Optimal Reconstruction of a Sequence from its Probes

Alan M. Frieze Franco P. Preparata Eli Upfal

Abstract

An important combinatorial problem, motivated by DNA sequencing in molecular biology, is the reconstruction of a sequence over a small finite alphabet from the collection of its probes (the sequence *spectrum*), obtained by sliding a fixed sampling pattern over the sequence. Such construction is required for Sequencing-by-Hybridization (SBH), a novel DNA sequencing technique based on an array (SBH chip) of short nucleotide sequences (*probes*). Once the sequence spectrum is biochemically obtained, a combinatorial method is used to reconstruct the DNA sequence from its spectrum.

Since technology limits the number of probes on the SBH chip, a challenging combinatorial question is the design of a smallest set of probes that can sequence an arbitrary DNA string of a given length. We present in this work a novel probe design, crucially based on the use of universal bases (bases that bind to any nucleotide [**?**]) that drastically improves the performance of the SBH process and asymptotically approaches the information-theoretic bound up to a constant factor. Furthermore, the sequencing algorithm we propose is substantially simpler than the Eulerian path method used in previous solutions of this problem.