HIROSHIMA MATH. J.
22 (1992), 597–629

An efficient method for searching characteristic patterns of a subset in a large set of character sequences

Kahoru FUTAGAMI

(Received September 3, 1991)

1. Introduction

It is important to search similarities between two character sequences or characteristic patterns of a subset in a large set of sequences, in the areas of molecular biology, computer science and so on. For simplicity, we call sequences instead of character sequences.

The problem of searching similarities between two sequences has been formulated as the one of searching the longest common subsequence of two sequences under certain deletion/insertion constraints. This problem can be modified so as to search an optimum alignment under certain scoring rules, such as $+1$ for a base match and $-g$ for a gap. These problems have been studied by many authors. For global search methods, see Fitch [4], Dayhoff [1], Lipman and Pearson [13], Needleman and Wunsch [17], Sellers [20], Sankoff [19], and Wilbur and Lipman [23, 24]. For local search methods, see Hirschberg [10], Sellers [21], Smith and Waterman [22], and Goad and Kanehisa [9].

With the development of large database of sequences such as genes or images, it is necessary to compare several sequences. Relating to this problem, Korn et al. [12] developed a program for searching subsequences common to all of several sequences. In this paper we consider the problem of searching characteristic patterns of a subset in a large set of sequences. We formulate this problem as follows:

Let Z be a finite set of some alphabet, and let S and P be two finite sets of sequences whose units are composed from Z, such that $S \supsetneqq P$. Then, we are interesting in a sequence $a = (a_1, \ldots, a_k)$ with $a_i \in Z \cup \{0\}$, $i = 1, \ldots, k$ satisfying the following conditions (1) $\sim$ (4):

(1) $a \neq (0, \ldots, 0)$,
(2) $k \leqq \min \{\ell(b) \mid b \in P\}$ ($\ell(b)$ denotes the length of b),
(3) For any $b = (b_1, \ldots, b_h) \in P$, there exists an integer i_0 such that

$$0 \leqq i_0 \leqq h - k \quad \text{and} \quad a_i = b_{i+i_0} \text{ if } a_i \neq 0,$$