MACA-MCC-DA: A Fast MACA with Modified Clonal Classifier Promoter Region Prediction in Drosophila and Arabidopsis

Pokkuluri Kiran Sree, Inampudi Ramesh Babu

ABSTRACT
DNA is a very important component in a cell, which is located in the nucleus. DNA contains lots of information. For DNA sequence to transcribe and form RNA, which copies the required information, we need a promoter. So a 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 a 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 for Drosophila and Arabidopsis. We have used three important features like TATA box, GC box and CAAT box for developing this classifier. We have also used 3-mers and 6-mers for predicting the promoters. The proposed classifier is tested with datasets from Berkeley Drosophila database for drosophila and TAIR Arabidopsis thaliana database. We have achieved an average classifier accuracy more than 89.6% for Drosophila and 92.6 for Arabidopsis.

Keywords: Cellular Automata (CA), multiple attractor cellular automata (MACA), Clonal Classifier (CC), promoter.

1. Introduction
Most of the problems in bioinformatics can be addressed 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.

Fig1: Promoter Region in a DNA sequence

Fig: 1 shows the location of promoter and protein coding region in the untranslated region (UTR). The 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 fig 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. Von Neumann and Stanislaw Ulam initially proposed the model of Cellular Automata in 1940. Stephen Wolfram did a detailed study of 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 the 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 physical.

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

2. Literature Survey

Vladimir B. Bajic et al. 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, Promoter Find, Promoter Scan, TATA, TSSG, TSSW, IMC, SPANN, SPANN2 for True Positives and False Positives. Jih-Wei Hung has developed an effective forecast calculation that can expand the recognition (power =1 - false negative) of the promoter. The authors introduce two strategies that utilize the machine force to ascertain all conceivable examples which are the conceivable characteristics of promoters. The primary strategy, they 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 (Non fixed 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 TSS into attention. By the exploratory effects, our expectation has a higher right rate than different past systems.

Marshall S.Z. Horwitz et al. have chosen an assembly of Escherichia coli promoters from irregular DNA groupings by swapping 19 base sets at the -35 promoter area of the tetracycline safety gene 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 successes advertise translation more determinedly than the local promoter.

2. 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 these cells change their states according to the local evaluation function which is a function of its state and its neighboring states.
The general design of MACA [7, 8, 9] 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. The input processing unit will process sequences three at a time as three MACA basins are calculated as per the instructions of proposed algorithm and an inverter tree as in Fig 3,4 named as AIS multiple neighborhood cellular automata is considered for processing DNA sequences. The rule generator will transform the complemented and non-complemented rules in the form of a matrix, so that we can apply the rules to the corresponding sequence positions very easily. Attractor cellular automata is formed which can predict the class of the input after all iterations.

**Fig 3:** MACA-Modified Clonal Classifier Tree with basin 1, .75, 1

**Fig 4:** MACA-Modified Clonal Classifier Tree with basin 1, .25, 1
The algorithms take 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.

4. Experimental Results
Drosophila Promoter sequences are taken from Berkeley Drosophila database [11]. We have extracted 1,844 base pairs [-250, +50] around TSS from that database for training and testing. Non promoter sequences of exons, introns are taken from the same database. We have extracted 292 exons and 542 introns from the databases. All the sequences for both promoter and non-promoter region prediction are of length 251 base pairs.

Arabidopsis Promoter sequences are taken from TAIR Arabidopsis thaliana database [12]. We have extracted 306 base pairs [-200, +50] around TSS from that database for training and testing. Non promoter sequences of exons, introns are taken from the same database. We have extracted 292 exons and 542 introns from the databases. All the sequences for both promoter and non-promoter region prediction are of length 251 base pairs.

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

Using the above measures following are calculated
- Actual Positives (AP) = TP + FN
- Actual Negatives (AN) = TN + FP
- Predicted Positives (PP) = TP + FP
- Predicted Negatives (PN) = TN + FN
- Sensitivity (SN) = TP / (TP + FN)
- 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. 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 the promoter region in .7 nanoseconds for a DNA sequence of length 252.

<table>
<thead>
<tr>
<th>Table 1: Comparison of MACA-MCC-DA with existing approaches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Promoter Inspector</td>
</tr>
<tr>
<td>Dragon Promoter Finder</td>
</tr>
<tr>
<td>Promo Predictor</td>
</tr>
<tr>
<td>CNN-Promoter</td>
</tr>
<tr>
<td>IMC</td>
</tr>
<tr>
<td>MACA-Modified CC</td>
</tr>
</tbody>
</table>
5. Conclusion
We have successfully developed and tested the MACA based modified Clonal Classifier for predicting promoter regions in Drosophila and Arabidopsis. 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 0.4 micro seconds. A sensitivity of 84.5% and specificity of 92.7 were reported.

6. References