Predicting DNA-Binding Proteins with A Hybrid Architecture of GRU, BI- LSTM, and Capsule Networks

Rahu Sikander[1,2]*, Ali Ghulam[3], Mujeebu Rehman[4]

[1]Department of Computer Science, Suzhou University of Technology, Changshu, China.

[2]Jinnah University for Women, Karachi, Pakistan.

[3]Information Technology Centre, Sindh Agriculture University, Tandojam Sindh, Pakistan.

[4]School of Information and Communication Engineering, Guilin University of Electronic Technology, Guilin, China.

*E-mail: sikander.rahu@szut.edu.cn

ABSTRACT

Proper computational determination of the DNA-binding proteins (DBPs) based on their simple sequence remains a major challenge in genomics. DBPs play a key role in essential cellular processes (including: DNA replication, transcription, splicing and repair) and are essential targets of innovative therapy programs to treat cancer and immunological diseases. Experimental methods of DBP characterizing are, however, infamously tedious, time- consuming, and expensive to carry out. Hence, there is a necessity to develop effective and accurate models of computational forecasting. It contains an exemplar of a deep learning framework that is a hybrid and involves predicting DNA-binding proteins (DBPs) by using simple sequences. The proposed model includes the three main parts: a GRU, a Bi-LSTM, and a 1D-CapsNet. Protein sequences are first converted to two different representations, a one-hot encoding of the GRU layer and a distributed word2vec embedding of the Bi-LSTM layer. The representations of features acquired using the GRU and Bidirectional LSTM are then combined and sent to a 1D-CapsNet. This final module will aim to summarize all the complicated hierarchical associations amidst the characteristics, and eventually act as the effective classifier. Practical evidence confirms that the framework offered has better performance which beats most of the existing benchmark techniques used to predict DNA-binding proteins.

Keywords: Bidirectional LSTM, Capsule Network, Gated Recurrent Unit.

INTRODUCTION

Proper identification of DNA-binding proteins (DBPs) is crucial towards understanding DNA-protein interactions and fundamental processes such as replication and transcription (Barukab *et al*., 2022). DBPs are important therapeutic targets, nuclear receptors are important in cancer treatment, glucocorticoid receptors are involved in the development of autoimmune drugs, The association between Inhibitor of DNA-binding (ID) proteins and oncogenesis has been established in various cancer types (Chowdhury *et al*., 2017; Sandman, *et al*., 1998). However, Traditional techniques for their study, including X-ray crystallography and ChIP are limited by time consuming, expensive and laborious procedures that prevents high-throughput research (Overington *et al*., 2006; Hudson, *et al*., 2018).

To overcome this, computational techniques have been established with the prevalent one being the Position-Specific Scoring Matrices (PSSM) being used to store the evolutionary data to be used in prediction (Wang *et al*., 2020). Nevertheless, these approaches are computationally expensive and frequently based on the conventional machine learning models that are not efficient and scalable with large-scale data and heavy engineering of features manually, which is restrictive to their performance and scalability (Elnaggar *et al*., 2021). Deep learning has had a significant influence in bioinformatics in recent years, which provides new methods of prediction of DBP. Its main advantage over the conventional machine learning lies in that it automatically acquires hierarchical features directly out of sequence data. LMST and Bi-LMST architectures

are useful in representing time. However, Convolutional Neural Networks need large-scale data, and Recurrent Neural Networks cannot address long- term dependencies.

This work introduces a new hybrid deep learning model that will employ a GRU and Bi-LSTM network or CapsNet to predict DNA-binding proteins (DBPs). This model is based on a dual-input representation whereby one-hot representation is used on the GRU and word2vec representation on the Bi-LSTM to represent the change in contextual features in amino acid sequences. The obtained feature maps of these recurrent networks are thereupon concatenated and processed by 1D-CapsNet, which is a high-level feature interrelationship modeler and the ultimate classifier. The main contribution of this work is the substitution of computationally-intensive Evolutionary Information (EI) by the more computationally-efficient one-hot and word2vec encodings thereby automating the feature selection procedure and eliminating the manual curation.

OBJECTIVE

1. We propose a deep learning model that integrates GRU, Bidirectional LSTM, and a 1D-Capsule Network to predict DNA-binding proteins from protein sequences.
2. To replace Evolutionary Information (EI) whose computational demands are high with more efficient feature representations approaches, The feature engineering approaches of one-hot encoding and the learned embeddings from Word2Vec model.
3. To automate the selection of features and eliminate the need to perform the search of the relevant protein sequence properties manually.
4. In order to effectively describe spatial and temporal connections in protein sequences based on the synergistic benefits of GRU (contextual features), Bi-LSTM (past and future contextual information), and 1D-CapsNet (spatial hierarchies and classification).
5. To achieve improved performance on DBP prediction on different benchmark datasets (PDB1075, PDB186, PDB14189, PDB2272) in comparison with the existing state-of-the-art methodologies as measured by performance indices such as accuracy, sensitivity, specificity, and MCC.

METHODOLOGY

An effective predictive model needs to have reliable benchmark data. In this study, the researchers use four accepted datasets to deliver a fair evaluation. Training and cross-validation are performed using PDB1075 and PDB14189, and PDB186 and PDB2272 are used as an independent test set. All the datasets have been carefully filtered to remove highly similar sequences and ensure the quality of data as indicated.

The suggested approach to the prediction of DNA-binding proteins (DBPs) implies a systematic deep learning model, illustrated in Figures **1**. The process starts with the input datasets of two sizes (big and small), which are separated into training and independent evaluation. In the preprocessing stage, the protein sequences undergo one-hot encoding to convert them into numerical data to serve as the input of GRU and Word2Vec to convert them into numerical data to serve as the input of the Bi-LSTM. Such a dual-encoding methodology includes a great variety of contextual amino acid data. GRU and Bi-LSTM data are merged and then handled by a 1D-CapsNet layer of the features that identifies complex spatial structures in order to perform the final classification. After the hyperparameter optimization, the generalizability of the model is carefully tested on independent test set with predefined performance measures.

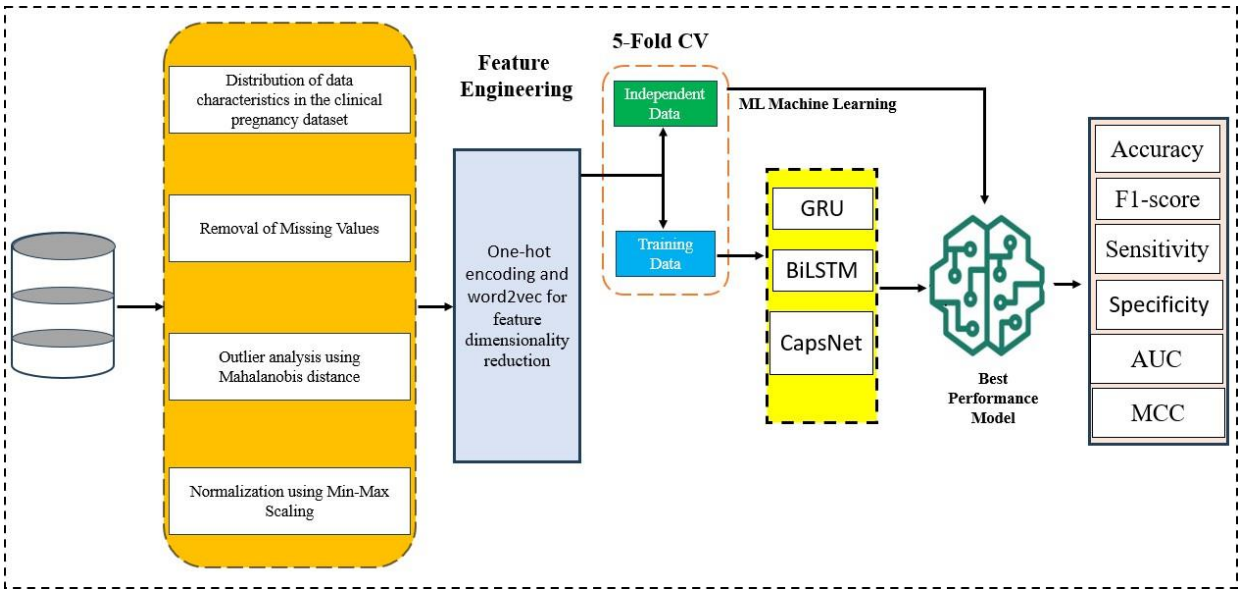

Figure1. Proposed the farmwork model.

RESULT

Although the proposed approach had a variety of optimized hyperparameters, including the number of epochs, batch size, learning rate, count of hidden units, choice of optimizer, and capsule network-specific parameters. All experiments were implemented using the Keras framework on a system equipped with an NVIDIA Tesla T4 with 16 GB of VRAM.

The comparative analysis of optimizers revealed that RMSprop has obtained the best performance among the compared optimizers, with an F1-score of 82.25, accuracy of 83.33, sensitivity of 93.01, AUC of 83.88, and MCC of 0.58. This performance profile confirms its efficacy in correctly classifying DBPs while maintaining a strong equilibrium between minimizing false positives and false negatives. On the other hand, Adamax showed the worst performance, and Nadam showed intermediate performance. Thus, it was discovered that RMSprop is the most effective optimizer of the proposed architecture.

Ablation experiments were done to determine the impact of each component in the proposed model. shows that one- hot encoding outperformed word2vec in all of the combinations considered. The combination of the two encoding methods provided a considerable improvement to the alternative models and the previously existing BiCaps-DBP benchmark, which employed one-hot encoding alongside a Bi-LSTM and 1D-CapsNet architecture (Mursalim *et al*., 2023). GRU, 1D-CapsNet, and Bi-LSTM were combined to give the best overall performance, with improvements of 6.38 percent and 9.84 percent in F1-score and accuracy, respectively, compared to BiCaps-DBP. Additional increases in sensitivity, specificity, and AUC were observed, which confirms the effectiveness of the proposed hybrid architecture.

CONCLUSION

In this study, an analysis of deep learning combining a multi-layer structure composed of a gated recurrent unit and bidirectional LSTM and 1D-CapsNet was proposed to predict DBP. The model was also highly accurate, sensitive, and specific, without the use of evolutionary information with one-hot encoding and word2vec. It performed better than other techniques, as it was able to capture spatial and contextual information in protein sequences, in which the gated recurrent mechanisms of GRU and Bi-LSTM were able to extract contextual features, and 1D-CapsNet was able to optimize feature hierarchies into improved classification. Nevertheless,

it has not been experimentally validated, has no structural and functional annotations, and its applicability and interpretability are limited by the absence of explainable AI integration.

REFERENCES

1. Barukab, Omar, *et al*. "DBP-CNN: Deep learning-based prediction of DNA-binding proteins by coupling discrete cosine transform with two-dimensional convolutional neural network." Expert Systems with Applications 197 (2022): 116729.
2. Chowdhury, Shahana Yasmin, Swakkhar Shatabda, and Abdollah Dehzangi. "iDNAProt-ES: identification of DNA-binding proteins using evolutionary and structural features." Scientific reports 7.1 (2017): 14938.
3. Sandman, K., S. L. Pereira, and J. N. Reeve. "Diversity of prokaryotic chromosomal proteins and the origin of the nucleosome." Cellular and Molecular Life Sciences CMLS 54.12 (1998): 1350-1364.
4. Overington, John P., Bissan Al-Lazikani, and Andrew L. Hopkins. "How many drug targets are there?" Nature reviews Drug discovery 5.12 (2006): 993-996.
5. Hudson, William H., *et al*. "Cryptic glucocorticoid receptor-binding sites pervade genomic NF-κB response elements." Nature communications 9.1 (2018): 1337.
6. Wang, Jun, *et al*. "PredDBP-Stack: Prediction of DNA-Binding Proteins from HMM Profiles using a Stacked Ensemble Method." BioMed research international 2020.1 (2020): 7297631.
7. Elnaggar, Ahmed, *et al*. "Codetrans: Towards cracking the language of silicon's code through self-supervised deep learning and high-performance computing." arXiv preprint arXiv:2104.02443 (2021).
8. M.K.N. Mursalim, T.L.E.R. Mengko, R. Hertadi, A. Purwarianti, M. Susanty, BiCaps-DBP: predicting DNA-binding proteins from protein sequences using Bi-LSTM and a 1D-capsule network, Comput. Biol. Med. 163 (2023) 107241.