The Degenerate Primer Design Problem

Linhart, C. and Shamir, R.
School of Computer Science, Tel Aviv University

A degenerate primer is a primer sequence, in which some positions contain more than one possible base. Degenerate primers can be used in PCR procedures in order to amplify a variety of similar sequences. We define the Degenerate Primer Design problem (DPD, in short) as a combinatorial optimization problem and prove that various restricted versions of it are NP-complete. We develop an approximation algorithm for one of these variants, and discuss its properties. Finally, we describe an experimental scheme for deciphering the human olfactory subgenome, in which the first step is the design of degenerate primers using a heuristic based on our approximation algorithm.

Given a set of DNA sequences, we wish to design a pair of degenerate primers, so that the primers match and amplify (in the PCR sense) as many of the input sequences as possible. The degeneracy of a degenerate primer is the product of the number of possible bases in each position. In order to reduce the probability of amplifying non-related sequences, we require that each degenerate primer has a degeneracy of no more than some pre-defined constant. We focus on a variant of DPD, called Maximum Cover DPD, in which the input strings and the primer are the same length, and we wish to maximize the number of strings that are matched by the primer. We develop an approximation algorithm for this variant of DPD.

DPD was studied and implemented as part of an experimental scheme for analyzing the composition of a large family of genes with conserved regions. Given a subset of known genes, we design degenerate primer pairs, which are used in PCR procedures to amplify fragments of known, as well as unknown, genes of the same family. The fragments are cloned, spotted on a high-density membrane and oligo-fingerprinted. Another novel algorithm, called CLICK, then clusters the clones into groups of similar genes according to their fingerprints. Finally, representatives from each cluster are sequenced and compared to the existing database. The experimental process was implemented for the human olfactory subgenome, which contains some 1% of all human genes. This project is a joint work with T. Fuchs, M. Khen, G. Glusman, T. Pilpel, D. Lancet (Weizmann), B. Malecova, U. Radelof, J. O'Brien, R. Herwig, H. Lehrah (Max-Planck, Berlin), R. Sharan, and D. Shmulevich (Tel-Aviv).

We have implemented a 3-phase DPD algorithm for designing good primers for the amplification of a family of genes. In the first step, we extract non-degenerate primer candidates from the input DNA sequences using an entropy score. In the second phase, we design primers using an efficient heuristic, which is based on our approximation algorithm. Finally, a greedy hill-climbing function improves the primers we obtain. Our work on the olfaction genes indicates that the DPD algorithm is an efficient heuristic that produces satisfactory results for biological data.