Visualization of DNA Sequence Features Based on Cellular Automata

Qingnan Huang, Xuanqi Wang, Huili Li, Feng He, and Xiaoming Wu

Abstract. Visualization of special patterns in biological sequences can assist revealing important roles in gene regulation and other basic molecular activities of the sequence. The visualization method needs to highlight interesting sequence patterns while suppressing trivial aspects. A biology sequences visualization scheme based on cellular automata is developed in this study. Features such as alleles of a DNA sequence were extracted and mapped into a grid in a two-dimensional plane, creating an initial pattern. Then, two-dimensional cellular automata were iteratively executed according to predefined rules and turned the initial pattern into a two-dimensional pattern, forming the fingerprint of the sequence. This fingerprint can be served as a representation of the sequence and can be used to make sequences comparing.

1 Introduction

Important patterns or features in a biological sequence are closely related to the biological function of the sequence. In DNA sequence analysis, the bases in specific locations are usually used to make comparisons, so as to find their roles and functions by combining phenotypic information and other data. Visualization is important in studying sequence functions, because it is inconvenient when observing a sequence in a string of characters, particularly when the sequence is long.
A proper visualization method showing only the essential features of the sequence, can help researchers carrying out in-depth studies.

The approaches to visualize biological sequence can be divided into four categories. The first directly displays the three-dimensional structure of the corresponding macromolecule of the DNA sequence, when the structure is known. Software such as Rasmol and Cn3D are this kind of tools.

The second category visualizes the statistical characters of a sequence. A DNA molecule can be represented by a character string from 4 alphabets of A, C, G, T. Software such as SeqVISTA[1] can display a variety of sequence features graphically. Other features, such as base composition, dimmer distribution, and codon usage, can also be displayed graphically by some software such as BioEdit.

The third category uses curves to represent biological sequences. For example, the H curve has been used to represent DNA sequence [2]. In an H curve, a DNA sequence is mapped into a three dimensional space; repeating sequence and other sequence patterns can be distinguished according to the shape of the curve [3]. DNA walking is another method that represents each letter in the sequence as a movement in the 2D plane. The Z curve is a three-dimensional curve that uniquely represents a given DNA sequence and can provide insight into the understanding of gene replication mechanisms through the visualization result [4].

The 4th category uses a two-dimensional plane to show images generated by some rules. In these kinds of methods, each pixel in the image comes from some statistic of the whole sequence. A full image shows the feature distribution of the whole sequence. For example, a complete genome can be shown as an image of the K-string distribution [5].

Although many methods have been developed to visualize biological sequences, we particularly investigate cellular automata for genomic sequences visualization. In order to show the relationship between energy consumption and gene mutation of the Hepatitis B virus, Shao et al. represented base pairs of a DNA sequence with 4 different colors, and then used cellular automata to generate two-dimensional images[6]. When changes occur in the sequence, its two-dimensional image will change accordingly. By comparing the differences between images, one can evaluate sequence variations and the inspect sequences evolution.

The genome of the SARS virus can also be visualized by cellular automata. It can be realized by turning the bases into a string of 0 and 1, and then use iterative cellular automata rules to create images. Different texture occurs when the sequence of SARS and non-SARS sequence are turned into images[7]. This method is capable of visualizing a whole genomic sequence.

Besides the whole sequence, many features such as genetic markers, functional sites are very informative and need to be visualized. In this paper, these interesting sites were extracted from DNA sequences to create an initial image, and then two-dimensional cellular automata were used to generate representative images of the sequence. The resulted image was used to compare heredity information of STR sites in this study.