AIS-PRMACA: Artificial Immune System based Multiple Attractor Cellular Automata for Strengthening PRMACA, Promoter Region Identification

Pokkuluri Kiran Sree*, Inampudi Ramesh Babu** & S.S.S.N. Usha Devi Nedunuri***

*Research Scholar, Department of Computer Science & Engineering, Jawaharlal Nehru Technological University Hyderabad, INDIA. E-Mail: profkiransee@gmail.com
**Professor, Department of Computer Science & Engineering, Acharya Nagarjuna University, INDIA.
***Assistant Professor, Department of Computer Science & Engineering, Jawaharlal Nehru Technological University Kakinada, INDIA.

Abstract—Identifying promoter region is an important problem in bioinformatics. Transcription of particular gene of DNA is initiated by a promoter. So many techniques are employed in identifying these promoter regions but all of them will attract an average accuracy of 84%. A new algorithm based on artificial immune system AIS-PRMACA was proposed to strengthen and improve the accuracy of promoter prediction system PRMACA. The proposed classifier is tested with BDGP and ENCODE data sets has produced an accuracy of 90.5% which is a considerable improvement when compared with the accuracy of conventional methods available.

Keywords—Artificial Immune System; Cellular Automata; MACA; Promoter Region.

Abbreviations—Cellular Automata (CA); Dependency Vector (DV); Dependency String (DS); Genetic Algorithm (GA); Multiple Attractor Cellular Automata (MACA).

I. INTRODUCTION

Promoters are responsible for most important biochemical functions such as nutrient transportation, making muscle fibers and important bio chemical functions which occupy macro regions in human body. Specifically, the Promoters are chains of amino acids and DNA sequences, of which there are 20 different types, coupled by peptide bonds [Miskimins et al., 1985]. The structural hierarchy possessed by Promoters is typically referred to as primary and tertiary region. Promoter Region Predication from sequences of amino acid gives tremendous value to biological community. This is because the higher-level and secondary level [P. Kiran Sree et al., 2013], [Miskimins et al., 1985] regions determine the function of the Promoters and consequently, the insight into its function can be inferred from that.

II. RELATED WORKS IN PROMOTER REGION IDENTIFICATION

Nearest neighbor techniques attempt to predict the region of a central residue, within a segment of amino acids [Kiran Sree & Dr. Inampudi Ramesh Babu, 2013; 2013A], based on the known regions of homologous segments. Steen Knudsen al [Bauer et al., 1988] has used statistical classifiers to identify promoter regions. Techniques for region identification include, but are not limited to, constraint programming methods, statistical approaches to predict the probability of an amino acid being in one of the structural elements, and Bayesian network models. We also proposed an algorithm to identify the promoter regions using CA with text clustering with a lesser accuracy. We have also proposed an algorithm [Pokkuluri Kiran Sree et al., 2013] with true MACA with considerable accuracy [Abagyan et al., 1997; Maji & Chaudhuri, 2004].

III. cellular automata

Cellular Automata (CA) is a simple model of a spatially extended decentralized system, made up of a number of individual components (cells). The communication among constituent cells is limited to local interaction. Each individual cell is in a specific state that changes over time depending on the states of its neighbors. From the days of Von Neumann who first proposed the model of Cellular Automata (CA) [Reese, 2001; Debasis Mitra & Michael Smith, 2004], to Wolfram’s recent book ‘A New Kind of
Science’, the simple and local neighborhood region of CA has attracted researchers from diverse disciplines. It has been subjected to rigorous mathematical and physical analysis [Kiran Sree & Dr. Inampudi Ramesh Babu, 2008; 2010] for past fifty years and its application has been proposed in different branches of science - both social and physical.

Definition: CA is defined a four tuple \( \langle G, Z, N, F \rangle \)

Where \( G \rightarrow \) Grid (Set of cells)
\( Z \rightarrow \) Set of possible cell states
\( N \rightarrow \) Set which describe cells neighborhoods
\( F \rightarrow \) Transition Function (Rules of automata)

The evolution process is directed by the popular Genetic Algorithm (GA) with the underlying philosophy of survival of the fittest gene. This GA framework can be adopted to arrive at the desired CA rule region appropriate to model a physical system. The goals of GA formulation are to enhance the understanding of the ways CA performs computations and to learn how CA may be evolved to perform a specific computational task and to understand how evolution creates complex global behavior in a locally interconnected system of simple cells.

IV. ARTIFICIAL IMMUNE SYSTEM (AIS) LEARNING

AI learning involves the task of learning from a series of examples. The end objective is to deal with any general types of data, including cases where the number and type of attributes may vary, and where additional layers of learning are superimposed, with hierarchical structure of attributes and classes. AIS learning aims to generate classifying expressions simple enough to be understood easily by human. It must mimic human reasoning sufficiently to provide insight into the AIS process. Background knowledge may be exploited in AIS learning development, but operation is assumed without human intervention.

4.1. Artificial Immune System (AIS) Tree

AIS tree performs multistage hierarchical decision making. Pattern recognition based on AIS trees was motivated by the need to interpret images from remote sensing satellites. AIS trees in particular and induction methods in general, are associated with AIS learning to avoid the knowledge acquisition bottleneck for expert systems. An overview of work on AIS trees in pattern recognition can be found in.

4.2. AIS-PRMACA Tree Building

Input: Training set \( S = \{S_1, S_2, \ldots, S_K\} \)
Output: AIS-PRMACA Tree.

Partition\( (S, K) \)

Step 1: Generate a AIS-PRMACA with \( k \) number of attractor basins.
Step 2: Distribute training set \( S \) into \( k \) attractor basins (nodes).
Step 3: Evaluate the distribution in each attractor basin.

Step 4: If all the examples \( (S') \) of an attractor basin (node) belong to a single class, then label the attractor basin.
Step 5: If examples \( (S') \) of an attractor basin belong to \( K' \) number of AIS-MACA classes, then, Partition \( (S', K') \).
Step 6: Stop.

A high level comparative perspective on the classification literature in pattern recognition and artificial intelligence can be found in. Tree induction from a statistical perspective, is reviewed in. A majority of work on AIS trees [Kiran Sree & Dr. Inampudi Ramesh Babu, 2009] in AIS learning is an offshoot of Breiman’s work and Quinlan’s algorithm. Moret provided an overview of the work on representing boolean functions as AIS trees and diagrams and its application in pattern recognition. Safavin and Landgrebe surveyed the literature on AIS tree classifiers, almost entirely from a pattern recognition perspective.

The model is built describing a predefined set of data classes. A sample set from the database, each member belonging to one of the predefined classes, is used to train the model. The training phase is also termed as supervised learning of the classifier. Each member may have multiple features/attributes. The classifier is trained based on a specific feature/metric. Subsequent to training, the model performs the task of prediction in the testing phase. Prediction of the class of an input sample is done based on some metric, typically distance metric.

4.3. Random Generation of Initial Population

To form the initial population, it must be ensured that each solution randomly generated is a combination of an \( n \)-bit DS with \( 2m \) number of attractor basins (Classifier #1) and an \( m \)-bit DV (Classifier #2). The chromosomes are randomly synthesized according to the following steps.

1. Randomly partition \( n \) into \( m \) number of integers such that
   \( n_1 + n_2 + \cdots + nm = n \).
2. For each \( ni \), randomly generate a valid Dependency Vector (DV).
3. Synthesize Dependency String (DS) through concatenation of \( m \) number of DVs for Classifier #1.
4. Randomly synthesize an \( m \)-bit Dependency Vector (DV) for Classifier #2.
5. Synthesize a chromosome through concatenation of Classifier #1 and Classifier #2.

4.4. Mutation Algorithm

The mutation algorithm emulates the normal mutation scheme. It makes some minimal change in the existing chromosome of PP (Present Population) to form a new chromosome for NP (Next Population). The chromosome in the current GA formulation is mutated at a single point. Figure 1 shows a simple mutation example.
AIS-PRMACA Output

31958321; length 12001
Length of sequence: 12001
4 promoter/enhancer(s) are predicted
Promoter Pos: 6078 LDF: +1.629 TATA box at
Promoter Pos: 1363 LDF: +5.235 TATA box at
Promoter Pos: 7068 LDF: +1.165 TATA box at
Promoter Pos: 9650 LDF: +1.051 TATA box at

Table 1: Predictive Accuracy of Promoter Prediction

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>DNA Sequence</th>
<th>Amino Acid Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesian</td>
<td>51%</td>
<td>46%</td>
</tr>
<tr>
<td>Normal ID-CA</td>
<td>68%</td>
<td>72%</td>
</tr>
<tr>
<td>Neural Networks</td>
<td>74%</td>
<td>68%</td>
</tr>
<tr>
<td>PRMACA</td>
<td>79%</td>
<td>72.5%</td>
</tr>
<tr>
<td>AIS-PRMACA</td>
<td>89.6%</td>
<td>91%</td>
</tr>
</tbody>
</table>

VI. CONCLUSION

AIS-PRMACA was trained to identify promoter regions from sequences of DNA as well as amino acid. Through analysis and extensive testing/training are conducted for improving the proposed classifier accuracy over several data sets. The experiments results over ENCODE and BDGP datasets indicate an average accuracy of the classifier is 90.5%. The technique of artificial immune system was employed earlier on strengthening protein coding regions and protein structure prediction also, seems promising to improve the classifier accuracy in many other problems related to bioinformatics.

REFERENCES


V. EXPERIMENTAL RESULTS

Experiments are conducted on EDCODE and BDGP data sets. The output and result (Table 1) were represented below.


**Prof P. Kiran Sree** is working as Professor in department of CSE in BVC Engineering College. He has published forty technical papers in international journals and conferences. His areas of interests include parallel algorithms, artificial intelligence, and compiler design and computer networks. He was the reviewer for many IEEE Society conferences and Journals in artificial intelligence and networks. His bibliography is listed in Marquis Who’s Who in the World, 29th Edition (2012), USA.

**Dr Inampudi Ramesh Babu** is working as a Professor in the Department of Computer Science & Engineering, Acharya Nagarjuna University. Also, he has been an Academic Senate member of the same university since 2006. He held many positions in the Acharya Nagarjuna University as Head, Director of the Computer Centre, Chairman of the PG Board of studies, Member of the Executive Council, Special Officer, Additional Convener of ICET Examinations and Convener of MCA Admissions. He is currently supervising ten PhD students who are working in different areas of image processing and artificial intelligence. He has published 100 papers in international journals and conferences.

**Smt S. S. S. N. Usha Devi** is working as Assistant Professor in department of CSE in University College of Engineering, JNTU Kakinada. She has published five papers in international journals and four in international conferences. She is the member of IAENG.