Chapter 6
A New Approach for DNA Sequence Similarity Analysis Based on Triplets of Nucleic Acid Bases

Dan Wei
Xiamen University, China & Fujian Key Laboratory of the Brain-like Intelligent Systems (Xiamen University), China

Qingshan Jiang
Xiamen University, China

Sheng Li
Xiamen University, China

ABSTRACT

Similarity analysis of DNA sequences is a fundamental research area in Bioinformatics. The characteristic distribution of L-tuple, which is the tuple of length L, reflects the valuable information contained in a biological sequence and thus may be used in DNA sequence similarity analysis. However, similarity analysis based on characteristic distribution of L-tuple is not effective for the comparison of highly conservative sequences. In this paper, a new similarity measurement approach based on Triplets of Nucleic Acid Bases (TNAB) is introduced for DNA sequence similarity analysis. The new approach characterizes both the content feature and position feature of a DNA sequence using the frequency and position of occurrence of TNAB in the sequence. The experimental results show that the approach based on TNAB is effective for analysing DNA sequence similarity.

INTRODUCTION

DNA sequence similarity, the degree of similarity among finite sets of strings of nucleic bases, is a basic problem in Bioinformatics, and the resulting information can be used to deduce structures, functions and evolutionary relationships of genes. Therefore, research in this realm has become an important topic in the field of Bioinformatics (Jiang et al., 2002).
There are a number of computational and statistical methods for similarity analysis developed over the past decade. The traditional methods are mainly based on alignment of the strings including insertions, deletions and gaps in sequence (Waterman, 1995), and hence their time and space complexities expand with increasing sequence length. In addition, these methods involve some subjective factors, including determining optimal criteria and assigning scores during alignment. Thus, there are some difficulties associated with these methods in comparing DNA sequences.

In recent years, some alignment-free methods, such as those based on graphical representation (Liao et al., 2005), matrices (Randic, 2000), vectors (Vinga & Almeida, 2003), and so on, have been proposed to overcome critical limitations of sequence analysis by alignment. Among them, methods based on vectors have been widely studied and applied. By constructing vectors for DNA primary sequences and calculating the distance between two vectors one can obtain the similarity of two DNA sequences. The distance between two vectors can be computed in three ways: Euclidean (Blaisdell, 1986), the cosine of the correlation angle (Stuart et al., 2002), and correlation coefficient (Petrilli, 1993).

The basic rationale for using vectors is that similar sequences will share word composition to some extent, which is then quantified by the distance among such vectors. Generally, sequence comparison is only based on the statistics of word frequency. The statistical and probabilistic properties of words in sequences (Reinert et al., 2000) were studied, with emphasis on the deductions of exact distributions and the evaluation of its asymptotic approximations. The first usage of k-word counts for biological sequence comparison was implemented by Blaisdell (1986). Two vectors of relative frequencies of k-words over a sliding window from two given DNA sequences were computed to analyze the sequences (Wu et al., 2001). Later, word-based comparisons were reviewed by Vinga and Almeida (2003). Recently, a sixteen-component vector of the relative frequencies of the dual nucleotides has been applied to characterize and compare the coding sequences (Luo et al., 2008). Based on the short word composition of biological sequences, a new distance metric was presented (Wang & Zheng, 2008). These word-based studies determined the similarity of DNA sequences by ignoring the position of each word within a sequence and noting only its frequency. However, the position of occurrence of word in the sequence has a close relation with gene transposition, translocation and converse. Considering the positions of L-tuple, which is a DNA sequence segment with the length L, the Characteristic Distribution of L-tuple (CDLt) was presented by Liu et al. (2007). In addition, CDLt was used to compare DNA sequences (Liu, 2008). Unfortunately, the methods based on CDLt failed to measure similarity among conserved sequences.

In this chapter, the frequency and position of occurrence of Triplets of Nucleic Acid Bases (TNAB) in a DNA sequence are discussed. Based on the frequency and position, this chapter presents a new approach which can extract more information from the DNA sequence. Moreover, a new algorithm SATNAB (Similarity Analysis based on TNAB) for sequence comparison is outlined. The experimental result suggests that this new approach is suitable for the analysis of DNA sequences.

This chapter is organized as follows. First, similarity analysis based on CDLt is described. The approach based on TNAB is presented next. Afterwards, we present the experiments and the performance results, followed by the conclusions.

**SIMILARITY ANALYSIS**

The characteristic distribution of L-tuple was studied to extract position feature from DNA sequences by Liu et al. (2007).