World Academy of Science, Engineering and Technology
International Journal of Biotechnology and Bioengineering
Vol:19, No:04, 2025

Identifying Protein-Coding and Non-Coding Regions in Transcriptomes

Authors : Angela U. Makolo

Abstract : Protein-coding and Non-coding regions determine the biology of a sequenced transcriptome. Research advances have shown that Non-coding regions are important in disease progression and clinical diagnosis. Existing bioinformatics tools have been targeted towards Protein-coding regions alone. Therefore, there are challenges associated with gaining biological insights from transcriptome sequence data. These tools are also limited to computationally intensive sequence alignment, which is inadequate and less accurate to identify both Protein-coding and Non-coding regions. Alignment-free techniques can overcome the limitation of identifying both regions. Therefore, this study was designed to develop an efficient sequence alignment-free model for identifying both Protein-coding and Non-coding regions in sequenced transcriptomes. Feature grouping and randomization procedures were applied to the input transcriptomes (37,503 data points). Successive iterations were carried out to compute the gradient vector that converged the developed Protein-coding and Non-coding Region Identifier (PNRI) model to the approximate coefficient vector. The logistic regression algorithm was used with a sigmoid activation function. A parameter vector was estimated for every sample in 37,503 data points in a bid to reduce the generalization error and cost. Maximum Likelihood Estimation (MLE) was used for parameter estimation by taking the log-likelihood of six features and combining them into a summation function. Dynamic thresholding was used to classify the Protein-coding and Non-coding regions, and the Receiver Operating Characteristic (ROC) curve was determined. The generalization performance of PNRI was determined in terms of F1 score, accuracy, sensitivity, and specificity. The average generalization performance of PNRI was determined using a benchmark of multi-species organisms. The generalization error for identifying Protein-coding and Non-coding regions decreased from 0.514 to 0.508 and to 0.378, respectively, after three iterations. The cost (difference between the predicted and the actual outcome) also decreased from 1.446 to 0.842 and to 0.718, respectively, for the first, second and third iterations. The iterations terminated at the 390th epoch, having an error of 0.036 and a cost of 0.316. The computed elements of the parameter vector that maximized the objective function were 0.043, 0.519, 0.715, 0.878, 1.157, and 2.575. The PNRI gave an ROC of 0.97, indicating an improved predictive ability. The PNRI identified both Protein-coding and Non-coding regions with an F1 score of 0.970, accuracy (0.969), sensitivity (0.966), and specificity of 0.973. Using 13 non-human multi-species model organisms, the average generalization performance of the traditional method was 74.4%, while that of the developed model was 85.2%, thereby making the developed model better in the identification of Protein-coding and Non-coding regions in transcriptomes. The developed Protein-coding and Non-coding region identifier model efficiently identified the Protein-coding and Non-coding transcriptomic regions. It could be used in genome annotation and in the analysis of transcriptomes.

Keywords : sequence alignment-free model, dynamic thresholding classification, input randomization, genome annotation
Conference Title : ICBCMB 2025 : International Conference on Bioinformatics and Computational Molecular Biology
Conference Location : London, United Kingdom
Conference Dates : April 22-23, 2025