Journal of Animal Research: v.14 n.02, p. 125-130. April 2024

DOI: 10.30954/2277-940X.02.2024.4

Gene Prediction in Rumen Metagenomic Reads of Cattle Using Machine Learning Based Approach

Safeer M. Saifudeen[1]*, Anilkumar K[1], T.V. Aravindakshan[1], Jamuna Valsalan[2], Ally K.[3] and Gleeja V.L.[4]

[1]Department of Animal Genetics and Breeding, Kerala Veterinary and Animal Sciences University, INDIA
[2]Centre for Advanced Studies in Animal Genetics and Breeding, Kerala Veterinary and Animal Sciences University, INDIA
[3]Department of Animal Nutrition, Kerala Veterinary and Animal Sciences University, INDIA
[4]Department of Statistics, Kerala Veterinary and Animal Sciences University, INDIA

**Corresponding author: SM Saifudeen; E-mail: safeermsaifudeen@gmail.com*

Received: 21 Dec., 2023 **Revised:** 10 Feb., 2024 **Accepted:** 16 Feb., 2024

ABSTRACT

The present study was focused to build a predictive model for protein coding genes from the rumen metagenomic data utilising most promising machine learning (ML) tools. We classified the sequence reads into coding genes and non-coding sequences, converted the sequences into k-mers of various sizes (k = three to six) and extracted features named k-mer count that were representative of the sequence reads. ML classifiers were trained using 16 genomes consisted of 13 bacterial kingdom and 3 archaeal kingdom selected from diverse environment and various systems. Among the five ML models for gene prediction, artificial neural network (ANN) performed best with maximum accuracy 89 per cent for k-mer three. We observed that logistic regression and SVMtook only reasonable computational time when compared to ANN.DNA was isolated from rumen liquor of crossbred cattle and were used for metagenomic sequencing. Annotated rumen metagenomic sequences was used to validate the ML models created. Logistic regression performed best with 85 per cent accuracy on minimum feature count itself (unigram) for k-mer four. Out of 8718 coding sequences provided to logistic regression classifier, 8073 sequences correctly predicted as genes (true positives) and remaining 645 coding sequences were predicted as non-coding (false negatives). We concluded that machine learning models created namely artificial neural network, support vector machine and logistic regression shows strong, robust and powerful ability for classification of coding and non-coding genes and it represents an intriguing and promising avenue for predicting rumen metagenomic genes.

HIGHLIGHTS

- Classification of sequence into coding and non-coding based on k-mers.
- Machine learning models for gene prediction in metagenomic DNA fragments.
- Validation of the models using bovine rumen metagenomic sequences.

Keywords: Rumen metagenomics, sequence, gene prediction, k-mer, machine learning

The microbial diversity in most environments exceeds the biodiversity of plants and animals by orders of magnitude (Hoff *et al.,* 2009). Relative to eukaryotic genomes, prokaryotic genomes are small and structurally simple, with ninety per cent of their DNA typically devoted to protein-coding genes (Sommer and Salzberg, 2021). The ruminal microbial population was characterized by complexity of interactions and dominated in number by bacteria (McSweeney *et al.,* 2001). It has been estimated that only a small proportion of organisms in nature can be cultured using standard cultivation methods (Wooley *et al.,* 2010). In order to facilitate the study of uncultivated micro-organisms a new field known as 'metagenomics',

How to cite this article: Saifudeen, S.M., Anilkumar, K., Aravindakshan, T.V., Valsalan, J., Ally, K. and Gleeja, V.L. (2024). Gene Prediction in Rumen Metagenomic Reads of Cattle Using Machine Learning Based Approach. *J. Anim. Res.,* **14**(02): 125-130.

Source of Support: None; **Conflict of Interest:** None

has emerged in the area of genetic research (Thomas *et al.,* 2012).

One of the largest challenges in the field of bioinformatics is the need for a mechanism to transform the raw data into a format that could be classified by a machine learning algorithm in an accurate way (Kaehler, 2017). Machine learning has been used broadly in biological studies for prediction and discovery. Massive and rapid advancements in both biological data generation and machine learning methodologies are promising for the analysis and discovery from complex biological data (Xu and Jackson, 2019). Over the past decade, there had been a steady increase in studies utilizing machine learning algorithms for various aspects of functional prediction, because these algorithms were able to integrate large amounts of heterogeneous data and detect patterns inconspicuous through rule-based approaches (Mahood *et al.,* 2020). Machine learning techniques played an important role in solving metagenomic problems such as gene prediction and comparative metagenomics analysis (Soueidan and Nikolski, 2017).

Gene prediction is the problem of identifying the portions of DNA sequence that are biologically functional. The first step towards successful genome annotation is gene prediction (Goel *et al.,* 2013. Gene prediction from bovine rumen metagenomics is a necessary step to fully understand the functions, activities and roles of microbial genes in cellular processes. Accurate gene prediction in metagenomes is more complicated than in isolated genomes (Hyatt *et al.,* 2012). The source genomes of the metagenomic fragments are always unknown or totally new, which brings challenge on statistical model construction and feature selection (Liu *et al.,* 2013). In this context, the present study was focused to build a predictive model for protein coding genes from the rumen metagenomic data utilising most promising machine learning tools.

MATERIALS AND METHODS

Our approach was to classify the sequence reads into either coding genes or non-coding sequences by converting the sequences into k-mers of various sizes (k = three to six) and extracting features named k-mer count that are representative of the sequence reads. For a particular k-mer, different order n-grams were included as a part of feature extraction. N-grams started from (1, 1) which was unigram, (1, 2) which included both unigram and bigram, (1, 3) included combination of unigram, bigram and trigram. Likewise up to N-gram (1, 6) were included in current study to obtain the continuity of features and to found out the optimum N-gram suitable for the ML models.

Machine learning approach

Our machine learning approach for gene prediction in metagenomic DNA fragments is based on learning the characteristics of coding and non-coding regions from 16 fully sequenced prokaryotic genomes and their GenBank annotation for protein coding genes. These 16 genomes consisted of 13 bacterial kingdom and 3 archeal kingdom selected from diverse environment and various systems. In the current study features were the k-mer counts and the target attribute (label) was discrete values namely zero and one where zero represented non-coding sequences and one represented protein coding genes. Sequence features were extracted using different combination of n-gram up to hexagram. Our goal was to empirically evaluate how well the standard machine learning algorithms perform in classifying metagenomic data.

Rumen metagenome sequence data

Here we used rumen metagenome sequence data to validate the ML models created. For that, rumen samples were collected from five HF crossbred cattle as detailed below which were maintained on standard ration (forage: concentrate ratio of 50:50).

DNA isolation

Rumen liquor (both solid and liquid fractions) were collected from the rumen of each crossbred cattle, three hours after morning feeding. The DNA was isolated from rumen liquor of crossbred cattle using cetyl trimethyl ammonium bromide (CTAB) based buffer for cell lysis followed by purification with phenol: chloroform: isoamyl alcohol. The isolated metagenomic DNA samples were pooled and were used for metagenomic sequencing.

The *de novo* assembly of the adapter trimmed fastq files was carried out using MetaSPAdes (v 3.10.1), which could

be utilised optionally and independently for different processing and assembly steps. The contigs obtained from the assembly were used as input to software tool prokka (v 1.14.6) for the prediction of open reading frames (ORFs), which was used to identify the coding regions and to distinguish them from noncoding DNA. Prokka used Prodigal for the ORF predictions. Total 51,430 contigs (contig length varies from 39463 bp to 918 bp) were used for gene prediction. Using prokka, 70,631 sequences were predicted as coding sequences.

Training datasets

Out of 31,208 sequence data fetched from RefSeq databases, 15,580 were coding sequences. 15,628 ORF like but non coding sequences were generated from the inter-genic region so that a balanced data of coding and non-coding sequences were made. We used an 80-20 split of the sequence data obtained from RefSeq databases, where 80% of the data was used for training and 20% of the data was used as the test set to found the efficiency of the machine learning model created. Out of 31,208 total sequences, 24966 (80 percent) sequences were used for training the model and 6242 (20 percent) for testing the model. The machine learning classifier is evaluated on how well it classified the test set based on the learned concept.

Validation of test datasets

For validating the ML classifiers, annotated rumen metagenomic sequences obtained were used. Out of 70631 coding sequences obtained 8718 coding sequences were selected randomly. A total of 17456 sequences including 8738 non coding sequences were used for validation.

Model selection and evaluation

We used three well-known supervised learners- logistic regression, support vector machine (SVM) and artificial neural network (ANN) and their performances in predicting coding sequences were assessed. In this study, we have used the metrics namely accuracy, precision, recall and F score to evaluate the model performance. True positives (TP) are the number of positive examples classified as positive. False negatives (FN) are the number of positive examples which are classified as negatives. True negatives (TN) are the number of negative examples which are classified as negative. False positives (FP) are the number of negative examples which are classified as positive.

$$\text{Accuracy} = TP + TN / (TP + FN + TN + FP)$$

RESULTS AND DISCUSSION

Training data evaluation

Machine learning approach for gene prediction in metagenomic DNA fragments was based on learning the characteristics of coding and non-coding regions from the 16 fully sequenced prokaryotic genomes which was used as training data for ML. Models were evaluated in terms of accuracy to know whether the selected models had good performance or not before validating the model using rumen metagenomic sequenced data. We have computed the experiments for different k ranging from three to six. For a particular k-mer, number of features depends on different combination of n-grams.

Table 1: Machine learning model evaluation - Comparison of different k-mer for particular ML

ML tool	K-mer size	No. of features (Range)	Accuracy (Range)
Logistic regression	three	64 - 65536	0.84 - 0.88
	four	**256 - 349425**	**0.85 - 0.88**
	five	1024 - 1394501	0.85 - 0.87
	six	4096 - 5249988	0.83 - 0.87
Support vector machine	three	64 - 65536	0.83- 0.88
	four	**256 - 349425**	**0.85 - 0.88**
	five	1024 - 1394501	0.84 - 0.87
	six	4096 - 5249988	0.83 - 0.87
Artificial neural network	**three**	**64 - 65536**	**0.86 - 0.89**
	four	256 - 349425	0.85 - 0.88
	five	1024 - 1394501	0.84 - 0.88
	six	4096 - 5249988	0.85 - 0.88

Comparison of different k-mer for particular ML

For the training data, the machine learning classifiers were evaluated on how well they classified the test set based on the learned concept. Accuracies of the five different

machine learning models created along with k-mers used were represented in table 1. Number of features varied for different k-mers. As k-mer size increased, number of features also increased. Among the five ML models, artificial neural network performed best with maximum accuracy 89 per cent for k-mer three. Current study showed that maximum accuracy could be attained on k-mer size three itself with minimum feature count on artificial neural network. Advantage was that we could overcome the computational issues created by more and more feature counts. Feature count could be increased by both increase in k-mer size and by more number of n-grams. In performance, just behind artificial neural network was logistic regression and support vector machine both of which had maximum accuracy of 88 percent for k-mer size four.

Validation of ML models with rumen whole metagenomic sequenced data

A total of 51,430 contigs were obtained from the assembly of reads obtained from rumen whole genome sequenced data. Metagenomic sequences were annotated to obtain 70,631 protein coding sequences. The annotation information obtained from rumen whole metagenomic sequence data were used as an input for the validation study of the machine learning models built.

Comparison of different k-mer for particular ML

Accuracies of different k-mers ($k = 3$ to 6) versus different combination of n-grams were plotted in line graphs for three machine learning models to found out the best k-mers and their corresponding n-gram combinations on gene prediction.

Comparison of different k-mer for Logistic Regression

In case of logistic regression, overall accuracy ranged from 76 per cent to 85 per cent. While comparing different k-mers, overall performance was in the order k-mer three to six where k-mer three performed best. For k-mers three, four and five, unigram performed best when compared to other n-gram combinations. There was no point in increasing the feature count by more n-gram combination in case of logistic regression. Extraction of more features such as pentagram, hexagram etc. and running the ML was comparatively tedious and time consuming. Here the advantage was that computational issues were minimum. 85.10 percent accuracy was the highest accuracy provided by logistic regression. This was obtained in k-mer four at unigram. K-mer six was found not much efficient for logistic regression in the given study.

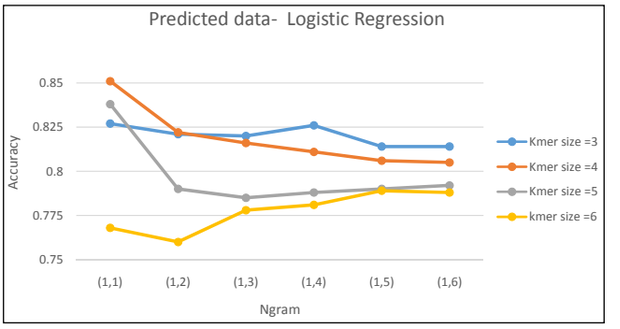

Fig. 1: Accuracies of Logistic regression classifier for different k-mer sizes

Comparison of different k-mer for support vector machine

For support vector machine, overall accuracy ranged from 75.80 per cent to 84 per cent. Accuracy was highest for k-mer size three at (1, 3) and lowest for k-mer size six at n-gram combination (1, 2). K-mer size four performed well in n-gram (1, 3) having accuracy of 81.20 percent. K-mer size five and six also had good accuracy in unigram itself (82.80 per cent and 81.30 per cent).

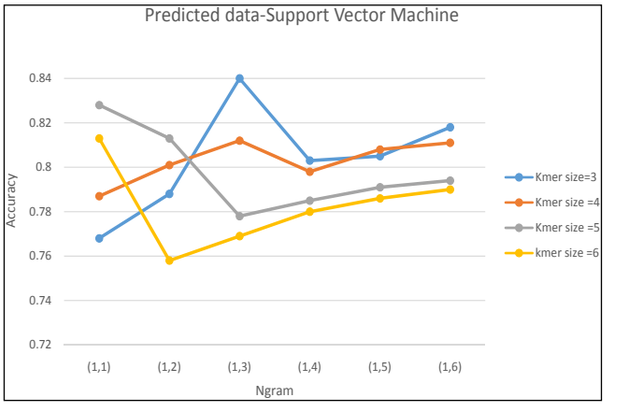

Fig. 2: Accuracies of support vector machine for different k-mer sizes

So computationally it was found very effective. We were able to achieve better accuracies with minimum feature count in case of k-mer five and six. Here for SVM classifier, it was difficult to say which k-mer performed best in gene prediction because all the four k-mers (k = 3 to 6) had their own good performance either in unigram (1, 1) or in combination up to trigram (1, 3).

Comparison of different k-mer for Artificial Neural Network

For artificial neural network, overall accuracy ranged from 75 per cent to 82.50 per cent. Accuracy was highest for k-mer size three at (1, 4) and lowest for k-mer size six at n-gram combination (1, 6). For k-mer size three n-gram combination (1, 4) was found better option for gene prediction and for k-mer size four, (1, 2) had the highest accuracy (81.60 percent). In the case of k-mer five and six, minimum feature frequency itself produced better accuracies with values 81.60 per cent and 82.40 per cent respectively. In k-mer six, huge decline in accuracy noticed in (1, 2) when compared to (1, 1). There was no meaning in increasing features by more and more n-gram combinations. As we know, artificial neural network needed more computational power and time when compared to others. So better performance in unigram itself (k-mer five and six) as showed in the current study will save computational power and time for the gene prediction problems.

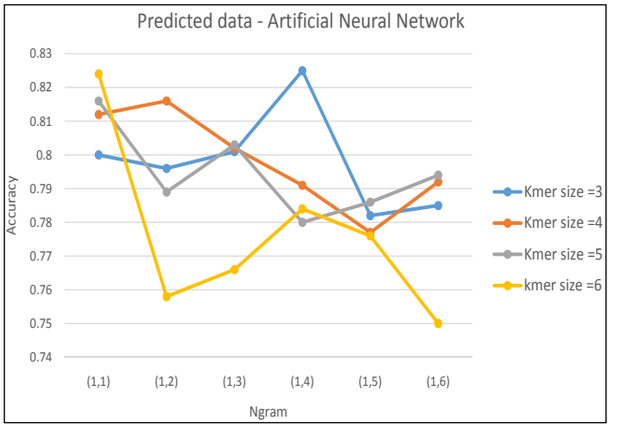

Fig. 3: Accuracies of Artificial Neural Network for different k-mer sizes

CONCLUSION

The motivation behind this work is to see how various machine learning classifiers performs by predicting genes in metagenomic DNA fragments. We addressed an important problem in the field of bioinformatics, which is to discriminate open reading frame (ORF) from the non-coding sequences. Identification of genes directly from metagenomic fragments is an important task in annotating metagenomes. In the current study, we used a diverse approach in classification of the sequence reads by converting the sequences into k-mers and extracted features named k-mer count. Various feature size were used in the study by applying different combination of continuous n-grams along with single n-gram for k-mers. Feature extraction and feature selection are important step toward enhancing the gene prediction process. Our future works will investigate the application of more and more diverse features for gene prediction from metagenomic DNA fragments. Large scale machine learning methods are well-suited for gene prediction in metagenomic DNA fragments. We concludes that machine learning algorithms namely artificial neural network, support vector machine and logistic regression shows strong, robust and powerful ability for classification of coding and non-coding genes and it represents an intriguing and promising avenue for predicting rumen metagenomic genes.

ACKNOWLEDGEMENTS

Authors are thankful to the Dean, College of Veterinary and Animal Sciences, Thrissur and other officials of Kerala Veterinary and Animal Sciences University for providing financial support and necessary facilities for conduction of the current research work.

REFERENCES

Goel, N., Singh, S. and Aseri, T.C. 2013. A review of soft computing techniques for gene prediction. *Int. Sch. Res. Notices*., **2**: 15-25.

Hoff, K.J., Tech, M., Lingner, T. and Daniel, R. 2008. Gene prediction in metagenomic fragments: a large scale machine learning approach. *BMC Bioinform*., **9**: 1-14.

Hyatt, D., LoCascio, P.F. and Hauser, L.J. 2012. Gene and translation initiation site prediction in metagenomic sequences. *J. Bioinform.,* **28**: 2223-2230.

Kaehler, R. 2017. K-mer analysis pipeline for classification of DNA sequences from metagenomic samples. *Graduate Student Theses.* University of Montana. Missoula.

Libbrecht, M.W. and Noble, W.S. 2015. Machine learning applications in genetics and genomics. *Nat. Rev. Genet*., **16**: 321-332.

Liu, Y., Guo, J. and Hu, G. 2013. Gene prediction in metagenomic fragments based on the SVM algorithm. *BMC Bioinform.,* **14**: 1-12.

Lo Bosco, G. and Pinello, L. 2014. A new feature selection strategy for K-mers sequence representation. *Proc. CIBB*. Pp. 1-10.

Mahamuda, V., Man Chon, U. and Rasheed, K. 2010. Application of Machine Learning Algorithms for Binning Metagenomic Data. *InProceedings of the International Conference on Bioinformatics and Computational Biology*, pp. 68-74.

Mahood, E.H., Kruse, L.H. and Moghe, G.D. 2020. Machine learning: A powerful tool for gene function prediction in plants. *Appl. Plant Sci*., **8**: 11-31.

McSweeney, C.S., Aminov, R. and Mackie, R.I. 2001. Rumen. In *Encyclopedia of life science Nature Publishing Group*.

Mello, A. 2021. A comparison of machine learning algorithms to investigate methylation profiles and predict type 1 diabetes. *Master's thesis*. Norwegian University of Science and Technology.

Sommer, M.J. and Salzberg, S.L. 2021. Balrog: A universal protein model for prokaryotic gene prediction. *PLoSComput. Biol.,* **17**: 25-35.

Soueidan, H. and Nikolski, M. 2017. Machine learning for metagenomics: methods and tools. *Metagenomics*., **1**: 50-75.

Thomas, T., Gilbert, J. and Meyer, F. 2012. Metagenomics-a guide from sampling to data analysis. *Microb. Inform. Exp*., **2**: 13-23.

Wang, Z., Chen, Y. and Li, Y. 2004. A brief review of computational gene prediction methods. *Genom. Proteom. Bioinform.,* **2**: 216-221.

Wooley, J.C., Godzik, A. and Friedberg, I. 2010. A primer on metagenomics. *PLoSComput. Biol.,* **6**: 15-25.

Xu, C. and Jackson, S.A. 2019. Machine learning and complex biological data. *Genome Biol.,* **20**: 76-79.