Journal of Fiber Bioengineering and Informatics 8:3 (2015) 539–546
doi:10.3993/jfbim00154

Research on DNA Sequence Homology Based on Second Order Markov Model [⋆]

Junyan Zhang [a,b,*], Chenhui Yang [a,b]

[a] *College of Computer Science, Chengdu University, Chengdu 610106, China*

[b] *Key Laboratory of Pattern Recognition and Intelligent Information Processing of Sichuan
Chengdu University, Chengdu 610106, China*

Abstract

DNA sequence homology is a critical and fundamental problem in bioinformatics. In this paper, we solve this problem by use of the second order Markov modal instead of traditional sequence alignment because DNA character sequence meets the Markov properties. Hence, the characteristics of DNA sequences are represented by using their two-step transition probabilities matrices. The similarity degree measurement between two different DNA sequences is defined. Our DSHM algorithm is put forward which is implemented by MyEclipse. The contrast experiments are done between DSHM and other two methods. The experimental results show that DSHM algorithm can determine DNA sequence homology correctly in the more effective way.

Keywords: DNA Sequence Homology; Similarity Degree; Second Order Markov Model

1 Introduction

Generally, a DNA sequence is treated as a long string of characters with a four-character set Σ=A, C, G, T. Thus, any one DNA sequence S$\in$ $\Sigma*$ [1]. DNA sequence homology is the most fundamental and important problem in the field of bioinformatics, which refers to the string comparison between or among two or more DNA sequences by a mathematical algorithm and searching for the similarities and dissimilarities, which is used for detecting their links in respect of the structure, function and evolution [2].

There are many ways to solve this problem, such as: (1) Graphical representation of DNA sequences. 2-D or 3-D graphics are employed to represent corresponding DNA sequences to analyse the relationship among DNA sequences [3, 4], which has better visual effect but higher computation complexity. Especially, graphics can only represent the characteristics of the content of the base, so most of the results can only demonstrate part of the features of DNA sequences. Therefore, the accuracy of this comparison method is not satisfactory. (2) Numerical representation of

[⋆] Project supported by the Education Department of Sichuan Province of China (No. 15ZB0373).

[*] Corresponding author.
Email address: `67306683@qq.com` (Junyan Zhang).

DNA sequences. It means mapping four-character set Σ of DNA sequences to the corresponding numerical sequences in accordance with certain rules, and comparing and analysing the DNA sequences by the characteristics thereof [5, 6], which can make us obtain better predictive effect but lack of a unified measurement. By this way, certain information may be lost in the process of mapping by the mathematical algorithms. This method is still under developing, and there remain many problems need to be solved. (3) String comparison and text compression methods, according to which DNA sequences can be considered as text information, and then be compared and analysed by the relative compression ratio of information [7]. This method only focuses on content of the base but ignoring its structural order. Clearly, the methods of compression can be introduced to improve speed, but some redundant sequences still exist. (4) Non-alignment methods are put to use for analysing features of DNA sequences in order to improve efficiency though the segmentation and positioning are difficult to be achieved [8, 9]. The number of DNA sequences is usually very large and their structures are very complicated. Therefore, the existed algorithms have their own advantages and disadvantages respectively. To overcome the above-mentioned disadvantages, we determine DNA sequence homology based on their similarity degree by use of second order Markov model.

The remainder of this paper is organized as follows. In the next section, some relative concepts and definitions are given. Section 3 describes the problem-solving ideas and put forward to our DSHM algorithm. In Section 4, we discuss the performance of the proposed algorithm in term of the results of contrast experiments. Finally, Section 5 concludes this paper.

2 Concepts and Definitions

Markov model [10] is one of the most important stochastic processes, and it is widely applied to modern biology, physics, business, geology, atmospherics, and so on.

Definition 1 Let $X = \{x_n,\ n{\in}T\}$ be a stochastic process, where x_n denotes the state of the system at time n. We assume the random variable x_n takes on values in set I. The set I is called the state space. The stochastic process X is a discrete-time Markov chain with state space I if $p\{x_{n+1}{=}i_{n+1}|x_0{=}i_0,\ x_1{=}i_1,\ \cdots,\ x_n{=}i_n\}$ holds for all $i_0,\ i_1,\ \cdots,\ i_n,\ i_{n+1}$ in I [11].

Definition 2 Conditional probability $p_{ij}(n){=}p\{x_{n+1} = j|x_n = i\}$ $(i,\ j{\in}I,\ n{\geq}0)$ is called one-step transition probabilities of Markov chain with state space I. Let P be the one-step transition probabilities matrix, which is made of p_{ij}, that is to say:

$$P = (p_{ij}) = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} & \cdots \\ p_{21} & p_{22} & \cdots & p_{2n} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix} \tag{1}$$

Definition 3 Conditional probability $p_{ij}^{(2)}{=}p\{x_{m+1}{=}k|x_m{=}g\}$ $(k,\ g{\in}I,\ m{\geq}0)$ is called two-step transition probabilities of Markov chain with the state space I. And $P^{(2)}{=}(p_{ij}^{(2)})$ is named two-step transition probabilities matrix, where $p_{ij}^{(2)}{\geq}0$ for $\forall i,\ j{\in}I$, and $\sum\limits_{j\in I} p_{ij}^{(2)}{=}1$ for $\forall i{\in}I$.