World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:19, No:03, 2025

Assessment of DNA Sequence Encoding Techniques for Machine Learning Algorithms Using a Universal Bacterial Marker

Authors : Diego Santibañez Oyarce, Fernanda Bravo Cornejo, Camilo Cerda Sarabia, Belén Díaz Díaz, Esteban Gómez Terán, Hugo Osses Prado, Raúl Caulier-Cisterna, Jorge Vergara-Quezada, Ana Moya-Beltrán

Abstract : The advent of high-throughput sequencing technologies has revolutionized genomics, generating vast amounts of genetic data that challenge traditional bioinformatics methods. Machine learning addresses these challenges by leveraging computational power to identify patterns and extract information from large datasets. However, biological sequence data, being symbolic and non-numeric, must be converted into numerical formats for machine learning algorithms to process effectively. So far, some encoding methods, such as one-hot encoding or k-mers, have been explored. This work proposes additional approaches for encoding DNA sequences in order to compare them with existing techniques and determine if they can provide improvements or if current methods offer superior results. Data from the 16S rRNA gene, a universal marker, was used to analyze eight bacterial groups that are significant in the pulmonary environment and have clinical implications. The bacterial genes included in this analysis are Prevotella, Abiotrophia, Acidovorax, Streptococcus, Neisseria, Veillonella, Mycobacterium, and Megasphaera. These data were downloaded from the NCBI database in Genbank file format, followed by a syntactic analysis to selectively extract relevant information from each file. For data encoding, a sequence normalization process was carried out as the first step. From approximately 22,000 initial data points, a subset was generated for testing purposes. Specifically, 55 sequences from each bacterial group met the length criteria, resulting in an initial sample of approximately 440 sequences. The sequences were encoded using different methods, including one-hot encoding, k-mers, Fourier transform, and Wavelet transform. Various machine learning algorithms, such as support vector machines, random forests, and neural networks, were trained to evaluate these encoding methods. The performance of these models was assessed using multiple metrics, including the confusion matrix, ROC curve, and F1 Score, providing a comprehensive evaluation of their classification capabilities. The results show that accuracies between encoding methods vary by up to approximately 15%, with the Fourier transform obtaining the best results for the evaluated machine learning algorithms. These findings, supported by the detailed analysis using the confusion matrix, ROC curve, and F1 Score, provide valuable insights into the effectiveness of different encoding methods and machine learning algorithms for genomic data analysis, potentially improving the accuracy and efficiency of bacterial classification and related genomic studies.

Keywords : DNA encoding, machine learning, Fourier transform, Fourier transformation
Conference Title : ICMCSSE 2025 : International Conference on Mathematical, Computational and Statistical Sciences and Engineering
Conference Location : Madrid, Spain
Conference Dates : March 17-18, 2025