A NOVEL MACA BASED CLONAL CLASSIFIER FOR PROTEIN CODING AND PROMOTER REGION PREDICTION: AN INTEGRATED APPROACH

Pokkuluri Kiran Sree 1, Inampudi Ramesh Babu 2, SSSN Usha Devi 3

1. Research Scholar, Dept of CSE, JNTU Hyderabad, profkiransree@gmail.com
2. Professor, Dept of CSE, Acharya Nagarjuna University, Guntur.
3. Assistant Professor, Dept of CSE, JNTU Kakinada.

Abstract

Most of the problems in bioinformatics are now the challenges in computing. This paper aims at building a classifier based on Multiple Attractor Cellular Automata (MACA) which uses fuzzy logic. It is strengthened with an artificial immune system technique, Clonal algorithm for identifying a protein coding and promoter region in a given DNA sequence. To obtain good fitness rules the basic concept of Clonal selection algorithm was used. The proposed classifier can handle DNA sequences of lengths 54, 108, 162, 252, 354. This classifier gives the exact boundaries of both protein and promoter regions with an average accuracy of 89.6%. This classifier was tested with 97,000 data sets which were taken from Fickett and Toung, E.Coli, and other sequences from a renowned medical university. This proposed classifier can handle huge data and can find protein and promoter regions even in mixed and overlapped DNA sequences. This paper introduces a novel concept to combine CA with artificial immune system to produce a better classifier which can address major problems in bioinformatics. This will be the first integrated algorithm which can predict both promoter and protein coding regions. This work also aims at identifying the logicality between the major problems in bioinformatics and tries to obtaining a common frame work for addressing major problems in bioinformatics like protein structure prediction, RNA structure prediction, predicting the splicing pattern of any primary transcript and analysis of information content in DNA, RNA, protein sequences and structure. This work will attract more researchers towards application of CA as a potential pattern classifier to many important problems in bioinformatics.
KeyWords:

MACA(Multiple Attractor Cellular Automata), CA(Celluar Automata), AIS(Artificial Immune System), Clonal Algorithm.

I. Introduction

An essential venture in genomic annotation is to distinguish protein coding areas of genomic successions, which is a testing issue particularly in the investigation of eukaryote genomes. In a eukaryote genome, protein coding areas (exons) are typically not nonstop, yet are flanked by noncoding areas (introns). Because of the absence of self-evident arrangement emphasizes between exons and introns, adequately recognizing protein coding areas from non coding areas is a testing issue in bioinformatics. Identifying a protein coding regions plays a vital role in understanding the genes. If we identify the protein coding regions and analyze we can extract lot of information like reason for a chronic disease, how once cell is going to control another cell.

As the Human Genome Project enters its extensive scale sequencing stage, routines for the ID of genes and administrative components in silico has gotten to be greatly significant. In the previous decade, numerous solid in silico routines have been created and utilized effectively to recognize promoter regions, yet lamentably in any case, solid routines for recognizing either the 5'- or 3'- finishes of gene transcript units is as of now the need of the hour. The failure to recognize the 5'- and 3'- finishes of genes utilizing computational systems intensely constrains our capacity to differentiate one gene from an alternate when examining multi-gene parts. Despite the fact that the 3'-finishes of numerous transcripts may be distinguished via looking the EST (Expressed Sequence Tags), where 3'-closure arrangements are enhanced, no fast and precise 5'-close gene sequencing system is yet accessible. Obviously an efficient change of computational promoter distinction systems is really in necessity. If we identify the promoter regions we can extract information regarding cancer causing symptoms and information about endocrine diseases (P.Maji, 2004).
II. Literature Survey

Changchuan Yin et al have proposed a system (Eric E. Snyder, 2002) to foresee protein coding locales dependent upon the way that a large portion of exon successions have a 3-base periodicity, while intron groupings don’t have this remarkable characteristic. The system processes the 3-base periodicity and the foundation commotion of the stepwise DNA (Horwitz, 1986) sections of the target DNA arrangements utilizing nucleotide conveyances as a part of the three codon positions of the DNA groupings. Exon and intron arrangements might be recognized from patterns of the degree of the 3-base periodicity to the foundation commotion in the DNA successions. Stromo at al have proposed a neural network with dynamic programming to predict the protein coding regions. Datta et al have proposed DFT (Datt, 2005) based algorithm for predicting the protein coding regions.

Marshall S.Z. Horwitz et al have chosen an assembly (E E Snyder, 1993) of Escherichia coli promoters from irregular DNA groupings by reinstating 19 base sets at the -35 promoter area of the tetracycline safety gene te" of the plasmid pbr322. Substitution of 19 base sets with synthetically blended irregular groupings brings about a most extreme of 419 (in the ballpark of 3 x 1011) conceivable displacement arrangements. From a populace of something like 1000 microscopic organisms harboring plasmids with these arbitrary substitutions, tetracycline choice has uncovered a few useful -35 promoter groupings. These promoters have held just fractional. Homology to the -35 promoter agreement arrangement. In three of these promoters, the accord executor moves 10 nucleotides downstream, permitting the RNA polymerase to distinguish an alternate Pribnow box from inside the definitive pbr 322 succession. Two of the groupings advertise translation more unequivocally than the local promoter. This procedure may have requisition for the choice of extra DNA successions with shifted natural movement. Jih-Wei Huang has proposed (Jih-Wei, 2003) a general algorithm to predict the promoter region.

III. Design of MACA – Clonal Based Classifier

MACA is a special class in cellular automata which can use fuzzy logic to process the real values in order to report the boundaries of corresponding regions. The following diagram gives the design of the entire system. The input to the entire system is a DNA (Krisna Kumar, 2002) sequence and this input is processed in the terms of three. So we need a transformation rule which is applied in the form of a matrix. Then use the algorithm one to dissipate the given sequence into number of attractor basins. Algorithm two will be used to know the class of the corresponding sequence. Algorithm three is an integrated algorithm based on which the protein and promoter regions are predicted. Fig 2 shows the complexity of MACA.

Algorithm 1 is used for creating of MACA tree (P. Maji, 2004), (Yin Changchuan, 2007). This tree will dissipate the DNA sequence into respective leaves of the tree. If the sequence is falling into two or more class labels, the algorithm will recursively partition in such a way that
all the sequences will fit into one of the leaves. Every leaf will have a class. Algorithm 2 will be used for getting the class as well as the required transition function. The best fitness rules with a score more than .5 is considered. Algorithm 3 uses, algorithm 1, 2 for predicting the protein and promoter coding regions. Fig 2 shows the processing of the DNA sequences with respect to a complicated structure (Kiran Sree Pokkuluri, 2013).

**Algorithm 1:**

Input: Training Set S= \{S1, S2, ............., Sx\} with P classes

Output: AIS-MACA tree

Partition(S, P)

1. Generate a AIS-MACA with x attractor basins
2. Distribute S in x attractor basins (Nodes)
3. Evaluate the patterns distributed in each attractor basin.
4. If all the patterns say S’ which are covered by the attractor basin belong to only one class, then label the attractor basin (Leaf Node) as the class
5. If S’ of an attractor basin belong to more than one class partition (S’, P’)
6. Stop

**Algorithm 2:**

Input: Training set S = \{S1, S2, · · · , SK\}, Maximum Generation (Gmax).

Output: Dependency matrix T, F, and class information.

begin

Step 1: Generate 100 new chromosomes for IP.
Step 2: Initialize generation counter GC=zero; PP← IP.
Step 3: Compute fitness Ft for each chromosome of PP according to Equation
Step 4: Store T, F, and corresponding class information for which the fitness value Ft>0.5
Step 5: If number of chromosomes with fitness more than 0.5 are 50 then go to 12.
Step 6: Rank chromosomes in order of fitness.
Step 7: Increment generation counter (GC).
Step 8: If GC > Gmax then go to Step 11.
Step 9: Form NP by selection, cloning and mutation.
Step 10: PP← NP; go to Step 3.
Step 11: Store and output T, F, and corresponding class information for which the fitness value is maximum.
Step 12: Stop.
Algorithm 3:

1. Uses the AIS-MACA Tree construction Algorithm 1
2. Uses the AIS-MACA Evolution Algorithm 2
3. Trace the corresponding attractor
4. Travel back from attractor to the starting node
5. Identify the start codon
6. Identify the stop codon
7. Report the boundaries of protein coding region.
8. From first codon of the sequence to start codon Search for TAATAA.
9. Report the promoter boundary located at upstream
Experiments were conducted by using Fickett and Toung data sets (P. Flocchini, 2000). The figure 3, 4, 5 shows the developed interface. Table 1 reports the time taken for processing and giving output. Table 2 show the number of data sets for handled (Pokkuluri Kiran Sree, 2013)
A Novel Integrated MACA based Clonal Classifier for Protein Coding and Promoter Region Prediction

Fig 3. Interface

Fig 4. Training Interface
The time taken for predicting the protein and promoter region is considerable when trained and tested with 50,000+ data sets as reported in Table 1.

Table 1: Accuracy Time Reporting

<table>
<thead>
<tr>
<th>Size of Data Set</th>
<th>Prediction Time of Integrated Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>5000</td>
<td>1064</td>
</tr>
<tr>
<td>6000</td>
<td>1389</td>
</tr>
<tr>
<td>10000</td>
<td>2002</td>
</tr>
<tr>
<td>20000</td>
<td>2545</td>
</tr>
</tbody>
</table>

The time taken for predicting the protein and promoter region is considerable when trained and tested with 50,000+ data sets as reported in Table 1.
The figure 6 shows that the integrated algorithm gives the best output when tested with 354 length DNA sequence. Figure 7 also shows that promoter prediction is considerable in variable lengths (Pokkuluri Kiran Sree, 2013).

Table 2: Number of Data Sets Used

<table>
<thead>
<tr>
<th>Data Set DNA Sequences</th>
<th>54 Human</th>
<th>108 Human</th>
<th>162 Human</th>
<th>252 Human</th>
<th>354 Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Set Coding</td>
<td>21,203</td>
<td>7,452</td>
<td>3,520</td>
<td>2,201</td>
<td>1,003</td>
</tr>
<tr>
<td>Training Set Non Coding</td>
<td>22,563</td>
<td>35,256</td>
<td>2,560</td>
<td>2,056</td>
<td>506</td>
</tr>
<tr>
<td>Testing Set Coding</td>
<td>22,569</td>
<td>21,023</td>
<td>16,564</td>
<td>3,002</td>
<td>689</td>
</tr>
<tr>
<td>Testing Set Non Coding</td>
<td>32,562</td>
<td>28,568</td>
<td>12,056</td>
<td>2,006</td>
<td>700</td>
</tr>
</tbody>
</table>

Fig 6: Predictive Accuracy for Protein Coding Regions
Fig 6,7 shows the predictive accuracy of the promoter region and protein region prediction. This is the first algorithm to integrate these two logical problems (Pokkuluri Kiran Sree, 2013).

IV. Conclusion

We have successfully developed a logical classifier designed with MACA and strengthened with AIS technique. The accuracy of the classifier is considerably more when compared with the existing algorithms which are 84% for protein coding and 90% for promoter region prediction. The proposed classifier can handle large data sets and sequences of various lengths. This is the first algorithm to process DNA sequences of length 252,354. This novel classifier framework can be used to address many problems in bioinformatics like RNA structure prediction, DNA sequencing and many more.

V. References


**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. N** 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.