Received: 30 November 2021 | Revised: 2 January 2022 | Accepted: 17 January 2022 | Accepted article online: 3 February 2022

DOI: 10.1002/pmic.202100197

Proteomics
Proteomics and Systems Biology

REVIEW

Protein–DNA/RNA interactions: Machine intelligence tools and approaches in the era of artificial intelligence and big data

Feifei Cui[1,2] | Zilong Zhang[1,2] | Chen Cao[2] | Quan Zou[1,2] | Dong Chen[3] | Xi Su[4]

[1] Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China

[2] Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, China

[3] College of Electrical and Information Engineering, Quzhou University, Quzhou, China

[4] Foshan Maternal and Child Health Hospital, Foshan, Guangdong, China

Correspondence
Dong Chen, College of Electrical and Information Engineering, Quzhou University, Quzhou 324000, China.
Email: peakgrin@outlook.com
Xi Su, Foshan Maternal and Child Health Hospital, Foshan, Guangdong, China.
Email: xisu_fsfy@163.com

Abstract

With the development of artificial intelligence (AI) technologies and the availability of large amounts of biological data, computational methods for proteomics have undergone a developmental process from traditional machine learning to deep learning. This review focuses on computational approaches and tools for the prediction of protein–DNA/RNA interactions using machine intelligence techniques. We provide an overview of the development progress of computational methods and summarize the advantages and shortcomings of these methods. We further compiled applications in tasks related to the protein–DNA/RNA interactions, and pointed out possible future application trends. Moreover, biological sequence-digitizing representation strategies used in different types of computational methods are also summarized and discussed.

KEYWORDS
artificial intelligence, biological data, deep learning, feature representation, machine learning, protein–DNA/RNA interaction, proteomics

1 | INTRODUCTION

The interactions among biomolecules comprise a universal theme in living organisms, and an analysis of this is not only the focus of many disciplines but also the focus of applied research. Biomolecular interactions mainly include nucleic acid–nucleic acid interactions, nucleic acid–protein interactions, and protein–protein interactions, etc. Among them, the interactions between proteins and nucleic acids (general terms of DNA and RNA) are widely prevalent in the regulation of life activities, such as gene replication, transcription, translation, modification, and other processes that are inseparable from the interactions between DNA/RNA and protein [1–4]. By understanding the function of proteins in interactions, we can further understand the mechanisms of related cellular process mechanisms, such as viral infections, or the design of new drug targets. For example, in the initial research phase of the new type of coronavirus pneumonia (COVID-19) caused by SARS-CoV-2 infection, one of the most urgent tasks is to understand the mechanisms underlying the molecular interactions between protein and viral RNA, which might promote virus replication or facilitate host defense in infected cells [5,6]. Therefore, it is important to understand the interactions between proteins and DNA/RNA.

Research methods to study the interactions between proteins and DNA/RNA include experimental and computational techniques. Experimental approaches for protein–DNA/RNA interactions, such as electrophoretic mobility shift assay [7,8], chromatin immunoprecipitation (ChIP) [9], X-ray diffraction crystallography, and UV-crosslinking immunoprecipitation, are usually time-consuming and costly [10,11]. Therefore, computational methods have become increasingly popular in the field of bioinformatics since the 1990s owing to the development of machine intelligence methods in the era of artificial intelligence (AI).

AI has been rapidly and radically changing in various areas of industry and our lives as AI technology is flourishing and demanding. The field of bioinformatics, which aims to improve our understanding of biological data by developing methods, tools, and software, is a major benefactor of the recent advancements in AI. Because of the application of AI technologies to bioinformatics [12,13], computational methods for the prediction of interactions between proteins and DNA/RNA have undergone a developmental process from traditional machine learning to deep learning. This review focuses on computational approaches and tools using machine intelligence techniques to understand protein-DNA/RNA interaction. We summarized the

machine learning (deep learning) approaches that can be utilized to predict interactions between protein and DNA/RNA, including their advantages and shortcomings, applications in tasks related to the protein–DNA/RNA interactions, and possible future application trends.

2 | COMPUTATIONAL METHODS IN THE ERA OF AI: FROM MACHINE LEARNING TO DEEP LEARNING

2.1 | Development process for computational methodsin the AI era

Computational approaches can be classified into sequence and structure-based methods. Because three-dimensional (3D) protein structure is generally more conserved and has more information compared to the amino acid sequences (primary structures), structural similarity is a good indicator of similar functions among proteins. Moreover, structure-based methods usually perform better than sequence-based methods. Structural information, such as accessible surface area (or solvent-accessible surface area), secondary structure (including hydrogen-bonding potential, helical conformation, etc.), can be utilized as feature vectors for protein function prediction. These structural feature data are typically obtained through DSSP algorithm and protein databases such as Protein Data Bank (PDB). Structural features separately or combined with sequence-based features are commonly employed for protein function prediction models including statistical models and machine learning-based models that uses computational algorithms such as statistics singular value decomposition algorithms [14] and machine learning algorithms (e.g., naive Bayesian method, random forest) [15,16]. However, structures are not available for most proteins. When the structure of a target protein has not yet been experimentally determined, computational structure prediction methods can be applied to model the 3D structure. However, this approach cannot obtain good accuracy when homologous templates are not available at the early stage. Thus, many sequence-based computational methods have been developed.

In the era of AI, the sequence-based computational methods in proteomics have roughly gone through a process from homology-based methods and machine learning-based methods to deep learning-based methods. Homology-based methods predict the function of a protein by computing the similarity between biological sequences (sequence identity). Sequences with high similarity are generally considered to have similar functions. Sequence similarity can be computed using multiple sequence alignment tools (e.g., BLAST [17]) or by searching databases, such as the UniProtKB database [18], Gene Ontology database [19], Pfam [20], and PROSITE [21], in which the entries can be used for protein functional prediction. However, for some specific tasks, the homology-based method is not realistic when few databases of target sequences are available.

Owing to the continued development of AI techniques and the increasing maturity of machine learning algorithms, machine learning algorithms have become increasingly popular in sequence-based computational methods for the prediction of protein–DNA/RNA interactions. In the 2000s, traditional machine learning methods, such as support vector machine (SVM), decision trees, naïve Bayes classifier, random forest [22,23], and logistic regression, were most popular. These algorithms cannot be directly applied to raw data (biological sequences). Instead, a preprocessing step called feature extraction is required. This step requires human experts to have a detailed knowledge of the task, adaptation, testing, and refinement over several iterations. Therefore, machine learning is considered more dependent on manual human intervention.

Due to the limitations of traditional machine learning methods and arrival of the big data era, deep learning methods, a subset of machine learning, has gradually surpassed traditional machine learning methods. Since 2005, deep learning has brought large changes to AI; in general, deep learning has driven many AI applications and services that improve automation by performing analytical and physical tasks without human intervention [24,25]. Deep learning is essentially a neural network with several neural network layers (three or more layers) and large-scale training datasets. Precisely because of the deeper neural network layers and larger training datasets, deep learning-based applications typically outperform traditional machine learning-based models. In the past decade, as a large amount of biological data have become available [26,27], deep learning approaches have been widely utilized in proteomics [28,29]. In one study [28], convolutional neural networks (CNNs) were used to predict the sequence specificities of DNA- and RNA-binding proteins using primary biological data. Another study [30] presented a prediction algorithm for prediction protein subcellular localization using CNN and a recurrent neural network (RNN) only from sequences. Hashemifar et al. [31] presented a deep learning framework for predicting protein–protein interactions based on amino acid sequences, in which protein sequence pairs are projected onto a representation by convolutional modules. Kulmanov et al. [32] developed a novel classifier for protein functional prediction based on sequence and interaction information, in which sequence data were represented as trigrams of amino acids and embedded through CNN layers.

2.2 | Categorization of computational methods

Classical machine learning (deep learning) can be categorized into four basic approaches, unsupervised learning, semi-supervised learning, supervised learning, and reinforcement learning [33]. The type of algorithm you chosen depends on the type of data available. In supervised learning, labeled inputs and targets (desired outputs) are required to train the algorithm. Supervised learning algorithms are usually suitable for classification tasks (binary or multiclass classification), regression problems (prediction of continuous values), and ensemble models (combinations of multiple machine learning models). Frequently used algorithms in supervised learning include SVM, random forest, and logistic regression. They are popular in protein functional prediction, such as diverse types of protein binding [34–36]. Unsupervised

learning refers to training with unlabeled data to group data into subsets by sifting through unlabeled input data. It is good for the clustering tasks (grouping dataset based on similarity) and dimensionality reduction, among other tasks. Dimensionality reduction algorithms (unsupervised learning), such as principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE), are commonly used for protein sequence feature visualization [37–39]. Semi-supervised learning refers to the training of a model with a small amount of labeled data, which is then used for the new unlabeled data. Semi-supervised learning algorithms can not only avoid the disadvantage of labeling data (expensive and cost time), but also be used to implement the improvements in algorithm performance brought about by training with labeled data. The size of protein database shows an explosive growth because of new sequencing technologies. However, the development of annotated subsets lags far behind due to the high cost of acquiring meaningful labels and annotations. Therefore, semi-supervised protein representation learning has emerged as an import pattern in protein modeling [40]. Reinforcement learning, which is a goal-oriented learning method rather than classifying or clustering data, is usually utilized with deep learning. It can be considered as an embedded framework in deep learning. For example, the famous AlphaGO [41] and AlphaFold [42], provided by DeepMind lab, have demonstrated that reinforcement deep learning (combining reinforcement learning with deep learning) is applicable in game playing and protein folding problems.

3 | COMPUTATIONAL APPROACHES UTILIZED IN PROTEIN–DNA/RNA INTERACTIONS-RELATED TASKS

The process of predicting protein–DNA/RNA interactions using computational methods can be roughly summarized as follows: (1) sequence data digitization, (2) model construction and training, and (3) prediction. In the era of AI and big data, machine learning (deep learning) methods have been widely used in tasks related to the DNA/RNA–protein interactions (Figure 1).

The computational model for prediction should be designed according to the principles of the algorithms introduced previously herein and the characteristics of the available data. Generally speaking, traditional machine learning method-based schemes are chosen when the amount of data is small; deep learning can be utilized when large amounts of data are available. This is because the accuracy usually increases with an increasing amount of training data in deep learning models, and models can be scaled better with a larger amount of data, whereas traditional machine learning models might stop improving after a saturation point [43]. In addition, with the development AI and machine learning (deep learning), diverse frameworks and AI tools have been established, such as Scikit Learn [44], Keras [45], PyTorch [46], Tensorflow [47], Theano [48], and Weka software [49], among others. These frameworks can be used to construct task models using machine learning (deep learning) algorithms. Further, they can be chosen depending on what best meets the desired requirements.

Sequence digitizing refers to converting the biological sequences to numerical data (e.g., vectors) that can be used as input for computational models. In traditional machine learning-based prediction models, the extracted features are usually selected to compose the digitized representation of sequences. For deep learning-based prediction models, digitized sequences are acquired by sequence representation learning (sequence encoding or embedding), either together with the training of the model or by utilizing a deep neural network-based learning model. Sequence data digitization is the first necessary step. In this section, we focus mainly on the types of sequence digitizing schemes.

3.1 | Feature representation for machine learning models

In traditional machine learning models, such as SVMs, decision trees and random forests, feature extraction should be preliminarily conducted. In proteomics, composition information, physicochemical properties (such as hydrophobicity, polarizability, volume, helix probability, sheet probability, isoelectric point, and steric parameters) of amino acids and protein evolutionary information (e.g., position-specific scoring matrix, PSSM) can be manually acquired through the third-party applications [50,51]. For example, Hwang et al. [34] developed a DNA-binding residue predictor by using three machine learning approaches, including SVM and two types of logistic regression methods and PSSM to form feature vectors. Ballester et al. [52] presented a novel computational method for predicting protein-ligand binding affinity using random forest and intermolecular interaction features. Kong et al. [53] proposed an SVM-based classifier to assess the protein-coding potential of a transcript using features extracted from the nucleotide sequence of the transcript. Then, statistical measures were determined to be required as follows: assign a score to each of the extracted features, rank the features according to their scores, and decide whether to keep the form feature vectors or remove them. The extracted features are independent and univariate, and if feature selection fails, the machine learning model will not perform well.

3.2 | Deep learning-based feature learning

The process of protein-related prediction using a deep neural network-based model can be roughly described in three parts as follows: feature representation of sequence data (sequence encoding or embedding); model construct and training; and prediction outputs. The primary biological sequences (such as amino acid sequences) can be directly be used as inputs for the deep learning model, instead of relying entirely on third-party applications to extract features to form a numerical representation of the biological data, as in traditional machine learning methods. The representation learning methods of biological sequences can be categorized into several groups as follows: end-to-end learning, LSTM-based representation learning (non-contextual sequence embedding), and transformer-based sequence representation learning [54]. Except for graph computation-based

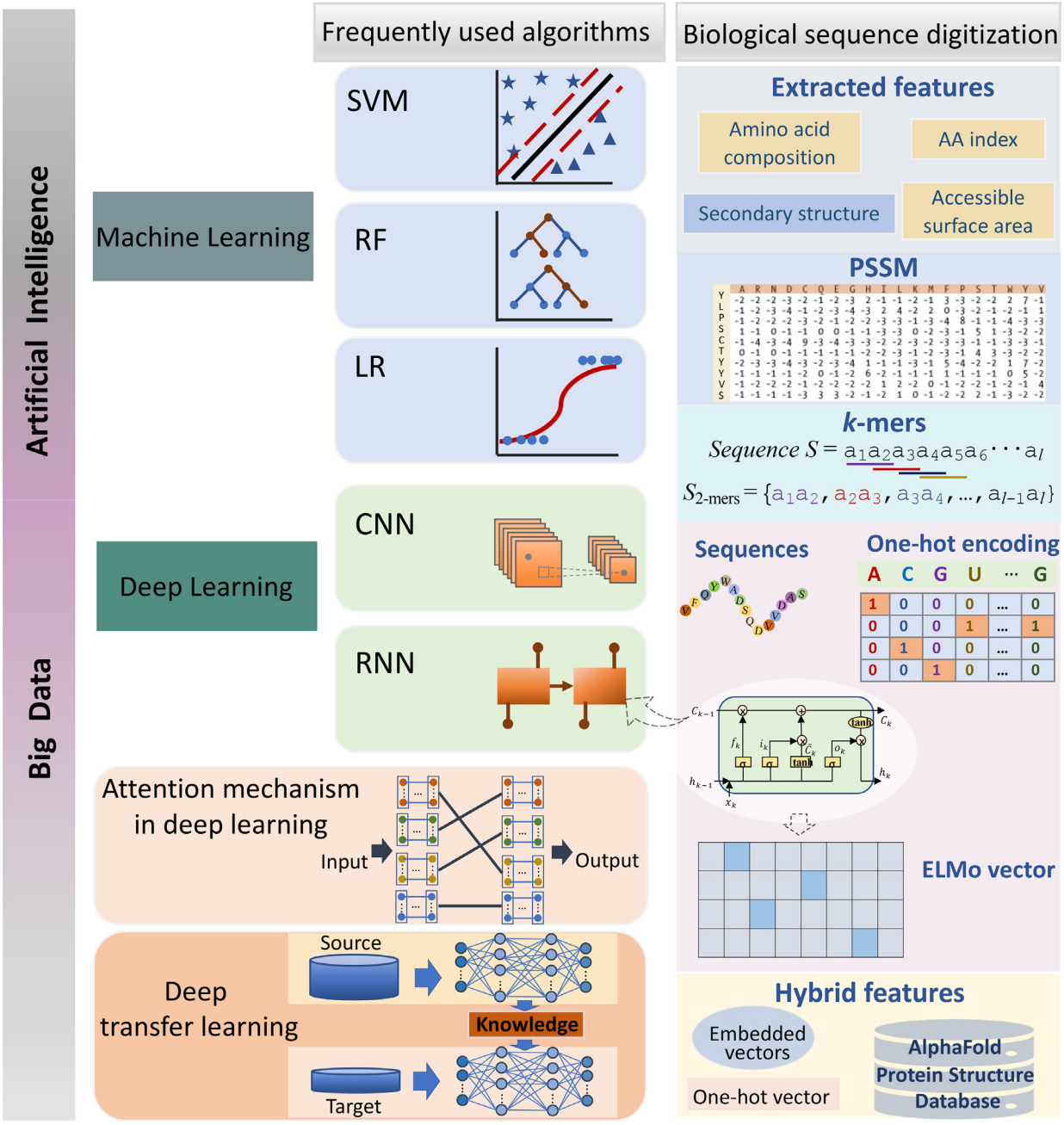

FIGURE 1 Machine learning algorithms frequently used in research related to the DNA/RNA–protein interactions. Abbreviations: AA, amino acid; CNN, convolutional neural network; LR, Logistic regression; PSSM, position specific scoring matrix; RF, random forest; RNN, recurrent neural network; SVM, support vector machine.

representation and mixed representation, the other three representation methods have been applied in bioinformatics over time according to their emergence in the field of natural language processing. One-hot encoding as a typical encoding method of the end-to-end model has been widely used for proteomics, such as the prediction of protein function [32], prediction of protein subcellular localization [30], and identification of disordered regions [55]. Subsequently, the ELMo model and its variant models that are based on LSTM have been used in proteomics research for sequence representation learning. For example, Heinzinger et al. [37] utilized a bidirectional LSTM-based model to learn sequence data for the prediction of protein function. Alley et al. [39] proposed a unified representation for amino acid sequences learned from a multiplicative LSTM-based model, and it was proven that the representation model has the ability to significantly improve the efficiency of protein engineering tasks. More recently, the transformer-based language model for natural language processing, namely the bidirectional encoder representation from transformers (BERT) [56,57], has attracted substantial attention because of its excellent performance. The protein prediction tasks based on primary sequences are similar to the tasks in the NLP field;

therefore, the application of the BERT model to bioinformatics is worth exploring.

In general, from traditional machine learning to deep neural networks, the flexibility and performance of the computational model have gradually improved. However, there are still some problems in deep learning-based biological sequence representation, such as the lack of interpretability, which is currently one of the mainstream research directions of deep learning and is widely regarded as a key part of the next generation of AI technology. Therefore, representation learning based on deep learning for biological data should consider strategies for an interpretable deep neural network. Deep neural networks are habitually considered black box models, from which model parameters and high-fit identification results are acquired. However, in addition to the final identification results, the output results based on the knowledge that the model can learn from the data should also be considered. Two types of interpretation strategies for deep learning can be considered, specifically post hoc interpretability analysis and ad hoc interpretable modeling. The former refers to interpreting the trained model, whereas the latter refers to the design of an interpretable model from scratch. For biological data representation in proteomics, feature analysis techniques, such as the visualization of learned features (characteristics of neurons and layers of neural networks) as post hoc interpretability analysis, are more suitable. Aiming for an ad hoc interpretable modeling strategy (interpretable representation) to represent biological data, it can be suggested that mixed data representations, such as sequence and structure information, are used for data representation, integrating traditional graph-based technologies into existing deep neural networks (such as CNN).

In recent years, the emergence of AlphaFold [58] has resulted in considerable advances in protein structure prediction, and a series of highly reliable protein structure computational models make structural information available [59–61]. Recently, free access to the AlphaFold protein structure database [42] (developed by DeepMind and EMBL-EBI) has been provided. The AlphaFold database covers the most complete and accurate high-quality predictions of the whole protein structure of the human proteome and 20 other key organisms. Therefore, highly reliable protein structure data acquired from the database can be used for proteomics tasks, especially for the feature representation of biological data, which might have important implications for protein engineering.

4 | APPLICATIONS IN TASKS RELATED TO PROTEIN–DNA/RNA INTERACTIONS

Protein–DNA/RNA interaction-related research mainly includes identification of proteins that bind to DNA/RNA, prediction of binding sites, and analysis of sequence specificities of DNA- and RNA-binding proteins. In this section, we focus mainly on tools and approaches developed on these three research topics, including types of biological sequence representation (feature extraction), machine learning (deep learning) algorithms used in these tools, and web servers or code available provided by researchers, as shown in Table 1.

For model algorithms, shallow learning algorithms were used in the early research works, among which SVM is almost the most widely used method, followed by RF and logistic regression, etc. For instance, RPISeq [62], mRMR-IFS-SVM [35], RPI-Pred [10], and StackDPPred [11] utilize SVM algorithm to identify DNA/RNA-binding proteins; SVM classifier and its variants are employed for the prediction of binding sites, such as BindN [63], DP-Bind [64], DISIS [65], PiRaNhA [36], and TriPepSVM [66]. Studies in recent years focus on deep learning algorithms, among which CNN and LSTM (bidirectional LSTM) are extremely popular. Popular examples are the combinations of CNN and LSTM, such as DeeperBind [67], DanQ [68], iDeep [69], and DeepGRN [70]. In addition, a few deep learning algorithms, such as transfer learning, attention mechanism-based deep neural network, and capsule neural network, have captured researchers' attention lately. It is noteworthy that CNN also can combine with other deep neural networks such as capsule neural network (CapsNet) and deep transfer learning network. For example, iProDNA-CpasNet [71] conducts a computational model that consists CNN layers and CapsNet layers for identifying protein-DNA binding residues.

Regarding the sequence representation (feature extraction), in traditional machine learning models, protein sequences are usually digitized using extracted features, such as PSSM (i.e., PWM), physicochemical and biochemical properties (AA index), amino acid composition, secondary structure information, and n-mer peptide (dipeptide and tripeptide, i.e., $n = 2$ and 3, are commonly used); RNA sequences are generally digitized as secondary structure and k-mer frequencies. n-mer peptide and k-mer frequencies are actually the same matter, while n-mer peptide (e.g., tripeptide) is frequently employed after grouping amino acids. Whereas, in deep learning models, the overwhelming digital representation of biological sequences is one-hot encoding which is utilized independently or combined with PSSM and secondary structure information. Hybrid feature representation might be deeply explored in further research.

5 | CONCLUSIONS AND OUTLOOK

In the era of AI and big data, the power of big data technology and AI progress hand in hand with the development of computational methods in bioinformatics. With the development of AI technologies and the availability of large amounts of biological data, computational methods for proteomics have undergone a developmental process from traditional machine learning to deep learning. This review focused on these computational approaches in proteomics, especially for prediction tasks regarding protein–DNA/RNA interactions. We summarized machine learning (deep learning) methods used in proteomics, including advantages and shortcomings, and the types of tasks they are suitable for. Moreover, biological sequence-digitizing representation strategies used in different types of computational methods are also summarized, and problems existing in biological data representation methods that use deep learning were pointed out. Besides, we compiled DNA/RNA–protein interaction-related prediction tools and approaches, especially sequence representation (feature

TABLE 1 Applications in tasks related to the protein–DNA/RNA interactions

Approaches	Task type	Tools	Year	Sequence representation (Feature extraction)	Model algorithms	Web server (Code available)
Traditional machine learning	DNA/RNA- protein interaction identification	Boosted trees [72]	2010	LEAC	Boosted DT, ADTree	http://proteomics.bioengr.uic.edu/malibu [a]
		RPISeq [62]	2011	*k*-mers	SVM, RF	http://pridb.gdcb.iastate.edu/RPISeq/
		lncPro [73]	2013	SS, hydrogen-bonding and Vander Waal's interaction of RNA sequence; physicochemical properties of protein sequences	Fisher's LDA	http://cmbi.bjmu.edu.cn/lncpro [a]
		mRMR-IFS-SVM [35]	2013	OCTD, ACC, SAA	SVM, ensemble learning	None
		RPI-Pred [10]	2015	RNA SS and RNA sequence; normalized frequencies of protein sequence	SVM	http://ctsb.is.wfubmc.edu/projects/rpi-pred [a]
		DisoRDPbind [74]	2015	AAC, AA indices, SS, disorder information, sequence complexity,	LR, sequence alignment	http://biomine.ece.ualberta.ca/DisoRDPbind/ [a]
		StackDPPred [11]	2019	PSSM, RCEM, PSSM-DT, RPT, EDT	SVM	http://bmll.cs.uno.edu/add
	Sequence specificities of DNA/RNA-binding proteins	MEMERIS [75]	2006	Integrated RNA SS information	MLE	http://www.bioinf.uni-freiburg.de/~hiller/MEMERIS/ [a]
		RNAcontext [76]	2010	RNA sequence and annotation profiles	Motif-base mathematical	http://www.cs.toronto.edu/~hilal/rnacontext/
		C2H2-ZF [77]	2014	Domain, PWM	SVM	http://zf.princeton.edu
		GraphProt [78]	2014	Graph kernel features	SVM, SVR	http://www.bioinf.unifreiburg.de/Software/GraphProt [a]
Traditional machine learning	Prediction of binding sites	SeqPredNet [79]	2004	ASA, AAC, binding density	NN	www.netasa.org/dbs-pred/ [a]
		BindN [63]	2006	Side chain pK value, HI, molecular mass	SVM	http://bioinformatics.ksu.edu/bindn/
		DP-Bind [64]	2007	PSSM	SVM, KLR, PLR	http://lcg.rit.albany.edu/dp-bind
		RNABindR [80]	2007	Distance-based cutoff	Naïve Bayes	http://bindr.gdcb.iastate.edu/RNABindR
		DISIS [65]	2007	SS, SA, homologs	NN, SVM	http://cubic.bioc.columbia.edu/services/disis [a]
		DBindR [81]	2009	SS, OBV, evolutionary information	RF	http://www.cbi.seu.edu.cn/DBindR/DBindR.htm [a]
		PiRaNhA [36]	2010	PSSM, hydrophobicity, interface propensity, accessibility	SVM	http://www.bioinformatics.sussex.ac.uk/PIRANHA [a]

(Continues)

TABLE 1 (Continued)

Approaches	Task type	Tools	Year	Sequence representation (Feature extraction)	Model algorithms	Web server (Code available)
		mCarts [82]	2013	Distance, accessibility, conservation information	HMM	http://zhanglab.c2b2.columbia.edu/index.php/MCarts
		DRNApred [83]	2017	AA indices, SS, SA, intrinsic disorder, HMM profile	LR	http://biomine.cs.vcu.edu/servers/DRNApred/ [a]
		TriPepSVM [66]	2019	*k*-mers ($k = 3$)	SVM	https://github.com/marsicoLab/TriPepSVM
Deep learning	DNA/RNA- protein interaction identification	Zeng et al. [84]	2016	One-hot vector	CNN	http://cnn.csail.mit.edu
		RPI-SAN [85]	2018	*k*-mers, PSSM	SAE, RF	None
		ELM [86]	2019	RMR, PSSM	CNN, ELM	None
Deep learning	Sequence specificities of DNA/RNA-binding proteins	DeepBind [28]	2015	High-throughput assay data	CNN	http://tools.genes.toronto.edu/deepbind/ [a]
		DeeperBind [67]	2016	PBM data one-hot encoding	CNN and LSTM	https://github.com/hassanzadeh/DeeperBind
		DanQ [68]	2016	One-hot encoding	CNN and LSTM	http://github.com/uci-cbcl/DanQ
		DeFine [87]	2018	One-hot vector	Deep CNN	http://define.cbi.pku.edu.cn [a]
		iDeep [69]	2018	one-hot encoding	CNN, BiLSTM	https://github.com/xypan1232/iDeepS
		NucleicNet [88]	2019	physicochemical properties	DRN	http://www.cbrc.kaust.edu.sa/NucleicNet/
		Differential DL [89]	2020	*k*-mer	CNN, TL, LR	https://doi.org/10.5281/zenodo.2635463
	Prediction of binding sites	Deepnet-RBP [90]	2016	SS, *k*-mers ($k = 6$)	DBN	https://github.com/thucombio/deepnet-rbp
		iProDNA-CpasNet [71]	2019	PSSM	CNN, CapsNet	https://github.com/ngphubinh/iProDNA-CapsNet
		DeepGRN [70]	2021	CNN, BiLSTM	Attention-based CNN and LSTM	https://github.com/jianlin-cheng/DeepGRN

Abbreviations: AA, amino acid; AAC, amino acid composition; ACC, autocross-covariance; ADTree, alternating decision tree; ASA, solvent accessibility or accessible surface area; BiLSTM, bidirectional long short-term memory; CapsNet, capsule neural network; CNN, convolutional neural network; DBN, deep belief network; DL, deep learning; DRN, Deep residual network; DT, decision tree; ELM, extreme learning machine classifier; HI, hydrophobicity index; HMM, hidden Markov model; KLR, kernel logistic regression; LEAC, local environment amino acid composition; LDA, linear discriminant analysis; LR, Logistic regression; MLE, Maximum likelihood estimation; NN, neural network; OBV, orthogonal binary vector; OCTD, overall composition-transition-distribution; PLR, penalized logistic regression; PSSM, position specific scoring matrix; PSSM-DT, PSSM-distance transformation; RCEM, reside wise contact energy matrix; RMR, RNA matrix representation; RPT, residue probing transformation; PSW, POSITION-SPECIFIC WEIGHT; RF, random forest; SA, solvent accessibility; SAA, split amino acid; SAE, stacked autoencoder; SS, secondary structure; SVM, support vector machine; SVR, support vector regression; TL, transfer learning.

[a] Link not working at the time of writing.

extraction) and model algorithms used in these applications. In the era of AI and big data, further research might explore the combination of different approaches for AI, such as combining deep learning with reinforcement learning, deep learning with evolutionary methods, as well as hybrid feature representation strategies as we suggested previously herein.

ACKNOWLEDGEMENT

The work was supported by the National Natural Science Foundation of China (No. 62131004, N0.62101100), the Sichuan Provincial Science Fund for Distinguished Young Scholars (2021JDJQ0025), and the Special Science Foundation of Quzhou (2021D004).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID


Quan Zou https://orcid.org/0000-0001-6406-1142

REFERENCES

1. Sagendorf, J. M., Markarian, N., Berman, H. M., & Rohs, R. (2020). DNAproDB: An expanded database and web-based tool for structural analysis of DNA-protein complexes. *Nucleic Acids Research*, *48*(D1), D277-D287. https://doi.org/10.1093/nar/gkz889.
2. Yan, Y., Zhang, D., Zhou, P., Li, B., & Huang, S. -. Y. (2017). HDOCK: A web server for protein-protein and protein-DNA/RNA docking based on a hybrid strategy. *Nucleic Acids Research*, *45*(W1), W365-W373. https://doi.org/10.1093/nar/gkx407.
3. Stingele, J., Bellelli, R., & Boulton, S. J. (2017). Mechanisms of DNA-protein crosslink repair. *Nature Reviews Molecular Cell Biology*, *18*(9), 563–573. https://doi.org/10.1038/nrm.2017.56.
4. Klattenhoff, C. A., Scheuermann, J. C., Surface, L. E., Bradley, R. K., Fields, P. A., Steinhauser, M. L., Ding, H., Butty, V. L., Torrey, L., Haas, S., Abo, R., Tabebordbar, M., Lee, R. T., Burge, C. B., … Boyer, L. A. (2013). Braveheart, a long noncoding RNA required for cardiovascular lineage commitment. *Cell*, *152*(3), 570–583. https://doi.org/10.1016/j.cell.2013.01.003.
5. Schmidt, N., Lareau, C. A., Keshishian, H., Ganskih, S., Schneider, C., Hennig, T., Melanson, R., Werner, S., Wei, Y., Zimmer, M., Ade, J., Kirschner, L., Zielinski, S., Dölken, L., Lander, E. S., Caliskan, N., Fischer, U., Vogel, J., Carr, S. A., Bodem, J., & Munschauer, M. (2021). The SARS-CoV-2 RNA-protein interactome in infected human cells. *Nature Microbiology*, *6*(3), 339–353. https://doi.org/10.1038/s41564-020-00846-z.
6. Zhang, Z., Cui, F., Cao, C., Wang, Q., & Zou, Q. (2021). Single-cell RNA analysis reveals the potential risk of organ-specific cell types vulnerable to SARS-CoV-2 infections. *Computers in Biology and Medicine*, *140*, 105092. https://doi.org/10.1016/j.compbiomed.2021.105092.
7. Jones, S., Van Heyningen, P., Berman, H. M., & Thornton, J. M. (1999). Protein-DNA interactions: A structural analysis. *Journal of Molecular Biology*, *287*(5), 877–896. https://doi.org/10.1006/jmbi.1999.2659.
8. Jones, S. (2003). Using structural motif templates to identify proteins with DNA binding function. *Nucleic Acids Research*, *31*(11), 2811–2823.
9. Kono, H., & Sarai, A. (1999). Structure-based prediction of DNA target sites by regulatory proteins. *Proteins: Structure, Function, and Bioinformatics*, *35*(1), 114–131.
10. Suresh, V., Liu, L., Adjeroh, D., & Zhou, X. (2015). RPI-Pred: Predicting ncRNA-protein interaction using sequence and structural information. *Nucleic Acids Research*, *43*(3), 1370–1379. https://doi.org/10.1093/nar/gkv020.
11. Mishra, A., Pokhrel, P., & Hoque, M. T. (2019). StackDPPred: A stacking based prediction of DNA-binding protein from sequence. *Bioinformatics*, *35*(3), 433–441.
12. Cao, C., Kwok, D., Edie, S., Li, Q., Ding, B., Kossinna, P., Campbell, S., Wu, J., Greenberg, M., & Long, Q. (2021). kTWAS: Integrating kernel machine with transcriptome-wide association studies improves statistical power and reveals novel genes. *Briefings in Bioinformatics*, *22*(4), bbaa270.
13. Dou, L., Yang, F., Xu, L., & Zou, Q. (2021). A comprehensive review of the imbalance classification of protein post-translational modifications. *Briefings in Bioinformatics*, *22*(5), bbab089. https://doi.org/10.1093/bib/bbab089.
14. Franceschini, A., Lin, J., Von Mering, C., & Jensen, L. J. (2016). SVD-phy: Improved prediction of protein functional associations through singular value decomposition of phylogenetic profiles. *Bioinformatics*, *32*(7), 1085–1087. https://doi.org/10.1093/bioinformatics/btv696.
15. Taherzadeh, G., Zhou, Y., Liew, A. W.-C., & Yang, Y. (2018). Structure-based prediction of protein–peptide binding regions using Random Forest. *Bioinformatics*, *34*(3), 477–484.
16. Li, N., Ainsworth, R. I., Wu, M., Ding, B., & Wang, W. (2016). MIEC-SVM: Automated pipeline for protein peptide/ligand interaction prediction. *Bioinformatics*, *32*(6), 940–942.
17. Mcginnis, S., & Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research*, *32*(suppl_2), W20-W25.
18. Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bansal, P., Bridge, A. J., Poux, S., Bougueleret, L., & Xenarios, I. (2016). *UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: How to use the entry view Plant Bioinformatics* (pp. 23–54). Springer.
19. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556.
20. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1), D412-D419. https://doi.org/10.1093/nar/gkaa913.
21. Sigrist, C. J. A., De Castro, E., Cerutti, L., Cuche, B. A., Hulo, N., Bridge, A., Bougueleret, L., & Xenarios, I. (2013). New and continuing developments at PROSITE. *Nucleic Acids Research*, *41*(Database issue), D344-D347. https://doi.org/10.1093/nar/gks1067.
22. Ao, C., Zhou, W., Gao, L., Dong, B., & Yu, L. (2020). Prediction of antioxidant proteins using hybrid feature representation method and random forest. *Genomics*, *112*(6), 4666–4674.
23. Jiao, S., Zou, Q., Guo, H., & Shi, L. (2021). iTTCA-RF: A random forest predictor for tumor T cell antigens. *Journal of Translational Medicine*, *19*(1), 449. https://doi.org/10.1186/s12967-021-03084-x.
24. Dimiduk, D. M., Holm, E. A., & Niezgoda, S. R. (2018). Perspectives on the impact of machine learning, deep learning, and artificial intelligence on materials, processes, and structures engineering. *Integrating Materials and Manufacturing Innovation*, *7*(3), 157–172.
25. Ravi, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., & Yang, G.-Z. (2016). Deep learning for health informatics. *IEEE Journal of Biomedical and Health Informatics*, *21*(1), 4–21.

26. Zhang, Z., Cui, F., Lin, C., Zhao, L., Wang, C., & Zou, Q. (2021). Critical downstream analysis steps for single-cell RNA sequencing data. *Briefings in Bioinformatics*, *22*(5), bbab105. https://doi.org/10.1093/bib/bbab105.
27. Zhang, Z., Cui, F., Wang, C., Zhao, L., & Zou, Q. (2021). Goals and approaches for each processing step for single-cell RNA sequencing data. *Briefings in Bioinformatics*, *22*(4), bbaa314.
28. Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, *33*(8), 831–838. https://doi.org/10.1038/nbt.3300.
29. Lv, Z., Cui, F., Zou, Q., Zhang, L., & Xu, L. (2021). Anticancer peptides prediction with deep representation learning features. *Briefings in Bioinformatics*, *22*(5), bbab008. https://doi.org/10.1093/bib/bbab008.
30. Almagro Armenteros, J. J., Sønderby, C. K., Sønderby, S. K., Nielsen, H., & Winther, O. (2017). DeepLoc: Prediction of protein subcellular localization using deep learning. *Bioinformatics*, *33*(21), 3387–3395. https://doi.org/10.1093/bioinformatics/btx431.
31. Hashemifar, S., Neyshabur, B., Khan, A. A., & Xu, J. (2018). Predicting protein–protein interactions through sequence-based deep learning. *Bioinformatics*, *34*(17), i802-i810.
32. Kulmanov, M., Khan, M. A., & Hoehndorf, R. (2018). DeepGO: Predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics*, *34*(4), 660–668. https://doi.org/10.1093/bioinformatics/btx624.
33. Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.
34. Hwang, S., Gou, Z., & Kuznetsov, I. B. (2007). DP-Bind: A web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. *Bioinformatics*, *23*(5), 634–636. https://doi.org/10.1093/bioinformatics/btl672.
35. Zou, C., Gong, J., & Li, H. (2013). An improved sequence based prediction protocol for DNA-binding proteins using SVM and comprehensive feature analysis. *Bmc Bioinformatics [Electronic Resource]*, *14*, 90-90. https://doi.org/10.1186/1471–2105–14-90.
36. Murakami, Y., Spriggs, R. V., Nakamura, H., & Jones, S. (2010). PiRaNhA: A server for the computational prediction of RNA-binding residues in protein sequences. *Nucleic Acids Research*, *38*(suppl_2), W412-W416.
37. Heinzinger, M., Elnaggar, A., Wang, Y., Dallago, C., Nechaev, D., Matthes, F., & Rost, B. (2019). Modeling aspects of the language of life through transfer-learning protein sequences. *Bmc Bioinformatics [Electronic Resource]*, *20*(1), 1–17.
38. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., & Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(15).
39. Alley, E. C., Khimulya, G., Biswas, S., Alquraishi, M., & Church, G. M. (2019). Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods*, *16*(12), 1315–1322.
40. Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P., & Song, Y. S. (2019). Evaluating protein transfer learning with TAPE. *Advances in Neural Information Processing Systems*, *32*, 9689–9701.
41. Granter, S. R., Beck, A. H., & Papke, D. J., Jr. (2017). AlphaGo, deep learning, and the future of the human microscopist. *Archives of Pathology & Laboratory Medicine*, *141*(5), 619–621. https://doi.org/10.5858/arpa.2016-0471-ED.
42. Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A. A., … Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. *Nature*, *596*(7873), 590–596.
43. Gupta, S., Agrawal, A., Gopalakrishnan, K., & Narayanan, P. (2015). Deep learning with limited numerical precision. Paper presented at the International conference on machine learning.
44. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.
45. Ketkar, N. (2017). *Introduction to keras Deep learning with Python* (pp. 97–111). Springer.
46. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., … Antiga, L. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, *32*, 8026–8037.
47. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y. & Zheng, X. … (2016). Tensorflow: A system for large-scale machine learning. Paper presented at the 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16).
48. Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I., Bergeron, A., Bouchard, N., Warde-Farley, D. & Bengio, Y. … (2012). Theano: New features and speed improvements. arXiv preprint arXiv:1211.5590.
49. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10–18.
50. Ao, C., Yu, L., & Zou, Q. (2021). Prediction of bio-sequence modifications and the associations with diseases. *Briefings in Functional Genomics*, *20*(1), 1–18. https://doi.org/10.1093/bfgp/elaa023.
51. Cui, F., Zou, Q., Ma, Q., Wei, L., Tang, J., & Mrozek, D. (2021). IEEE access special section editorial: Feature representation and learning methods with applications in large-scale biological sequence analysis. *IEEE Access*, *9*, 33110–33119.
52. Ballester, P. J., & Mitchell, J. B. O. (2010). A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics*, *26*(9), 1169–1175. https://doi.org/10.1093/bioinformatics/btq112.
53. Kong, L., Zhang, Y., Ye, Z.-Q., Liu, X.-Q., Zhao, S.-Q., Wei, L., & Gao, G. (2007). CPC: Assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Research*, *35*(suppl_2), W345-W349.
54. Cui, F., Zhang, Z., & Zou, Q. (2021). Sequence representation approaches for sequence-based protein prediction tasks that use deep learning. *Briefings in Functional Genomics*.61–73,.
55. Tang, Y.-J., Pang, Y.-H., & Liu, B. (2020). IDP-Seq2Seq: Identification of intrinsically disordered regions based on sequence to sequence learning. *Bioinformatics*, *36*(21), 5177–5186.
56. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
57. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240.
58. Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, *577*(7792), 706–710.
59. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate

protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589.
60. Hiranuma, N., Park, H., Baek, M., Anishchenko, I., Dauparas, J., & Baker, D. (2021). Improved protein structure refinement guided by deep learning based accuracy estimation. *Nature Communication*, *12*(1), 1–11.
61. Xu, J., Mcpartlon, M., & Li, J. (2021). Improved protein structure prediction by deep learning irrespective of co-evolution information. *Nature Machine Intelligence*, *3*, 601–609.
62. Muppirala, U. K., Honavar, V. G., & Dobbs, D. (2011). Predicting RNA-protein interactions using only sequence information. *Bmc Bioinformatics [Electronic Resource]*, *12*, 489–489. https://doi.org/10.1186/1471–2105–12–489.
63. Wang, L., & Brown, S. J. (2006). BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Research*, *34*(suppl_2), W243-W248.
64. Hwang, S., Gou, Z., & Kuznetsov, I. B. (2007). DP-Bind: A web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. *Bioinformatics*, *23*(5), 634–636.
65. Ofran, Y., Mysore, V., & Rost, B. (2007). Prediction of DNA-binding residues from sequence. *Bioinformatics*, *23*(13), i347-i353.
66. Bressin, A., Schulte-Sasse, R., Figini, D., Urdaneta, E. C., Beckmann, B. M., & Marsico, A. (2019). TriPepSVM: de novo prediction of RNA-binding proteins based on short amino acid motifs. *Nucleic Acids Research*, *47*(9), 4406–4417.
67. Hassanzadeh, H. R., & Wang, M. D. (2016, December 15–18). DeeperBind: Enhancing prediction of sequence specificities of DNA binding proteins. Paper presented at the 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM).
68. Quang, D., & Xie, X. (2016). DanQ: A hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Research*, *44*(11), e107-e107. https://doi.org/10.1093/nar/gkw226.
69. Pan, X., Rijnbeek, P., Yan, J., & Shen, H.-B. (2018). Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. *BMC Genomics [Electronic Resource]*, *19*(1), 511. https://doi.org/10.1186/s12864-018–4889-1.
70. Chen, C., Hou, J., Shi, X., Yang, H., Birchler, J. A., & Cheng, J. (2021). DeepGRN: Prediction of transcription factor binding site across cell-types using attention-based deep neural networks. *BMC Bioinformatics [Electronic Resource]*, *22*(1), 38. https://doi.org/10.1186/s12859-020-03952-1.
71. Nguyen, B. P., Nguyen, Q. H., Doan-Ngoc, G.-N., Nguyen-Vo, T.-H., & Rahardja, S. (2019). iProDNA-CapsNet: Identifying protein-DNA binding residues using capsule neural networks. *BMC Bioinformatics [Electronic Resource]*, *20*(23), 1–12.
72. Langlois, R. E., & Lu, H. (2010). Boosting the prediction and understanding of DNA-binding domains from sequence. *Nucleic Acids Research*, *38*(10), 3149–3158. https://doi.org/10.1093/nar/gkq061.
73. Lu, Q., Ren, S., Lu, M., Zhang, Y., Zhu, D., Zhang, X., & Li, T. (2013). Computational prediction of associations between long non-coding RNAs and proteins. *BMC Genomics [Electronic Resource]*, *14*(1), 651. https://doi.org/10.1186/1471–2164-14–651.
74. Peng, Z., & Kurgan, L. (2015). High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. *Nucleic Acids Research*, *43*(18), e121-e121. https://doi.org/10.1093/nar/gkv585.
75. Hiller, M., Pudimat, R., Busch, A., & Backofen, R. (2006). Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Research*, *34*(17), e117-e117. https://doi.org/10.1093/nar/gkl544.
76. Kazan, H., Ray, D., Chan, E. T., Hughes, T. R., & Morris, Q. (2010). RNAcontext: A new method for learning the sequence and structure binding preferences of RNA-binding proteins. *Plos Computational Biology*, *6*(7), e1000832-e1000832. https://doi.org/10.1371/journal.pcbi.1000832.
77. Persikov, A. V., & Singh, M. (2014). De novo prediction of DNA-binding specificities for Cys2His2 zinc finger proteins. *Nucleic Acids Research*, *42*(1), 97–108. https://doi.org/10.1093/nar/gkt890.
78. Maticzka, D., Lange, S. J., Costa, F., & Backofen, R. (2014). GraphProt: Modeling binding preferences of RNA-binding proteins. *Genome Biology*, *15*(1), R17-R17. https://doi.org/10.1186/gb-2014–15-1-r17.
79. Ahmad, S., Gromiha, M. M., & Sarai, A. (2004). Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, *20*(4), 477–486.
80. Terribilini, M., Sander, J. D., Lee, J.-H., Zaback, P., Jernigan, R. L., Honavar, V., & Dobbs, D. (2007). RNABindR: A server for analyzing and predicting RNA-binding sites in proteins. *Nucleic Acids Research*, *35*(suppl_2), W578-W584.
81. Wu, J., Liu, H., Duan, X., Ding, Y., Wu, H., Bai, Y., & Sun, X. (2009). Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. *Bioinformatics*, *25*(1), 30–35.
82. Zhang, C., Lee, K.-Y., Swanson, M. S., & Darnell, R. B. (2013). Prediction of clustered RNA-binding protein motif sites in the mammalian genome. *Nucleic Acids Research*, *41*(14), 6793–6807.
83. Yan, J., & Kurgan, L. (2017). DRNApred, fast sequence-based method that accurately predicts and discriminates DNA-and RNA-binding residues. *Nucleic Acids Research*, *45*(10), e84-e84.
84. Zeng, H., Edwards, M. D., Liu, G., & Gifford, D. K. (2016). Convolutional neural network architectures for predicting DNA–protein binding. *Bioinformatics*, *32*(12), i121-i127. https://doi.org/10.1093/bioinformatics/btw255.
85. Yi, H.-C., You, Z.-H., Huang, D.-S., Li, X., Jiang, T.-H., & Li, L.-P. (2018). A deep learning framework for robust and accurate prediction of ncRNA-protein interactions using evolutionary information. *Molecular Therapy. Nucleic Acids*, *11*, 337–344. https://doi.org/10.1016/j.omtn.2018.03.001.
86. Wang, L., You, Z.-H., Huang, D.-S., & Zhou, F. (2018). Combining high speed ELM learning with a deep convolutional neural network feature encoding for predicting protein-RNA interactions. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *17*(3), 972–980.
87. Wang, M., Tai, C., Weinan, E., & Wei, L. (2018). DeFine: Deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional non-coding variants. *Nucleic Acids Research*, *46*(11), e69-e69.
88. Lam, J. H., Li, Y., Zhu, L., Umarov, R., Jiang, H., Héliou, A., Sheong, F. K., Liu, T., Long, Y., Li, Y., Fang, L., Altman, R. B., Chen, W., Huang, X., & Gao, X. (2019). A deep learning framework to predict binding preference of RNA constituents on protein surface. *Nature Communication*, *10*(1), 4941. https://doi.org/10.1038/s41467-019-12920-0.
89. Phuycharoen, M., Zarrineh, P., Bridoux, L., Amin, S., Losa, M., Chen, K., Bobola, N., & Rattray, M. (2020). Uncovering tissue-specific binding features from differential deep learning. *Nucleic Acids Research*, *48*(5), e27-e27.
90. Zhang, S., Zhou, J., Hu, H., Gong, H., Chen, L., Cheng, C., & Zeng, J. (2016). A deep learning framework for modeling structural features of RNA-binding protein targets. *Nucleic Acids Research*, *44*(4), e32-e32.

How to cite this article: Cui, F., Zhang, Z., Cao, C., Zou, Q., Chen, D., & Su, X. (2022). Protein–DNA/RNA interactions: Machine intelligence tools and approaches in the era of artificial intelligence and big data. *Proteomics*, *22*, e2100197. https://doi.org/10.1002/pmic.202100197