Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume7-Issue1, January 2024
Paper: 19

Protein Sequence Classification Using Deep Learning Models

Malavika S[1],Dr. Devi Kannan[2],Nithika[3], Pratheeksha M C[4], Navya Ganapati Hegde[5]

[1,2,3,4,5]Atria Institute of Technology, Bangalore
[1]malvikas230@gmail.com, [2]devi.kannan@atria.edu

***Abstract*— Proteins are the multiplexed textures that can be found in everything. They consist of many peptide-bonded amino acids and are macromolecular polypeptides. It is composed of numerous amino acids that are joined to create long chains in a specific order. Similar to human language, protein structure is a linear sequence that is typically represented by character threads. There are 20 amino acids that are shared by every protein chain. Processing and analysing vast volumes of natural language data is the focus of natural language processing, a branch of linguistics and artificial intelligence. For many NLP (natural language processing) approaches, proteins make perfect sense. These techniques will analyse proteins and display text-based protein information. This study utilises machine learning and natural language processing to classify and analyse proteins in order to identify diseases. BERT is an NLP model that shows linguistic modelling of amino acid sequences using intelligent deep self attention based transformers. For proteins with missense mutations, this forecasts the likelihood of illness. In this paper we make comparison between LSTM's, a type of artificial neural networks and TCN's, a variation of convolutional neural networks.**

***Keywords*—Natural language processing, Bert, ALBert, XLNet, LSTM(Long-short term memory), TCN(Temporal convolutional networks), PLUS-RNN, CNN.**

I. INTRODUCTION

Proteins are necessary, and comprehending their structure can help us grasp how they work mechanistically. Forecasting a protein's three-dimensional structure based only on its amino acid sequence. The availability of enormous databases on protein sequences, as well as computational methods for protein sequence analysis, has expanded the frontiers of protein and related domain study. The protein sequence is concerned with identifying and analysing the polypeptic chains and protein structures found in protein databases. Automated text and language analysis is the focus of the computer science discipline known as natural language processing. Protein information is codified as text, which is subsequently analysed using NLP techniques as well as more recent tools like Bert and its variants. We give a current perspective on the use of NLP tools to the investigation of protein sequences in this review. In natural language problems, self-supervised deep language modelling has demonstrated unparalleled performance. Existing models and pre-training approaches, on the other hand, are built and optimised for text analysis.

II. NATURAL LANGUAGE PROCESSING AND PROTEINS

The goal of natural language processing is to enable computers to comprehend spoken and written language in a manner similar to that of humans. NLP combines machine learning, deep learning, and statistical models with computational rule-based, or linguistics modelling of human language. Computers can now read human language in the form of text or audio data and completely "understand" what is being written or said, including the writer's or speaker's thanks and intents, emotions to the employment of these technologies. Computer programmes that translate text between languages, reply to spoken commands, and quickly summarise vast amounts of text—even in real time—are all powered by NLP. We have probably used NLP in the form of voice-activated GPS devices, digital assistants, speech-to-text dictation programmes, customer service chatbots, and other consumer conveniences. The use of NLP in corporate solutions, however, is expanding as a means of streamlining company operations, boosting worker productivity, and

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume7-Issue1, January 2024
Paper: 19

streamlining mission-critical business procedures. Unknowingly, we may employ NLP technology in our daily lives. Chatbots that filter our requests as well as Siri, Alexa, and Hello Google are a few examples. In this sense, technology may be seen as a real-time link between computers and people, improving company operations and procedures to boost overall productivity.

In order to analyse and comprehend human languages and, in certain situations, to foretell a human's intention and purpose, NLP approaches rely on Deep Learning and algorithms. Unstructured data, including speech and text, is ingested by deep learning models, which transform it into structured and practical data insights. By breaking down the language into words and inferring context from the relationships between these words, the technology is able to extract meaning.. We can accurately index data and separate it into different groups or classes by using NLP in this way. These categories can comprise, among other things, mood, intent, and price data.

Several NLP techniques naturally fit proteins, that may be illustrated as strings of letters with different amino acids. We examine the theoretical parallels and discrepancies linking proteins and languages, as well as a variety of protein-associated machine learning challenges. We cover approaches for textual protein information encoding and NLP analysis, reviewing both traditional ideas like k-mers/n-grams, bag-of-words, and text search, in addition to contemporary approaches like word encapsulated, contextualised embedding, deep learning, and neural language models.

III. RELATED WORKS

To examine the theoretical parallels and discrepancies between language and proteins, as well as to assess a variety of protein-related activities accessible to machine learning, a study on nlp, machine learning, and protein sequences was conducted. Different approaches were discussed for text-based protein information encoding and analysis using NLP techniques, reviewing time-tested ideas like text search, k-mers/n-grams, and bag-of-words furthermore cutting-edge approaches like word embedding, contextualised embedding, neural language and deep learning models [1].

It has been demonstrated that transformer architectures may learn helpful representations for protein synthesis and categorization tasks. The interpretability of these representations is complicated, though. They have shown a series of techniques for looking at protein Transformer models via the prism of attention in one of their research papers. They show that attention connects amino acids particularly spatially separated in the elemental sequence but adjacent together in the protein folding structure, focuses on binding sites, a crucial part of proteins' functionality, and, as layer depth increases, pays more attention to increasingly complex biophysical properties. We discover that this behavior is continuous across two distinct protein datasets and three Transformer designs (BERT, ALBERT, and XLNet) [2].

Self-supervised deep language modelling has recently been applied to biological sequences and has demonstrated extraordinary effectiveness across natural language domains. Nonetheless, pre-training techniques and existing models are created and enhanced for text analysis. ProteinBERT, a deep language model created especially for proteins, is described in a study. Our pre-training method combines the unique job of Gene Ontology (GO) annotation prognosis with language modelling. In spite of applying a much minor and quicker model than ambitious deep-learning techniques, ProteinBERT achieves performance that is close to, and occasionally even surpasses, the state-of-the-art on numerous benchmarks housing various protein attributes (along with protein structure, post-translational modifications, including biophysical attributes) [4].

In representations of the protein sequence, another study introduces Learning Utilizing Structural Information, or PLUS, a novel pre-training strategy. PLUS is made up of a protein-specified pre-training task termed same-family prediction and masked language modelling. In six of seven regularly applied protein biology tasks, their trial results reveal that PLUS-RNN outcompete alternative models of equal proportions that were simply pre-trained using language modelling [3].

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume7-Issue1, January 2024
Paper: 19

IV. TRANSFORMERS AND NEURAL NETWORKS

The self-attentional process is used by a deep learning model known as a transformer, which weights the significance of each incoming data element differently.Natural language processing (NLP) and computer vision are two discipline that heavily employ it (CV). Similar to recurrent neural networks (RNNs), transformers are outlined to analyse successive input facts, including natural language, and have uses in translation and text summarization. Transformers operates the full instantly, unlike RNNs, with the help of the attention mechanism, any location in the input series can have context.

By resolving sequence-to-sequence actions, Transformers aims to manage long-range dependencies with simplicity.It uses neither convolution nor sequence-aligned RNNs to construct depictions of its input or output, instead only relying on self-attention. Self attention, also called as intra-attention, is a focus technique that connects many spots in a single sequence to create a depiction of the sequence.To put it another way, personal attention enables us to make comparable connections inside a single sentence.

Transformers was created in 2017 by a team at Google Brain, and it is now the preferred model for NLP problems, taking the place of RNN models like long short-term memory. Larger datasets may be trained using the extra parallelization for training. As a result, pretrained systems like BERT (Bidirectional Encoder Representations from Transformers) as well as GPT (Generative Pre-trained Transformer) were created.These systems, that were trained using sizable language datasets like the Wikipedia Corpus in addition to Common Crawl, may be tailored for certain purposes. In this essay, our primary emphasis is on using BERT to achieve our goals.

The long short-term memory (LSTM) artificial neural network is used in both deep learning and artificial intelligence.As opposed to traditional feed-forward neural networks, the LSTM has feedback connections.This type of recurrent neural network (RNN) can analyze whole data sequences in addition to single data points, such as images. This characteristic makes LSTM networks ideal for managing and anticipating data.For instance, LSTM may be used for tasks like voice recognition, machine translation, speech activity detection, robot control, video games, and healthcare.Applications like connected, unsegmented handwriting identification are also possible with it.

By fusing elements of RNN and CNN architectures, Temporal Convolutional Networks, or simply TCN, is a variant on Convolutional Neural Networks for sequence modelling applications.A straightforward convolutional architecture outruns conventional recurrent networks like LSTMs over a wide range of functions and datasets although exhibiting elongated effectual memory, according to preliminary empirical evaluations of TCNs.TCNs are distinguished by the fact that the architectural convolutions are causal, i.e., there is no data "leakage" from the forthcoming to the antiquity.Similar to an RNN, the architecture may accept any length input sequence and plan it to an output series of the uniform dimensions.TCNs combine very deep networks (enhanced by residual layers) with dilated convolutions to achieve very large effective history sizes, or the capacity to view distant into the prior in order to generate a prediction. The categorization of proteins using LSTMs and TCNs will be compared in this research.

V. LSTM'S AND TCN'S IN PROTEIN SEQUENCE CLASSIFICATION

For the purpose of classification using the above methods we use deep learning and a protein database called 'Pfam' to classify amino acid sequences into their protein families. Different components in the database includes, the sequence; there are 4 unusual amino acids and 20 uncommon amino acids (frequency > 1,000,000): The letters U, B, O, Z, and X stand for an unidentified or generic amino acid, the family_accession; these are the labels (outputs) of the model and family_id; one word name for protein family.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume7-Issue1, January 2024
Paper: 19

The first step would be to analyse the most prevalent/common sequence length, which will be helpful in pre-processing our sequences. In pre-processing we assign 1 letter code for 20 natural amino acids as shown below.

{'A': 1, 'C': 2, 'D': 3, 'E': 4, 'F': 5, 'G': 6, 'H': 7, 'I': 8, 'K': 9, 'L': 10, 'M': 11, 'N': 12, 'P': 13, 'Q': 14, 'R': 15, 'S': 16, 'T': 17, 'V': 18, 'W': 19, 'Y': 20}

Now we will encode our sequences made of letters into sequences of integers. In this 20 Common amino acids (and X 'wild-card') are taken into consideration. Remaining 4 uncommon amino acids are categorized as 0 (so not included). Now we must ensure all sequences are the same length since our sequence model can only take in fixed-size inputs. We pad the sequences using Keras pad_sequences function, using the max_length the was determined in previous steps. We use 'post' padding to pad with 0 if the total sequence length is less than max_length or truncate

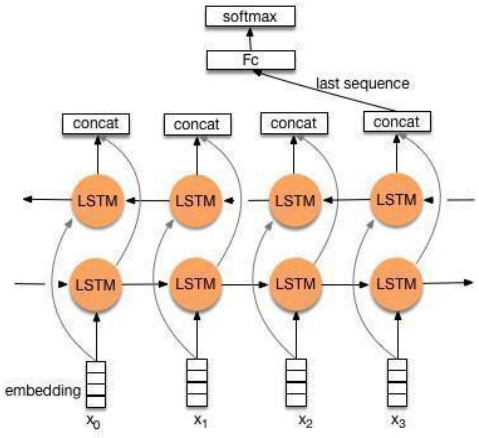

Fig. 1. Bidirectional LSTM

the sequence (if longer) to max_length. Now we encode input sequences and output labels using one-hot encoding (zeros everywhere except for at location of the appropriate amino acid code) using Keras to_categorical function.Now we implement the bidirectional LSTM usingKeras Functional or Sequential API. Unlike RNNs which only take input from the previous timestep, LSTMs can remember information from the distant past via gating mechanisms.Bidirectional LSTMs can incorporate contextual information from both past and future, compared to unidirectional LSTMs on sequentialdata.

After that we train our model using Kerasmodel.fit function. We can add the given early stopping as a parameter to our fit function to prevent overtraining. The validation and test accuracy was found out to be approximately, 0.9794 and 0.9775 respectively.

Now we will assess the interclass variability of embeddings. We will plot t-SNE clusters within and between family_accession samples. Extract the weights from the embedding layer of your model1. The shape of your weight matrix should make sense - it should be (sequence_length x embedding_dim). Within 5 classes, we can compute the embeddings for each sample in the validation set, storing them as you go to a new data matrix (X).Now we use t-SNE to fit and transform the X matrix to 2 components for visualization. The plot in fig. 2 shows the lower-dimensional representations of samples, colored by family_accession.

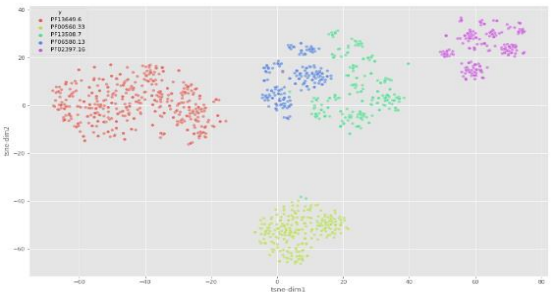

Fig. 2. The lower dimensional representatioms of samples colored by family_accession

For the most part, each cluster for a different protein class is located within a distinct region of the plot (with just a few outliar data points). This suggests that the embeddings of the various amino acids (AAs) sequence samples within a given class are rather similar to one another, further suggesting the importance of contextual information (i.e. the location of a given AA with respect to its surrounding AAs in a given sequence). The projection of an AA into a continuous vector space is represented as a dense vector in an embedding. An AA's location inside the vector space is

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume7-Issue1, January 2024
Paper: 19

determined by its surrounding AAs in the data and is learnt through sequences [6].

As the embeddings can be reduced to relatively low-dimensional spaces to represent high-dimensional vectors (confirmed by the distinct clusters after reducing to just 2 components), they can help an LSTM better and more efficiently learn the protein classifications from large input data. Embeddings can be learned and reused across models, adding to their benefits.

Now we will repeat the protein sequence classification task using a ResNet based temporal (1D) convolution network with dilation. We have used a Keras functional API to design two convolutional blocks and then use the

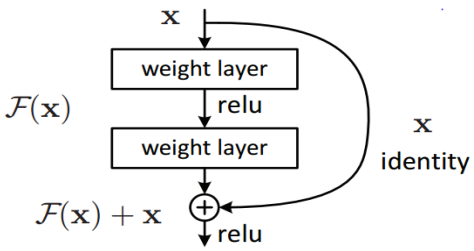

Fig. 3. Residual neural network

Add() function to combine our input (data) with the output of our 2nd convolutional block.

We design a simple 1D temporal convolution network around the residual networks (called below using the definition you implemented in part a) with Keras Functional API. Then we include an input layer and convolution layer before the residual networks and max pooling, dropout, flatten, and output dense layers after them. Then we train our model with early stopping using 10 epochs and a batch size of 256 and validate with our one-hot-encoded validation data from before. The validation and test accuracy was found out to be approximately, 0.9913 and 0.9917 respectively.

VI. CONCLUSION

Upon early stopping after 35/40 epochs, results from bidirectional LSTM model (model1), the train accuracy was 0.9862, the val accuracy was 0.9762 and the test accuracy was 0.9757. Upon early stopping after 6/10 epochs, results from ResNet-based convolutional model (model2), the train accuracy was 0.9984, the Val accuracy was 0.9908 and the test accuracy was 0.9914.

The ResNet-based model achieved superior train, val, and test accuracies after much less epochs as compared to the LSTM model (6 vs. 35 epochs). This suggests a spatial representation of the amino acid sequences might be more appropriate for a model to learn features rather than temporal-based representations. Utilization of dilated convolutions in the ResNet-based model seemed to help the model better learn more complex features, as the dilations enable larger receptive fields. The skip connections helped the model retain important spatial information from preceding layers. Overall, these components, as well as batch normalization which standardizes the data, helped to stabilize while accelerating the training process, significantly reducing the number of epochs necessary to acheive acceptable results.

REFERENCES

[1] Ofer, D., Brandes, N., &Linial, M. (2021). Thelanguage of proteins: NLP, machine learning &protein sequences. Computational and StructuralBiotechnology Journal, 19, 1750-1758.

[2] Vig, J., Madani, A., Varshney, L. R., Xiong, C.,Socher, R., & Rajani, N. F. (2020). BERTologymeets biology:interpreting attention in proteinlanguage models. arXiv preprint arXiv:2006.15222.

[3] Min, S., Park, S., Kim, S., Choi, H. S., Lee, B.,& Yoon, S. (2021). Pre-training of deepbidirectional protein sequence representations withstructuralinformation. IEEE Access, 9, 123912-123926.

[4] Brandes, N., Ofer, D., Peleg, Y., Rappoport, N.,&Linial, M. (2022). ProteinBERT: a universaldeep-learning model of protein sequence andfunction. Bioinformatics, 38(8), 2102-2110.

[5] Bileschi, M. L., Belanger, D., Bryant, D.,Sanderson, T., Carter, B., Sculley, D., ... & Colwell,L. J. (2019). Using deep learning to

annotate the protein universe. BioRxiv, 626507.

[6] Yuan, L.,Ma, Y., & Liu, Y. (2023). Ensemble deep learningmodels for protein secondary structure predictionusing bidirectional temporal convolution andbidirectional long short-term memory. Frontiers inBioengineering and Biotechnology, 11.

[7] Cheng, J., Liu, Y., & Ma, Y. (2020). Proteinsecondary structure prediction based on integrationof CNN and LSTM model. Journal of VisualCommunication and Image Representation, 71,102844.

[8] Zhang, Y., Ma, Y., & Liu, Y. (2022).Convolution-Bidirectional Temporal ConvolutionalNetwork for Protein Secondary StructurePrediction. IEEE Access, 10, 117469-117476.

[9] Soleymani, F., Paquet, E., Viktor, H. L.,Michalowski, W., &Spinello, D. (2023). ProtInteract: A deep learning framework forpredicting protein–proteininteractions. Computational and StructuralBiotechnology Journal, 21, 1324-1348.

[10] Guo, J. (2021, September). TCN-HBP: A DeepLearning Method for Identifying Hormone-BindingProteins from Amino Acid Sequences Based on aTemporal Convolution Neural Network. In Journalof Physics: Conference Series (Vol. 2025, No. 1, p.012002). IOP Publishing.

[11] Bepler, T., & Berger, B. (2019). Learningprotein sequence embeddings using informationfrom structure. arXiv preprint arXiv:1902.08661.

[12] Berman, H. M., Westbrook, J., Feng, Z.,Gilliland, G., Bhat, T. N., Weissig, H., ... &Bourne,P. E. (2000). The protein data bank. Nucleic acidsresearch, 28(1), 235-242.

[13] Khurana, D., Koli, A., Khatter, K., & Singh, S.(2023). Natural language processing: State of the art, current trends and challenges. Multimedia tools andapplications, 82(3), 3713-3744.

[14] Xu, Y., Verma, D., Sheridan, R. P., Liaw, A.,Ma, J., Marshall, N. M., ... & Johnston, J. M. (2020).Deep dive into machine learning models for proteinengineering. Journal of chemical information and modeling, 60(6), 2773-2790.

[15] Weinert, W. R., & Lopes, H. S. (2004). Neuralnetworks for protein classification. AppliedBioinformatics, 3, 41-48.