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.

Qingnan Huang
The 705 Research Institute, China Shipbuilding Industry Corporation, Xi'an, P.R. China

Xuanqi Wang
Department of Cardiology, The Fourth People's Hospital of Shaanxi Province, Xi'an, P.R. China

Huili Li · Feng He · Xiaoming Wu
The Key Laboratory of Biomedical Information Engineering of Ministry of Education, Xi'an Jiaotong University, Xi'an, 710049, P.R. China
e-mail: wxm@mail.xjtu.edu.cn

Z. Qian et al. (Eds.): Recent Advances in CSIE 2011, LNEE 129, pp. 77–82.
springerlink.com

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.

2 Methods

Cellular automataon (CA) is a discrete dynamical system model. Currently, it is wildly used in complex system simulation, cryptography, mathematics chemical reaction simulation, microstructure modeling and etc. Probably the best known example of a cellular automaton is Conway's "Game of Life" introduced by Gardner [8], and has been studied extensively by Wolfram.

A cellular automaton consists of a grid of cells, each in one of a finite number of states and usually formed in one or two dimensions. The state of each cell takes discrete values from a finite set. The state of a cell at time t is a function of the states of a finite number of neighborhood cells at time $t-1$. Cellular automaton includes five parts: cell space, states, neighborhood structure, update function, and boundary condition. A cellular automata system A can be represented by:

$$A = (L_d, S, N, f, B) \tag{1}$$

Here, L_d is the cell arrangement lattice of some cellular automata, where d is the dimension of the cellular automata. S is the discrete state of each cell. N is the collection of all cells in a neighborhood of a center cell. If a neighborhood has n cells, they form a vector written as $N = (s_1, s_2, ..., s_n)$, $s_i \in S, i \in \{1,2,...n\}$. f, the update function, is a rule that defines the state for each cell at time t as a function of the state of the neighborhood cells at time $t-1$. Usually, it is denoted by a table containing neighborhood cell's states and cell's new state. B is the boundary condition describing the boundary cell's state conversion criteria.

As to two dimensional cellular automata, integer value $a_{i,j}(t) \in \{0,1,2,...\}$ are assigned to discrete states of cell (i, j), $i \in \{0,1,...,N\}$, $j \in \{0,1,...,N\}$, where $t \in \{0,1,2,...\}$ are time points. The value of cell (i, j) at time t is calculated by

$$a_{i,j}(t) = f[a_{i-1,j}(t-1), a_{i+1,j}(t-1), a_{i,j-1}(t-1), a_{i,j+1}(t-1), a_{i,j}(t-1)] \tag{2}$$

where f is the update function, or called rules of the cellular automata. In this case, the state of the cell at time t is determined by the states of 4 neighboring cells and its own state value at time $t-1$.

If the states are Boolean, and the neighborhood cells are defined by von Neumann neighborhood, which is a group of 5 cells composing a middle cell with four cells in up, down, left and right directions, particularly. The total states combination number of all the five cells is 2^5=32, since each cell has only two states of 0 and 1. The state of the middle cell in time t is determined by the states of all the 5 neighborhood cells in time $t-1$. A truth table of 32 rows with five input cell states and one output cell states can describe the state transformations. When the output value permuted according to the value of input, a 32bit binary string can be formed. For example, if the string is 01101101101101101111101011001000, it means when the states of up, down, left, right cells and middle cells is 00000 in time $t-1$, the state of the middle cell is the first bit in the 32-bit string, which is "0". This represents a cellular automata rule named "Crystal3a" by Suzudo T [9].

Similarly, if a cell's state has 3 values in {0,1,2}, the von Neumann neighborhood can generate 3^5=243 states combination of the 5 cells. Each state combination would define a new state of the center cell in one of 3 states, and the result would form a tertiary states string length at 243. In this study, a rule named "bird" will be used, whose corresponding string is

01011202011211222202022202011211222211210222022222020202022202022222020202020202011210222210200020022220020210200020000001202020002000022220020220002000020200020002022000022022202000002000022022022222222202022202000002000002022202000002000

In fact, when the cellular automata run, and each cell's states are represented by different colors in 2D plane, the image represent the result would change each cycle. In this case, the "Crystal3a" rule can show a crystal growing effect, and the "bird" rule can show a self-organize effect of flying birds. Both the rules are member of the "Neumann" family [9].

3 Results

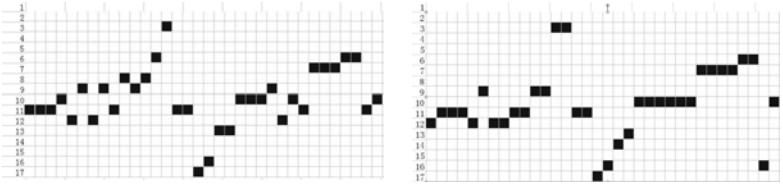
In genetics, the number of Short Tandem Repeats (STR) at the same locus may differ among individuals. When a number of such sites are considered, different individual can be identified correctly. However, by converting the sequence to a two-dimensional grid image, then using cellular automata iteratively, we can visualize the sequences and make comparisons easily.

We mapped the STR into a two-dimensional grid as follows: for N STR sites, a 2N column, two-dimensional grid was created. The state of each cell in the grid was set to 0 at the beginning. The sites were placed one by one from the left to the right; each site has two alleles, and occupies two columns; the vertical location in the column of each allele corresponds to the number of STR repeats. That is, the state of the cell whose vertical position corresponding to the number of repeats of the allele was set to 1. If the cells with the state 0 were marked as white and the cells with the state 1 were marked as black, a group of STRs would form a mosaic pattern in the plane, and an initial image was created. For example, for individual A and B, their genetic information in 17 sites is in table 1:

Table 1 The STR information of two individuals

Allel name:	Repeate times		Allel name	Repeate times	
	individual A	individual B		individual A	individual B
D16S539	11, 11	12, 11	HPRTB	13, 13	14, 13
D7S820	11, 10	11, 11	F13B	10, 10	10, 10
D13S317	12, 9	12, 9	FABP	10, 9	10, 10
CSF1PO	12, 9	12, 12	LPL	12, 10	10, 10
TPOX	11, 8	11, 11	CYAR04	11, 7	7, 7
TH01	9, 8	9, 9	CD4	7, 7	7, 7
F13A01	6, 3	3, 3	GPP3A09	6, 6	6, 6
FES/FPS	11, 11	11, 11	D8S1179	11, 10	16, 10
vWFA31/A	17, 16	17, 16			

The initial image can be generated according to the above steps, see Fig. 1.



(a) (b)

Fig. 1 Initial images created according to STR information of two individuals: (a) The initial image of individual A. (b) The initial image of individual B.

Using the two rules mentioned in Method section, the initial images of (a) and (b) in figure 1 can be converted to final images shown in Fig. 2. We can see when the heritage sites of individuals were different, the visualization results were dissimilar also, and can be distinguished easily.

(a) (b)

Fig. 2 The visualization result using STR data: (a) final result of individual A. (b) final result of individual B.

4 Discussion

A biological sequence is a linear and simple structure that concatenates bases out of a small alphabet and is often very long. Appropriate visualization methods can help understand the various features of a sequence by revealing the characteristics and rules hidden in the sequence. DNA sequences can be highly similar to each other except in the polymorphism sites, which may indicate significant functional differences. Most existing visualization methods display an entire sequence, but the approach in this study is able to extract features from the sequence, and use these features as a characteristic fingerprint of the original sequence. Our scheme involves a step to map the biological characteristics into a two-dimensional plane using iterative process according to cellular automata rules. Different sequences would produce distinctive mapping results, which can act as a sequence icon, a

much more informative way to represent a sequence. The salient patterns on the images can be used to compare sequences and make important features such as differences in genotypes stand out.

It is also necessary to hide the original genetic information when the right to privacy is involved. Our approach can be considered as an encryption method and visualization of biological inheritance information. Each feature in the original DNA sequence is mapped to a different plane's location, and then participates in an iterative process of cellular automata. Although the final image is derived from the original genetic information, the original sequence cannot be readily reconstructed from the final image. Visualization can assist understanding large scale and complex data such as biological sequences. We have developed an approach to visualize important biological sequence features, by generating images from the original sequences based on cellular automata. We have demonstrated that through an iterative process, important sequence features such as short tandem repeats or single nucleotide polymorphism in a pair of biological sequences can be converted to a pair of images with visually easy to identify patterns, offering a new way for sequence visualization and comparison. When different rules were used in the process, or different times of iteration were run, the result could be distinctively different. In fact, other features such as bases distribution, motif interval can also be visualized by our method.

Acknowledgments. This study was supported by the National Natural Science Foundation of China (60601017), Scientific Research Foundation of Shaanxi Provincial Office of Health, P.R. China(No. 2010D21,2010E06), and Fundamental Research Funds for Xi'an Jiaotong University.

References

1. Hu, Z., Frith, M., Niu, T., Weng, Z.: SeqVISTA: a graphical tool for sequence feature visualization and comparison. BMC Bioinformatics 4, 1 (2003)
2. Hamori, E., Varga, G., LaGuardia, J.J.: HYLAS: program for generating H curves. Comput. Appl. Biosci. 5(4), 263–269 (1989)
3. Roy, A., Raychaudhury, C., Nandy, A.: Novel techniques of graphical representation and analysis of DNA sequences — A review. J. Biosci. 23(1), 55–71 (1998)
4. Guo, F.B., Ou, H.Y., Zhang, C.T.: ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. Nucleic Acids Res. 31(6), 1780–1789 (2003)
5. Deng, X., Havukkala, I., Deng, X.: Large-scale genomic 2D visualization reveals extensive CG-AT skew correlation in bird genomes. BMC Evol. Biol. 7, 234 (2007)
6. Shi-huang, S.: Visualization of Gene Mutation Complicated Pattern of Hepatitis B Virus Based on Cellular Automata. Journal of Donghua University 22(1), 4 (2005)
7. Xiao, X., Shao, S., Ding, Y., Huang, Z., Chen, X., Chou, K.C.: Using cellular automata to generate image representation for biological sequences. Amino Acids 28(1), 29–35 (2005)
8. Gardner, M.: Mathematical Games-The Fantastic combinations of John Conway's new solitaire game "life". Scientific American 223, 120–123 (1970)
9. Suzudo, T.: Crystallisation of Two-dimensional Cellular Automata. Complexity International 6, 1–11 (1999)