International Journal of Genetics
ISSN: 0975-2862 & E-ISSN: 0975-9158, Volume 9, Issue 8, 2017, pp.-291-295.
Available online at http://www.bioinfopublication.org/jouarchive.php?opt=&jouid=BPJ0000226

Research Article

DISCRIMINATION OF CODING FROM NON-CODING REGIONS IN CATTLE BASED ON METHYLATION AND DEAMINATION PATTERNS

VARSHNEY NITIN*, PAUL A.K. AND WAHI S.D.
ICAR-Indian Agricultural Statistics Research Institute, Pusa, New Delhi, 110012
*Corresponding Author: Email- jadamprakash@gmail.com

Received: August 01, 2017; Revised: August 05, 2017; Accepted: August 06, 2017; Published: August 28, 2017

Abstract- Coding regions (CDS) are the part of DNA that helps in protein synthesis and non-coding region (Introns) do not code for proteins. So, discrimination of coding region from the non-coding region is of prime importance for genome annotation. In this study discrimination of Coding and Non-coding regions of cattle genome has been done based on epigenetic mechanism. The features of DNA methylation and spontaneous deamination were used in this approach. Five different indices namely Deviation of nucleotide (NUF), Deviation of dinucleotide (DNF), Intensity of methylation (IME), Triplet avoidance index (TAI), and Tendency of Poly-purine and Poly-pyrimidine (PPI) have been used in this study. These indices were used to encode exon and intron sequences. The proposed method has been compared with the methods based on LDA (Linear Discriminant Analysis) and QDA (Quadratic Discriminant Analysis) based on the area under the ROC (Receiver Operating Characteristic) and the PR (Precision Recall) curves, whereby the proposed method proved to be better as compared to LDA and QDA based methods

Keywords- Coding and noncoding regions, DNA methylation, Epigenetic mechanism, Content sensors.

Citation: Varshney Nitin, *et al*., (2017) Discrimination of Coding from Non-Coding Regions in Cattle Based on Methylation and Deamination Patterns**.** International Journal of Genetics, ISSN: 0975- 2862 & E-ISSN: 0975-9158, Volume 9, Issue 8, pp.-291-295.

Academic Editor / Reviewer: Prof Ramesh S. Bhat

Introduction

The whole transcriptional output from the genome can be roughly divided into two major parts: protein coding m-RNAs and non-coding RNAs [1]. The coding region of a gene is the portion of DNA or RNA that consists of exons. This region is also recognised as the coding DNA sequence (CDS). Coding region have a vast importance in molecular biology as it contains the information about protein synthesis and non-coding regions are the portion of DNA that do not have any information about protein synthesis. Some non-coding DNA is transcribed into functional non-coding RNA molecules like t-RNA, r-RNA and regulatory RNAs. A large fraction of non-coding DNA had no known biological function and referred as junk DNA or introns. An important step in genomic annotation is to discriminate coding regions from non-coding regions of a genome and this is a challenging problem especially in eukaryotic genome. In eukaryotic genomes, protein coding regions called, as exons, are usually not continuous, but are flanked by non-coding regions called introns. Due to lack of obvious sequence features between exons and introns, effective distinction of protein coding regions from non-coding regions is a pernicious problem. Moreover, attempts have been made in the past to discriminate coding regions from non-coding regions but not much work has been done through characterization of epigenetic mechanisms. Epi-genetics is the mechanism which explains the interaction between the genes and there surrounding within a multicellular organisms to produce a phenotype without changes in the genotype. The fact that the differentiated cells in a multicellular organism express only those genes that are absolutely essential for its own activity, can only be explained by epigenetics. Higher organisms, as well as plants, use three systems to initiate and maintain epigenetic gene regulation.

i. DNA methylation
ii. Histone modification
iii. RNA-interference.

Hence, these mechanisms are extremely important all higher organisms. Earlier works have not used epigenetic mechanisms for discrimination of coding and non-coding sequence. Hence in this study an attempt has been made to discriminate the coding and non-coding sequences using epigenetic gene regulation mechanisms. Determination of coding (exonic) and non-coding (intronic and inter-genic) regions are important to predict gene structure as well as to identify the protein coding regions of the gene. Epigenetic mechanisms [2] like DNA methylation play a key role in gene regulation in many vertebrate species. It is a biochemical process particularly occurs in vertebrate genomes, with its main function being tissue specific gene regulation. DNA methylation and spontaneous deamination induces differential substitution patterns in coding and non-coding sequences. For example, nucleotide frequencies at three different positions of the codon are often different between coding and non-coding regions. DNA methylation is a heritable epigenetic enzymatic modification resulting the addition of a methyl group (CH_3) at fifth carbon position (C_5) of cytosine. In prokaryotes, the site of DNA methylation can be cytosine and adenine bases but in multicellular eukaryotes, the DNA methylation confined to cytosine bases [3].

Deamination is the process in which an amine group is removed from a molecule. Deaminases enzymes catalyze the process of deamination. In the situation of excess protein intake, amino acids are broken down in the process of deamination. The amino group is removed from the amino acid and converted into ammonia. Spontaneous deamination is hydrolysis reaction in which Cytosine is converted into Uracil [Fig-1] and ammonia is released.

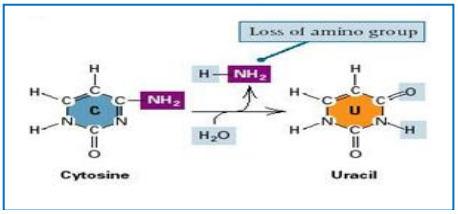


Fig-1 Conversion of Cytosine to Uracil due to deamination

This can take place invitro using bisulfite which covert only cytosine but not 5-methyl cytosine. Spontaneous deamination of 5-methyl cytosine results in thymine [Fig-2] and it is the single nucleotide mutation.

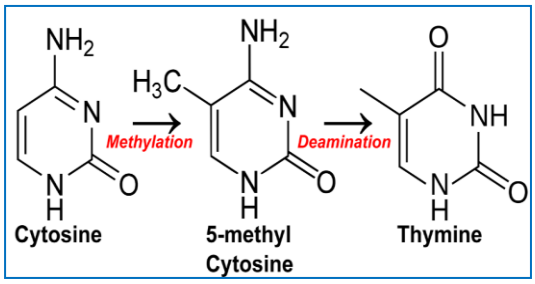


Fig-2 Conversion of Cytosine to Thymine due to methylation and deamination

In coding sequences, various mutations and selection processes create heterogeneity in nucleotide frequencies at three different positions in a codon. For example, DNA methylation and spontaneous deamination tend to change NCG codons (N stands for any of the four nucleotides) to NTG and NCA codons, where the former change is non-synonymous and the latter is synonymous. Because, non-synonymous substitution [4] is generally deleterious, their occurrence is rare as compared to the synonymous substitutions and hence NCG→NCA mutations are more often than NCG→NTG mutations. This tends to increase the frequency of A at the third codon position. Following the similar arguments as in case of nucleotide frequencies at triplet positions, the dinucleotide frequencies at triplet positions (1, 2), (2, 3) and (3, 1) are expected to be like each other in non-coding region but different in coding region.

DNA methylation and spontaneous deamination [5] decrease the CG containing triplets and increase the UG and CA containing triplets i.e., CGN→TGN, CGN→CAN and NCG→NTG. However, their effect is stronger in non-coding region than in the coding. Since, NCG→NTG mutations are non-synonymous, they tend to be selected more in non-coding sequences. Therefore, the intensity of methylation effects is expected to be larger in non-coding region than in coding region. Based on DNA methylation and spontaneous deamination decrease, Xia (2005) developed five different indices [6] for discriminating coding regions from noncoding regions in Human genome. In this study, we used these five indices for discriminating coding region from non-coding region in cattle genome.

Materials and Methods

In the present study, intron sequences of cattle have been collected from UCSC genome browser, available at https://genome.ucsc.edu. Whereas, the CDS (coding DNA sequences) sequences of cattle has been collected from http://asia.ensembl.org. The collected intron and CDS sequences were put in six categories i.e., L_a, L_b, L_c, L_1, L_2, L_3 depending upon their length. The summary of the dataset is provided in [Table-1].

Table-1 *Different classes of sequences and corresponding the number of CDS and Introns*

Category	Classes	CDS	Introns
L_a	200-500 bp	2452	10,650
L_b	500-1000 bp	6526	12,134
L_c	1000-2000 bp	7843	13,930
L_1	2000-3000 bp	2750	7200
L_2	3000-4000 bp	1135	4120
L_3	4000-5000 bp	505	2735

Methodology

The methodology for discrimination of coding region from non-coding region of cattle genome involves two steps approach:

i. Extraction of features for both coding and non-coding sequences.
ii. Classification of the data set with extracted features by using classificatory techniques.

Since CDS and intron are often short in vertebrate genome, it is required to extract the feature that differs substantially among CDS and intron. In this study, we used five different indices (Deviation of nucleotide, Deviation of dinucleotide, Intensity of methylation effect, Triplet avoidance index, Measure of tendency of Poly-purine and Poly-pyrimidine [7]) to encode coding and non-coding regions. These indices have been developed by Xia (2005) to capture differential substitution patterns in coding and non-coding sequences based on DNA methylation and Spontaneous deamination.

Proposed Methodology

Let π_1 is the population of CDS and π_2 is the population of intron sequences. 80% of sequences in CDS and intron population were considered as training set and remaining as test set. Based on the performance of five indices in CDS training set, dispersion matrix was estimated. PCA was performed and PC scores were computed for all the training sequences in CDS population.

Similarly, PC scores were computed from the training set of intron population. For classifying an unknown sequence (test instance) into CDS or Intron population, initially PC scores for it were calculated from the derived PCs of CDS as well as intron populations. Two sets of Euclidian distances were computed:

(i) First set having *N* distances where each distance represents the distance between test instance PC scores and the individual PC scores in the CDS training set
(ii) Second set having *M* distances where each distance represents the distance between test instance PC scores and the individual PC scores in the intron training set.

The mean of first set of distances is denoted as $\bar{d}_1$ and that of second set is denoted as $\bar{d}_2$.

Classification Rule

$$f(\mathbf{x_0}) \begin{cases} \bar{d}_1 > \bar{d}_2 & \mathbf{x_0} \in \pi_2 \\ \bar{d}_1 < \bar{d}_2 & \mathbf{x_0} \in \pi_1 \end{cases}$$

Criteria of comparison with five-fold cross validation

In the present study, we used mainly two criteria of comparison of proposed method with the methods based on LDA and QDA i.e. AUC-ROC and AUC-PR. Prediction assessment is essential for evaluating the prediction accuracy as well as the generalization abilities of statistical predictors. ROC analysis [8] is an effective and widely used method to assess the performance of a predictors or classifiers. In the present study, we used ROC curve and estimate of AUC-ROC to assess and compare the prediction accuracy of different classifiers. ROC curves are drawn by taking false positive rate (α) in the ordinate and true positive rate (1-β) in the abscissa for different cutoff values. The AUC-ROC was computed using trapezoidal integration as suggested by Bradley (1997) and the formula is given by

$$AUC = \sum_i \left\{ (1-\beta_i).\Delta\alpha + \frac{1}{2}\left[\Delta(1-\beta).\Delta\alpha\right] \right\},$$

Where, $\Delta(1-\beta) = (1-\beta_i) - (1-\beta_{i-1})$, $\Delta\alpha = \alpha_i - \alpha_{i-1}$ and i=1,2, ..., N (number of test instances). To compare any two ROC curves statistically, the Standard Error (SE) of AUC-ROC is computed using the formula

$$SE = \sqrt{\{\theta(1-\theta) + (N^{CDS}-1)(Q_1-\theta^2) + (N^{INTRON}-1)(Q_2-\theta^2)\}/\{N^{CDS}.N^{INTRON}}$$

where, $Q_1 = \theta/(2-\theta)$, $Q_2 = 2.\theta^2/(1+\theta)$; $N^{(CDS)}$, $N^{(INTRON)}$ and θ are the number of CDS.

Results and Discussion

The proposed method of discrimination is to use the principal component analysis

based on the five indices to discriminate the coding and the non-coding sequences respectively.

Classes of sequences and the number of CDS and introns

The genome has been classified into mainly six categories based on the length of the genome, measured in base pairs (bp) *viz.* L_a, L_b, L_c, L_1, L_2, L_3 and the number of coding DNA sequences (CDS) as well as the number of introns has also been worked out for all the six categories of the sequences an exhaustive list of all the six categories of the sequences along with the number of CDS and introns contained in them has been illustrated in [Table-1].

From the [Table-1], it is very well evident that large number of CDS and introns are present between 1000-2000 bp (L_c) i.e. the numbers of CDS are 7843 and introns are 13,930. On the other hand, small number of CDS and introns are present between 4000-5000 bp length (L_3) category i.e. 505 and 2735, respectively. To visualize the distribution pattern of CDS and introns, histogram was plotted and is presented in [Fig-3, 4].

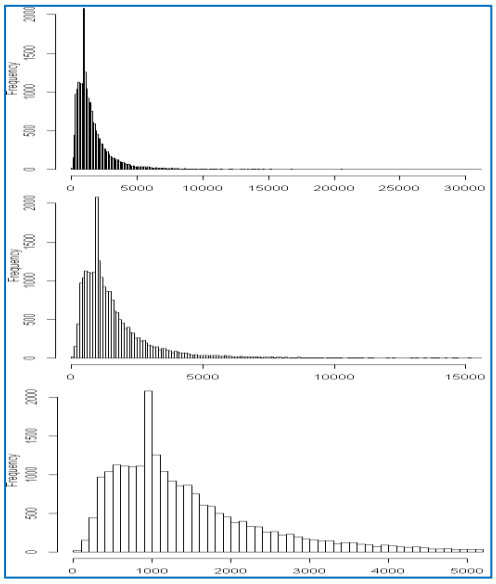


Fig-3 Frequency distribution of CDS with respect to length.

From [Fig-3], it is seen that most of the sequences are lying above 200 bp. Further it is observed that almost all the sequences are lying between 200-5000 bp and a very small number of sequences are lying above 5000 bp length. Similarly, for intron, most of the sequences are lying below 2000 bp and the numbers of sequences are decreases with increase in the length of sequences [Fig-4].

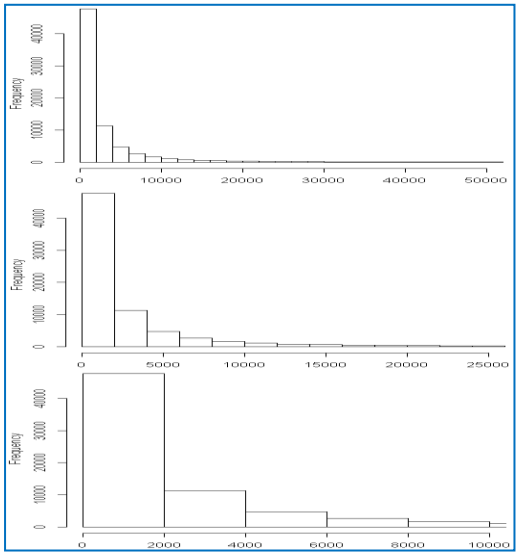


Fig-4 Frequency distribution of introns with respect to length.

Analysis of features of the five indices

Histograms was plotted for all the five indices for the CDS as well as the intron sequences for the three classes of sequences *viz.* L_1, L_2, L_3 to evaluate out of the five indices *viz.* Deviation of nucleotide (NUF), Deviation of dinucleotide (DNF), Intensity of methylation (IME), Triplet avoidance index (TAI), and Tendency of Poly-purine and Poly-pyrimidine (PPI) the one that is more abundant in CDS and the one that is more abundant in the intron sequences. The Histogram for the same has been illustrated in [Fig-3].

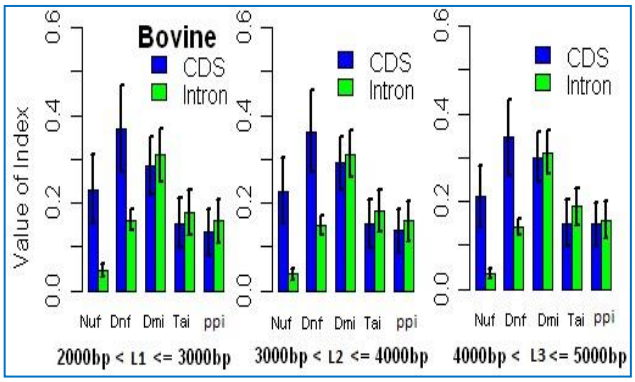
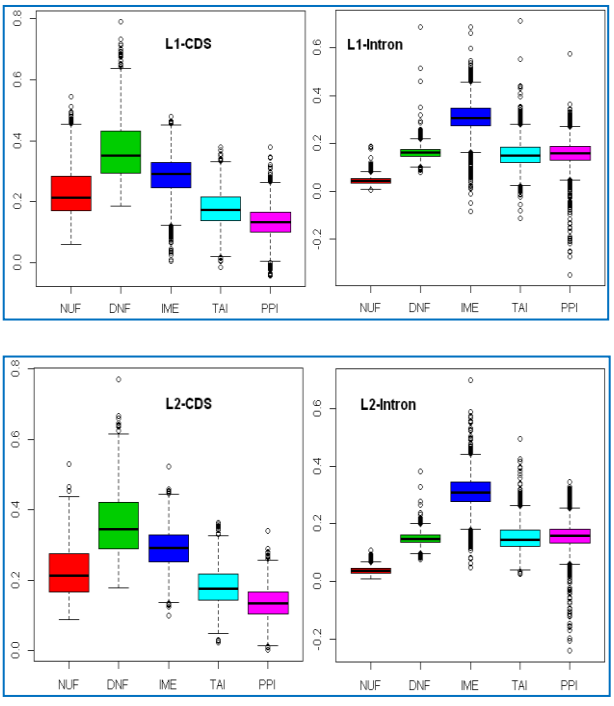


Fig-5 Features of Five Indices

It is very well evident from [Fig-5] that the first two indices *i.e.* Deviation of nucleotide, Deviation of dinucleotide are always much more abundant in the CDS as can be observed for all the three classes of sequences, whereas the remaining three indices *i.e.* Intensity of methylation, Triplet avoidance index, and Tendency of Poly-purine and Poly-pyrimidine are relatively much more abundant in the intron sequences for all the three sequences respectively. The above result satisfies the theory of five indices described in Xia (2007) [9]. We can also conclude that standard error is minimum in the measure of deviation of nucleotide for all the three categories of sequence length.

Variability in the CDS and the intron sequences

Box-plots has been plotted for both the CDS as well as the introns for all the three categories *viz.* L_1, L_2, L_3. In order to examine the variability existing among the CDS and the intron sequences. The Box-plot figures for the CDS and intron sequences have been illustrated in [Fig-6].



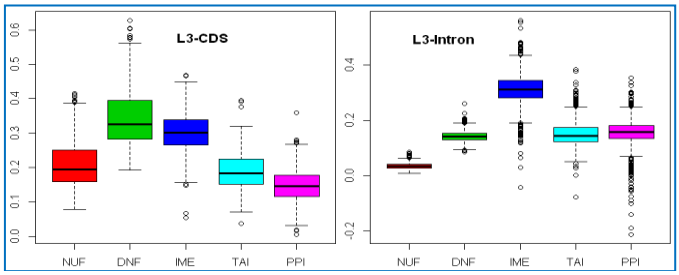


Fig-6 Features of five Indices by Box-plot diagram for different length category

From the box-plot figures it can very well be observed that the variability in the CDS for all the three categories of length L_1, L_2, L_3 is much higher than the corresponding sequences for the same length of introns based on the five indices. The variability in the CDS for the five indices are in the order of DNF > NUF> IME > TAI > PPI, whereas for the intron, it is in the order of IME > TAI > PPI > DNF > NUF. It can be further observed that the numbers of outlying observations are large in introns as compare to the CDS, in all the three length categories.

Area under ROC and PR curve for different length categories

The estimate of the total area under the Receiver operating curve (ROC) and the Precision recall curve (PR) has been obtained for all the six categories of the sequences *viz.* L_a, L_b, L_c, L_1, L_2, L_3. The results of the area under curve for the ROC and the PR curves for the sequences plus minus their standard error has been illustrated in [Table-2].

Table-2 Estimate of AUC-ROC and AUC-PR for proposed approach for different length category

Category	AUC-ROC ± S.E.	AUC-PR ± S.E.
L_a	83.94±.5455	89.47±.4515
L_b	96.66±.2373	98.55±.1558
L_c	99.59±.0633	99.8±.0440
L_1	99.93±.0343	99.89±.0439
L_2	99.99±.0159	99.98±.0216
L_3	99.99±.0071	99.99±.0116

From the above table, it is very well evident that as the length increases the area under the curve for both ROC as well as PR curve increases and reaches almost 99.99% in case of L_3 with smaller standard error among all the categories. Standard error decreases with the increment in the length of the sequence. Here minimum standard error is for L_3 category *i.e.*, 0.071 in ROC curve and 0.116 in PR curve. Moreover, it is observed that except first two length categories *i.e.*, L_a, L_b, the values of AUC-ROC and AUC-PR are more than 99%, this implies that the proposed approach can achieve higher accuracy with increase in the length of the sequence. Further, it is analyzed that the proposed approach can achieve >90% accuracy for the sequence length of ≥500 bp.

Comparison of the proposed approach with the approaches based on LDA and QDA

The proposed approach of discriminating the coding and the non-coding regions based on the five indices defined earlier has been compared with the approaches based on Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis based on the area under curve for the both ROC as well as PR curves. The estimates of AUC-ROC and AUC-PR for LDA, QDA and Proposed approach are obtained based on 5-fold cross validation, are given in [Table-3].

Table-3 *Comparison of proposed approach with LDA and QDA*

		L_1	L_2	L_3
LDA	AUC-ROC	99.84±0.0526	99.91±0.0627	98.91±0.1275
	AUC-PR	99.89±0.0430	99.96±0.0401	99.96±0.0299
QDA	AUC-ROC	99.84+0.0533	99.85+0.0802	99.85+0.1350
	AUC-PR	99.79+0.0610	99.73+0.1079	99.50+0.2311
PROPOSED	AUC-ROC	99.93±.0343	99.98±.0159	99.99±.0021
	AUC-PR	99.89±.0439	99.97±.0216	99.99±.0016

From above table, both AUC-ROC and AUC-PR of proposed approach are higher than that of LDA and QDA in all the three length categories. Further it is observed that the estimate of AUC-ROC and AUC-PR of LDA are higher than that of QDA in all the three length categories. It can further be seen that the estimate of AUC-ROC for LDA are higher in L_2 (99.91±0.0627) and the estimate of AUC-PR is higher in L_3 (99.96±0.0299). In QDA, the estimates of AUC-ROC were higher for L_2 (99.85±0.0802) and the AUC-PR is higher in L_1 (99.79±0.0610). However, both estimate of AUC-ROC and AUC-PR for the Proposed approach are higher in L_3 i.e. 99.99±.0021 and 99.99±.0016 respectively. This implies that in LDA and QDA, the accuracy is not increased with increase in the length of sequence which is in contrary to the proposed approach where the accuracy increases with increase in the length of sequence. Moreover, it is observed that the accuracy was > 99% for LDA, QDA and Proposed approach in all the three length categories with exception in AUC-ROC for L_3 in LDA. So, it can be concluded that the proposed approach can achieve higher accuracy as compared to LDA and QDA for the classification of CDS and introns by using the features based on DNA methylation and Spontaneous deamination. The classification accuracy can also be visualized from the ROC and PR curves of LDA, QDA and the proposed approach. Further the corresponding ROC and PR curves are plotted and are shown in the [Fig-7, 8, 9] for LDA, QDA and Proposed approach respectively.

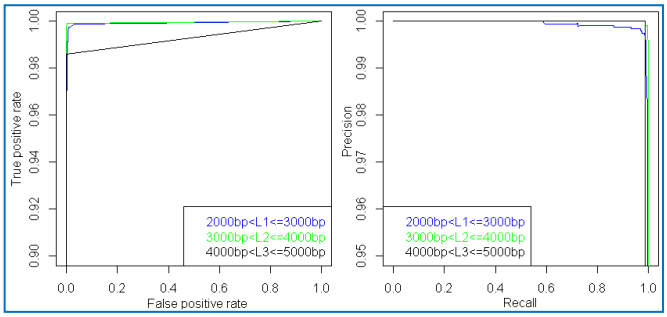


Fig-7 ROC and PR curve for LDA

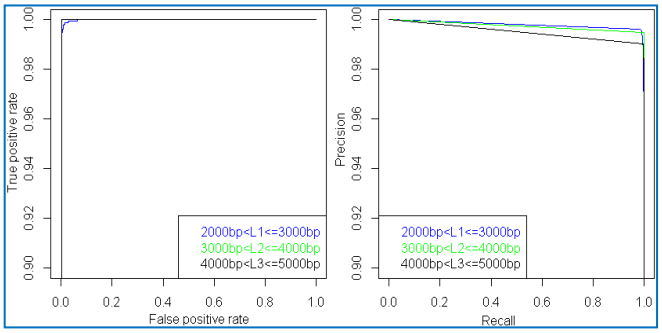


Fig-8 ROC and PR curve for QDA

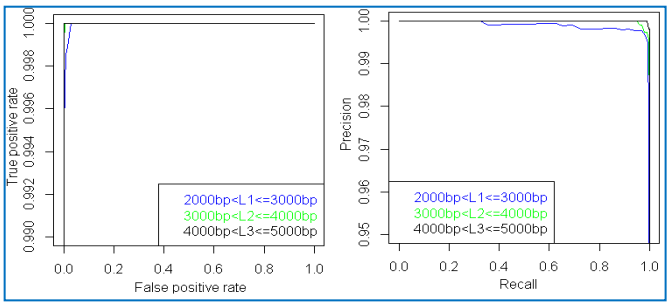


Fig-9 ROC and PR curve for proposed approach

Conclusion

Livestock sector plays a key role in the Indian economy which has 4.11% share in the GDP (Source: Central Statistics Office, Govt. of India). Cattle (*Bos taurus*) is the first livestock species whose whole genome was sequenced. An important

step in genomic annotation is to discriminate the coding regions (CDS) from the non-coding regions (Introns) of a genome and this is a challenging problem especially in eukaryotic genome. In this study, the coding and non-coding regions of cattle genome were classified through characterization of DNA methylation and spontaneous deamination using five indices and evaluation of the performance of the proposed method was compared with methods based on other classificatory techniques like those using linear discriminant analysis and quadratic discriminant analysis using suitable criteria of area under curves like ROC Curve and PR curve. Hence, it may be concluded that the proposed method performs better and is more suitable for discriminating the coding and non-coding regions in cattle based on Epigenetic mechanism and it shows higher classification accuracy than the traditional LDA and QDA based approaches. The results obtained from the present investigation regarding the classification of coding and non-coding regions corroborates the results obtained in case of humans [10].

Abbreviations

CDS- Coding DNA sequence
NUF- Deviation of nucleotide
DNF- Deviation of dinucleotide
IME- Intensity of methylation
TAI- Triplet avoidance index
PPI - Tendency of Poly-purine and Poly-pyrimidine
LDA- Linear Discriminant Analysis
QDA- Quadratic Discriminant Analysis
ROC- Receiver Operating Characteristic
PR- Precision Recall
AUC-ROC- Area under Curve- Receiver Operating Characteristic
AUC-PR- Area under Curve- Precision Recall
SE- Standard Error

Acknowledgment / Funding resource: This research work was conducted as the Post-Graduation dissertation. All the laboratory facility and software were provided by ICAR-Indian Agricultural Statistical Research Institute, New Delhi, is thankfully acknowledged.

Author Contributions: All author equally contributed

Ethical Approval

This article does not contain any studies with human participants or animals performed by any authors.

Conflict of Interest: None declared

References

[1] Fickett J.W. (1982) *Nucleic Acids Research,* 10(17), 5303-18.
[2] Russo V.E.A., Martienssen R.A. and Riggs A.D. (1996) *Cold Spring Harbor Monograph*, 32.
[3] Wilson G.G. and Murray N.E. (1991) *Annual Review of Genetics*, 25, 585-627.
[4] Xia X. (1998) *Molecular Biology and Evolution*, 15, 336-344.
[5] Xia X. (2004) *In Fourth International Conference on Bioinformatics of Genome Regulation and Structure, 1, Novosibirsk, Russia, IC&G, Novosibirsk,* 216-220.
[6] Xia X. (2005) *Bioinformatics of Genome Regulation and Structure II, Springer*, 21-29.
[7] Birnboim H.C., Sederoff R.R. and Paterson M.C. (1979) *European Journal of Biochemistry*, 98, 301-307.
[8] Bradley A.P. (1997) *Pattern recognition*, 30(7), 1145-1159.
[9] Xia X. (2007) *Bioinformatics and the Cell: Modern Computational Approaches in Genomics, Proteomics and Transcriptomics.* Springer Science & Business Media, U.S.
[10] Zhang M.Q. (1997) Identification of protein coding regions in the human genome by quadratic discriminant analysis, *Proceedings of the National Academy of Sciences of the United States of America, PNAS*, 94, 565-68.