Global Journal of Computer Science and Technology: G Interdisciplinary
Volume 14 Issue 2 Version 1.0 Year 2014
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

IN-AIS-MACA: Integrated Artificial Immune System based Multiple Attractor Cellular Automata for Human Protein Coding and Promoter Prediction of 252bp Length DNA Sequence

By Pokkuluri Kiran Sree, Inampudi Ramesh Babu & SSSN Usha Devi N
University College of Engineering, JNTU, India

Abstract- Gene prediction involves protein coding and promoter predictions. There is a need of integrated algorithms which can predict both these regions at a faster rate. Till date, we have individual algorithms for addressing these problems. We have developed a novel classifier IN-AIS-MACA, which can predict both these regions in genomic DNA sequences of length 252bp with 93.5% accuracy and total prediction time of 1031ms. This classifier will certainly create intuition to develop more classifiers like this.

GJCST-G Classification: *B.6.1, B.7*

Strictly as per the compliance and regulations of:

IN-AIS-MACA: Integrated Artificial Immune System based Multiple Attractor Cellular Automata for Human Protein Coding and Promoter Prediction of 252bp Length DNA Sequence

Pokkuluri Kiran Sree [α], Inampudi Ramesh Babu [σ] & SSSN Usha Devi N [ρ]

Abstract- Gene prediction involves protein coding and promoter predictions. There is a need of integrated algorithms which can predict both these regions at a faster rate. Till date, we have individual algorithms for addressing these problems. We have developed a novel classifier IN-AIS-MACA, which can predict both these regions in genomic DNA sequences of length 252bp with 93.5% accuracy and total prediction time of 1031ms. This classifier will certainly create intuition to develop more classifiers like this.

I. Introduction

DNA contains lots 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)'' . Identifying a new promoter in a DNA sequence will lead to find a new protein. If we identify the promoter region we can extract information regarding gene expression patterns, cell specificity and development. Promoters will regulate a gene expression. Some of the genetic diseases which are associated with variations in promoters are asthma, beta thalassemia and rubinstein-taybi syndrome. Promoter sequence can be used to control the speed of translation from DNA into protein. It is also used in genetically modified foods.

In vertebrates only five percentage of the gene is made up of exons. Genes mostly will have seven to eight exons with 145 bp length at an average. Introns have 3365 bp length at an average. Promoter comprises a small percentage of entire genome. The features of promoters are different from other functional regions like exons, introns and 3'UTRs. These facts make protein coding and promoter region predictions as very difficult tasks.

Author α: Dept of CSE, Sri Vishnu Engineering College for Women, Bhimavaram. e-mail: profkiransree@gmail.com
Authorv σ: Dept of CSE, Acharya Nagarjuna University, Guntur.
Author ρ: Dept of CSE, University College of Engineering, JNTU Kakinada.

II. Literature Review

Steven Salzberg [7] has used a decision tree algorithm for locating protein coding region. This algorithm is adoptable and can handle DNA sequences of length 54,108 and 162. P.Maji [8] et al. has developed neural network tree classifier for prediction of splice junction and coding regions in genomic DNA. A decision tree named as NNTree (Neural Network Tree) is constructed by dividing the training set with their corresponding labels to recursively generates a tree. Ying Xu [9] et al. has developed an improved system GRAIL II which is a hybrid AI system which can predict the number of exons in a human DNA sequence and also supports gene modeling. This process combines edge signal like accepter, donor, translation start site detection and coding feature analysis.

Eric E Snyder [10] et al. has applied dynamic programming and neural networks for predicting protein coding regions from a genomic DNA. They have developed a program Gene Parser which first scores the DNA sequences based on exon-intron specific measures like local compositional complexity, codon usage, length distribution, 6-tuple frequency and periodic asymmetry. Edward C Uberbacher [11] et al. has proposed a method which combines some set of sensor algorithms and neural network to predict the protein coding regions in eukaryotes. The programs developed will calculate the values of seven sensors that were considered by the authors. The measures are frame bias matrix, Fickett(three periodicity) , dinucleotide fractal dimension, coding six tuple word preferences, coding six tuple in frame preferences, word commonality and repetitive six tuple word preferences.

J. Pinho [12] et al. has proposed a three state model for protein coding region prediction. Authors have considered three base periodicity property. M.Q. Zhang [13] has used quadratic discriminant analysis method named as MZEF for identifying protein coding regions in genomic human DNA. David J. States [14] at el. proposed a computer program named BLASTC which

uses sequence similarity and codon utilization for predicting the protein coding regions.

Method [8] takes more time to construct a tree for sequences of length 162. The height of the trees is also a major concern for using this algorithm with DNA sequences of more length. Method [9] suffers with less accuracy due to more error rate at classifier nodes. Methods [10], [11], [12] depends more on the statistical information. After this literature survey the concern of a new classifier is to achieve a good classifier accuracy and develop a classifier which can handle DNA sequences of length more than 162 with a fewer nodes.

Jia Zeng [15] et al. has proposed a hierarchical promoter prediction system named as SCS where they have used signal, structure and context features .Xiomeng Li [16] et al. has proposed a method PCA-HPR (Principal Component Analysis-Human Promoter Recognition) to predict the promoters and transcription sites (TSS). Sridgar Hannenhalli [17] et al. tried to enhance the accuracy of promoter prediction by combining CpG island feature with information of independent signals which are biologically motivated and these cover most of the knowledge to predict the promoter in human genome.

Shuanhu Wu et al. have proposed a method [18] for enhancing the performance of human promoter region identification by selecting most important features of DNA sequence for each different functional region.Uwe Ohler et al. have proposed a model [19] which integrates physical properties of DNA into a probabilistic eukaryotic promoter prediction system.Goni J Ramon et al. has proposed a system ProStar[20] which uses structural parameters for promoter region identification. Authors only used descriptors derived from physical first principles.

Vladimir B. Bajic [21] et al. has developed new software for identifying promoters in a DNA sequence of vertebrates. This program takes input as DNA sequence and generates a list of predicted TSS (Transcription Stating Site).Michael Q.Zhang [22] has proposed a new program for predicting a core promoter in human gene named as CorePromoter. After the literature survey on promoter prediction, the main goal of proposed classifier is to reduce the false prediction rates and improve specificity and sensitivity values.

III. Design of In-Ais-Maca

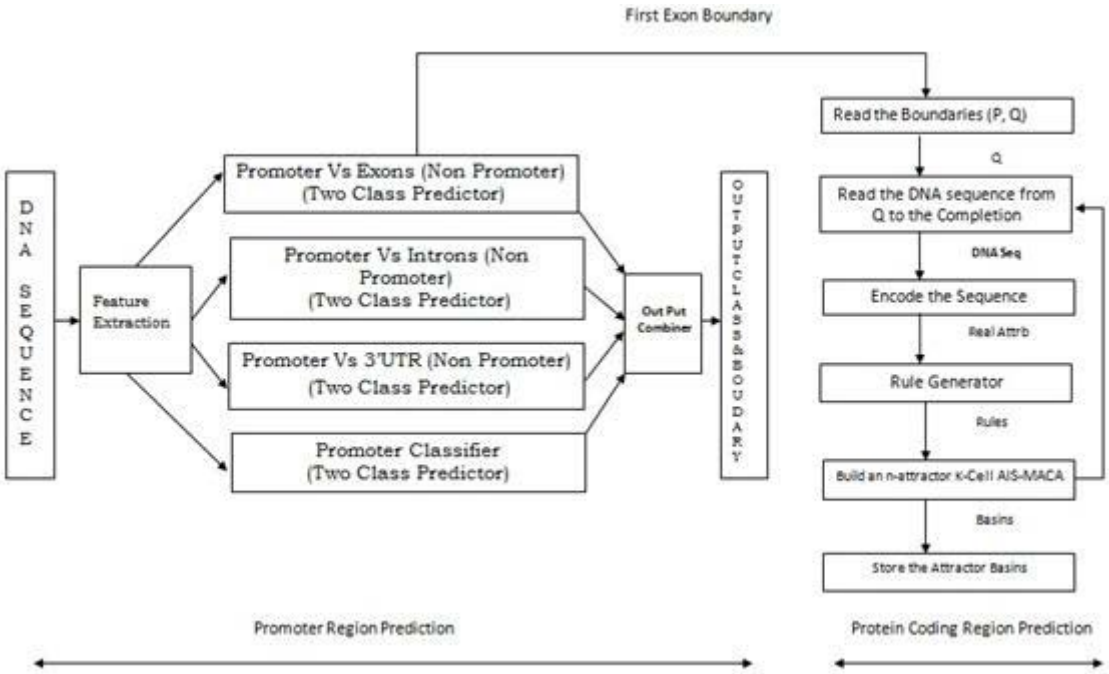

Figure 1 : IN-AIS-MACA design

IN-AIS-MACA partial design is shown in Fig: 1. IN-AIS-MACA takes a DNA sequence as input and extracts the features. Initially IN-AIS-MACA checks whether the given sequence belongs to an exon or not. If it belongs to an exon, the exact boundaries with non-promoter class will be displayed. These boundaries will be used to trace the protein coding region starting from that boundary. Since the first exon boundary is already predicted say (P, Q), this algorithm reads the encoded DNA sequence starting with Q to the end of the string say R. The IN-AIS-MACA tree is built only for a length R-Q for PCR prediction. If the input does not belong to exons then it is checked whether it is an intron or 3'UTR or a promoter. The corresponding class and boundary is displayed.

IV. Data Sets and Methods

Human promoter data sets are collected from DBTSS database consist of 30,966 of length 251. We have used 7,741 for constructing an IN-AIS-MACA tree and 7,741 for checking the accuracy of the tree. Rest of the 15,483 promoter sequences are used for testing the proposed classifier.

Human non-promoter data sets are collected from EID and UTRdb databases. We have extracted 75,438 exons from EID database, where 18,859 are used for constructing an IN-AIS-MACA tree and 18,860 data components are used for checking the accuracy of the constructed tree. Rest of 37,719 data components are used for testing the classifier. We have extracted 53,684 introns from EID database, where 13,421 are used for constructing the tree and 13,421 are used for checking the accuracy of the constructed tree, rest of the 26,842 are used for testing the classifier. We have extracted 80,538 3'UTRs from UTR dB. In that 21,134 are used for constructing the tree and 21,135 data components are used for checking the accuracy of the tree. The rest of 40,269 components are used for testing the classifier. IN-AIS-MACA allows 1bp error tolerance.

Data components for identifying the protein coding regions in human are taken from MMCRI database for length 252bp; 2,489 coding sequence examples are extracted for training. This training set is divided into 1229,1230 sets, where 1229 data components are used for building the IN-AIS-MACA tree and 1230 data components are used for testing the accuracy of the tree, 1895 data component are used for testing the coding regions.

20,002 non-coding sequence examples are extracted for training. This training set is divided into 10,000 , 10,002 sets, where 10,000 data components are used for building the IN-AIS-MACA tree and 10,002 data components are used for testing the accuracy of the tree; 15,456 data components are used for testing the non-coding regions.

DNA sequences of lengths 252 are taken from chromosome7, chromosome11 and GenBank. A total of 15,456 are extracted from the above data sets. We have extracted 1,300 examples of RSCS , where 650 examples are used for testing the classifier, 325 examples are used for constructing the tree and 325 examples are used for checking the accuracy of the constructed tree in training. We have extracted 16,456 data components from where 10,200 are used for testing the classifier, 3128 data components are used for constructing the tree and 3218 data components are used for checking the accuracy of the constructed tree in training.

No information regarding the reading frame is used in our study. We are going to predict both regions where nothing is known. Each window should belong to a single class (promoter/non-promoter, coding /non-coding). IN-AIS-MACA has created 4 best trees for predicting PCR and 8 best trees for predicting PR.

V. Learning of In-Ais-Maca

IN-AIS-MACA consists of five p state, 3 neighborhood AIS-MACA classifiers. Four classifiers are used for the predicting promoter regions and one for predicting protein coding regions. A total of 1, 43,158(1, 20,667 for promoters and 22,491 for protein coding regions) components are trained for predicting promoter and protein coding regions. IN-AIS-MACA algorithm will create five different set of trees and thirty attractor basins. This algorithm executes for five times (Exon, Intron, 3'UTR, Promoter, Protein Coding).

Algorithm:

Input: DNA Sequence

Output: Attractor Basins

Step 1: Read the DNA sequence in the multiples of three.

Step 2: Encode the sequence in the multiples of three

Step 3: Extract the features

Step 4: Construct a 3-cell, 6-attractor IN-AIS-MACA tree with 2 classes to be predicted.

Step 5: Save all the best IN-AIS-MACA trees. (Use Fitness Function)

Step 6: Store the basins (Be, Bi, Bu, Bp, Bpr).

Step 7: Repeat the steps 1 to 6 till the completion of input or individual attractor basins count is 6.

Step 8: Stop

Where Be represents the exon basins, Bi represents the intron basins, Bu represents the 3'UTR basins, Bp represents the promoter basins and Bpr represents the protein coding region basins.

VI. Testing of In-Ais-Maca

The accuracy of protein coding region prediction with IN-AIS-MACA depends on the accuracy of exon prediction. As the promoter prediction module has reported 96.5% accuracy, the protein coding region prediction accuracy gets improved. The main aim of this algorithm is to process the DNA sequence based on the features and distribute it into any one of the basin.

Algorithm:

Input: DNA Sequence

Output: Class of the sequence

Step 1: Read the DNA sequence in the multiples of three.

Step 2: Encode the sequence in the multiples of three

Step 3: Extract the features

Step 4: Check whether the input belongs to EXON class, if not, go to step 6. If it is found as EXON report the corresponding class and boundary.

Step 5: (a) Read the encoded DNA sequence starting with the upper bound to the end of the string.

(b) Choose best fitness rule to direct the sequence to the attractor basins of Bpr
(c) Report the respective class.

Step 6: Check whether the sequence belongs to intron, 3'UTR or promoter.

6a) Choose the best fitness rule to direct the sequence to the attractor basins of Bi,Bu,Bp
6c) Report the boundaries and respective class.

Step 7: Stop.

VII. Output & Experimental Results of In-Ais-Maca

The output1 shown below is a DNA sequence of length 252bp. The output of promoter prediction has indicated initial exon at 30 to 64. So the protein coding interface starts its processing from 64 to 251.The next internal and terminal exons are reported in both the strands.

Output 1:

DNA Sequence

GAATTCTTGTTGAGAAGGAATTGGGCTCAATGAAGTTCGGGGATATTCCAAGTGAATTATTCCAGTGAGTGTTATTCAG
CAATGGACGTGACTGTCGTTTGCCAGATCAGCAGAAGCCGAAAGGAATCCTTTCGGCTTCTGCTGATCTGGCAAAC
GACAGTCACGTCCATTGCTGAATAACACTCACTGGAATAATTCACTTGGAATATCCCCGAACTTCATTGAGCCCAATT
CCTTCTCAACAAGAATTC

Sequence Kiran_63jntuh Length = 252 bp

Sequence Kiran_63jntuh,			Human Promoter Prediction
Start	**End**	**Score**	**Non Promoter Sequence/Exon**
30	64	0.61	ATGAAGTTCGGGGATATTCCAAGTGAATTATTCC

Sequence Name	Program	Type of Exon	Boundary		Strand
Kiran_63jntuh	IN-AIS-MACA	First	82	189	+
Kiran_63jntuh	IN-AIS-MACA	First	82	207	+
Kiran_63jntuh	IN-AIS-MACA	First	82	222	+
Kiran_63jntuh	IN-AIS-MACA	First	198	207	+
Kiran_63jntuh	IN-AIS-MACA	First	198	214	+
Kiran_63jntuh	IN-AIS-MACA	First	198	222	+
Kiran_63jntuh	IN-AIS-MACA	First	198	226	+
Kiran_63jntuh	IN-AIS-MACA	First	198	232	+
Kiran_63jntuh	IN-AIS-MACA	First	198	207	+
Kiran_63jntuh	IN-AIS-MACA	Internal	53	222	+
Kiran_63jntuh	IN-AIS-MACA	Internal	66	87	+
Kiran_63jntuh	IN-AIS-MACA	Internal	80	199	+
Kiran_63jntuh	IN-AIS-MACA	Internal	80	207	+
Kiran_63jntuh	IN-AIS-MACA	Internal	80	222	+
Kiran_63jntuh	IN-AIS-MACA	Internal	106	132	+
Kiran_63jntuh	IN-AIS-MACA	Internal	106	207	+
Kiran_63jntuh	IN-AIS-MACA	Terminal	106	136	+
Kiran_63jntuh	IN-AIS-MACA	Terminal	106	197	+
Kiran_63jntuh	IN-AIS-MACA	Terminal	111	136	+
Kiran_63jntuh	IN-AIS-MACA	Terminal	111	197	+
Kiran_63jntuh	IN-AIS-MACA	Terminal	167	193	+
Kiran_63jntuh	IN-AIS-MACA	Terminal	167	197	+
Kiran_63jntuh	IN-AIS-MACA	Internal	151	249	-
Kiran_63jntuh	IN-AIS-MACA	Internal	151	249	-
Kiran_63jntuh	IN-AIS-MACA	Internal	130	249	-
Kiran_63jntuh	IN-AIS-MACA	Terminal	194	249	-

Kiran_63jntuh	IN-AIS-MACA	Terminal	76	249	-
Kiran_63jntuh	IN-AIS-MACA	Terminal	72	249	-

VIII. Comparison of the Performance of In-Ais-Maca

IN-AIS-MACA uses the strength of existing AIS-PRMACA design to predict both PR & PCR regions. The accuracy, Se, Sp and execution time of PR prediction with IN-AIS-MACA is same as of AIS-PRMACA reported in chapter 6. So we report the accuracy, Se and Sp of predicting PCR using this IN-AIS-MACA. The important challenge of IN-AIS-MACA is to reduce the total prediction time (TPT) of both PCR and PR which will be discussed in this section.

The performance of IN-AIS-MACA is measured with Se,Sp and accuracy as shown in table 1. We have extended the DT and NNtree to accommodate 252 length DNA sequences and compared the results with them. IN-AIS-MACA reports a high sensitivity, specificity, accuracy of 0.934, 0.925 and 0.93 respectively. This improved performance, when compared with AIS-MACA prediction for 252bp length DNA sequence is due to the classifier accuracy of AI-PRMACA.

Table 1 : IN-AIS-MACA Performance in PCR prediction

Method	Se	Sp	Se+Sp	Accuracy
IN-AIS-MACA	0.934	0.925	1.859	0.93
Decision Tree	0.851	0.879	1.73	0.865
Neural Network Tree	0.876	0.87	1.746	0.873

If the accuracy of AIS-PRMACA to predict the first exon is more, then the accuracy of predicting the PCR with IN-AIS-MACA is more. The accuracy of AIS-PRMACA prediction of exon is 94.5%, so there is a considerable improvement of PCR prediction with IN-AIS-MACA particularly in the 252bp length DNA sequences. IN-AIS-MACA maintains good balance between Se and Sp, Se+Sp ie 1.859. The performance of a decision tree in processing lengths of 252bp is poor due to the height of the tree build for predicting the PCR is more. Decision tree reports an accuracy of 86.5%. NNtree performs better compared with DT reports 87.3% accuracy. Performance of both classifiers suffers when processing a DNA sequence of length more than 162.

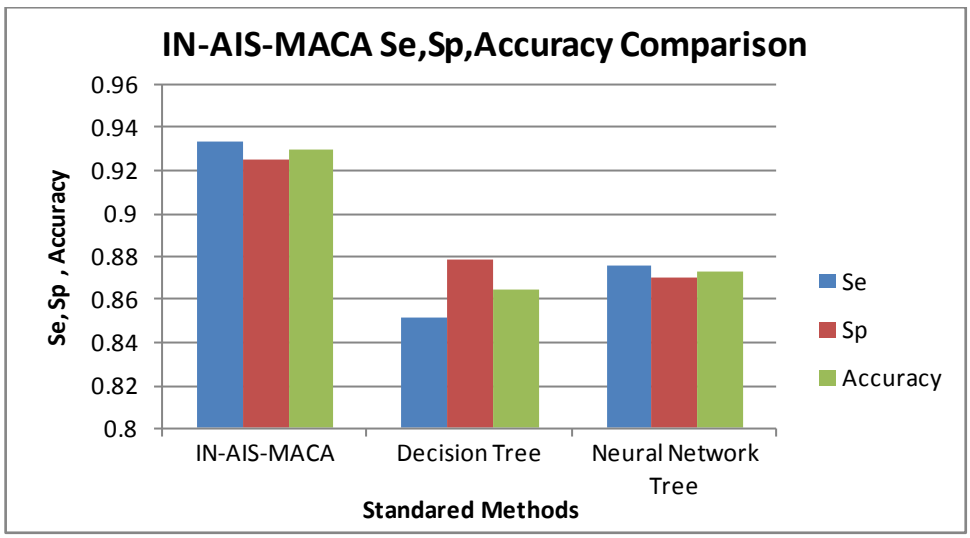

Figure 2 : IN-AIS-MACA Performance in PCR prediction VS Standard Methods

IX. Execution Time Comparisons with In-Ais-Maca

The aim of IN-AIS-MACA is to predict both PCR and PR in human DNA sequence of length 252bp. Since this is the first algorithm to handle predictions of both regions, we have chosen better algorithms in combination , to report the corresponding execution times of individual predictions and total predictions. In the first combination we have used classifiers AIS-MACA

and AIS-PRMACA which reports the total prediction time of 1827ms. In the second combination we have used classifiers AIS-MACA and SCS which reports the total prediction time of 1917 ms. In the third combination we have used classifiers AIS-MACA and McPromoter which reports the total prediction time of 1821 ms.

Table 2 : IN-AIS-MACA total prediction time comparison

Method	Execution time to predict PCR (ms)	Execution time to predict PR (ms)	Total Prediction Time(TPT) (ms)
IN-AIS-MACA	1031	1031	1031
AIS-MACA & AIS+PRMACA	796	1031	1827
AIS-MACA & SCS	796	1121	1917
AIS-MACA & McPromoter	796	1025	1821
DT & AIS-PRMACA	899	1031	1930
NNTree & AIS-PRMACA	866	1031	1897
Dicodon Usage & AIS-PRMACA	956	1031	1987

In the fourth combination we have used classifiers decision tree and AIS-PRMACA which reports the total prediction time of 1930 ms. In the fifth combination we have used classifiers NNtree and AIS-PRMACA which reports the total prediction time of 1897 ms. In the sixth combination we have used classifiers dicodon usage and AIS-PRMACA which reports the total prediction time of 1897 ms. The proposed classifier IN-AIS-MACA reports a total prediction time of 1031ms which is best among all the reported classifiers in table 2 and figure 3. Identifying both PCR and PR with a minimum execution time leads to a faster gene prediction.

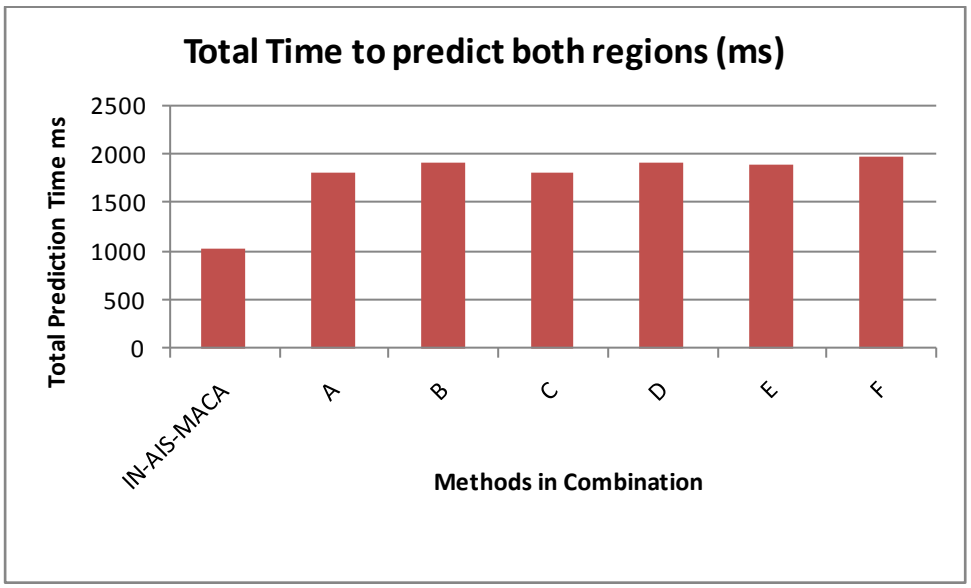

Figure 3 : IN-AIS-MACA total execution time Prediction Comparisons

A: AIS-MACA & AIS+PRMACA B: AIS MACA & SGS
C: AIS-MACA & McPromoter D: DT & AIS-PRMACA
E: NNTree & AIS-PRMACA F: Dicodon Usage & AIS PRMACA -

X. Parameters Manipulation for Higher Accuracies of In-Ais-Maca

For achieving higher accuracies with IN-AIS-MACA to predict protein coding regions and promoter regions, we have to analyze three important parameters. The first parameter is the number of generations. We have to extract higher accuracies with lesser generations. Figure 4 shows that the minimum number of generations that required to achieve a higher accuracy for PCR prediction is 75.

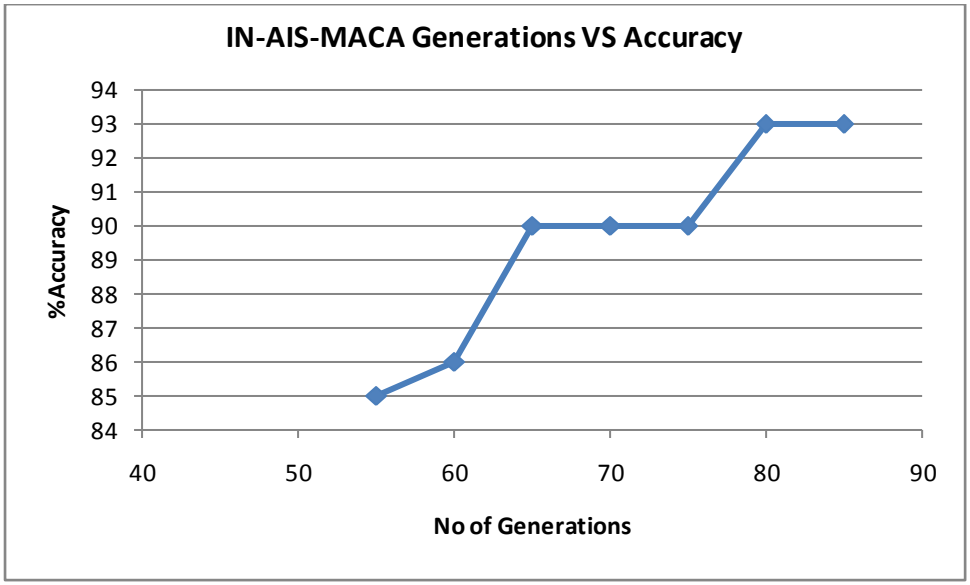

Figure 4 : IN-AIS-MACA Accuracy VS No of generations

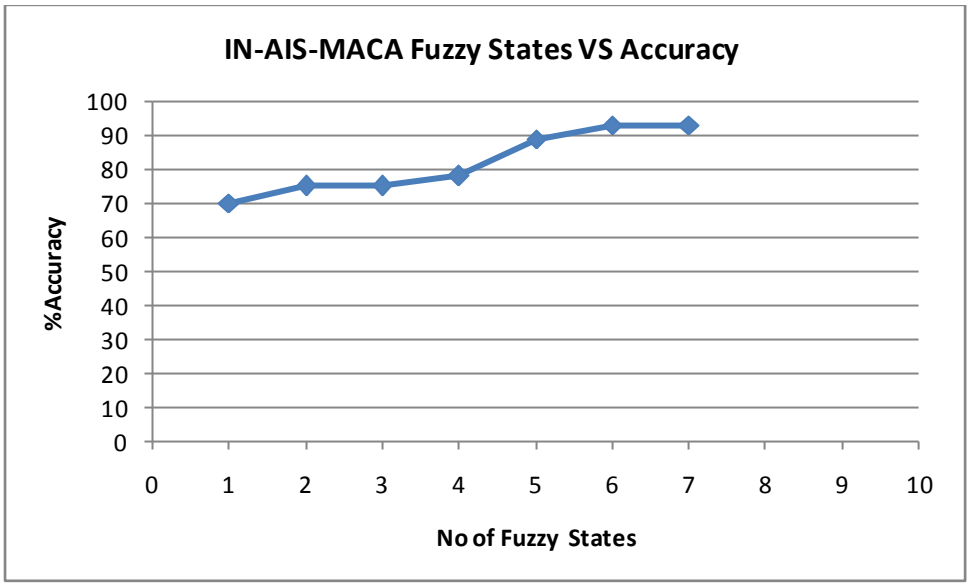

Figure 5 : IN-AIS-MACA Accuracy VS Fuzzy States

The second parameter is the number of fuzzy states. Depending on the fuzzy states also the performance of the classifier varies considerably. IN-AIS-MACA attains good accuracy with six fuzzy states as shown in figure 5. The third parameter to be considered is the clonal factor (β). The clonal factor plays an important role in attaining a higher accuracy. IN-AIS-MACA attains higher accuracy when clonal factor (β) is 0.5 as shown in fig 6.

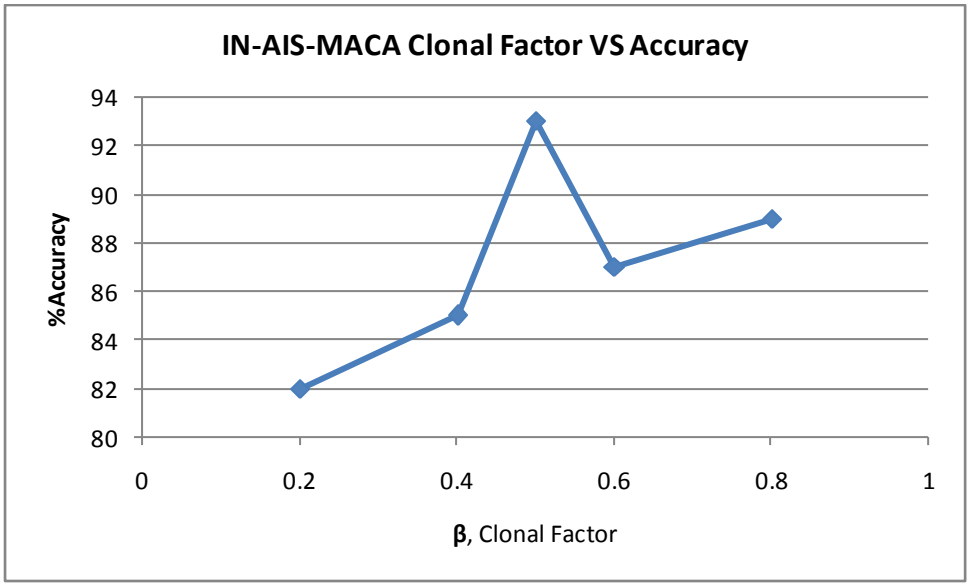

Figure 6 : IN-AIS-MACA Accuracy VS Clonal Factor

XI. Conclusion

We have successfully developed an integrated classifier which can predict both protein coding and promoter regions in human DNA of length 252bp. IN-AIS-MACA reports a Sensitivity (Se) of 0.934 ,Specificity(Sp) of 0.925 and accuracy of 93% which makes this as the best algorithm for predicting both PCR and PR. The important contribution of this classifier lies in predicting both these regions with an execution time of 1031ms, which will faster the gene perdition rate.

References Références Referencias

1. Attwood, Teresa K. "The Babel of bioinformatics." Science 290, no. 5491 (2000): 471-473.
2. Fickett, James W., and Chang-Shung Tung. "Assessment of protein coding measures." Nucleic acids research 20, no. 24, 1992:pp. 6441-6450.
3. Yamashita, Riu, Yutaka Suzuki, Hiroyuki Wakaguri, Katsuki Tsuritani, Kenta Nakai, and Sumio Sugano. "DBTSS: database of human transcription start sites, progress report 2006." Nucleic acids research 34, no. suppl 1 ,2006:pp. 86-89.
4. Saxonov, Serge, Iraj Daizadeh, Alexei Fedorov, and Walter Gilbert. "EID: the Exon–Intron Database—an exhaustive database of protein-coding intron-containing genes." Nucleic acids research 28, no. 1 ,2000:pp. 185-190.
5. Pesole, Graziano, Sabino Liuni, Giorgio Grillo, Flavio Licciulli, Flavio Mignone, Carmela Gissi, and Cecilia Saccone. "UTRdb and UTRsite: specialized databases of sequences and functional elements of 5′ and 3′ untranslated regions of eukaryotic mRNAs. Update 2002." Nucleic acids research 30, no. 1 , 2002: pp.335-340.
6. Sree, Pokkuluri Kiran, and Inampudi Ramesh Babu. "AIX-MACA-Y Multiple Attractor Cellular Automata Based Clonal Classifier for Promoter and Protein Coding Region Prediction." Journal of Bioinformatics and Intelligent Control 3, no. 1 (2014): 23-30.
7. Salzberg, Steven. "Locating protein coding regions in human DNA using a decision tree algorithm." Journal of Computational Biology 2, no. 3 ,1995:pp. 473-485.
8. Maji, Pradipta, and Sushmita Paul. "Neural Network Tree for Identification of Splice Junction and Protein Coding Region in DNA." In Scalable Pattern Recognition Algorithms, Springer International Publishing, 2014: pp. 45-66.
9. Xu, Ying, R. Mural, M. Shah, and E. Uberbacher. "Recognizing exons in genomic sequence using GRAIL II. " Genetic engineering 16 ,1993: pp. 241-253.
10. Snyder, Eric E., and Gary D. Stormo. "Identification of protein coding regions in genomic DNA." Journal of molecular biology 248, no. 1 ,1995:pp. 1-18.
11. Uberbacher, Edward C., and Richard J. Mural. "Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach." Proceedings of the National Academy of Sciences 88, no. 24 ,1991:pp. 11261-11265.
12. Pinho, Armando J., António JR Neves, Vera Afreixo, Carlos AC Bastos, and Paulo Jorge SG Ferreira. "A three-state model for DNA protein-coding regions." Biomedical Engineering, IEEE Transactions on 53, no. 11 ,2006:pp. 2148-2155.
13. Zhang, M. Q. "Identification of protein coding regions in the human genome by quadratic discriminant analysis." Proceedings of the National Academy of Sciences 94, no. 2, 1997:pp. 565-568.

14. Gish, Warren, and David J. States. "Identification of protein coding regions by database similarity search." Nature genetics 3, no. 3, 1993:pp. 266-272.
15. Zeng, Jia, Xiao-Yu Zhao, Xiao-Qin Cao, and Hong Yan. "SCS: Signal, context, and structure features for genome-wide human promoter recognition." Computational Biology and Bioinformatics, IEEE/ACM Transactions on 7, no. 3, 2010: pp.550-562.
16. Li, Xiaomeng, Jia Zeng, and Hong Yan. "PCA-HPR: A principle component analysis model for human promoter recognition." Bioinformation 2, no. 9 ,2008:pp. 373.
17. Hannenhalli, Sridhar, and Samuel Levy. "Promoter prediction in the human genome." Bioinformatics 17, no. suppl 1 2001:pp. 90-96.
18. Wu, Shuanhu, Xudong Xie, Alan Wee-Chung Liew, and Hong Yan. "Eukaryotic promoter prediction based on relative entropy and positional information." Physical Review E 75, no. 4 (2007): 041908.
19. Ohler, Uwe, Heinrich Niemann, Guo-chun Liao, and Gerald M. Rubin. "Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition." Bioinformatics 17, no. suppl 1 (2001): S199-S206.
20. Goñi, J. Ramon, Alberto Pérez, David Torrents, and Modesto Orozco. "Determining promoter location based on DNA structure first-principles calculations." Genome Biol 8, no. 12 (2007): R263.
21. Bajic, Vladimir B., Seng Hong Seah, Allen Chong, Guanglan Zhang, Judice LY Koh, and Vladimir Brusic. "Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoters." Bioinformatics 18, no. 1 ,2002:pp. 198-199.
22. Zhang, Michael Q. "Identification of human gene core promoters in silico." Genome research 8, no. 3, 1998:pp. 319-326.
23. http://www.mmchri.res.in

This page is intentionally left blank