Iraqi Journal for Computer Science and Mathematics

Journal Homepage: http://journal.esj.edu.iq/index.php/IJCM
e-ISSN: 2788-7421 p-ISSN: 2958-0544

More-SPEED: Enhancing Protein Activity Prediction from DNA Sequences

Samaher Al-Janabi* [1], **Zena A. Kadhuim**[2]

[1] Department of Computer Science, Faculty of Science for Women (SCIW), University of Babylon, Babylon, Iraq
[2]Department of Software/College of Information Technology, University of Babylon, Babylon, Iraq
*Corresponding Author

DOI: https://doi.org/10.52866/ijcsm.2023.04.04.005
Received July 2023; Accepted September 2023; Available online October 2023

ABSTRACT: This work presents More-SPEED, a novel model for accurately predicting protein activity while minimizing computational demands. Leveraging optimized structures and data preprocessing techniques, More-SPEED achieves high accuracy in protein activity prediction. The model incorporates the DC-3D layer, utilizing the GMP-FFGM algorithm for efficient preprocessing of complex DNA sequence datasets. Additionally, the DSN-WOA structure optimizes parameters of the BDLSTM model, reducing processing time and eliminating manual parameter selection. The BDLSTM layer plays a crucial role in matching codons and predicting protein names, reducing computational complexity without compromising accuracy. The Bi-Rule layer efficiently determines protein activity, especially in disease contexts, providing valuable insights in a shorter time compared to alternative approaches. Evaluation metrics validate the effectiveness of More-SPEED in accurately predicting protein activity, making it a promising solution for advancing protein research.

Keywords: Intelligent Data Analysis; Deep Learning; GMP-FFGM; DSN-WOA, BDLSTM, Bi-Rule; Active Proteins.

1. INTRODUCTION

Technological advancements have significantly improved various aspects of our daily lives through the invention of tools and devices that make life easier and faster. Communication technologies, for instance, have enabled individuals from around the world to communicate with ease. The recent outbreak of epidemics such as COVID-19 has further highlighted the importance of internet-based services, which have become essential for limiting physical interactions and reducing the spread of the virus (Tyagi et al., 2021). Diagnosing diseases associated with an organism requires an understanding of its DNA structure and the genes that generate proteins. However, predicting the proteins that promote or inhibit the presence of diseases is a complex process that necessitates studying the organism's DNA structure (Awad, 2022).

Intelligent Data Analysis (IDA) is a multidisciplinary field that employs techniques from artificial intelligence, high-performance computing, pattern recognition, and statistics to extract meaningful knowledge from data (Sarker, 2021). The IDA process involves three primary steps: problem identification and parameter understanding, model building using techniques such as clustering, classification, prediction, optimization, etc., and evaluation of the results. Finally, the results must be interpreted in a way that is understandable to both specialists and non-specialists (Sarker, 2021). The main benefits of IDA can be summarized as follows (Al-Janabi 2022): (a) Extraction of meaningful knowledge from data. (b) Multidisciplinary approach that combines techniques from various fields. (c) Identification and understanding of real-world problems. (d) Effective model building and evaluation techniques. (e) Interpretation of results in a way that is understandable to all specialists and non-specialists.

Data analysis is divided into four types: Descriptive Analysis, Diagnostic Analysis, Predictive Analysis, and Prescriptive Analysis. The goal of intelligent data analysis is to extract knowledge from data, which can be used to inform decision-making and improve outcomes. Prediction is a crucial data analysis task that involves estimating the value of a target feature that is not known. Prediction techniques can be classified into two main fields based on the scientific area: prediction techniques related to data mining and prediction techniques related to deep learning, such as neuron computing techniques (Bárbara et al., 2021). Various types of data analysis have been introduced, each with their unique advantages and applications in different fields. The aim of prediction is to analyze trends by making estimations for future events based on the impact of past and present data.

Bioinformatics is a sub-discipline that lies at the intersection of biology and computer science, dealing with the extraction, storage, analysis, and dissemination of biological data. The primary objective of bioinformatics is to manage data in a way that allows easy and efficient access to information and to submit new entries as they are produced. Moreover, bioinformatics involves the development of technological tools that help analyze biological data.

*Corresponding author: samaher@itnet.uobabylon.edu.iq