2013 4th International Conference on Biology, Environment and Chemistry
IPCBEE vol.58 (2013) © (2013) IACSIT Press, Singapore
DOI: 10.7763/IPCBEE. 2013. V58. 12

A New Cellular Automata Based Converter for Genetic Sequences

Jui-Ming Chen[1,2], Ying-chieh Lin[2], Meng-Hsiun Tsai[3+], Yu-Chen Lin[4], Hsiao-Wei Lin[4] and Sheng-Hsiung Chiu[5]

[1]Department of Endocrinology and Metabolism, Tungs' Taichung MetroHarbor Hospital, R.O.C
[2]Department of Biomedical Informatics, Asia University Taiwan, R.O.C
[3]Institute of Genomics and Bioinformatics, National Chung Hsing University Taiwan, R.O.C
[4]Department of Management Information System, National Chung Hsing University Taiwan, R.O.C
[5]Troilus Biotechnology Co., Ltd, Taiwan

Abstract. With the development of computer technology, the more information is obtained from biological experiments through computer analysis, and the term bioinformatics is even coined. The commonly used tool in bioinformatics is sequence alignment, which is to compare the sequence with unknown function to the sequence with known function, and identify the possible biological function for them. Instead of using traditional dotplot or dynamic programming to conduct sequence alignment, this thesis uses cellular automata (CA) theory as the research topic. The parallel computing characteristic of cellular automata makes the future expansion model tremendously decrease the massive sequence computing costs. This thesis modifies the originally defined rules of cellular automata in order to make it more appropriate for amino acid sequence alignment. In addition, a cloud computing for sequence alignment system is proposed in this thesis so that users can do sequence alignment with the transfer methods used in this thesis through the Internet.

Keywords: Bioinformatics, Sequence alignments, Cellular automata, Cloud computing

1. Introduction

In recent years, National Institutes of Health (NIH) has launched Human Genome Project (HGP) in 1989 [1]. The project has successfully collected a lot of Gene sequence data. Due to the rapid development of compute technology, Amino acid and Nucleotide sequence alignments can be digitized and stored in a computer through a simple procedure [2]. In addition, the ability of computer hardware advances by leaps and bounds, Amino acid or Nucleotide sequences are getting longer and longer.

Therefore, sequence alignment is performed by the computer instead of manual interpretation in recent years [3-11]. However, if these sequences can be converted to some graphs [12-14] so that some important functions can be automatically shown. We can roughly predict the sequence function before using the computer to do Sequence alignment and it will save a lot of time.

2. Material and Research Method

2.1. Sequence convert to binary sequence

Gene is a meaningful unit of heredity. Take the human body as an example; body's cells have cell membrane, cytoplasm, nucleus and other organelles. Nucleus has chromosomes that composed of a number of DNA. And chromosomes can carry the genetic code. When the species complete gene sequencing, we can know the genetic code and it can be expressed as an amino acid sequence. To human cells, for example, the protein consists of 20 amino acids, so that the genetic code can be converted to the amino acid sequences.

[+] Corresponding author. Tel.: + 886-4-22840864 ext 666.
E-mail address: mht@nchu.edu.tw

Therefore, after completion of the species gene sequencing, the genetic code can be stored in the form of a string fragment (amino acid composition of the arrangement). In order to make the genetic code be kept and stored more properly, a sequence can be converted to a binary sequence.

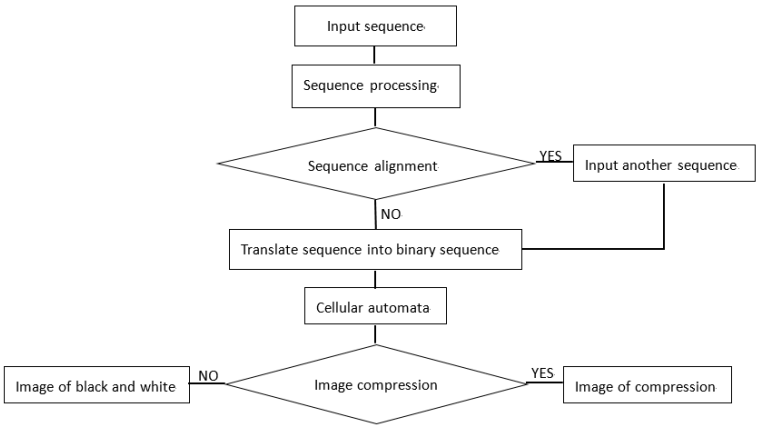

Fig. 1: The flowchart of the proposed method

2.2. Implement cellular automata image process

A cellular automaton is a discrete dynamic system. Each cell change their own situation based on the state of surrounding cells. Cellular automata can be regarded as a normal existence in the universe; each cell will only have a state at each time. Cellular automata can be re-designed in discrete time and space of macroscopic or microcosmic rules to observe the complex of space and rules. The application of cellular automata is for one-dimensional and two-dimensional currently. In this study, we used one-dimensional cellular automata.

One-dimensional cellular automata consists of a set of state variables according to the time change the composition S_t^i, i from 0 to N-1, N is the number of variables. Each variable can be placed in a grid, so that the composition is called a cell. Each cell has different states, and in the one-dimensional cellular automata are usually have two states, 0 and 1, which visualize the performance can be converted into black and white, so that the sequence of cellular automata array of genes can be expressed as a barcode. Among them, the sequence of the genetic code look-up table transformation method to convert a binary sequence, the process will let amino acids convert to binary sequence. After the completion of the binary sequence of conversion, the binary sequence will do rearrangements, where the Department of sequence alignment algorithms take the sequence of binary bits in each of three adjacent table of values to be under the control iteration conversion, when the completion of each of three adjacent bits of the iteration is changed, then move to the right one bit to continue iterative process, and finally generate a new sequence of a cellular automata, Complete rearrangement of the calculus sequence, select a plurality of side by side to form a binary sequence to form a selected number of rows, and then form a sequence of cellular automata array. This procedure can be simplified to the following equation:

$$D_{(i,j)} = F\left[D_{(i-1,j-1)}, D_{(i-1,j)}, D_{(i-1,j+1)}\right], If\ 1 \le j < S-1; 1 \le i < n \quad (1)$$

$$D_{(i,0)} = F\left[D_{(i-1,S-1)}, D_{(i-1,0)}, D_{(i-1,1)}\right], If\ j = 0; 1 \le i < n \quad (2)$$

$$D_{(i,S-1)} = F\left[D_{(i-1,S-2)}, D_{(i-1,S-1)}, D_{(i-1,0)}\right], If\ j = S-1; 1 \le i < n \quad (3)$$

$\boldsymbol{D_{(i,j)}}$ is on behalf of state value of converted the sequence of cellular automata array of (i, j) position of 0 or 1; i for the selected lines; S represents the binary sequence of length; and F refers to the rearrangement of the sequence rules of calculus functions. The sequence rearrangement calculus function F rules for the following table (Table 1).

Table 1: Iteration rule Table

rule	1	2	3	4	5	6	7	8
Iteration before	000	001	010	011	100	101	110	111
After iteration	1	0	1	1	1	0	0	0

In this research, we used the input genome [15] are:

- AAA96714 (ATPase [*Schistosoma mansoni*])
- AAC72756 (calcium ATPase 2 [*Schistosoma mansoni*])
- AAD09924 (plasma membrane calcium ATPase isoform 1 [*Homo sapiens*])
- AAF33531 (plasma membrane calcium ATPase isoform 3 [*Homo sapiens*])
- AF361357 (Schistosoma mansoni Ca-ATPase-like protein SMA3 mRNA, complete cds)
- CAA09303 (calcium ATPase [*Caenorhabditis elegans*])
- CAA09985 (calcium ATPase [*Caenorhabditis elegans*])
- CAA59762 (calcium transporting ATPase 1 [*Saccharomyces cerevisiae*])
- CAB04015 (C. elegans protein ZK256.1a, confirmed by transcript evidence [*Caenorhabditis elegans*].)
- KIAA0703 (Homo sapiens DNA (Celera Genomics) Homo sapiens genomic, genomic survey sequence)
- P20647 (Calcium-transporting ATPase sarcoplasmic reticulum type, slow twitch skeletal muscle isoform)

And AAC72756, CAA09985, P20647 and AAA96717 relatively similar; AF361357, CAB04015, KIAA0703 and CAA59762 can be attributed to a class; AAD09924, AAP33531 and CAA09303 can attributed to a class.

2.3. Compressed color graphics

Once the sequence length is longer, then converted the images width is also large, is often more than 2000 pixels, in order to facilitate the transmission it performs compression, compression is as follows, find the length, then divide by 8; find the width and divide by 3. If the number can't divide clearly and the number + 1, the equation can be expresses as:

$$W' = \frac{W}{8};\ If\ W\ mod\ 8 = 0 \quad (4) \qquad W' = \frac{W}{8} + 1;\ If\ W\ mod\ 8 \neq 0 \quad (5)$$

$$H' = \frac{H}{3}; If\ H\ mod\ 3 = 0 \quad (6) \qquad H' = \frac{H}{3} + 1;\ If\ H\ mod\ 3 \neq 0 \quad (7)$$

W represents the original image length, W ' on behalf of the length of the compressed image; H on behalf of the width of the original image, H' on behalf of the width of the compressed image, Each eight string is a group, this group of strings from the binary to decimal conversion of numbers, the equation as follows:

$$D'_{(i,j)} = F\left[D_{(i,j*8-7)}, D_{(i,j*8-6)}, D_{(i,j*8-5)}, \ldots, D_{(i,j*8)}\right] \quad (8)$$

For example, there is a string of "01000001", according to the equation converted to 65, so the length of the original image can be reduced to 1 / 8. When the length of the compression is completed and we can also do the width of the compression; according to H ', we can use H ' as the interval in same column, and set out a set of three numbers, this set of figures were set in the new image the same location R plane (red), G plane (green) and B plane (blue), by this compression, the original width can be reduce to 1 / 3, the equation as follows:

$$D'_{(i,j)} = F\left[D_{(i,j)}, D_{(i+H',j)}, D_{(i+2*H',j)}\right] \quad (9)$$

3. Results

3.1. Result of sequence translate into binary sequence

In the back-end system, if you input an access number, the system will determine the sequence is a protein database or a nucleotide database; if it is the protein database, the system will directly capture sequence form the NCBI database for conversion. If it is the nucleotide database, the complete data capture and search after the acquisition for the provision of translation part of the system to use.

When the sequence captured, they perform the conversion process, system will receive the sequence string into a binary sequences. After received sequence, each letter in turn will transfer into the corresponding 5-digit binary number.

3.2. Implement cellular automata image process

When the sequence from the string into a binary number, we can begin to convert the digital sequence of images by the specified rules for cellular automata iterations. According to different rules, converted the

images will not be the same. The image shows the bottom of this group under the AF361357 sequence and converted with different rules to get the images.

Fig. 2: Image of sequence AF361357 by rule 14

Fig. 3: Image of sequence AF361357 by rule 18

Fig. 4: Image of sequence AF361357 by rule 32

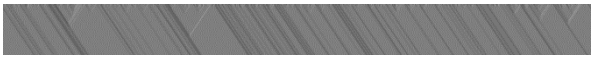

Fig. 5: Image of sequence AF361357 by rule 57

Fig. 6: Image of sequence AF361357 by rule 84

Fig. 7: Image of sequence AF361357 by rule 113

Fig. 8: Image of sequence AF361357 by rule 184

Fig. 9: Image of sequence AF361357 by rule 226

Because the experimental images are a lot, there are only a few selected out of representative comparison. According to the experimental results can be compared with that degree of identification rules were 14, 57, 184 and 226. In this research, the section 184 rules is the most identification of all, so we chose the rules of section 184 after the experiment as a sequence alignment using the basis. Under section 184 rules, we will enter the experimental group and converted to an image sequence. With these images, we can found that they have high similarity.

Fig.10: Image of sequence AF361357 by rule 184

Fig.11: Image of sequence CAB04015 by rule 184

AF361357 and CAB04015 are similar sequences after comparing, Fig.12 is the AF361357 protein fragment sequence image by CDS converted, the major conversion of the amino acid sequence capture contains 303~308 locations in the resulting image. In this fragment, the V-shaped images contain most of the right, and the black lines are more obvious. Both can be found than the latter (Fig.14), Fig.12 of the V-shaped stack of image-intensive, while the loose part of Fig.13, but AF361357303~308 for the HILELL, CAB04015276~281 for the NVVDMF, HILELL and NVVDMF the two are similar in chemical nature from the BLOSUM 62 table and more positive points in that, so the image are generated out of the V-shaped, but in different positions.

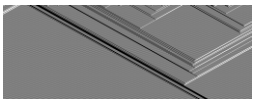

Fig. 12: Fragment image of AF361357

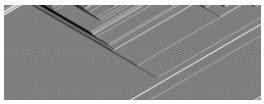

Fig. 13: Fragment image of CAB04015

Fig. 14: AF361357 (up) and CAB04015 (down) different in picture

3.3. Compressed color graphics

Due to the general display of black and white images are usually bigger than the normal images. Therefore, this paper uses the reduced length and width of the compression method, so that the length reduced. Compressed images (Fig.15), it is mainly in the transmission more convenient, but not appropriate under the naked eye.

Fig.15: Image of compression

4. Conclusion

Sequence alignment is a very basic tool of analyzing biological information. In this research, we used graphics conversion to replace traditional sequence alignment such as Dot plot and dynamic programming. We compared the different amino acid sequences which were converted to graphics. By this conversion, we can easily compare the similarity between the more complex sequences. The shortcoming of traditional dynamic programming is that the results can only be known from scores. But the implement system in this research can visually identify which sequences are more similar.

5. Acknowledgment

The authors would like to thank the reviewers for their valuable suggestions and comments that are helpful to improve the content and quality for this paper. This paper is supported by the Taichung Veterans General Hospital / National Chung Hsing University Joint Research Program, under the contract of TCVGH-NCHU1017613 and TCVGH-NCHU1027619 and the National Science Council of Taiwan, ROC, under the contract of NSC 100-2221-E-005-089- and NSC 101-2221-E-005-093-.

6. References

[1] M. D. Adams, J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, R. F. Moreno, and et al., "Complementary DNA sequencing: expressed sequence tags and human genome project," *Science*, vol. 252, no. 5013, pp. 1651-6, Jun 21, 1991.

[2] D. Wu, J. Roberge, D. J. Cork, B. G. Nguyen, and T. Grace, "Computer visualization of long genomic sequences." pp. 308-315.

[3] R. Stevens, C. Goble, P. Baker, and A. Brass, "A classification of tasks in bioinformatics," *Bioinformatics*, vol. 17, no. 2, pp. 180-8, Feb, 2001.

[4] K. W. C. a. J. I. Helfman, "Dotplot: A Program for Exploring Self-Similarity in Millions of Lines of Text and Code," *Journal of Computational and Graphical Statistics*, vol. 2, pp. 153-174, 1993.

[5] S. B. Needleman, and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *J Mol Biol*, vol. 48, no. 3, pp. 443-53, Mar, 1970.

[6] T. F. Smith, and M. S. Waterman, "Identification of common molecular subsequences," *J Mol Biol*, vol. 147, no. 1, pp. 195-7, Mar 25, 1981.

[7] S. F. Altschul, and B. W. Erickson, "Optimal sequence alignment using affine gap costs," *Bull Math Biol*, vol. 48, no. 5-6, pp. 603-16, 1986.

[8] G. S. a. H. Imai, "Finding K-best Alignments of Multiple Sequences," *Proceeding Genome Informatics Workshop IV*, pp. 120-129, 1993.

[9] A. M. Waterhouse, J. B. Procter, D. M. Martin, M. Clamp, and G. J. Barton, "Jalview Version 2--a multiple sequence alignment editor and analysis workbench," *Bioinformatics*, vol. 25, no. 9, pp. 1189-91, May 1, 2009.

[10] D. J. Lipman, S. F. Altschul, and J. D. Kececioglu, "A tool for multiple sequence alignment," *Proc Natl Acad Sci U S A*, vol. 86, no. 12, pp. 4412-5, Jun, 1989.

[11] A. Wirawan, C. K. Kwoh, N. T. Hieu, and B. Schmidt, "CBESW: sequence alignment on the Playstation 3," *BMC Bioinformatics*, vol. 9, pp. 377, 2008.

[12] M. Alston, C. G. Johnson, and G. Robinson, "Colour merging for the visualization of biomolecular sequence data." pp. 169-175.

[13] X. Xiao, S. Shao, Y. Ding, Z. Huang, X. Chen, and K. C. Chou, "Using cellular automata to generate image representation for biological sequences," *Amino Acids*, vol. 28, no. 1, pp. 29-35, Feb, 2005.

[14] E. Hamori, and J. Ruskin, "H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences," *J Biol Chem*, vol. 258, no. 2, pp. 1318-27, Jan 25, 1983.

[15] A. Da'dara, M. H. Tsai, L. F. Tao, K. A. Marx, C. B. Shoemaker, D. A. Harn, and P. J. Skelly, "Schistosoma mansoni: molecular characterization of a tegumental Ca-ATPase (SMA3)," *Exp Parasitol*, vol. 98, no. 4, pp. 215-22, Aug, 2001.