analysis of Algorithms
2. Recurrences
3. Generating Functions
4. Asymptotic Analysis

B. Pattern Matching

1. Exact Pattern Matching
2. Generalized Pattern Matching
3. Subsequence Pattern Matching
4. Digital Trees

C. Applications

1. Information Theory and Data Compression (Huffman, Lempel-Ziv, Tunstall)
2. Finding Motifs in DNA
3. Classification of `twitter` – Joint String Complexity
4. Finding a Needle in a Haystack – Massive Data Sets.

Books used in the course:

- W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*, Wiley, New York, 2001
- P. Jacquet and W. Szpankowski, *Analytic Pattern Matching: DNA, Lempel-Ziv and Trees*, Cambridge University Press, 2014, to appear.
- Online material: B. Sedgewick and P. Flajolet:
  http://aofa.cs.princeton.edu/lectures/
  http://ac.cs.princeton.edu/lectures/
  https://class.coursera.org/introACpartI-001

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