Chapter 3.12
DNA Sequence Visualization

Hsuan T. Chang
National Yunlin University of Science and Technology, Taiwan

ABSTRACT

This chapter introduces various visualization (i.e., graphical representation) schemes of symbolic DNA sequences, which are basically represented by character strings in conventional sequence databases. Several visualization schemes are reviewed and their characterizations are summarized for comparison. Moreover, further potential applications based on the visualized sequences are discussed. By understanding the visualization process, the researchers will be able to analyze DNA sequences by designing signal processing algorithms for specific purposes such as sequence alignment, feature extraction, and sequence clustering, etc.

INTRODUCTION

Recently, the great progress of biotechnology makes the deoxyribonucleic acid (DNA) sequencing more efficient. Huge amounts of DNA sequences of various organisms have been successfully sequenced with higher accuracies. By analyzing DNA sequences, the biological relationships such as homologous and phylogeny of different species can be investigated. However, the analysis of DNA sequences by the use of biological methods is too slow for processing huge amount of DNA sequences. Therefore, the assistance of computers is necessary and thus bioinformatics is extensively developed. Efficient algorithms and implemented computer-based tools are desired to deal with the considerable and tedious biomolecular data.

In general, DNA sequences are stored in the computer database system in the form of character strings. In a human somatic cell, its haploid nuclear DNA contains 2.85 billion base pairs (bps), in which a tremendous wealth of genetic information resides (Collins et al., 2004). Distinguishing the differences and similarities among DNA sequences has been a major task for biologists. Most of the sequences in their character strings are too long to be displayed on the computer screen and, therefore, are very hard to be extracted for any feature or characteristic.

Development of visualization techniques for presenting biological sequences has been widely attempted (Roy, Raychaudhury, & Nandy, 1998; Loraine & Helt, 2002). Mehta & Sahni (1994) proposed some efficient algorithms that make use of the compact symmetric directed acyclic word graph (csdawg) data structure. Blumer, Blumer, Haussler, McConnell, and Ehrenfeucht (1987) proposed the analysis and visualization of patterns in long string. Some previous studies (Anastassiou, 2001; Berger, Mitra, Carli, & Neri, 2002; Wang & Johnson, 2002) have shown various methods (such as discrete Fourier transform or wavelet transform) of transforming the symbolic DNA sequences to numeric sequences for further processing. With the methods described above, the periodic patterns existed in DNA sequences can be observed from the determined scalograms or spectrograms. On the other hand, some methodologies (Cheever & Searls, 1989; Cork & Wu, 1993; Wu, Roberge, Cork, Nguyen, & Grace, 1993) were proposed to depict symbolic sequences by two-dimensional (2-D) images, three-dimensional (3-D) curves, or graphs. The calculation in some methods was troublesome and required intensive computation. Efficient and direct mapping methods are desired to convert the symbolic sequences into the numeric sequences, and have them displayed in graphs.

Visualization (i.e., graphical representation) of DNA sequences provides corresponding pseudo shapes in a global view, which makes the sorting, comparison, and feature extraction based on the pseudo shape available. Visual recognition of differences among related DNA sequences by inspection can be made through sequence visualization. The graphical form can be viewed on a computer display or be printed on a piece of paper. Therefore, global and local characterizations/features of sequences can be quickly grasped in a perceivable form. Moreover, numerical characterizations (or features) of sequences can be determined from the visualized data (Chang, Xiao, & Lo, 2005). The extracted characterizations or features make the sequence data in a much more manageable fashion. Visualization is an alternative for DNA sequence representations. In addition to complementing the limitations of conventional symbolic-based methods, more powerful application tools would be expected in the near future.

There have been many researchers around the world working on this topic. The related bioinformatics tools and databases of visualized DNA sequence also have been accessible over the Internet. Therefore, the objective of this chapter is a literature review and summarization of the visualization methods for DNA sequences. Further applications based on the visualization methods will also be mentioned.

BACKGROUND

A DNA strand is a biomolecular polymer characteristic of four different bases, i.e., adenine (A), guanine (G), cytosine (C), and thymine (T). The length unit of a DNA sequence is in base pair (bp) for a double-stranded DNA, or in nucleotide (nt) for a single strand. With the rapid development and progress in bioinformatics and the completion of more genome sequencing projects, many sequence databases have been constructed and will be continuously constructed for sequence query.

It is not easy to directly access large amount of genomic DNA sequence data to perform mathematical analysis. That is, character-based representation of DNA sequences cannot provide immediately useful nor informative characterizations. To perform visualization, symbolic-based sequences should be transformed into numeric-based versions. Different transformation will provide various characterizations of the visualized sequences. Therefore, the transformation (or mapping, simply saying) is a critical issue in sequence visualization.

There are many transformation methods for mapping the symbolic characters A, T, C, G, to