Enliven Archive
Intensifying Thoughts

Research Article www.enlivenarchive.org Enliven: Bioinformatics

A Human Promoter Prediction using MACA-CLONAL Classifier

Pokkuluri Kiran Sree[1*] and Inampudi Ramesh Babu[2]

[1]Department of CSE, JNTU Hyderabad, India

[2]Department of CSE, ANU, Guntur, India

Abstract

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. So we propose a novel fast multiple attractor cellular automata (MACA) with modified Clonal classifier for promoter prediction in eukaryotes. We have used three important features like TATA box, GC box and CAAT box for developing this classifier. We have also used context future 6-mer for predicting the same. The proposed classifier is trained and tested with datasets from DBTSS, EID, UTRdb datasets . In training phase of the classifier 100% specificity was obtained. In testing phase 84.5% sensitivity and 92.7% specificity was achieved in an average. The time taken to predict the promoter region of length 252 in an average is 4 micro seconds.

Keywords: Cellular automata; Multiple attractor cellular automata; Clonal classifier; Promoter

Abbreviations: CA: Cellular Automata; MACA: Multiple Attractor Cellular Automata; CC: Clonal Classifier

*Corresponding author: Pokkuluri Kiran Sree, Professor, Department of CSE, JNTU Hyderabad, India, E-mail: profkiransree@gmail.com

Received Date: 12 April 2014
Accepted Date: 28 May 2014
Published Date: 31 May 2014

Citation: Kiran SP, Ramesh Babu I(2014) A Human Promoter Prediction Using MACA-CLONAL Classifier. Enliven: Bioinform 1(1): 002.

Introduction: Most of the problems in bioinformatics can be address through bioinformatics. Promoter prediction plays a vital role in protein formulation and DNA transcription. Some of the genetic diseases which are associated with variations in promoters are asthma, beta thalassemia and rubinstein-taybi syndrome. Promoter sequence [1] can be used to control the speed of translation from DNA into protein. It is also used in genetically modified foods.

Figure 1 shows the location of promoter and protein coding region in untranslated region (UTR). Promoter is located towards the upstream (5') of the DNA sequence. Promoter initiates the Transcription. The start codon (ATG) of the protein coding region and stop codon (TAG) were also indicated in the figure 1.

Cellular Automata (CA) is a basic model of a spatially developed decentralized system, made up of various unique components called Cells.

It is a computing model which can provide a good platform for performing complex computations with the available local information. Each cell in the system has a specific state which changes with over time depending on the neighboring states.

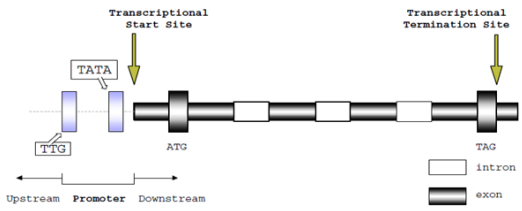

Figure 1: Promoter Region in a DNA sequence

CA is defined a four tuple <G, Z, N, F>

Where G ⟶ Grid (Set of cells)

Z ⟶ Set of possible cell states

N ⟶ Set which describe cells neighborhoods

F ⟶ Transition Function (Rules of automata)

A CA displays three basic characteristics – locality, infinite parallelism, and simplicity

Neumann V et al. [2] and Stanislaw Ulam initially proposed the model of Cellular Automata in 1940. Wolfram S [3] did a detailed study on one-dimensional CA (Elementary CA). He later published a book on "A New Kind of Science" in 2002 which dealt with basic and neighborhood structure of CA has pulled in scientists from different disciplines. It has been subjected to thorough numerical and physical dissection for most recent fifty years and its requisition has been proposed in diverse extensions of science – both social and physics.

One dimensional CA became popular due to vast simplicity which uses two states per cell. CA became very popular in the context of VLSI assuming zero and one as states which uses additive and linear CA. CA has been applied to multi dimension grid apart from one dimension and two dimensions. The rules can be applied to each cell either uniform or non uniform. Global transition rule and local transition rule can be represented accordingly.

So we apply a special class of CA [4] termed as multiple attractor cellular automata which uses fuzzy logic strengthened with modified Clonal classifier to predict the promoters efficiently and fastly.

The proposed classifier first encodes the corresponding DNA sequence which is attributed with real values between one and zero. This modified classifier now will process the input three in a sequence. The inverted tree is formed based on the three content features and one context feature. This will provide flexibility to map the frame work of this modified MACA with so many other species calculations of promoters also which is not there in earlier approaches.

Literature Survey

Vladimir B. Bajic et al. [1] have developed ANN (Artificial Neural Networks) based program for finding promoters using micro-structural promoter component recognition. Authors have considered features like TATA box, CCAAT box, Inr and GC box for promoter prediction. All these features are cascaded and every feature has a corresponding ANN developed. The output of all features will be given to the integration layer ANN to give the final output. Authors have compared their work with Audic, Autogene, Promoter 2.0, NNPP, Promter Find, Promoter Scan, TATA, TSSG, TSSW, IMC, SPANN, SPANN2 for True Positives and False Positives.

Hung JW et al. [5] 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. The primary strategy we exhibit FTSS (Fixed Transcriptional Start Site) utilizes the known TSS positions of promoter arrangements to prepare the score record that helps us in promoter forecast. The other strategy is NTSS (Nonfixed TSS). The TSS positions of promoter arrangements utilized as a part of NTSS are thought to be obscure, and NTSS won't take irrefutably the positions of Tsss into attention. By the exploratory effects, our expectation has higher right rate than different past systems.

Horwitz MSZ et al. [6] have chosen an assembly of Escherichia coli promoters from irregular DNA groupings by swapping 19 base sets at the -35 promoter area of the etracycline safety gene te" of the plasmid pbr322. Substitution of 19 base sets with artificially blended irregular groupings brings about a greatest of 419 (something like 3 x 1011) conceivable swap groupings. From a populace of in the ballpark of 1000 microscopic organisms harboring plasmids with these irregular substitutions, tetracycline choice has uncovered numerous practical -35 promoter successions. These promoters have held just halfway. Homology to the -35 promoter accord grouping. In three of these promoters, the agreement operator moves 10 nucleotides downstream, permitting the RNA polymerase to distinguish an alternate Pribnow box from inside the definitive pbr322 succession. Two of the successions advertise translation more determinedly than the local promoter.

Zeng J et al. [7] have used signal, structure and context features for predicting promoter regions in humans. Authors have observed 50% of mammalian promoters can be predicted with CpG signal so they have considered this signal feature for promoter prediction. Authors have considered n-mers and their statistics as a context feature for promoter prediction. They also considered the flexibility that comes from 3D structural feature which plays vital role in guiding transcription factors.

Oscar et al. [8] has presented a novel promoter prediction program named CNN-Promoter which is trained to predicts three types of promoters named core promoter, proximal promoter and distal promoter in eukaryotic organisms. Authors propose a consensus strategy which is implemented with neural networks. Authors have achieved 100% specificity during training phase. Authors have used three important features like TATA box, GC box and CAAT box for developing this classifier.

Design of MACA Based Modified Clonal Classifier

A Cellular Automata which uses fuzzy logic is an array of cells arranged in linear fashion evolving with time. Every cell of this array assumes a rational value in the interval of zero and one. All this cells changes their states according to the local evaluation function which is a function of its state and its neighboring states.

The general design of MACA [9-11] based Modified Clonal Classifier is indicated in the figure 2. Input to this 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 and non complemented rules in the form of matrix, so that we can apply the rules to the corresponding sequence positions very easily. MACA [12-14] basins are calculated as per the instructions of proposed algorithm and an inverter tree as in figures 3 and 4 named as AIS multiple attractor cellular automata is formed which can predict the class of the input after all iterations.

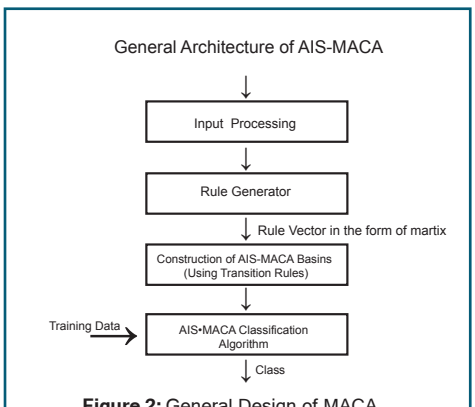

Figure 2: General Design of MACA

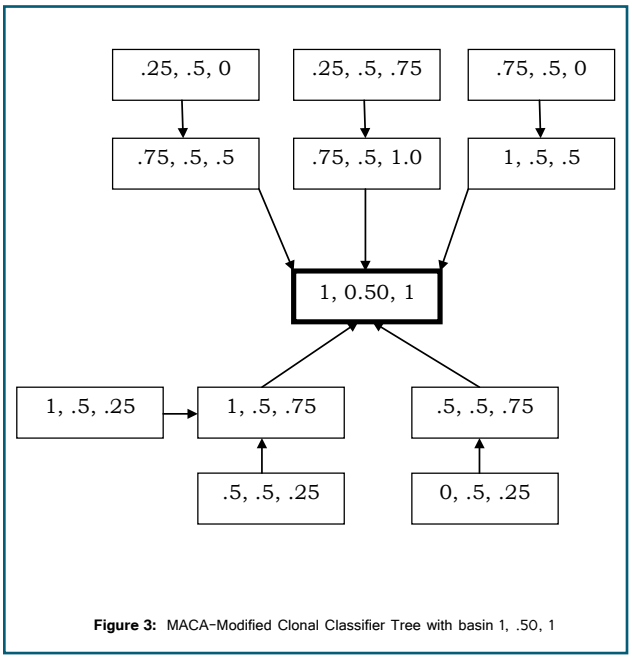

Figure 3: MACA-Modified Clonal Classifier Tree with basin 1, .50, 1

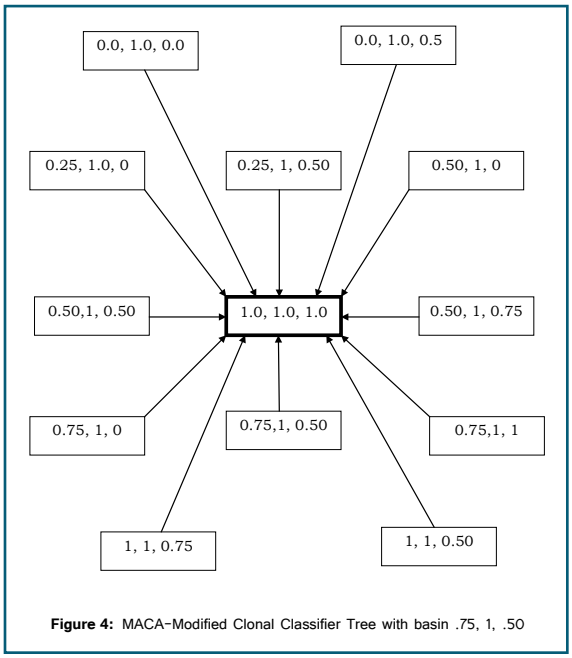

Figure 4: MACA-Modified Clonal Classifier Tree with basin .75, 1, .50

The algorithms takes input as DNA sequence and the maximum population and give output as the class, matrix representation and rule specification.

Input: S = {S1, S2, • • • , Sl}, Training Set, Maximum Population Mmax).

Output: Matrix Representation T, F, and information of the class

Begin

Step 1: Generate 500 new chromosomes for Initial Population.

Step 2: Initialize Maximum Population MM=zero; PP←IP.

Step 3: Compute fitness FF for each chromosome of PP according

Step 4: Store T, F, and corresponding class information for which the fitness value FF = 1.

Step 5: If FF = 1 for at least one chromosome of PP, then go to Stop.

Step 5a: Check the TATA box

Step 5b: Check the GC box

Step 5c: Check the CAAT box.

Step 6a: Construct the MACA-CC tree based on 5a, 5b, 5c.

Step 6: Order chromosomes in order of fitness.

Step 7: Increment Maximum Population (MM).

Step 8: If GC > Gmax then go to Step 11.

Step 9: Form NP by operations of Modified Clonal algorithm

Step 10: PP← NP; go to Step 3.

Step 11: Out Put and Store T, F, and corresponding class information for which the fitness value is maximum.

Step 12: Stop.

Human body consists of lot of cells, molecules and organs. No organ or cell or molecule can control the functioning of an immune system. The main aim of the immune system is to search and find the malfunctioning cells within the body and foreign elements which may cause diseases. The element which can be recognized by an immune system is named as antigen. Artificial Immune System have intelligent algorithms like negative selection algorithm, Clonal selection algorithm etc. We have taken the clonal algorithm and made significant modifications enough to map this to our prediction.

Modified Clonal Algorithm

1.Generate initial antibody population (AIS-MACA rules) randomly and call it as Ab. It consists of two subsets memory population Ab_m and reservoir population Ab_r.

2.Construct a set of Antigens population call it as Ag (DNA Sequence with Class/ Input).

3.Select an antigen Ag_j from Ag the antigen population.

4.Apply every member of antibody population to the selected antigen Ag_j and calculate affinity of the rule with the antigen via fitness equation in 4.1.

5.Select m highest affinity antibodies (AIS-MACA rules) from Ab and generate clones for each antibody, which will be proportional to the affinity as per the equation 4.2. Place the clones in the new population P_i.
6.Apply mutation to the newly formed population Pi where the degree is inversely proportional to their affinity as per equation 4.3. This produces a more mature population P_i^*

7.Re Calculate the affinity of the rule with the corresponding antigen as we did it in step four. Order the antibodies in descending order. (high fitness antibody will be on top)

8.Compare the antibodies from P_i^* with the antibodies population from Abm. Select the better fitness rules and remove them from Pi* and place them in Ab^m.

Experimental Results

An extensive testing of the classifier id done and the results are quite promising. The data sets are taken from DBTSS [10], EIP [16], UTRdb [17]. A total of 50% of each data set is used for training and 50% are used for testing the promoters.

Parameters for testing promoters

The important statistics to look at include:

1.True Positives (TP): Number of correctly predicted promoters.

2.False Positives (FP): Number of incorrectly predicted promoters

3.True Negatives (TN): Number of correctly predicted non promoters

4.False Negatives (FN): Number of incorrectly predicted non promoters

Using the above measures following are calculated

1.Actual Positives (AP) = TP + FN

2.Actual Negatives (AN) = TN + FP

3.Predicted Positives (PP) = TP + FP

4.Predicted Negatives (PN) = TN + FN

5.Sensitivity (SN) = TP / (TP + FN)

6.Specificity (SP) = TP / (TP + FP)

The proposed classifier is compared with standard promoter prediction programs like Promoter Inspector, Dragon Promoter Finder, Promo Predictor, CNN-Promoter, SPANN and IMC as shown in table 1. Table 2 shows the comparison of true positives and false positives with standard promoter prediction programs.

The developed front is reported in figure 5. This classifier has an inbuilt parameter for estimating the average time to predict promoters in a given DNA sequence. This classifier predicts promoter region in .7 nano seconds for a DNA sequence of length 252. Figure 6 shows the accuracy of promoter prediction.

	Sensitivity	Specificity
Promoter Inspector	56.9	46.9
Dragon Promoter Finder	62.3	59.3
Promo Predictor	65.3	66.9
CNN-Promoter	76.3	82.3
SPANN	68.9	84
IMC	76	86
MACA-Modified CC	84.5	92.7

Table 1: Comparison of MACA-Modified CC with existing approaches with Sensitivity Vs Specificity

	True Positives	False Positives
Promoter 2.0	12	45
Promoter Finder	8	30
Promoter Scan	4	10
TATA	10	40
TSSG	8	20
SPANN	11	12
MACA-Modified CC	06	08

Table 2: Comparison of MACA-Modified CC with existing approaches with True Positive Vs False Positives

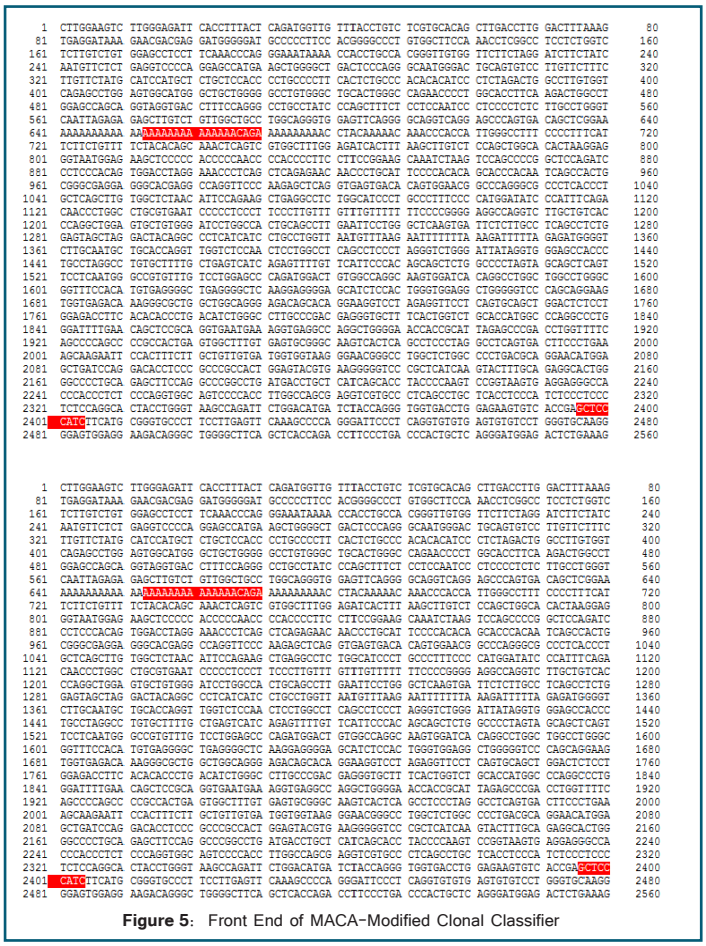

Figure 5: Front End of MACA-Modified Clonal Classifier

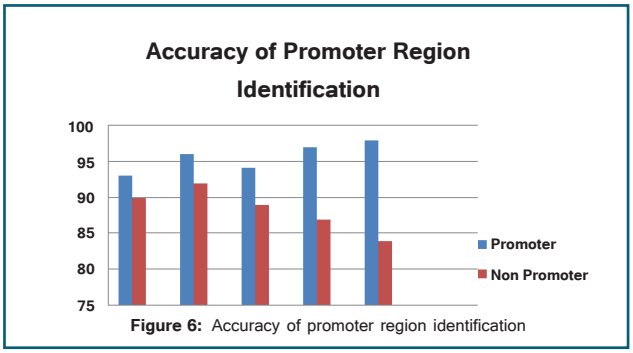

Figure 6: Accuracy of promoter region identification

Conclusion

We have successfully developed and tested the MACA based modified Clonal Classifier for predicting promoter regions in eukaryotes. The proposed classifier is tested for specificity and sensitivity. It is compared with important promoter programs available. The results obtained are found promising and comparable. This classifier is also observed and tested for the amount of time it will be taking to predict the promoter and it was found as .7 nano seconds. A sensitivity of 84.5% and specificity of 92.7 were reported.

References

1. Bajic VB, Tan SL, Suzuki Y, Sugano S (2004) Promoter prediction analysis on the whole human genome. Nat Biotechnol 22: 1467-1473.
2. John VN, John C (1988) John von Neumann. American Mathematical Soc. Oxtoby, Eds.
3. Stephen W (1994) Cellular automata and complexity: collected papers. Reading: Addison-Wesley.
4. Maji P, Shaw C, Ganguly N, Sikdar BK, Chaudhuri PP (2003) Theory and application of cellular automata for pattern classification. Fundamenta Informaticae 58: 321-354.
5. Yuan HY, Chen JJ, Lee MT, Wung JC, Chen YF, et al. (2005) A novel functional VKORC1 promoter polymorphism is associated with inter-individual and inter-ethnic differences in warfarin sensitivity. Hum Mol Genet 14: 1745-1751.
6. Horwitz MS, Loeb LA (1998) Method for producing novel DNA sequences with biological activity. U.S. Patent 5,824,469, issued October 20.
7. Zeng J, Zhu S, Yan H (2009) Towards accurate human promoter recognition: a review of currently used sequence features and classification methods. Brief Bioinform 10: 498-508.
8. Óscar B, Bustamante S (2011) Cnn-Promoter, New Consensus Promoter Prediction Program Based on Neural Networks. Revista EIA 15.
9. Kiran S, Babu R (2010) Identification of Promoter Region in Genomic DNA Using Cellular Automata Based Text Clustering. International Arab Journal of Information Technology 7.
10. Kiran SP, Ramesh Babu I, Usha Devi N (2013) PSMACA: An Automated Protein Structure Prediction using MACA (Multiple Attractor Cellular Automata). Journal of Bioinformatics and Intelligent Control 2: 211-215.
11. Kiran SP, Ramesh Babu I (2014) AIX-MACA-Y Multiple Attractor Cellular Automata Based Clonal Classifier for Promoter and Protein Coding Region Prediction. Journal of Bioinformatics and Intelligent Control 3: 23-30.
12. Usha Devi NSSSN, Ramesh Babu I, Kiran SP (2013) Multiple Attractor Cellular Automata (MACA) for Addressing Major Problems in Bioinformatics. Review of Bioinformatics and Biometrics 2: 3.
13. Suzuki Y, Yamashita R, Nakai K, Sugano S (2002) DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. Nucleic Acids Res 30: 328-331.
14. Kiran SP, Ramesh Babu I (2013) An extensive report on Cellular Automata based Artificial Immune System for strengthening Automated Protein Prediction. Advances in Biomedical Engineering Research 1: 45-51.
15. Kiran SP, Ramesh Babu I, Usha Devi NSSSN (2014) PRMACA: A Promoter Region Identification Using Multiple Attractor Cellular Automata (MACA). In ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India I: 393-399.
16. Philipp B, Boveresses ChD (1991) Eukaryotic promoter database. NETSERVE@ EMBL-Heidelberg. DE.
17. Pesole G, Liuni S, Grillo G, Licciulli F, Mignone F, et al. (2002) UTRdb and UTRsite: specialized databases of sequences and functional elements of 5′ and 3′ untranslated regions of eukaryotic mRNAs. Update 2002. Nucleic Acids Res1: 335-340.