Near optimal solutions for the distinguishing string selection problem

Lusheng Wang
City U. of Hong Kong, Hong Kong
lwang@cs.cityu.edu.hk

Abstract

Consider two sets of strings, $\mathcal{B}$ (bad genes) and $\mathcal{G}$ (good genes), as well as two integers d_b and d_g ($d_b \leq d_g$). A frequently occurring problem in computational biology (and other fields) is to find a (distinguishing) substring s of length L that distinguishes the bad strings from good strings, i.e., for each string $s_i \in \mathcal{B}$ there exists a length-L substring t_i of s_i with $d(s, t_i) \leq d_b$ (close to bad strings) and for every substring u_i of length L of every string $g_i \in \mathcal{G}$, $d(s, u_i) \geq d_g$ (far from good strings).

We present a polynomial time approximation scheme to settle the problem, i.e., for any constant $\epsilon > 0$, the algorithm finds a string s of length L such that for every $s_i \in \mathcal{B}$, there is a length-L substring t_i of s_i with $d(t_i, s) \leq (1 + \epsilon)d_b$ and for every substring u_i of length L of every $g_i \in \mathcal{G}$, $d(u_i, s) \geq (1 - \epsilon)d_g$, if a solution to the original pair ($d_b \leq d_g$) exists. Since there are a polynomial number of such pairs (d_b, d_g), we can exhaust all the possibilibities in polynomial time to find an good approximation required by the correpsonging application problems.