

AIS-INMACA: A Novel Integrated MACA based Clonal Classifier for Protein Coding and Promoter Region Prediction

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 (AIS), Clonal algorithm for identifying a protein coding and promoter region in a given DNA sequence. The proposed classifier is named as AIS-INMACA 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. 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 & Toung , E.Coli, and other sequences from a renowned medical university. This proposed classifier can handle huge data sets and can find protein and promoter regions even in mixed and overlapped DNA sequences. This work also aims at identifying the logic between the major problems in bioinformatics and tries to obtaining a common framework 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.

Key Words: MACA(Multiple Attractor Cellular Automata) , CA(Cellular Automata) ,AIS(Artificial Immune System) ,Clonal Algorithm, AIS-INMACA(Artificial Immune System-Integrated Multiple Attractor Cellular Automata).

1. Introduction

1.1 Cellular Automata

Pattern classification encompasses development of a model which will be trained to solve a given problem with the help of some examples; each of them will be characterized by a number of features. The development of such a system is characterized as pattern classification. We use a class of Cellular Automata (CA)[18],[19] to develop the proposed classifier. Cellular automata consist of a grid of cells with a finite number of states. Cellular Automata (CA) is a computing model which provides a good platform for performing complex computations with the available local information.

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)

As Cellular Automata consists of a number of cells structured in the form of a grid. The transitions between the cells may depend on its own state and the states of its neighboring cells. The equation one states that if i^{th} cell have to make a transition, it has to depend on own state, left neighbor and right neighbor also.

$$q_i(t + 1) = f(q_{i-1}(t), q_i(t), q_{i+1}(t)) \quad \text{---- Equation 1.1}$$

1.2 Problems in Bioinformatics

Bioinformatics can be characterized as a collection of statistical, mathematical and computational methods for dissecting biological sequences like DNA, RNA and amino acid. It deals with the design and development of computer based technology that supports biological processing. Bioinformatics tools are aimed at performing lot of functions like data integration, collection, analysis, mining, management, simulation, visualization and statistics.

The central dogma shown in the Fig: 1.1 of molecular biology was initially articulated by Francis Crick [20] in 1958. It deals with point by point transfer of important sequential data. It states that data can't be exchanged back from protein to either protein or nucleic acid. So once data gets into protein, it can't stream again to nucleic acid. This dogma is a framework for comprehension about the exchange of sequence data between sequential data carrying biopolymers in living organisms. There are 3 significant classes of such biopolymers ie DNA, RNA and protein.

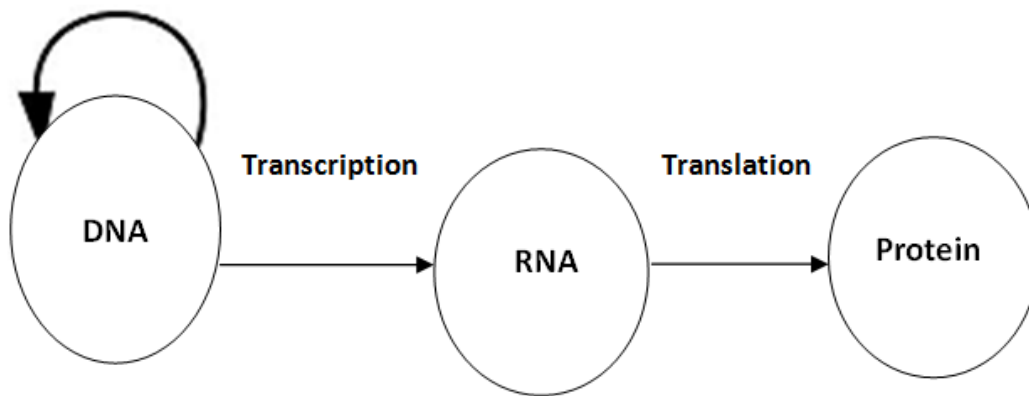


Figure 1.1 Central Dogma

There are 9 possible immediate exchanges of data that can happen between these. This dogma classifies these nine exchanges into three transfers (Normal, Special, and Never Happens). The normal flow of biological information is DNA is copied to DNA (DNA replication), DNA information is copied into RNA (Transcription) and protein are synthesized by using the protein coding region exists in DNA/RNA (Translation).

1.3 Protein Coding Region Identification

DNA is an important component of a cell and genes will be found in specific portion of DNA which will contain the information as explicit sequences of bases (A, G, C, T). These explicit sequences of nucleotides will have instructions to build the proteins. But the region which will have the instructions which is called as protein coding regions occupies very less space in a DNA sequence. The identification of protein coding regions plays a vital role in understanding the gens. We can extract lot of information like what is the disease causing gene, whether it is inherited from father or mother, how one cell is going to control another cell.

1.4 Promoter Region Identification

DNA is a very important component in a cell, which is located in the nucleus. DNA contains lot of information. For DNA sequence to transcript and form RNA which copies the

required information, we need a promoter. So promoter plays a vital role in DNA transcription. It is defined as “the sequence in the region of the upstream of the transcriptional start site (TSS)”. If we identify the promoter region we can extract information regarding gene expression patterns, cell specificity and development. Some of the genetic diseases which are associated with variations in promoters are asthma, beta thalassemia and rubinstein-taybi syndrome.

2. Literature Survey

Eric E. Snyder[1] at el. has developed a PC program, Geneparser, which distinguishes and verifies the protein coding genes in genomic DNA arrangements. This program scores all subintervals in a grouping for substance facts characteristic of introns and exons, and for destinations that recognize their limits. Jonathan H. Badger[21] at el. has proposed a protein coding region identification tool named CRITICA which uses comparative analysis. In the comparative segment of the investigation, regions of DNA are straightened with identified successions from the DNA databases; if the interpretation of the arranged groupings has more stupendous amino acid than needed for the watched rate nucleotide character; this is translated as proof for coding.

David J. States [22] proposed a PC project called BLASTX was formerly indicated to be viable in distinguishing and allotting putative capacity to likely protein coding districts by identifying huge likeness between a theoretically interpreted nucleotide query arrangement and parts of a protein grouping database.

Steven Salzberg [23] at el. has used a decision tree algorithm for locating protein coding region. Genes in eukaryotic DNA spread hundreds or many base sets, while the locales of the aforementioned genes that code for proteins may possess just a little rate of the succession. Eric E.Snyder [3] at el. has used dynamic programming with neural networks to address the protein coding region problem. Dynamic Programming (DP) is connected to the issue of exactly distinguishing inside exons and introns in genomic DNA arrangements. Suprakash Datta [2] at el. used a DFT based gene prediction for addressing this problem. Authors provided theoretical concept of three periodicity property observed in protein coding regions in genomic DNA. They

proposed new criteria for classification based on traditional frequency approaches of coding regions. Jesus P. Mena-Chalco [24] at el has used Modified Gabor-Wavelet Transform for addressing this issue. In this connection, numerous coding DNA model-free systems dependent upon the event of particular examples of nucleotides at coding areas have been proposed. Regardless, these techniques have not been totally suitable because of their reliance on an observationally predefined window length needed for a nearby dissection of a DNA locale.

Rakesh Mishra [25] at el. has worked on search and use of promoter region. Look for a promoter component by RNA polymerase from the to a great degree extensive DNA base arrangement is thought to be the slowest and rate-confirming for the regulation of interpretation process. Few immediate investigations we portrayed here which have attempted to accompany the robotic suggestions of this promoter look. Christoph Dieterich [26] at el. made an extensive study on promoter region. The robotized annotation of promoter districts joins data of two sorts. To begin with, it recognizes cross-species preservation inside upstream districts of orthologous genes. Pair wise too a various arrangement examinations are processed. Vetriselvi Rangnanan[27] at el. made a dissection of different anticipated structural lands of promoter districts in prokaryotic and in addition eukaryotic genomes had prior shown that they have a few normal characteristics, for example, lower steadiness, higher curve and less bendability, when contrasted and their neighboring areas. Jih-Wei Hung [4] has developed an effective forecast calculation that can expand the recognition (power =1 - false negative) of promoter. Authors introduce two strategies that utilize the machine force to ascertain all conceivable examples which are the conceivable characteristics of promoters.

3. AIS-INMACA(Artificial Immune System- Integrated Multiple Attractor Cellular Automata)

Multiple Attractor Cellular Automata which is used in this report is a special class of fuzzy cellular automata which was introduced thirty years ago .It uses fuzzy logic[11] to handle real value attributes. The development process / implementation of Multiple Attractor Cellular Automata is administered by AIS technique, a Clonal Algorithm with the underlying theory of survival of the fittest gene.

Artificial Immune System is a novel computational intelligence technique with features

like distributed computing, fault /error tolerance, dynamic learning, adaption to the frame work, self monitoring, non uniformity and several features of natural immune systems. AIS take its motivation from the standard immune system of the body to propose novel computing tools for addressing many problems in wide domain areas. Some features of AIS which can be mapped with bioinformatics framework are chosen and used in the thesis to strengthen the proposed CA classifier.

This paper introduces the integration of AIS with CA which is first for its kind to produce a better classifier which can address major problems in bioinformatics. This proposed classifier named artificial immune system based multiple attractor cellular automata classifier (AIS-INMACA) uses the basic frame work of Cellular Automata (CA) and features of AIS like self monitoring and non uniformity which is potential, versatile and robust. This is the basic motivation of the entire research.

The objectives of AIS based evolution of MACA[12],[13],[15] is

1. To improve the conception of the ways CA performs calculations.
2. To figure out how CA may be advanced to perform a particular computational job.
3. To follow how advancement makes complex global behavior into locally interconnected cells on a grid.
4. To extract innate classification potential in CA and use Clonal algorithm for producing better rules with fitness.

3.1 Design of AIS-INMACA

The proposed AIS based CA classifier uses fuzzy logic can address major problems in bioinformatics like protein coding region identification and promoter region prediction .Even though some scientists have proposed different algorithms, all of these are specific to the problem. None of them have worked towards proposing a common classifier whose frame work can be useful for addressing many problems in bioinformatics.

The general design of AIS-INMACA is indicated in the figure 3.1. Input to AIS-INMACA algorithm and its variations will be DNA sequence and Amino Acid sequences. Input processing unit will process sequences three at a time as three neighborhood cellular automata is considered for processing DNA sequences. The rule generator will transform the complemented (Table 3.3) and non complemented rules (Table 3.2) in the form of matrix, so that we can apply the rules to the corresponding sequence positions very easily. AIS-INMACA basins are

calculated as per the instructions of proposed algorithm and an inverter tree named as AIS multiple attractor cellular automata is formed which can predict the class of the input after all iterations.

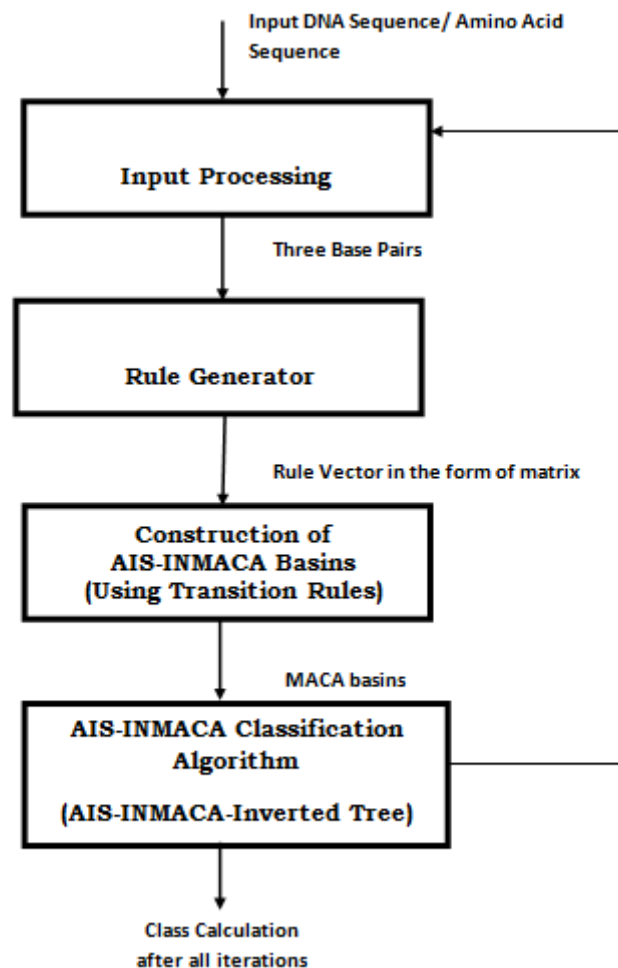


Figure 3.1 Design of AIS-INMACA

Algorithm 3.1 is used for creating of AIS-INMACA tree .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 3.3 uses, algorithm 3.1, 3.2 for predicting the protein and promoter coding regions.

3.2 Rules of AIS-INMACA

The decimal equivalent of the next state function, as defined as the rule number of the CA cell introduced by Wolfram [19], is. In a 2-state 3-neighborhood CA, there are 256 distinct next state functions, among 256 rules, rule 51 and rule 254 are represented in the following equations.

$$\text{Rule 51 : } q_i(t+1) = \overline{q_i(t)} \quad (1)$$

$$\text{Rule 254 : } q_i(t+1) = q_{i-1}(t) + q_i(t) + q_{i+1}(t) \quad (2)$$

Table3.1 : Transition Function

Neighbourhood	111	110	101	100	011	010	001	000	Rule
Next State (Rule 51)	0	0	1	1	0	0	1	1	51
Next State (Rule 254)	1	1	1	1	1	1	1	0	254

Table: 3.2 Non complemented Rules

SNO	Rule Number	General Representation
1.	254	$q_{i-1} + q_i + q_{i+1}$
2.	252	$q_{i-1} + q_i$
3.	238	$q_i + q_{i+1}$
4.	250	$q_{i-1} + q_{i+1}$
5.	204	q_i
6.	240	q_{i-1}
7.	170	q_{i+1}
8.	0	0

The transition function table 3.1 was shown for equations 1,2. The tables 3.2, 3.3 show the complemented and non complemented rules.

Table 3.3 Complemented Rules

SNO	Rule Number	General Representation
1.	1	$\overline{q_{i-1} + q_i + q_{i+1}}$
2.	3	$\overline{q_{i-1} + q_i}$
3.	17	$\overline{q_i + q_{i+1}}$
4.	5	$\overline{q_{i-1} + q_{i+1}}$
5.	51	$\overline{q_i}$
6.	15	$\overline{q_{i-1}}$
7.	85	$\overline{q_{i+1}}$
8.	255	1

Algorithm 3. 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 patters 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 3. 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 $F_t > 0.5$

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 $GC > G_{max}$ then go to Step 11.

Step 9: Form NP by selection, cloning and mutation.

Step 10: $PP \leftarrow NP$; 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 3.3:

1. Uses the AIS-INMACA Tree construction Algorithm 3.1
2. Uses the AIS-INMACA Evolution Algorithm 3. 2
3. Trace the corresponding attractor
4. Travel back form 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 fist codon of the sequence to start codon Search for TAATAA.
9. Report the promoter boundary located at upstream

4. Experimental Results & Discussions

Experiments were conducted by using Fickett and Toung data [28] for predicting the protein coding regions. All the 21 measures reported in [28] were considered for developing the classifier. For promoter region E Coli promoter collection data sets from UCI Machine Learning Repository was taken for testing and training. Fig 4.1 shows the interface developed. Fig 4.2 is the training interface with rules, sequence and real values. Fig 4.3 shows the testing interface. Table 4.1 show the execution time for predicting both protein and promoter regions which is very promising. Table 4.2 shows the number of datasets handled by AIS-INMACA. Fig 4.4 and Fig 4.5 shows the accuracy of prediction separately which is the important output of our work.

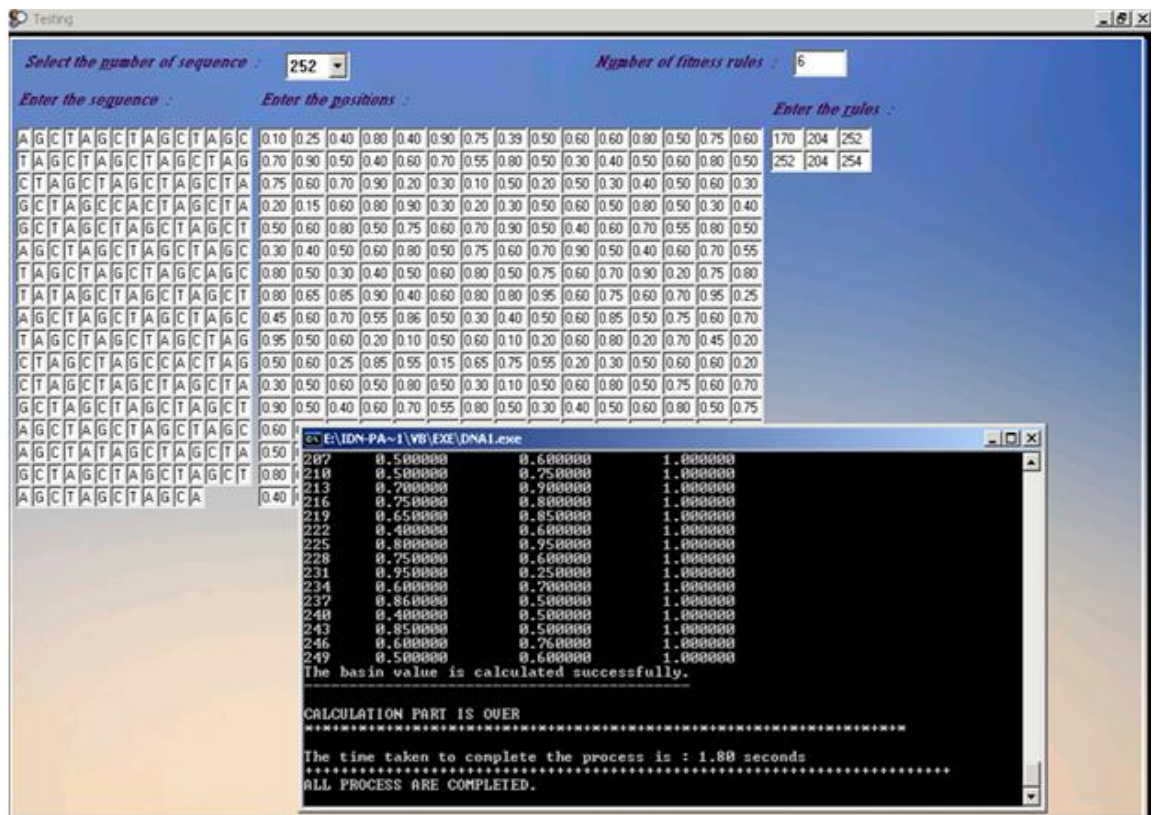


Fig 4.3: Testing Interface

Size of Data Set	Prediction Time of Integrated Algorithm
5000	1064
6000	1389
10000	2002
20000	2545

Table 4.1: Execution Time for prediction of both protein and promoter regions.

Number of Data Sets Used for Classification

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

Table 4.2: Number of Data Sets Used

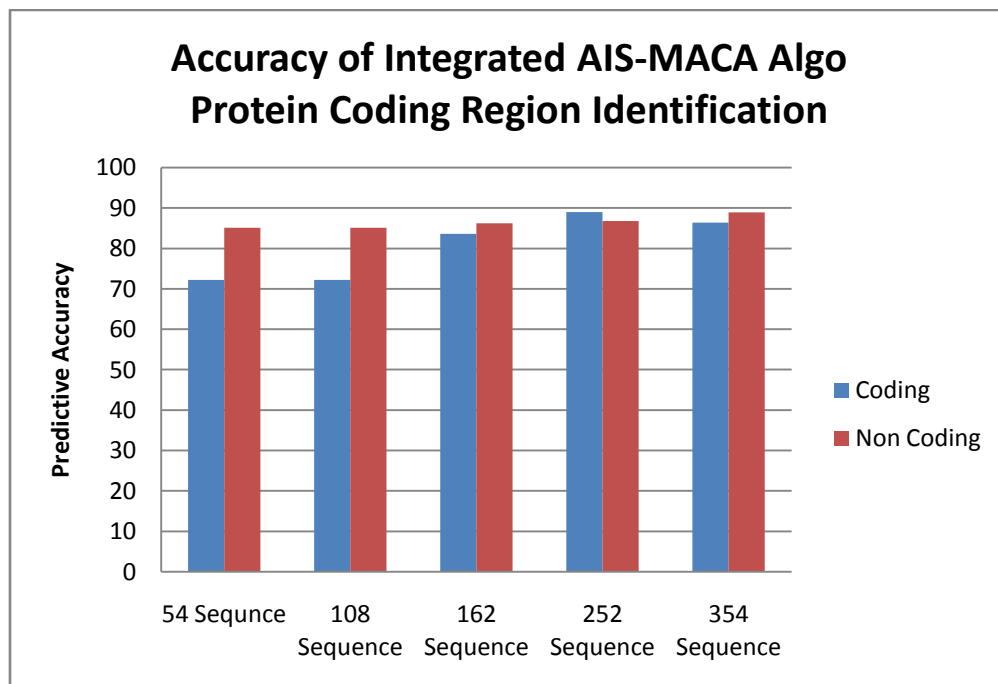


Fig 4.4: Predictive Accuracy for Protein Coding Regions

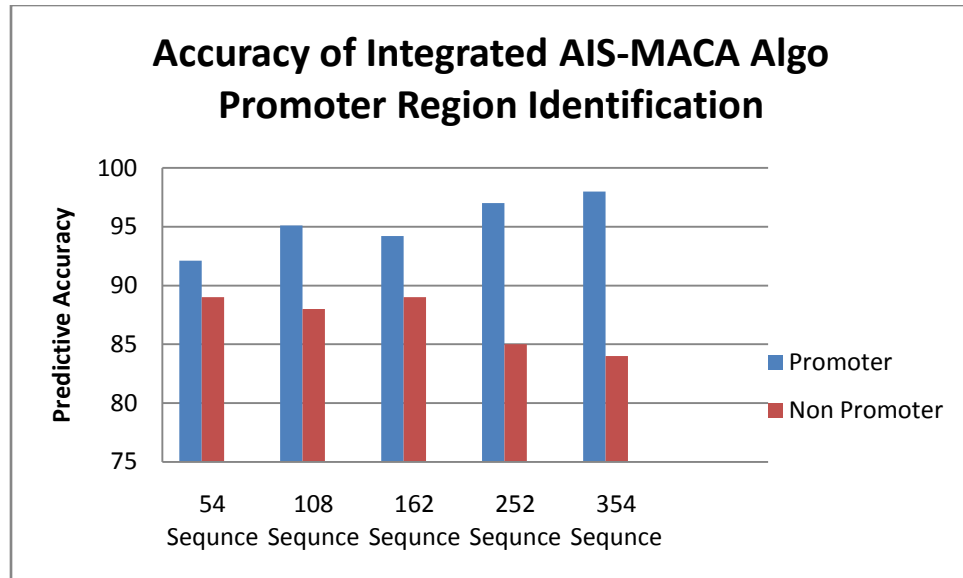


Fig 4.5: Predictive Accuracy for Promoter Regions

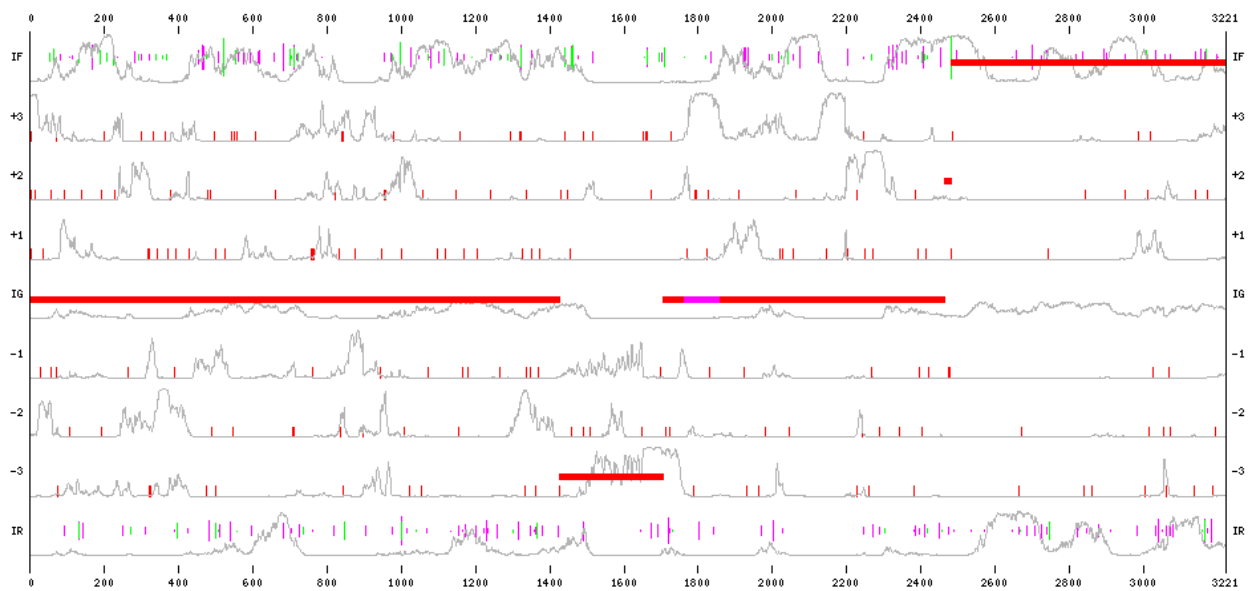


Fig 4.6: Exons Prediction

Gene number	Element number	Exons/UTR	Strand	Left end	Right end	Length
1	0	Utr3	+	1	1402	1402
2	0	Utr3	-	1404	1425	22
2	1	Single	-	1426	1704	279
2	0	Utr5	-	1705	1760	56
3	0	Utr5	+	1861	2464	604
3	1	Initial	+	2465	2482	18

Fig 4.7: Exons Boundaries Reporting

Start	End	Score	Promoter Sequence
268	313	0.99	ATACTTTGGACAGAAAGAGGGTCCCAGTATAGCTTGGGCGACCTCTGAAA
866	911	0.84	CACCCTCTTGATTTTTTTTTTTCCTTTCCAATTTATTCGGATCTGCTCTGGA
1139	1184	0.80	TGGAATGACAGTCCCTAGCTTGGACCCTTAGACTGCCCGTTCTCATCCA
1447	1492	0.86	AGCCCTTGACAGGTCAGGTCCGCGGGACCCAGAATTCGCTTTGAGCCTAT
1914	1959	0.91	GGAGTTGCAGACTCAGGAATCAGGGAGCGGATACTGGGCGTGGACCCCGT

Fig 4.8: Predictive Accuracy for Promoter Regions

Fig 4.6 gives the prediction of exons from the given input graphically. Fig 4.7 gives the boundary of the location of exons. Fig 4.8 gives the promoter region prediction and boundary reporting.

5. Conclusion

We have successfully developed a logical classifier designed with MACA and strengthened with AIS technique. The accuracy of the AIS-INMACA 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 integrated algorithm to process DNA sequences of length

252,354. This novel classifier frame work can be used to address many problems in 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 and many more. We have successfully developed a frame work with this classifier which will lay future intuition towards application of CA in number of bioinformatics applications.

6. References

1. Eric E. Snyder, Gary D. Stormo, "Identification of Protein Coding Regions in Genomic DNA". ICCS Transactions 2002.
2. Datta, Suprakash, and Amir Asif. "A fast DFT based gene prediction algorithm for identification of protein coding regions." In *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, vol. 5, pp. v-653. IEEE, 2005.
3. E E Snyder and G D Stormo,"Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks "Nucleic Acids Res. 1993 February 11; 21(3): 607–613.
4. Huang, Jih-Wei. "Promoter Prediction in DNA Sequences." PhD diss., National Sun Yat-sen University, 2003.
5. Horwitz, M. S., and Lawrence A. Loeb. "Promoters selected from random DNA sequences." *Proceedings of the National Academy of Sciences* 83, no. 19 (1986): 7405-7409.
6. KrishnaKumar, K., Kaneshige, J., and Satyadas, A., "Challenging Aerospace Problems for Intelligent Systems," *Proceedings of the von Karman Lecture series on Intelligent Systems for Aeronautics*, Belgium, May 2002.
7. Pokkuluri Kiran Sree , I.Ramesh Babu , "Investigating an Artificial Immune System to Strengthen the Protein Structure Prediction and Protein Coding Region Identification using Cellular Automata Classifier" *International Journal of Bioinformatics Research and Applications* ,Vol 5,Number 6,pp 647-662, ISSN : 1744-5493.
8. Pokkuluri Kiran Sree , I.Ramesh Babu,"PSMACA: An Automated Protein Structure Prediction using MACA (Multiple Attractor Cellular Automata)". *Journal of Bioinformatics and Intelligent Control (JBIC)* in Volume 2 Number 3, pp 211-215

- ,2013Pokkuluri Kiran Sree et al ,Identification of Promoter Region in Genomic DNA Using Cellular Automata Based Text Clustering. The International Arab Journal of Information Technology (IAJIT),Volume 7,No 1,2010,pp 75-78.
9. Pokkuluri Kiran Sree , I.Ramesh Babu,"An extensive report on Cellular Automata based Artificial Immune System for strengthening Automated Protein Prediction". Advances in Biomedical Engineering Research (ABER) Volume 1 Issue 3, September 2013, pp 45-51.
 10. Pokkuluri Kiran Sree , I.Ramesh Babu,"A Novel Protein Coding Region Identifying Tool using Cellular Automata Classifier with Trust-Region Method and Parallel Scan Algorithm (NPCRITCACA)". International Journal of Biotechnology & Biochemistry (IJBB) Volume 4, 177-189 Number 2 (December 2008).
 11. Pokkuluri Kiran Sree , I.Ramesh Babu, "HMACA: Towards proposing Cellular Automata based tool for protein coding, promoter region identification and protein structure prediction". International Journal of Research in Computer Applications & Information Technology, Volume 1 Number 1, pp 26-31,2013.
 12. Pokkuluri Kiran Sree , I.Ramesh Babu," Multiple Attractor Cellular Automata (MACA) for Addressing Major Problems" in Bioinformatics in Review of Bioinformatics and Biometrics (RBB) Volume 2 Issue 3, September 2013, pp70-76. Pokkuluri Kiran Sree et al ,Hybrid Attractor Cellular Automata for Addressing Major Problems in Bioinformatics in Research and Reviews: Journal of Engineering and Technology, Volume 2 Issue 4, October-2013,pp 42-48.
 - 10 Pokkuluri Kiran Sree , I.Ramesh Babu,"AIS-PRMACA: Artificial Immune System based Multiple Attractor Cellular Automata for Strengthening PRMACA, Promoter Region Identification", in The SIJ Transactions on Computer Science Engineering & its Applications (CSEA), The Standard International Journals (The SIJ), Vol. 1, No. 4, Pp. 124-127.(2013)
 - 11 P. Flocchini, F. Geurts, A. Mingarelli, and N. Santoro (2000), "Convergence and Aperiodicity in Fuzzy Cellular Automata: Revisiting Rule 90," Physica D.
 - 12 P. Maji and P. P. Chaudhuri (2004), "FMACA: A Fuzzy Cellular Automata Based Pattern Classifier," Proceedings of 9th International Conference on Database Systems, Korea, pp. 494–505, 2004.
 - 13 P. Maji and P. P. Chaudhuri, "Fuzzy Cellular Automata for Modeling Pattern Classifier,"

IEICE, (2004).

- 14 P.Kiran Sree, I .Ramesh Babu,” Identification of Protein Coding Regions in Genomic DNA Using Unsupervised FMACA Based Pattern Classifier” in International Journal of Computer Science & Network Security with ISSN: 1738-7906 Volume Number: Vol. 8, No. 1, 2008.
- 15 P. Maji and P. P. Chaudhuri (2004),“FMACA: A Fuzzy Cellular Automata Based Pattern Classifier,” Proceedings of 9th International Conference on Database Systems, Korea, pp. 494–505, 2004.
- 16 Snyder, Eric E., and Gary D. Stormo. "Identification of protein coding regions in genomic DNA." *Journal of molecular biology* 248, no. 1 (1995): 1-18.
- 17 Yin, Changchuan, and Stephen S-T. Yau. "Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence." *Journal of theoretical biology* 247, no. 4 (2007): 687-694.
- 18 J. Von. Neumann, ”The Theory of Self-Reproducing Automata,” A. W. Burks, Ed.University of Illinois Press, Urbana and London, 1966.
- 19 Wolfram, Stephen. "Universality and complexity in cellular automata." *Physica D: Nonlinear Phenomena* 10, no.1 ,1984: pp. 1-35.
- 20 Crick, Francis. "Central dogma of molecular biology." *Nature* 227, no. 5258 (1970): pp. 561-563.
- 21 Badger, Jonathan H., and Gary J. Olsen. "CRITICA: coding region identification tool invoking comparative analysis." *Molecular Biology and Evolution* 16, no. 4 (1999): 512-524.
- 22 Gish, Warren, and David J. States. "Identification of protein coding regions by database similarity search." *Nature genetics* 3, no. 3 (1993): 266-272.
- 23 Salzberg, Steven. "Locating protein coding regions in human DNA using a decision tree algorithm." *Journal of Computational Biology* 2, no. 3 (1995): 473-485.
- 24 Mena-Chalco, Jesús P., Helaine Carrer, Yossi Zana, and Roberto M. Cesar. "Identification of protein coding regions using the modified Gabor-wavelet transform." *Computational Biology and Bioinformatics, IEEE/ACM Transactions on* 5, no. 2 (2008): 198-207.

- 25 Mishra, Rakesh K., Vijaya Gopal, and Dipankar Chatterji. "Correlation between the DNA supercoiling and the initiation of transcription by *Escherichia coli* RNA polymerase in vitro: role of the sequences upstream of the promoter region." *FEBS letters* 260, no. 2 (1990): 273-276.
- 26 Dieterich, Christoph, Steffen Grossmann, Andrea Tanzer, Stefan Röpcke, Peter F. Arndt, Peter F. Stadler, and Martin Vingron. "Comparative promoter region analysis powered by CORG." *BMC genomics* 6, no. 1 (2005): 24.
- 27 Rangannan, Vetriselvi, and Manju Bansal. "Identification and annotation of promoter regions in microbial genome sequences on the basis of DNA stability." *Journal of biosciences* 32, no. 5 (2007): 851-862.
- 28 Fickett, James W., and Chang-Shung Tung. "Assessment of protein coding measures." *Nucleic acids research* 20, no. 24 1992: pp. 6441-6450.