2014

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

Konstantinidis, K.

IEEE (Computer Society)

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