Randomized Algorithms for Motif Detection

Lusheng Wang[1], Liang Dong[2], and Hui Fan[3]

[1] Department of Computer Science, City University of Hong Kong
Kowloon, Hong Kong
[2] Department of Computer Science, Peking University, Beijing 100871, P.R. China
[3] School of Information and Electronic Engineering,
Institute of Shandong Business and Technology, Yantai, Shandong, P.R. China
lwang@cs.cityu.edu.hk, dongl@theory.cs.pku.edu.cn, fanlinw@263.net

Abstract. Motivation: Motif detection for DNA sequences has many important applications in biological studies, e.g., locating binding sites and regulatory signals, and designing genetic probes etc. In this paper, we propose a randomized algorithm, design an improved EM algorithm and combine them to form a software.
Results: (1) We design a randomized algorithm for consensus pattern problem. We can show that with high probability, our randomized algorithm finds a pattern in polynomial time with cost error at most $\epsilon \times l$ for each string, where l is the length of the motif and ϵ can be any positive number given by the user. (2) We design an improved EM (Expectation Maximization) algorithm that outperforms the original EM algorithm. (3) We develop a software MotifDetector that uses our randomized algorithm to find good seeds and uses the improved EM algorithm to do local search. We compare MotifDetector with Buhler and Tompa's PROJECTION which is considered to be the best known software for motif detection. Simulations show that MotifDetector is slower than PROJECTION when the pattern length is relatively small, and outperforms PROJECTION when the pattern length becomes large.
Availability: Free from http://www.cs.cityu.edu.hk/~lwang/software/motif/index.html, subject to copyright restrictions.

1 Introduction

Motif detection for DNA sequences is an important problem in bioinformatics that has many applications in biological studies, e.g., locating binding sites [4], finding conserved regions in unaligned sequences, designing genetic probes [15, 17], etc. *Motif detection problem* can be defined as follows: given n sequences, each is of length m, and an integer l, where $l \leq m$, find a center string s of length l such that s appears (with some errors) in each of the n given sequences. If no error is allowed, the problem is easy. However, in practice, the occurrence of the center string s in each of the given sequences has mutations and is not exact. The problem becomes extremely hard when errors are allowed. Many mathematic models have been proposed. The following two are important.

R. Fleischer and G. Trippen (Eds.): ISAAC 2004, LNCS 3341, pp. 884–895, 2004.

The Consensus Pattern Problem: Given n DNA sequences $\{s_1, s_2, \ldots, s_n\}$, each is of length m, and an integer l, the *consensus pattern problem* asks to find a center string s of length l and a substring t_i of length l in s_i such that

$$\sum_{i=1}^{n} d(s, t_i)$$

is minimized.

The Closest Substring Problem: Given n DNA sequences $\{s_1, s_2, \ldots, s_n\}$, each is of length m, and an integer l, the *closest substring problem* asks to find a center string s of length l and a substring t_i of length l in s_i such that

$$d = \max_{i=1}^{n} d(s, t_i)$$

is minimized.

d here is called the *radius*. Other measures include the *general consensus score* [8] and SP-score.

Other than mathematic models, motif representation is another important issue. There are three representations, *consensus pattern*, *profile*, and *signature* [7]. Here we focus on consensus patterns and profiles. Let $t_1, t_2, \cdots, t_n$ be n strings of length l. Each t_i is an occurrence of a motif. The *consensus pattern* of the n occurrences is obtained by choosing the letter that appears the most in each of the l columns. The *profile* of the n occurrences is a $4 \times l$ matrix W, each cell $W(i, j)$ is a number indicating the occurrence rate of letter i in column j. Figure 1 gives an example.

```
                caaccca          a   0   1   0.4   0  0  0  0.4
                caacccc          c   1   0   0.2   1  1  1  0.2
                catcccg          g   0   0   0.0   0  0  0  0.2
                catccct          t   0   0   0.4   0  0  0  0.2
                cacccca
          --------------------                (b)
consensus pattern    caaccca
another con. pattern catccca
                (a)
```

Fig. 1. (a) The 5 occurrences of the motif and the consensus patterns. (b) the profile matrix

To evaluate those mathematic models, representations or programs, Pevzner and Sze [16] proposed a challenge problem, which has been studied by Keich and Pevzner [9, 10]. We randomly generate $n (n = 20)$ sequences of length $m (m = 600)$. Given a center string s of length l, for each of the n random sequences, we randomly choose d positions for s, randomly mutate the d letters from s and implant the mutated copy of s into the random sequence. The problem here is to find the implanted pattern. The pattern thus implanted is called