Contents lists available at ScienceDirect

Materials Today: Proceedings

journal homepage: www.elsevier.com/locate/matpr

CNN-LSTM based classification of polo like kinase family of Proteins: An emerging cancer drug target

Chinju John, Oommen K. Mathew, Jayakrushna Sahoo

Department of Computer Science and Engineering, Indian Institute of Information Technology, Kottayam 686635, India

ARTICLE INFO

Article history:
Available online 3 March 2022

Keywords:
Deep Learning
Convolutional Neural Network
Recurrent Neural Network
Sequence Classification
Cancer drug target

ABSTRACT

Artificial intelligence and deep learning are becoming an inevitable part of our life. Deep learning models are giving contributions to the identification of genetic causes behind various diseases affecting the human community. The prognosis and diagnosis of such diseases can be recommended with the help of artificial intelligence. Here we are proposing a novel deep learning model that employs the legendary deep learning architectures such as convolutional neural network and the variant of recurrent neural network known as long short-term memory network to classify the protein sequences belonging to the five-member polo like kinase family, a subclass of Serine-Threonine kinases, which is considered as an active anti-cancer drug target these days. The proposed deep learning model was trained and tested on sequences collected from biological sequence databases and classified the new sequences to their corresponding classes with an accuracy of 97.6%. Furthermore, this model could untangle the efforts associated with sequence annotation and classification, which is always a tedious and exigent task.

1. Introduction

Cancer is one among the life-threatening diseases that still remains a dilemma to mankind. According to the statistics, it is the second leading cause of death in United States [1]. Advancements in technology and medicine had influenced the mortality and early diagnosis of cancer to some extent [2–4]. Being the hot and trending research area, Artificial Intelligence has started to extend its arena spanning the health care industry too. Scientific community started to utilize the exclusive deep learning models in cancer detection, diagnosis, prognosis and treatment [4].Deep Learning is considered as a subset of Artificial Intelligence, which is designed to mimic the capabilities of human brain to certain limits. A well-structured and trained Deep Learning model can perform prediction and classification of cumbersome data with high accuracy and in a short span of time [5].

The recent advances in data science employs Deep Learning for image processing and pattern identification from large volume of data. Biological sequences are enriched sources of hidden patterns that can unlock the secrets to diseases such as cancer. These sequences are constructed with limited number of alphabets but can write the entire life story of a living organism. Differentiating such sequences with human intelligence is a tedious and vigorous task. There are gold standard tools such BLAST [6] for performing the similarity search of these biological sequences stored in the databases and identifying their phylogeny. Combination of Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) are proven to be outperforming in the text classification tasks [7].CNNs are capable of identifying the n-grams of alphabets from the input data with significant spatial–temporal relations through convolutional operations and pooling process[8].A variant of RNN, called, the Long short term memory network(LSTM) [9] was invented to relinquish the gradient exploding problem found in the traditional RNN. It is employed in sequence-based classification tasks with varying sequence lengths and even in the document level classification [10].

Polo Like Kinase belongs to the class of Serine/Threonine kinases that performs the regulation of cell cycles in living cells [11].This family of proteins consist of five members naming from PLK1 to PLK5.These macromolecules carry out the cellular division process with the aid of check-points. Overexpression of PLKs can lead to immature cell divisions which is considered as the basic trait of cancer. Studies had revealed the significance of targeting these proteins for controlling proliferation of cancerous cells, considering that each member of the family is identified to be taken

E-mail address: chinjuj.phd201001@iiitkottayam.ac.in (C. John)

https://doi.org/10.1016/j.matpr.2022.02.395
2214-7853/Copyright © 2022 Elsevier Ltd. All rights reserved.
Selection and peer-review under responsibility of the scientific committee of the International Conference on Artificial Intelligence & Energy Systems.

part in different types of cancers[12].Identifying the mutated protein sequences and thereby revealing the genetic backgrounds, is a strategy followed in the novel cancer therapies [13].

In this work, we employ a CNN-LSTM architecture for the classification of protein sequences which belong to the polo like kinase family. Being from the same family, these sequences are similar in pattern conservation and difficult to distinguish such regions [14]. Here the sequences are represented as vectors using the embedding layer[15]. These are given as input to the convolution layers for extracting the features from them. The following LSTM layer takes these high-level representations as input while preserving the order of these inputs. The input sequences for the model were obtained from biological databases like non-redundant proteins database and UniprotKB [16]. Our proposed model is capable of classifying the protein sequences to their corresponding class of family with a promising value of accuracy.

2. Related works

When the machine learning models faced the hurdles of feature extraction, deep learning came up with unsupervised feature extraction strategy, which had helped the data science community in many ways. Systematic research associated with biological information always deals with huge amount of data and hidden features. Biological sequences can be broadly classified as genomic sequences and protein sequences based on the molecules from which they are encoded from, Fig. 1. DNA sequences are the genetic codes that are transcribed to proteins. The constitutional residues of proteins and nucleic acids along with their corresponding alphabetical notations are listed in Table 1. There are four nucleotides that codes DNA whereas twenty amino acid residues are the basic units of a protein.

Table 1
Residue names of DNA and Protein.

Residue name	**Alphabet**
Adenine	A
Thymine	T
Guanine	G
Cytosine	C
Residue name	Alphabet
Alanine	A
Cystine	C
Aspartic acid	D
Glutamic acid	E
Phenylalanine	F
Glycine	G
Histidine	H
Isoleucine	I
Lysine	K
Leucine	L
Methionine	M
Asparagine	N
Proline	P
Glutamine	Q
Arginine	R
Serine	S
Threonine	T
Valine	V
Tryptophan	W
Tyrosine	Y

Convolutional Neural Networks are employed in classification of DNA sequences [17], for identifying the patterns in them [19] that are expressed as phenotypic characteristics. Bosco *et. al* conducted a comparison of deep learning architectures for the automated classification of bacteria species. Recognition of anti-microbial peptides were performed using a combinatorial architecture of CNN and RNN [26]. They have used protein sequences which were converted to padded vectors before giving as input to the convolutional layer. Length of the numerical vector was chosen based on the peptide with maximum amino acid count. CNN was employed for classifying protein families from raw protein sequences [17,18]. Here the protein sequences were represented as matrices of vectors and used for feature extraction without any prior domain knowledge. Protein folds can be identified using CNN, DeepSF [20] classifies the folds recognized from the protein sequences to one among the 1195 known folds. Advancements in next generation sequencing has contributed an immense growth in the volume of new biological sequences, annotation and labelling of such sequences are hectic and using machine

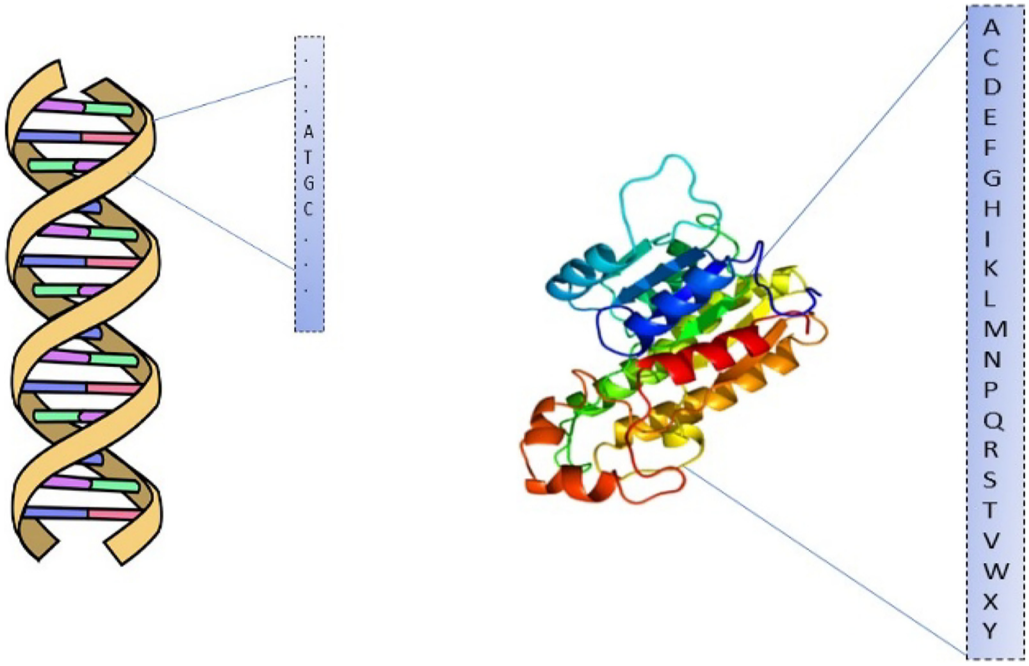


Fig. 1. Alphabets coding DNA and protein.

intelligence that could be performed easily. Seo *et.al* [21] introduced an alignment free method called DeepFam to capture the conserved regions of protein families using deep learning. The deep learning model proposed by Da Zhang and Mansur R. Kabuka uses only the CNN module for classification of protein families [22].

3. Materials and methods

3.1. Dataset

We build our dataset over the globally collected sequences from the biological databases, UniprotKB available at (*https://www.uniprot.org/*) and non-redundant protein sequences available at (*https://www.ncbi.nlm.nih.gov/refseq/*) respectively. Protein sequences collected based on the similarity search tool BLAST [6] was subjected to the removal of redundant sequences using the software Cd-hit, which works based on the clustering algorithm [23].The detailed steps followed for data preprocessing and cleaning is given below.

- Reviewed and annotated reference sequences for each class of PLK family were chosen from UniprotKB/swiss-prot using the class name as keyword.
- BLAST was used for identifying the similar sequences from UniprotKB database with restriction parameters like query coverage as 70 and percentage of identity as 60.
- BLAST was used for identifying the similar sequences from non-redundant protein database with restriction parameters of query coverage as 70 and percentage of identity as 60.
- Obtained sequences were combined and gave as input to Cd-hit with identity parameter as 100% in order to fetch then unique sequences.

Even though these sequences belong to the same family their lengths were not the same, hence the length of the largest sequence from the input was considered for padding. Summary of the input sequences are given in Table 2.

3.2. Deep learning architecture

The proposed deep learning architecture consist of five modules: the encoding module, the embedding module, the CNN module with activation layers, the LSTM module and the fully connected module with multi class classifier as depicted in the Fig. 3.

3.2.1. Encoding module

Each input sequence is constructed with alphabets that are corresponding to one of the twenty amino acid residues. These alphabets need to be numerated for giving as input for convolution operation. Tokenizer module from Keras(*https://github.com/fchollet/keras*) library was used for encoding each alphabet to a number. In order to make sure that all the sequences are of same length, padding with zeros was performed. Here the maximum length was considered as 1296 since the longest sequence had that many residues. The labels of the corresponding sequences were converted to one-hot representation using the label-binarize (*https://github.com/scikit-learn/scikit-learn/blob/844b4be24/sklearn*) module. The sequences are expected to be classified into five classes; hence they were converted to a binary vector, with each vector corresponding to the output class.Fig. 2. gives the part of an input sequence which was converted to the corresponding numerical representation using encoding module, numbers are unique to the corresponding alphabet.

The common encoding schemes [24] used in bioinformatics incorporate the domain knowledge, whereas one hot-encoding does not require any prior information. After encoding the input sequences will be converted to numerical vectors with each having a length of 1296.

3.2.2. Embedding module

Embedding module is responsible for converting the indices of amino acids into a representation of a fixed vector.Word-level embeddings are encoded using column vectors of the input matrix. A word *w* is transformed into its word-level embedding r^{wrd} using matrix–vector product [25]. Where υ^w is a vector of size $|V^{wrd}|$ which has value 1 at index *w* and zero in all other positions.

$$r^{wrd} = W^{wrd}\upsilon^w \tag{1}$$

We have given the length of embedding dimension as eight, since our dataset is not large volume and increased embedding dimensions can include more features of each amino acid [24]. and the process was done using the keras embedding layer. Embedding allows us to use the continuous metric notions, that can resemble the semantic characteristic of each amino acid.

3.2.3. CNN module

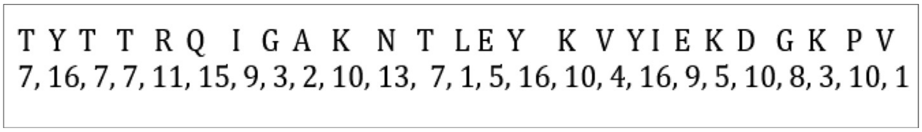
CNNs are composed of convolutional layers, non-linear layers, pooling layers and fully connected layers. Convolutional layers are capable of identifying temporal relationship of the input data - feature extraction, and pooling layers are responsible for representing input to the next level. Protein sequences now converted to a matrix format, will be given to the convolution layer of the CNN module. The neuron layers of this module can perform the data traversing to identify the local sequential patterns of significance. Here we have performed 1D convolutions three times with number of convolutions as 64 ,32 and 16 respectively, on the input data which is in numerical format. The feature detectors known as kernels will traverse through the entire sequence input to fetch the points that contribute to the pattern. We have used ReLU as the activation function:

$$y = \underset{i}{maxReLU}(f[i] + b) \tag{2}$$

Where *f[i]* is the output of convolution layer, the feature maps. Maxpooling layer following the convolution layer allows for the deeper training of the model. These layers down sample the feature maps by summarizing the presence of significant patterns or features in form of small patches. The maxpooling mechanism which we have used can select the optimal feature by taking the maximum value of multiple features.

Table 2
Summary of input data.

Class name	NR database	UniportKB	Cd-hit output	Final count	Average length
PLK1	609	50	114	112	594
PLK2	893	4	76	76	678
PLK3	696	3	57	56	661
PLK4	795	9	105	105	957
PLK5	135	1	79	79	579



T Y T T R Q I G A K N T L E Y K V Y I E K D G K P V
7, 16, 7, 7, 11, 15, 9, 3, 2, 10, 13, 7, 1, 5, 16, 10, 4, 16, 9, 5, 10, 8, 3, 10, 1

Fig. 2. Part of input sequence after converting into numbers.

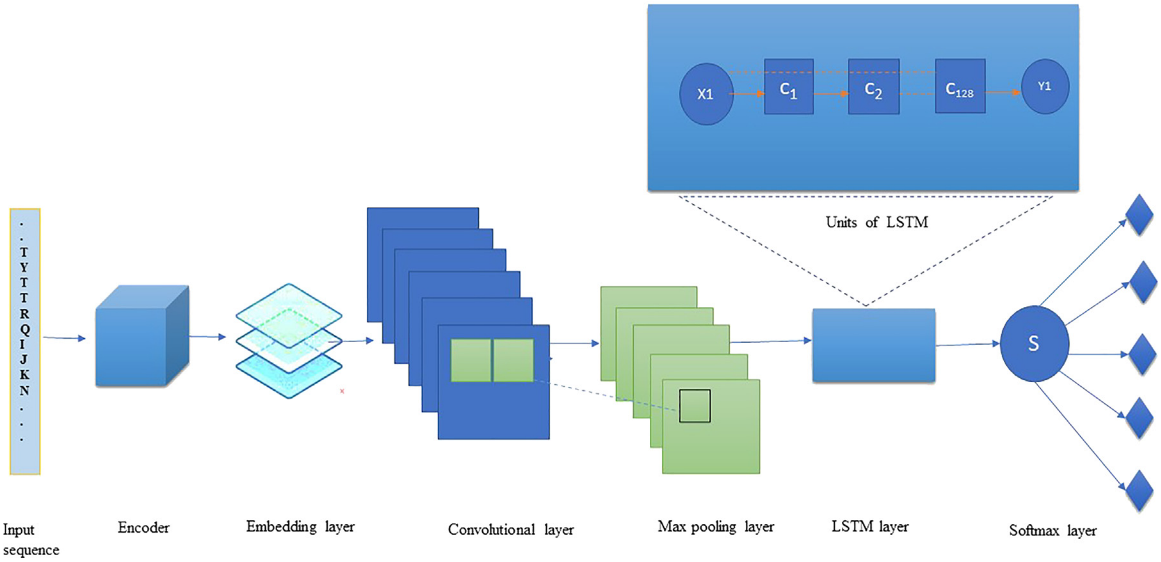


Fig. 3. Architecture of proposed deep learning model.

3.2.4. LSTM module

LSTM was designed to overcome the drawbacks associated with back-propagation in RNNs [9].It consists of small modules known as memory cells, through with each character of input data will be passed and processed. The input gate i_t of LSTM protects the memory content of cell from perturbation due to other inputs. Similarly, the output gate protects from irrelevant memory contents. Input gate of one memory cell uses the inputs of other cells to decide whether the information is relevant enough to be stored. Each LSTM unit is comprised of the following gates, where W and U are parameter matrices and b is a bias vector as outlined in [9]. The cell activation gate C_t considers the previous cell $(c\ _t)$ and information is forwarded by f_t to alter the hidden states h_t and final output vector o_t [26]. We have employed an LSTM layer with 128 units that takes input from the subsequent max pooling layer of CNN module.

$i_t = \sigma(W_i x_i + U_i h_{t-1} + b_i)$ input gate

$C_t = \tanh(W_c x_t + U_c h_{t-1} + b_c)$ candidate cell states

$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f)$ forget gate

$C_t = i_t o C_t + f_t o C_{t-1}$ cell activation gate

$o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o C_t + b_o)$ output gate

$$h_t = o_t o \tanh(C_t) \text{hiddengate} \tag{3}$$

3.2.5. Fully connected module

The output from LSTM module is concatenated and flattened to a single vector and fed it to a fully connected layer, which employ Softmax function for classification.

$$\sigma\left(\vec{z}\right)_i = \frac{e^z i}{\sum_{J=1}^{k} e^{z_j}} \tag{4}$$

Where K is the number of classes which we have set as five for the five members of PLK family.

4. Experimental results

4.1. Model evaluation metrics

The evaluation parameters which we have followed in our study are the metrics such as Accuracy, Recall, F1-score and Precision. Below given are the mathematical definition of each metric.

$$Accuracy = \frac{(TP + TN)}{TP + TN + FP + FN} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$F1 - Score = \frac{2 * recall * precision}{recall + precision} \tag{8}$$

We have also used receiver-operating characteristic (ROC) curve for showing the performance of the model based on a threshold value [27].

4.2. Model construction and fine-tuning

The proposed model was constructed using the deep learning frame work API Keras *vr.2.6.0* with a GPU-based TensorFlow *vr.2.6.0* at the back-end. The hyper parameters (Table 4.) were identified using the library function GridsearchCV. Grid Search is an exhaustive search that looks through all combinations of hyperparameters [28].Based on the results we have chosen the batch size as 16 and number of epochs as 60.In order to reduce the overfitting tendency of model we have introduced drop out layers after each max-pooling layer [29].We have studied how the model is varying

Table 4
Hyperparameters of the model.

Number of epochs	60
Drop out values	0.3 ,0.2
Embedding dimension	8
FC layer size	128
Batch size	16
Optimizer	Adam

Table 5
Classification efficiency comparison of the model with respect to number of convolutional layers.

Class	Precision	Recall	F1-score
Number of convolutional layers = 1 Accuracy = 88.37			
PLK1	0.70	1.00	0.82
PLK2	0.92	0.71	0.80
PLK3	0.92	0.80	0.80
PLK4	0.94	1.00	0.97
PLK5	1.00	0.90	0.95
Number of convolutional layers = 2 Accuracy = 94.13			
PLK1	0.96	0.93	0.95
PLK2	0.86	0.92	0.89
PLK3	1.00	0.90	0.95
PLK4	0.94	1.00	0.97
PLK5	0.95	0.90	0.95
Number of convolutional layers = 3 Accuracy = 97.67			
PLK1	1.00	0.96	0.98
PLK2	0.92	0.92	0.92
PLK3	0.93	1.00	0.96
PLK4	1.00	1.00	1.00
PLK5	1.00	1.00	1.00

in terms of efficiency subjected to the change in the number of convolutional layers. The results are depicted in Table 5. A 10-fold cross validation was performed to ensure the efficiency of the model while handling new input data[30].For the optimization of network weights during the training of the model we have used Adam[31].

4.3. Results

Once the parameters were fixed, we have constructed models by changing the number of convolution layers. Since biological data is more vulnerable to overfitting [32], we kept the number of layers to the optimum. A table comparing the values of precision, recall and F1-score of the models are given (Table no. 5). Classification accuracy of the models were changing proportionally with respect to the number of convolutional layers, the deep learning model with three convolutional layers was able to provide an accuracy of 97.67, which comparatively a good value. The ROC curves

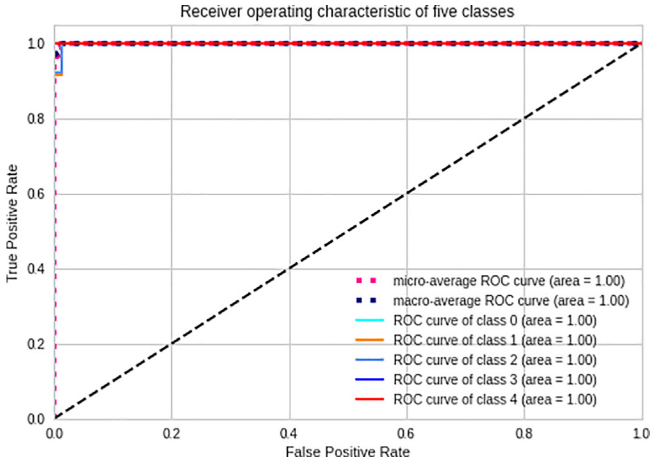


Fig. 4. ROC scores of five classes.

representing the classification accuracy of the chosen deep learning model is given in Fig. 4. It is visible from the plot that

the sensitivity and specificity parameters are good enough to rely the model for the assigned task. One-vs-One ROC AUC scores:0.998801 (macro),0.999014 (weighted by prevalence) One-vs-Rest ROC AUC scores: 0.998801 (macro),0.999014 (weighted by prevalence).

5. Conclusion and future work

In this work, we have designed and implanted a hybrid deep learning model by combining the legendary CNN and LSTM architectures. Our model takes protein sequences as input and determine whether they are a member of PLK family, then classify it under the corresponding member. Deep learning models are employed nowadays for solving significant biological data manipulations, our model contributes to the classification paradigm in a novel way by leveraging the advancements in Artificial Intelligence. In future we will be taking our deep learning model to perform the classification of protein sequences that could be derived from the entire genetic material present in an organism.

CRediT authorship contribution statement

Chinju John: Conceptualization, Writing – original draft. **Oommen K. Mathew:** Methodology, Data curation, Supervision. **Jayakrushna Sahoo:** Writing – review & editing, Validation, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] R.L. Siegel, K.D. Miller, H.E. Fuchs, A. Jemal, Cancer Statistics, 2021, CA. Cancer J. Clin. 71 (1) (2021) 7–33, https://doi.org/10.3322/caac.21654.

[2] K. Munir, H. Elahi, A. Ayub, F. Frezza, and A. Rizzi, “Cancer diagnosis using deep learning: A bibliographic review,” *Cancers*, vol. 11, no. 9. 2019, doi: 10.3390/cancers11091235.

[3] L. Shen, L.R. Margolies, J.H. Rothstein, E. Fluder, R. McBride, W. Sieh, Deep Learning to Improve Breast Cancer Detection on Screening Mammography, Sci. Rep. 9 (1) (2019) 1–12, https://doi.org/10.1038/s41598-019-48995-4.

[4] A. Kleppe, O.-J. Skrede, D. J. Kerr, and H. E. Danielsen, “in cancer diagnostics,” *Nat. Rev. Cancer*, vol. 21, no. March, 2021, doi: 10.1038/s41568-020-00327-9.

[5] Ian Goodfellow,Yoshua Bengio and Aaron Courville.Deep learning. MIT press, 2016.

[6] S. McGinnis, T.L. Madden, BLAST: at the core of a powerful and diverse set of sequence analysis tools, Nucleic Acids Res. 32 (Web Server) (2004) W20–W25.

[7] C. Zhou, C. Sun, Z. Liu, and F. C. M. Lau, “A C-LSTM Neural Network for Text Classification,” 2015.

[8] M. Denil, A. Demiraj, N. Kalchbrenner, P. Blunsom, and N. de Freitas, “Modelling, Visualising and Summarising Documents with a Single Convolutional Neural Network,” pp. 1–10, 2014.

[9] S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Comput. 9 (8) (1997) 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735.

[10] D. Tang, B. Qin, T. Liu, Document modeling with gated recurrent neural network for sentiment classification, Conf. Proc. - EMNLP 2015 Conf. Empir. Methods Nat. Lang. Process. no. September (2015) 1422–1432, https://doi.org/10.18653/v1/d15-1167.

[11] S.-Y. Lee, C. Jang, K.-A. Lee, Polo-Like Kinases (Plks), a Key Regulator of Cell Cycle and New Potential Target for Cancer Therapy, Dev. Reprod. 18 (1) (2014) 65–71, https://doi.org/10.12717/dr.2014.18.1.065.

[12] Y. Degenhardt, T. Lampkin, Targeting polo-like kinase in cancer therapy, Clin. Cancer Res. 16 (2) (2010) 384–389, https://doi.org/10.1158/1078-0432.CCR-09-1380.

[13] H.F.M. Kamel, H.S.A.B. Al-Amodi, Exploitation of Gene Expression and Cancer Biomarkers in Paving the Path to Era of Personalized Medicine, Genom. Proteom. Bioinforma. 15 (4) (2017) 220–235, https://doi.org/10.1016/j.gpb.2016.11.005.

[14] D.M. Lowery, D. Lim, M.B. Yaffe, Structure and function of Polo-like kinases, Oncogene 24 (2) (2005) 248–259, https://doi.org/10.1038/sj.onc.1208280.

[15] T. Demeester, T. Rocktäschel, and S. Riedel, "Lifted rule injection for relation embeddings," *EMNLP 2016 - Conf. Empir. Methods Nat. Lang. Process. Proc.*, pp. 1389–1399, 2016, doi: 10.18653/v1/d16-1146.

[16] D. Zou, L. Ma, J. Yu, Z. Zhang, Biological databases for human research, Genom. Proteom. Bioinforma. 13 (1) (2015) 55–63, https://doi.org/10.1016/j.gpb.2015.01.006.

[17] N.G. Nguyen, V.A. Tran, D.L. Ngo, D. Phan, F.R. Lumbanraja, M.R. Faisal, B. Abapihi, M. Kubo, K. Satou, DNA Sequence Classification by Convolutional Neural Network, J. Biomed. Sci. Eng. 09 (05) (2016) 280–286, https://doi.org/10.4236/jbise.2016.95021.

[18] J. Khanal, H. Tayara, K.T. Chong, Identifying Enhancers and Their Strength by the Integration of Word Embedding and Convolution Neural Network, IEEE Access 8 (2020) 58369–58376, https://doi.org/10.1109/ACCESS.2020.2982666.

[19] R. Umarov, H. Kuwahara, Y. Li, X. Gao, and V. Solovyev, "Promoter analysis and prediction in the human genome using sequence-based deep learning models," *Bioinformatics*, vol. 35, no. 16, pp. 2730–2737, 2019, doi: 10.1093/bioinformatics/bty1068.

[20] J. Hou, B. Adhikari, and J. Cheng, "DeepSF: Deep convolutional neural network for mapping protein sequences to folds," *arXiv*, 2017.

[21] S. Seo, M. Oh, Y. Park, S. Kim, DeepFam: Deep learning based alignment-free method for protein family modeling and prediction, Bioinformatics 34 (13) (2018) i254–i262, https://doi.org/10.1093/bioinformatics/bty275.

[22] D. Zhang, M. Kabuka, Protein Family Classification from Scratch: A CNN based Deep Learning Approach, IEEE/ACM Trans. Comput. Biol. Bioinforma. 14 (8) (2020) 1, https://doi.org/10.1109/tcbb.2020.2966633.

[23] "cd-hit.pdf." .

[24] H. Elabd, Y. Bromberg, A. Hoarfrost, T. Lenz, A. Franke, M. Wendorff, Amino acid encoding for deep learning applications, BMC Bioinform. 21 (1) (2020) 1–14, https://doi.org/10.1186/s12859-020-03546-x.

[25] C. N. Dos Santos and B. Zadrozny, "Learning character-level representations for part-of-speech tagging," *31st Int. Conf. Mach. Learn. ICML 2014*, vol. 5, no. 2011, pp. 3830–3838, 2014.

[26] D. Veltri, U. Kamath, and A. Shehu, "Deep learning improves antimicrobial peptide recognition," *Bioinformatics*, vol. 34, no. 16, pp. 2740–2747, 2018, doi: 10.1093/bioinformatics/bty179.

[27] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1) (1982) 29–36, https://doi.org/10.1148/radiology.143.1.7063747.

[28] D.K. Barupal, O. Fiehn, Generating the blood exposome database using a comprehensive text mining and database fusion approach, Environ. Health Perspect. 127 (9) (2019) 2825–2830, https://doi.org/10.1289/EHP4713.

[29] M. Melucci, Relevance Feedback Algorithms Inspired by Quantum Detection, IEEE Trans. Knowl. Data Eng. 28 (4) (2016) 1022–1034, https://doi.org/10.1109/TKDE.2015.2507132.

[30] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," *Int. Jt. Conf. Artif. Intell.*, no. March 2001, 1995.

[31] "Adam_ A Method for Stochastic Optimization _ Enhanced Reader.pdf." .

[32] C. Angermueller, T. Pärnamaa, L. Parts, O. Stegle, Deep learning for computational biology, Mol. Syst. Biol. 12 (7) (2016) 878, https://doi.org/10.15252/msb.20156651.