Computers and Electronics in Agriculture 129 (2016) 66–73

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Discriminating coding from non-coding regions based on codon structure and methylation-mediated substitution: An application in rice and cattle

Prabina Kumar Meher [a], Tanmaya Kumar Sahu [b], A.R. Rao [b,*], S.D. Wahi [a]

[a] Division of Statistical Genetics, ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110012, India
[b] Centre for Agricultural Bioinformatics, ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110012, India

ARTICLE INFO

Article history:
Received 30 November 2015
Received in revised form 17 September 2016
Accepted 22 September 2016

Keywords:
Content sensors
DNA methylation
Coding region
Random forest
DCDNC

ABSTRACT

Coding regions are the fragments of DNA sequence that codes for protein through the process of transcription and translation respectively. On the other hand, the non coding regions do not give rise to any protein. Discrimination of coding regions from the non coding regions is essential for genome annotation. In this study, an attempt has been made to develop a random forest based computational approach for discriminating coding regions (CDS) from non-coding regions (introns). The features based on codon structure and methylation mediated substitutions were used in this approach. The developed approach achieved high classification accuracy, while tested on two agriculturally important species *i.e.,* rice and cattle. The proposed approach is believed to complement the other prediction methods. Based on the proposed approach, an online prediction server 'DCDNC' has also been developed for easy prediction by the users. The prediction server is freely available at http://cabgrid.res.in:8080/DCDNC.

1. Introduction

Coding sequence (CDS) is the part of combined exons in mature mRNA that excludes the un-translated regions (John et al., 2014). This is the sequence that is fully translated to a complete protein. Hence, the importance of CDS cannot be ignored as they codes for the most functional components of the organism that is proteins. Thus, in eukaryotes, distinguishing protein-coding from non-protein coding sequence is the first and most crucial step in gene prediction and genome annotation (Washietl et al., 2011). Several computational approaches have been developed for discriminating CDS from non coding sequence (introns). Most of these approaches are based on statistical methods that required training dataset from known coding and non-coding sequences to compute prediction functions (Yin and Yau, 2007).

DNA methylation and spontaneous deamination occur in coding region, which lead to the substitution and deletion of certain nucleotides. For example, in DNA methylation and spontaneous deamination, the NCG codon changes to NTG and NCA codon, where the former change is non-synonymous and the latter one is synonymous. Since, synonymous substitutions occur more frequently than non-synonymous in coding regions, NCG → NCA mutations are supposed to happen more often as compared to NCG → NTG mutations (Xia and Li, 1998). Based on the concept of DNA methylation and spontaneous deamination, Xia (2005) developed five different indices and used them for distinguishing CDS from intron in human chromosome 22, by linear discriminant analysis (LDA; Yu and Yang, 2001). Since, supervised learning approaches have been provided higher accuracies than that of classical approaches, the predictive ability of the machine learning based approaches need to be tested for discrimination of coding and non-coding sequences. Among supervised learning techniques, application of random forest (RF; Breiman, 2001) for prediction purposes is widely seen in biological studies. Dehzangi et al. (2010) demonstrated that the RF achieved high prediction accuracy as well as reduced the time consumption of the prediction task in

Abbreviations: CDS, coding sequence; LDA, linear discriminant analysis; QDA, quadratic discriminant analysis; RF, random forest; ROC, receiving operating characteristics; PR, precision recall; AUC-ROC, area under ROC curve; AUC-PR, area under PR curve; TPR, true positive rate; TNR, true negative rate; Acc, accuracy; Nuft, nucleotide frequencies by triplet sites; Dnft, dinucleotide frequencies by triplet sites; Dmi, differential methylation intensity; Tai, triplet avoidance index; Popi, polypurine and polypyrimidine index; TP, true positive; TN, true negative; FP, false positive; FN, false negative.

* Corresponding author.
E-mail addresses: meherprabin@yahoo.com (P.K. Meher), tanmayabioinfo@gmail.com (T.K. Sahu), rao.cshl.work@gmail.com (A.R. Rao), sdwahi@iasri.res.in (S.D. Wahi).

http://dx.doi.org/10.1016/j.compag.2016.09.013
0168-1699/

protein folding. Khalilia et al. (2011) used RF to predict disease risk for eight disease categories and showed that the RF outperformed SVM, Bagging and Boosting classifiers. Moreover, most of the studies on discrimination of coding from non-coding regions are based on human data, and hence it is required to do this job in agricultural species also.

In this study, we have developed a computational approach for discriminating CDS from intron. For each CDS and intron, a feature vector of length 5 was obtained based on five different indices developed by Xia (2005). By treating CDS and intron as positive and negative instances, a binary classifier was constructed; where RF supervised learning approach was employed for classification. The proposed approach was tested on two agricultural species *i.e., Bos taurus* (Cattle) and *Oriza sativa* (Rice). The proposed approach achieved high prediction accuracy in terms of estimates of area under receiving operating characteristic (ROC; Fawcett, 2006) curve (AUC-ROC) and area under precision-recall (PR; Powers, 2011) curve (AUC-PR; Davis and Goadrich, 2006). Based on the proposed approach, an online web interface has also been developed for easy classification of CDS and intron by the user.

2. Materials and methods

2.1. Collection and processing of intron and CDS sequences

The CDS sequences of cattle were collected from http://asia.ensembl.org/, whereas the intron sequences were obtained from UCSC genome browser (https://genome.ucsc.edu/). In rice, both intron and CDS sequences were collected from FTP site of Rice Genome Annotation Project (ftp://ftp.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_7.0/). The collected intron and CDS sequences were divided into four categories *i.e.,* L_1, L_2, L_3 and L_4, depending upon their length. Summary of the dataset is given in Table 1.

2.2. Feature extraction

Each CDS and intron sequence was transformed into a numeric vector (of length five) based on five different indices *i.e.,* Nucleotide frequencies by triplet sites (Nuft), Di-nucleotide frequencies by triplet sites (Dnft), Differential methylation intensity (Dmi), Triplet avoidance index (Tai) and Polypurine and polypyrimidine index (Popi). For computing the values of these indices we have written code in R-programming language. The formulae for computing these indices are provided in Supplementary file 1. However, for a detailed study on these indices one can refer Xia (2005).

2.3. Random Forest

RF is an ensemble of several classification and regression trees (CART; Breiman et al., 1984), where each one is constructed upon a bootstrap sample of the original training data. CARTs are binary classification trees that are constructed by splitting the data into daughter nodes repeatedly, starting with the root node that contains the whole learning sample (Breiman et al., 1984). In each tree-based classifier of RF, searches are made across a randomly selected subset of input variables (m) out of p variable to determine the split (Khalilia et al., 2011). Further, each tree in RF casts a vote for any test instance and the output is determined by majority voting scheme. RF can handle high dimensional data, robust to noise and use a large number of trees in the ensemble (Breiman, 2001).

Table 1
Summary of CDS and intron sequences collected from public domain.

Length of sequence	CDS		Intron	
	Rice	Cattle	Rice	Cattle
1000 bp < $L_1 \leqslant$ 2000 bp	12,000	7840	12,000	13,910
2000 bp < $L_2 \leqslant$ 3000 bp	8130	2750	3325	7200
3000 bp < $L_3 \leqslant$ 4000 bp	1625	1135	330	4120
4000 bp < $L_4 \leqslant$ 5000 bp	3310	505	985	2735

bp – base pair.

2.4. Optimum parameter setting in Random Forest

There are two parameters *i.e.,* number of classification tress (*ntree*) and number of input variables to be chosen at random out of total variables at each node for splitting (*mtry*) need to be optimized to get the optimum classification accuracy (Breiman, 2001). We used 50% of dataset for each species under each length category to optimize *mtry* and *ntree*. Initially, *ntree* was kept as 500 (default) and optimum value of *mtry* was determined out of 5 different *mtry i.e.*, *mtry* = 1, *mtry* = 2, *mtry* = 3, *mtry* = 4 and *mtry* = 5 (p) on the basis of lowest Out-of-Bag (OOB) error rate. Then, the optimum *ntree* was determined on the basis of stable OOB error rate by keeping *mtry* as "optimum *mtry*".

2.5. Implementation and validation

The final classification was made using full dataset through fivefold cross validation procedure (Henderson et al., 1992). For 5-fold cross validation, CDS and intron sequences were divided into five subsets, each subset containing approximately equal number of instances. Then, five sets were created with each set containing a subset of CDS and a subset of intron. Four out of five sets were used for training and the remaining one set was used for testing. This step was repeated five times in such a way that each set was used once for testing. For implementation of RF, *randomForest* package (Liaw and Wiener, 2002) of R-software was used, where the function *randomForest()* was used to execute RF classifier. The RF model was trained with *mtry* = "optimum mtry", *ntree* = "optimum ntree" as explained in Section 2.4. A flow diagram showing the steps involved in the proposed approach is shown in Fig. 1.

2.6. Performance measure

The estimates AUC-ROC and AUC-PR were used to assess the performance of the classifier. Since AUC-ROC is independent of class ratios, the AUC-PR provides a better measure for assessing the performance for the unbalanced class distribution (Sonnenburg et al., 2007). The values of AUC-ROC and AUC-PR were computed using trapezoidal rule as suggested by Bradley (1997). Further, the standard error (SE) of AUC-ROC (or AUC-PR) (Bradley, 1997) was computed as $SE = \sqrt{\frac{\theta(1-\theta)+(N^{(C)}-1)(Q_1-\theta^2)+(N^{(I)}-1)(Q_2-\theta^2)}{N^{(C)}\cdot N^{(I)}}}$, where $Q_1 = \theta/(2-\theta)$, $Q_2 = 2\cdot\theta^2/(1+\theta)$; $N^{(C)}$, $N^{(I)}$ and θ are the number of CDS, intron in the test dataset and estimate of AUC-ROC (or AUC-PR) respectively. For estimating the values of AUC-ROC and AUC-PR, we have written R-code.

2.7. Comparison with linear and quadratic discriminant analysis

The performance of RF was also compared with that of most commonly used classical statistical approaches *i.e.,* LDA and quadratic discriminant analysis (QDA; Hastie et al., 2001) using the same dataset that was used for evaluating the performance of RF. For implementation of LDA and QDA model, *lda()* and *qda()* functions of R-software were respectively used. The performances of