The Chvátal–Sankoff problem: Understanding random string comparison through stochastic processes

Alexander Tiskin

Abstract. Given two equally long, uniformly random binary strings, the expected length of their longest common subsequence (LCS) is asymptotically proportional to the strings' length. Finding the proportionality coefficient γ, i.e. the limit of the normalised LCS length for two random binary strings of length $n \to \infty$, is a very natural problem, first posed by Chvátal and Sankoff in 1975, and as yet unresolved. This problem has relevance to diverse fields ranging from combinatorics and algorithm analysis to coding theory and computational biology. Using methods of statistical mechanics, as well as some existing results on the combinatorial structure of LCS combined with elementary probability theory, we link constant γ to the parameters of a certain stochastic particle process. These parameters are determined by a specific (large) system of polynomial equations, which implies that γ is an algebraic number. Using a computer program for exhaustive enumeration of configurations of the relevant stochastic process for a sufficient number of time steps, we solve our system numerically. Short of finding a closed-form solution for such a polynomial system, which appears to be unlikely, our approach essentially resolves the Chvátal–Sankoff problem, albeit in a somewhat unexpected way with a rather negative flavour.

Alexander Tiskin
Department of Mathematics and Computer Science
St. Petersburg State University
St. Petersburg, Russia