Identification and Retrieval of DNA Genomes Using Binary Image Representations Produced by Cellular Automata

K. Konstantinidis[1], A. Amanatiadis[2], S. A. Chatzichristofis[2], R. Sandaltzopoulos[3] and G. Ch. Sirakoulis[2]
[1]Centre for Research and Technology Hellas, Information Technologies Institute, Greece
[2]Department of Electrical and Computer Engineering, Democritus University of Thrace, Greece
[3]Department of Molecular Biology and Genetics, Democritus University of Thrace, Greece
konkonst@iti.gr, aamanat@ee.duth.gr, schatzic@ee.duth.gr, rmsandal@mbg.duth.gr, gsirak@ee.duth.gr

***Abstract*—We have developed a novel method for the identification and retrieval of DNA sequences which are represented as binary images. This type of representation emanates from the evolution of one-dimensional nucleotide arrays abiding to a set of Cellular Automaton rules. A thorough investigation of these rules was performed in order to determine their efficiency. The presented method has been applied on short nucleotide sequences as well as on eleven complete genes of various origins. The technology presented offers a novel approach for the rapid and efficient sequence identification of nucleotide sequences in database repositories. The proposed framework will be practically useful for applications involved in virus recognition and personalized medicine which rely heavily on the processing of huge volumes of nucleotide sequence data.**

I. Introduction

Advances in sequencing methodologies caused an unprecedented flux of information requiring even more efficient data handling approaches. The content of sequence databases such as GenBank and EMBL, increases at an exponential rate. Gene sequences are stored in the form of long sequences of characters. The mere reading of any long stretch of these sequences is meaningless as their bewildering complexity does not allow the extraction of a key characteristic. However, meaningful features could emerge and become distinguishable if a sequence was to be transformed into some kind of a diagram [1]. Thus, the visualization of nucleotide sequences is a very important issue [2], [3], [4].

Cellular Automata (CA) have been extensively used in the past for modeling biological systems [5], [6]. Following this trend, Xiao et al. [7] presented a method that transforms nucleotide sequences into binary images that result from the evolution of an array through the use of a Cellular Automata Representation Algorithm (CARA). This representation provides an inexpensive and extremely rapid genome visualization which in this work facilitates sequence information retrieval, recognition and comparison. Essentially, the contribution of this paper lies in the use of this representation in a dual retrieval process of DNA sequences. To the best of our knowledge it is the first time that a subset of a genome sequence is being used as a binary image representation to retrieve the original genome. An application of this process could be the identification of a virus when provided with only a small part of its nucleotide sequence. We applied our method on relatively short DNA sequences as well as on eleven full length gene sequences derived from a range of viruses.

We then evaluated the efficiency of all possible CA rules to transform nucleotide sequences into 2D binary images. For this purpose, we subjected all of the binary images derived from the transformation of a variety of nucleotide sequence of diverse length in extensive tests in order to identify those ones that yield the most useful results. Our analysis substantiated the reliability of our approach and illustrated the usefulness of the resulting binary images for the identification of nucleotide sequences a lot faster and easier in respect to other conventional methods.

The rest of the paper is organized as follows. Section II provides a brief description and analysis of the DNA image representation using the CA tool. The DNA image comparison and identification algorithms are presented in Section III. The experimental results are discussed in Section IV. Last, the conclusions are drawn in Section V.

II. DNA image representation using Cellular Automata

CAs were originally proposed by von Neumann [8] and Ulam [9] as a possible idealization of biological systems, with the particular purpose of modeling biological self-reproduction. They are dynamical systems in which space and time are discrete and operate according to local interaction rules [10]. In this section a formal definition of a CA will be presented. More specifically, in this paper, we focus on one-dimensional (1-d) CA of a regular uniform lattice, which may be of N *size* and expands in a space. Each site of this lattice is called *cell* and the corresponding variables of each cell are taking values from a discrete state resulting to the *state* of each cell. As proposed by [10] we consider two possible states per cell, i.e., $S = (0, 1)$. The CA lattice consists of identical cells, $\ldots, i-3, i-2, i-1, i, i+1, i+2, i+3, \ldots$, and the corresponding states of these cells are $S_{i-3}, S_{i-2}, S_{i-1}, S_i, C_{i+1}, S_{i+2}$ and S_{i+3}. The time evolution of CA in discrete time steps is described by the local transition/evolution *rule* f, which is usually a function $f : (0, 1)^n \rightarrow 0, 1$. Consequently, the possible change of CA cell state during time evolution is affected by the states of its neighboring cells and all the involved cells constitute CA cell's *neighborhood*. The neighborhood size n

978-1-4799-5220-5/14/$31.00 ©2014 IEEE

is usually taken to be $n = 2r + 1$ such that:

$$S_i(t+1) = f(S_{i-r}(t), \ldots, S_i(t), \ldots, S_{i+r}(t)) \quad (1)$$

where *r* (positive integer) is a parameter, known as the radius, representing the standard 1-d cellular neighborhood. We shall furthermore limit ourselves to the $r = 1$ case, i.e., so-called elementary CA, for which the neighborhood size is $n = 3$:

$$f : \{0,1\}^3 \rightarrow \{0,1\} \;\; S_i(t+1) = f(S_{i-1}(t), S_i(t), S_{i+1}(t)) \quad (2)$$

The domain of *f* is the set of all 2^3 3-tuples, which gives rise to 2^8 = 256 distinct elementary rules. We will use Wolfram's decimal numbering convention for describing these rules [11], e.g. $f(111) = 1, f(110) = 0, f(101) = 1, f(100) = 1, f(011) = 1, f(010) = 0, f(001) = 0, f(000) = 0$, is denoted rule 184. Having in mind that we are using Boolean variables to express the CA state, each of its cells can be considered to be black or white, respectively. For two-state CA a configuration of a size *N* grid at time t is a binary sequence $C(t)$. For a two-dimensional (2-d) CA, two neighborhoods are often considered, Von Neumann and Moore neighborhood. Von Neumann neighborhood is a diamond shaped neighborhood and can be used to define a set of cells surrounding a given cell (x_0, y_0). Equation 3 defines the Von Neumann neighborhood of range r.

$$N^{\nu}_{(x_0,y_0)} = \{(x,y) : |x - x_0| + |y - y_0| \leq (r)\} \quad (3)$$

For a given cell (x_0, y_0) and range r, Moore neighborhood can be defined by the following equation:

$$N^{M}_{(x_0,y_0)} = \{(x,y) : |x - x_0| \leq (r), |y - y_0| \leq (r)\} \quad (4)$$

The local rule, *f*, in all cases determines the way in which each cell of the 2-d CA is updated. Every cell's state is affected by the cell values in its neighborhood and its value on the previous time step, according to the transition rule or a set of rules.

Molecular biologists identify the nucleotide sequence of the genes in an organism's genome in order to deduce the aminoacid synthesis of the proteins encoded by these genes and the interspecies evolutionary relationships. It is easy to figure that DNA can be modeled as a 1-d CA, where the phosphate chain corresponds to the CA lattice and the deoxyribose sugars to the CA cells as follows: at each sugar molecule one of the four bases adenine (A), cytosine (C), guanine (G) and thymine (T) which is replaced by uracil (U) in RNA may bind. Consequently the cell state of the CA DNA model will now result as *S*=(A,C,T,G). Nevertheless, to enhance the CA performance we choose to represented the four bases by numbers. The most appropriate way is to correspond each one of them either to a respective number of the quaternary number system, which contains only four numbers, i.e. 0, 1, 2 and 3 or to two digit of the binary code, i.e. A=00, C=01, G=10, U=*T*=11. As already mentioned, it is now clear that a vast number of evolution rules can be applied to the CA DNA model [5], [12] More specifically, proteins are polymers of amino acids, hence each protein has a characteristic amino acid sequence. Taking into consideration that there are 20 amino acids and applying the rule of resemblance, the rule of complementarity, the theory of molecular recognition and the theory of information, a group of digital codes is formed for the representation of amino acids [12].

In this paper, we employ Xiao's et al. method for DNA image representation [7]. As a result, we use a CA based approach to transform gene sequences into binary images and evaluate the usefulness of the generated images as a means for information retrieval and genome identification. It is worth noting that CA showed to be a promising model for DNA sequence evolution [5], [6] as well combined with the retrieval process [13]. More specifically, based on the aforementioned 256 Wolfram 1-d CA evolution rules, for any given sequence, different CA rules generate distinct corresponding images which means that it is possible to create 256 different images for each sequence. In correspondence to Wolfram's categorization [10], [11], these images can be indexed into four categories. The first category leads to a homogeneous state, since the cell states quickly transform into uniform patterns where everything is stabilized to 0 or 1. The second category leads to a set of separated simple stable or periodic structures, while the third category leads to chaotic, aperiodic patterns, respectively. Finally, the fourth one is composite and produces persistent, complex patterns of localized structures. The rule of choice for the evolution of the visual representation should generate the most distinguishable characteristics so that it may be possible to distinguish gene sequences with a substantial degree of similarity. In this case, the bases in a nucleotide sequence behave as a single entity that defines the image to be generated.

III. Comparison and Identification Algorithm

Once a nucleotide sequence has been transformed into a binary image, the image is stored in a database. Since there are 3 coding alternatives and 256 available rules there is a dire need to index this database in order to be able to retrieve any matching sequence when a similar sequence is used as a query in a sequence search. Therefore, we have developed a retrieval system which searches the database and identifies the best matching images using a similarity ranking system.

A. Viral Genomes Identification based on image comparison

In case the query sequence is an entire viral genome, the retrieval algorithm is fairly simple. It compares the query image, i.e. the binary image derived from the transformation of the query sequence, to all the images in the database using the Sum of Absolute Differences (SAD) method (Eq. 2) and then ranks the SADs in increasing order. The image with the smallest SAD is the most similar one in respect to the query image.

$$SAD = \sum |Q(x,y) - D(x,y)| \quad (5)$$

where Q is the query image, D is the image from the database and x, y are the coordinates of each pixel.

Following a large number of tests, it was concluded that when the query image was a representation of the complete viral genome, the retrieval success was guaranteed, that is the queried virus came up always first in the ranking. However, as incomplete genome sequences are commonly used as queries, we modulated the retrieval process so that it may be applied to partial genome sequences as well.

B. Virus Identification using partial genome sequences

In case a part of a viral genome is presented as a query (top, middle or bottom part of the genome), the retrieval system algorithm is modified as follows: The part of the genome (query image) is cross-correlated with all the virus images in the database (Eq. 3). At the end of each cross-correlation, the greatest peak resulting for each image is retained and the final ranking for the full query is performed using this peak list.

$$C(x,y) = \sum_{m=0}^{Ma-1} \sum_{n=0}^{Na-1} Q(m,n) \cdot conj(D(m+x, n+y)) \quad (6)$$

where Q is the query image with dimensions (Ma, Na), D is the image from the database with dimensions (Mb, Nb), $0 \leq x < Ma + Mb - 1$ and $0 \leq y < Na + Nb - 1$. In contrast to the infallible retrieval results using the full genome image, in this case retrieval success may be compromised as a result of differences caused by the border conditions or even from the position of the genome part in the whole genome image. As expected, the effectiveness of the method is inversely proportional to the size of the query sequence. Extended examples and typical results are presented in the following section.

IV. Experimental Results

In order to test the validity and effectiveness of the employed CARA-DNA algorithm, the genome sequences of the following 11 viruses were selected at random from the Genbank database: The African horsesickness virus segment 7, the Bluetongue virus segment 4, the Boolarra virus RNA 2, the Burkholderia pseudomallei phage phi52237, the Carrot mottle mimic virus, the Chikungunya virus, the Citrus yellow mosaic virus (CMBV), the Ovine lentivirus, the Papaya leaf curl China virus, the Thogoto virus segment 4 and the Tomato yellow spot virus.

The images in Figure 1 resulted from the CARA-DNA algorithm using rule 184 and base coding 0 (i.e. A=00, G=01, C=10, U or T=11). It is obvious that there are distinguishable alterations in the entire length of the image and that the patterns

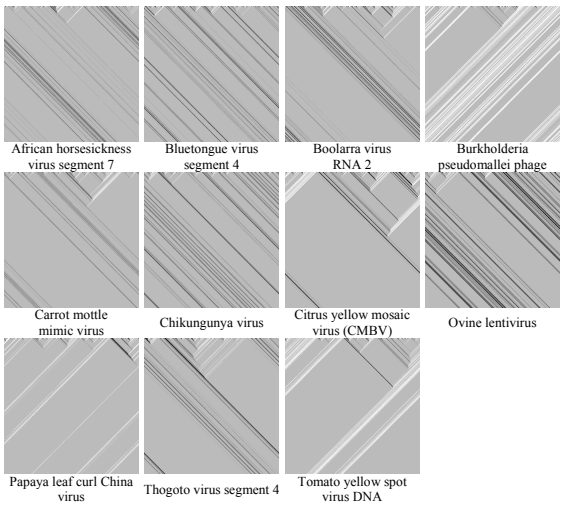

Fig. 1. DNA genome images of 11 viruses using the CARA-DNA algorithm. The rule 184 and coding 0 were applied.

in the image are not repeated thus rendering each image much more diverse than the next. Similar alterations are also present when base coding 1 and 2 are used.

The mere representation of the genes is not our main aim. The ultimate goal is the ability to insert a query image into a database of genome images and to retrieve the virus represented by that image. The retrieval experiments were performed using the viruses in Figure 1 on a database which was created using the representation of the viruses for all the available CA evolution rules and for all possible base codings. In order to find the most effective rule, the brute force trial and error method was employed. Identification tests were performed using all of the aforementioned 11 viruses for all three base codings. Moreover, given that our method should be able to correctly identify an organism even when only a small part of its genome sequence is available, small parts from the top, middle and bottom section of the sequences of each virus were used to perform a series of identifications thus leading to the most effective rule. Specifically, mini sequences ranging from 10 to 100 samples (with a step of 10) were extracted from three different parts of the sequence of each virus so as to create mini query images in order to strenuously test the employed algorithm. Hence, 23,040 mini databases (3 base codings $\times$ 256 rules $\times$ 10 sequence sizes $\times$ 3 different sequence areas) were created in order to test every aspect of the representation algorithm. As expected, the effectiveness of the method is inversely proportional to the size of the query sequence regardless to the position from which the part of the sequence was extracted as illustrated in Figure 2.

There are actually three aspects to be considered in respect to the effectiveness: the most effective base coding, the most effective rule and the most effective combination of base coding and rule. As follows, researching all three will provide the answer concerning the selection of parameters when implementing a virus identification system. Figure 3 presents the effectiveness of all three base codings through a histogram

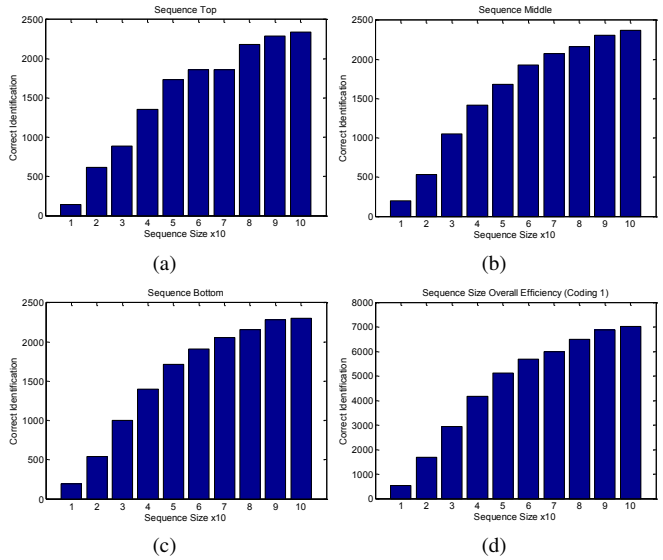

Fig. 2. Effectiveness of the representation method for all ten different sequence sizes from a) the beginning of the sequence, b) the middle and c) the end. Graph d) depicts the overall efficiency of the representation method using all possible rules and base coding 1.

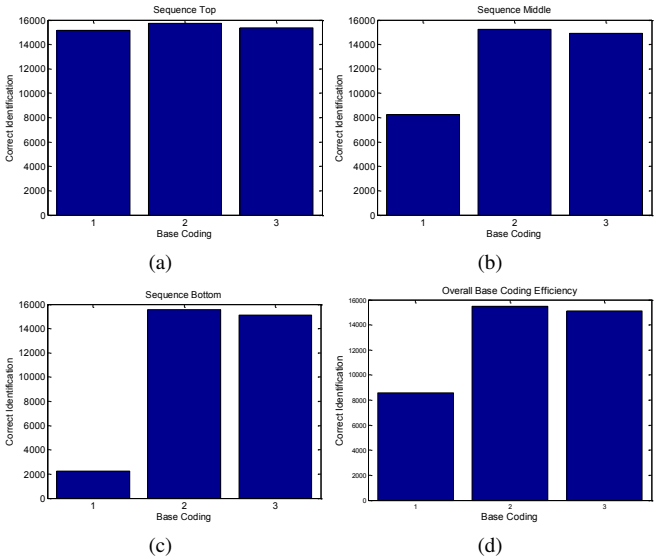

Fig. 3. Effectiveness comparison between base codings from the sequence a) top, b) middle and c) bottom. The overall sum for the three codings using every rule, for every mini sequence and for every virus is presented in d).

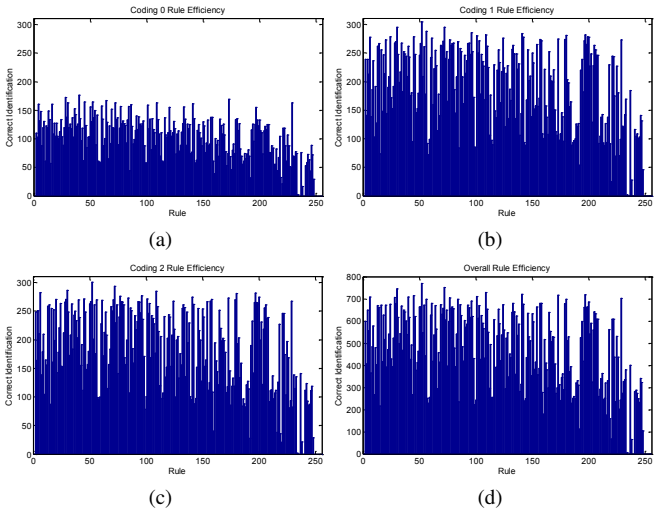

Fig. 4. Efficiency of all the rules for base codings; a) coding 0, b) coding 1, c) coding 2, d) the aggregate efficiency of each rule for all codings.

representing the sum of all correct identifications made for all possible combinations of rules, sequence lengths and viruses. It is easy to conclude that the most effective one is coding 1 (i.e. A=0, G=1, C=1, U or T=0). The first conclusion to be reached is that since not all coding can be used when a realistic virus identification system is to be utilized, coding 1 is the safest option.

The second aspect in respect to efficiency is that of the best performing rule. Figure 4 presents the efficiency of all the rules for all three base codings, as well as the aggregate efficiency of each rule pointing out rule 52 as the most efficient one. Moreover, Figure 4b provides the answer as to which combination is the most efficient and verifies the aggregate efficiency of base coding 1 as well as rule 52. Thus, it would be safe to say that this combination would present a fair result if a virus identification system were to be built.

V. Conclusions

A new method is presented in this paper for the identification and retrieval of gene sequences which relies on the binary image representation constructed via CAs. In particular, the aim is to use sections of DNA and of entire genes to construct query images in order to retrieve and identify the original whole genome in a fast and simple manner. The method was applied on eleven complete genomes from a range of viruses including the African horsesickness and the Bluetongue virus. The decision for the best CA rule was made through extensive efficiency tests using all of the binary genome images for all base codings and for a variety of sequence sizes. The cross correlation between the query image and the image dataset provides the means of identification. Future work in the identification part of the system could be made by extracting a descriptor from the images (e.g. moments). Such higher order features could increase the efficiency of the system rendering it faster. Moreover, it would be interesting to investigate the use of pseudo-color in the system in contrast to the binary black and white rules. Color would raise the dimensionality of the descriptor and thus the complexity of the algorithm but would also provide additional information which would prove useful in the genome representation process. The proposed method could also be used in the representation of the genome of more complex species.

References

[1] K.-C. Chou and C.-T. Zhang, "Predicting protein folding types by distance functions that make allowances for amino acid interactions." *Journal of Biological Chemistry*, vol. 269, no. 35, pp. 22 014–22 020, 1994.

[2] Z. Hu, M. Frith, T. Niu, and Z. Weng, "Seqvista: a graphical tool for sequence feature visualization and comparison," *BMC Bioinformatics*, vol. 4, no. 1, p. 1, 2003.

[3] Y. Liu, X. Guo, J. Xu, L. Pan, and S. Wang, "Some notes on 2-d graphical representation of DNA sequence," *Journal of chemical information and computer sciences*, vol. 42, no. 3, pp. 529–533, 2002.

[4] A. Mylläri, T. Salakoski, and A. Pasechnik, "On the visualization of the DNA sequence and its nucleotide content," *SIGSAM Bull.*, vol. 39, no. 4, pp. 131–135, Dec. 2005.

[5] G. C. Sirakoulis, I. Karafyllidis, C. Mizas, V. Mardiris, A. Thanailakis, and P. Tsalides, "A cellular automaton model for the study of DNA sequence evolution," *Computers in biology and medicine*, vol. 33, no. 5, pp. 439–453, 2003.

[6] C. Mizas, G. C. Sirakoulis, V. Mardiris, I. Karafyllidis, N. Glykos, and R. Sandaltzopoulos, "Reconstruction of DNA sequences using genetic algorithms and cellular automata: Towards mutation prediction?" *Biosystems*, vol. 92, no. 1, pp. 61–68, 2008.

[7] 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, 2005.

[8] J. V. Neumann, *Theory of Self-Reproducing Automata*, A. W. Burks, Ed. Champaign, IL, USA: University of Illinois Press, 1966.

[9] S. Ulam, "Random processes and transformations," in *International Congress of Mathematicans*, vol. 2, 1952, pp. 264–275.

[10] S. Wolfram, *Theory and Applications of Cellular Automata*. Singapore: World Scientific, 1986.

[11] ——, "Cellular automata as models of complexity," *Nature*, vol. 311, pp. 419–424, 1984.

[12] G. Sirakoulis, I. Karafyllidis, R. Sandaltzopoulos, P. Tsalides, and A. Thanailakis, "An algorithm for the study of DNA sequence evolution based on the genetic code," *Biosystems*, vol. 77, no. 13, pp. 11 – 23, 2004.

[13] K. Konstantinidis, G. C. Sirakoulis, and I. Andreadis, "Content-based image retrieval using cellular automata," in *Proceedings of 5th International Conference on Technology and Automation (ICTA05)*, 2005, pp. 371–375.