Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 24 (2013) 25 – 29

Procedia
Computer Science

17th Asia Pacific Symposium on Intelligent and Evolutionary Systems, IES2013

An iterative algorithm for motif discovery[☆]

Yetian Fan[a], Wei Wu[a,*], Rongrong Liu[a], Wenyu Yang[b]

[a]*School of Mathematical Sciences, Dalian University of Technology, Dalian, China 116023*
[b]*College of Science, Huazhong Agricultural University, Wuhan, China 430070*

Abstract

Analysis of DNA Sequence motifs is becoming increasingly important in the study of gene regulation, and the identification of motif in DNA sequences is a very complex problem in computational biology. In this paper, we propose a method that employs the general GA framework and computes the motifs from the shot motif length to the standard length with three operation in GA and a new operation called Addition proposed by us. The experiment results on simulated data and real biological data show that the obtained motifs are consistent with the real ones. Moreover, our method gets higher score than the other three methods: Gibbs Sampler, Genetic Algorithm (GA) and GARPS algorithm in terms of the data CRP. In addition, our algorithm is a parallel random search that is beneficial to implement parallel computing to increase computational efficiency of the algorithm.

Keywords: Bioinformatics; Motif discovery; DNA sequences; Transcription factors

1. Introduction

With the increasing volume of biological sequences in public databases, motif discovery has been one of the fundamental problems in computer science and molecular biology, which has important applications in locating regulatory sites and drug target identification. Genes are segments of the DNA that cooperate to produce different proteins for some particular functions. In order to start the protein decoding process (gene expression), transcription factors normally should bind to regulatory sites on DNA sequence proceeding the gene. These transcription factor binding sites on DNA sequence are called motifs, which are usually located in the upstream regions. And the actual instances of motifs on DNA sites corresponding has the same length, while they may be not exactly same with each other.

Motifs are generally relative short, recurring, conservative patterns in the regulatory regions. Compared with background distribution, the motif alignments contained in the data set are those whose letter distribution much more different. Accurate identification motifs is a challenging problem. Because the lengths of motifs are usually very short (up to 30 nucleotides), while that of the regulatory regions which contain motifs are very long (range from several

[☆] Research funded by National Natural Science Foundation of China (11171367) and the Fundamental Research Funds for the Central Universities of China.

[*] Corresponding author: School of Mathematical Sciences, Dalian University of Technology, Dalian, China 116023
E-mail address: wuweiw@dlut.edu.cn

doi:10.1016/j.procs.2013.10.024

hundreds to several thousands nucleotides). In addition, the mutations of the actual instances of motifs are adding to the burden.

So far, many algorithms are proposed to predict motifs. Some stochastic searching algorithms are very popular, such as Gibbs sampler and MEME[1]. They have many advantages, the consuming time of Gibbs sampler is lower, and MEME is superior to the other methods by its prediction accuracy. However, these algorithms have their drawback of dropping into local optimum easily. Therefore, other stochastic approaches introduced heuristics into their algorithms. Huo et al.[2] proposed an original algorithm (GARPS) in order to solve the problem of motif discovery, which combines Genetic Algorithm (GA) with Random Projection Strategy (RPS). In addition, there are many other heuristic methods to predict motifs, such as particle swarm optimization[3,4,5], tabu search algorithm[6] and simulated annealing[7].

In this paper, we propose a new approach that employs genetic algorithm to find motifs in DNA sequences. The approach started with short motifs whose length is only three, then added a site by three operators until the length of optimal motif is up to the standard length. The experiment results on simulated data and real biological data show that the obtained motifs are consistent with the real ones. Moreover, our method gets higher score than the other three methods: Gibbs Sampler[8], GA[9] and GARPS algorithm in terms of the data CRP.

This paper is divided into four sections. The new algorithm is described in Section 2. Supporting numerical experiments are presented in Section 3. In Section 4, a conclusion of the new algorithm is given.

2. Method

2.1. Population initialization

In this paper, we choose three as the length of initial individuals. As each site is chosen from $\{A, C, G, T\}$, there are totally 64 different initial individuals. Then the length of individuals will add one every epoch until it reaches to the standard length. And keep the population number of individuals of 64 throughout the algorithm.

2.2. Fitness score function

Definition 2.1. *We consider the fitness score function of one single sequence. For each region in the sequence S_m^k, it matches the given motif P_n and then has a fitness score which is calculated by the fitness score function, defined as follows:*

$$FS(P_n, S_m^k) = (\sum_{i=1}^{L} match(P_n^i, S_m^{ki}))/L, \qquad n = 1, 2, 3\cdots, and\ m = 1, 2, 3\cdots, M \tag{1}$$

where

$$match(P_n^i, S_m^{ki}) = \begin{cases} 1, if\ P_n^i = S_m^{ki} \\ 0, if\ P_n^i \neq S_m^{ki} \end{cases} \tag{2}$$

n is the index of motifs, M is the number of the sequences, m is the index of sequences, k is the position of matched regions in the sequence, i is index of the position within the motif and matched regions in the sequence, and L is the length of motifs.

Definition 2.2. *For a sequence S_m, the fitness score of motif pattern P_n, defined as follows:*

$$FS(P_n, S_m) = \max_{k=1}^{K}\{FS(P_n, S_m^k)\} \tag{3}$$

where there are K subsequences in sequence S_m, whose length is equal to the length of motif. We select the highest score as the fitness score of motif P_n.

Definition 2.3. *For the set of sequences $S = (S_1, S_2, \cdots S_M)$, the fitness score of motif P_n, defined as follows:*

$$TFS(P_n, S) = \sum_{m=1}^{M} FS(P_n, S_m)/M \tag{4}$$