Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 43, S. 80-85, Kasım 2022

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 43, pp. 80-85, November 2022

Research Article

Classification of Transcription Factor DNA in the Brassica Plant Species by Deep Learning

Ali Burak Öncül[1*]

[1*] Kastamonu University, Faculty of Engineering and Architecture, Departmant of Computer Engineering, Kastamonu, Turkey, (ORCID: 0000-0001-9612-1787), boncul@kastamonu.edu.tr

(4th International Conference on Applied Engineering and Natural Sciences ICAENS 2022, November 10 - 13, 2022)

(**DOI:** 10.31590/ejosat.1200680)

ATIF/REFERENCE: Öncül, A. B. (2022). Classification of Transcription Factor DNA in the Brassica Plant Species by Deep Learning. *European Journal of Science and Technology,* (43), 80-85.

Abstract

Determining the types of DNA and proteins, examining their similarities, etc., remains among the challenging problems in the research field. For this reason, the data obtained and the use of this data are also limited. In this study, we combined the power of computer science in data processing with biology. We classified the DNAs of transcription factor proteins found in cruciferous Brassica plants and identified the DNAs related to the synthesis of transcription factor proteins in the plant. We compiled the dataset from the Plant Transcription Factor Database (PlantTFDB). We used the code dictionary structure in the preprocessing part and provided a fast and successful model using Bidirectional LSTM and Bidirectional GRU networks. Our model has 90.40% test accuracy and 86.75% 5-fold cross-validation accuracy. Using LSTM in the layer with fewer units and GRU in the layer with more units in the model provided a shorter training time for the model. In addition, although the prepared model classifies the transcription factor DNAs of Brassica plants, it will also be successful at a certain level in the transcription factor DNAs of other plants. The prepared model stands out as an important innovation that has been added to the literature in terms of its field of study.

Keywords: Bioinformatics, DNA classification, Deep learning, Bidirectional, LSTM, GRU.

Brassica Bitki Türlerinde Transkripsiyon Faktörü DNA'sının Derin Öğrenme ile Sınıflandırılması

Öz

DNA ve protein türlerinin belirlenmesi, benzerliklerinin incelenmesi vb. araştırma alanındaki zorlu problemler arasında yer almaktadır. Bu nedenle elde edilen veriler ve bu verilerin kullanımı da sınırlıdır. Bu çalışmada bilgisayar biliminin veri işlemedeki gücünü biyoloji ile birleştirdik. Turpgillerden Brassica bitkilerinde bulunan transkripsiyon faktörü proteinlerinin DNA'larını sınıflandırdık ve bitkideki transkripsiyon faktörü proteinlerinin sentezi ile ilgili DNA'ları belirledik. Veri setini Bitki Transkripsiyon Faktörü Veritabanından (PlantTFDB) derledik. Önişleme kısmında kod sözlüğü yapısını kullandık ve Çift Yönlü LSTM ve Çift Yönlü GRU ağlarını kullanarak hızlı ve başarılı bir model sağladık. Modelimiz %90,40 test doğruluğuna ve %86,75 5-kat çapraz doğrulama doğruluğuna sahiptir. Modelde daha az birimli katmanda LSTM ve daha fazla birimli katmanda GRU kullanılması model için daha kısa eğitim süresi sağlamıştır. Ayrıca hazırlanan model Brassica bitkilerinin transkripsiyon faktör DNA'larını sınıflandırsa da diğer bitkilerin transkripsiyon faktör DNA'larında da belli bir düzeyde başarılı olacaktır. Hazırlanan model, çalışma alanı açısından literatüre katılmış önemli bir yenilik olarak öne çıkmaktadır.

Anahtar Kelimeler: Biyoinformatik, DNA sınıflandırma, Derin öğrenme, Çift yönlü, LSTM, GRU.

[*] Corresponding Author: boncul@kastamonu.edu.tr

1. Introduction

Deoxyribo Nucleic Acid (DNA) is the hereditary genetic information found in almost all living things. "The information in DNA is stored as a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T)." DNA is formed by the different sequencing of the A, G, C and T codons in every living thing. The DNA of almost every cell in the body of every living thing is the same. The ordering of these codons directly influences the creation and diversity of each organism, that is, each part of the organism. The order of these mentioned bases (A, G, C, T) ensures the diversity of the organism and each part of the organism and the survival of the organism. This structure is similar to the formation of different words with a different arrangement of letters and the formation of different sentences and texts with the arrangement of these different words (WATSON & CRICK, 1953).

The rules used to transform the information encoded in genetic material, namely DNA and mRNA, into proteins are called genetic code. The genetic code defines nucleotide triplet sequences called codons that determine which amino acid will be added during protein synthesis (Shu, 2017).

The gene is an essential physical and functional part of heredity. Genes formed from DNA through the genetic code, amino acids, and therefore the sequencing of amino acids serve as instructions for making molecules called proteins. In humans, genes range from a few hundred DNA bases to more than 2 million bases (WATSON & CRICK, 1953). The 3D structure of a DNA fragment is given in Figure 1. The helical structure in this 3D representation consists of the four chemical bases, A, G, C, and T, and are arranged reciprocally. This ordered chain structure is the structure that makes up DNA.

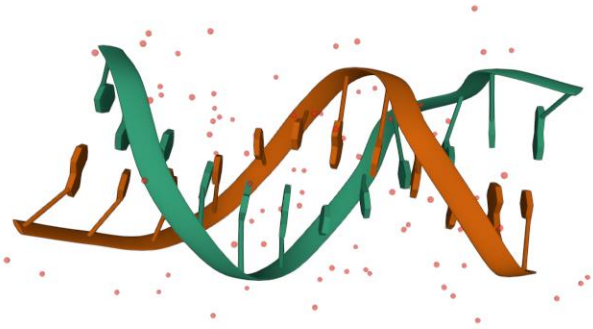

Fig. 1 An exemplary DNA 3D structure (Narayana et al., 1991)

The process, which is the first step of gene expression from the gene's DNA to the production of the primary RNA transcript, is called transcription. Transcription and subsequent process steps drive protein production. Here, the necessary information for protein production is provided from DNA. These transcription factors bind to specific DNA sequences in gene regulatory regions and control the transcription of the DNA sequences to which they bind. The function of transcription factors (TF) is to regulate the gene expression necessary for the survival of the cell and the organism (Karin, 1990; Latchman, 1993). Transcription factors manage many vital processes, such as development, growth, intercellular communication, and environmental response, together with DNA connection (Riaño-Pachón et al., 2007).

Since each different transcription factor protein will produce from a different DNA sequence, classifying the DNAs that produce these proteins will play an important role in the preliminary research of the proteins produced. The data processing power of computer science is frequently used to support biological studies on DNA and proteins. The first joint computer science and biology studies were statistics-based (Baldi & Brunak, 2001). The most important of these studies can be shown as Hidden Markov Model (HMM) (Eddy, 1996) based studies (Gromiha, 2010) and Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990). Some of these studies, in which predictions are made by subtracting the sequences of nucleotides or amino acids and their probabilities of being in certain positions, with the support provided by statistical science, require additional information such as various labels (Price et al., 2018; Strodthoff et al., 2020).

After the statistics-based studies, various artificial neural networks, machine learning applications, and computer science innovations have successfully found their place among the studies in the field. Examples of these studies are Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Naive Bayes Classifier-based studies (Huerta et al., 2000). In later studies, with the development of the concept of deep learning, models prepared with Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) based Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) come to the fore (Du et al., 2016).

Looking at the developments in the field, we designed a deep learning model based on Bidirectional LSTM and Bidirectional GRU to classify the DNA sequences of TFs of Brassica plants. Thanks to the model we designed, we classified the TF DNAs of Brassica plants without needing additional information except for DNA sequences, with a short training period. Thus, we created a piece of preliminary information about the TF proteins to be synthesized for the studies in the field of biology.

The remainder of the article is organized as follows: Chapter 2 includes a literature review, chapter 3 includes the methods used, chapter 4 the results of the experimental study, chapter 5 the discussion and conclusion, and finally, chapter 6 references.

2. Related Studies

Classification and analysis of previously discovered or newly discovered DNA sequences is still a challenging problem. When the literature is examined, it has been seen that various studies have been carried out for the classification of DNA sequences. Biological studies, experiments and analyzes constitute the basis of these studies. These biologically-based studies are relatively time-consuming, costly, and more prone to human error. After biological-based studies, it is statistical-based studies or artificial intelligence studies from the power combination of biology and computer science [12].

In one study, the prediction of DNA N6-methyladenosine regions among plant species was studied with a model prepared based on CNN and bidirectional LSTM (Tang et al., 2022). In another bidirectional GRU-based study, DNA N4-methylcytosine regions in the mouse genome were estimated by Jin et al. (Jin et al., 2022). In a Word2Vec-based deep learning study, it was aimed to identify DNA N4-methylcytosine regions(Fang et al., 2021). Support vector machines (SVM) were used in a study on DNA and amino acid approaches for human authentication with deep transfer learning (Sakr et al., 2022). In a study on predicting 3D chromatin interactions from DNA sequence using Deep Learning,

CNN, LSTM and GRU networks were studied and the transfer learning method was applied (Piecyk et al., 2022). In a study on the classification of viruses such as COVID, SARS, MERS, dengue, hepatitis and influenza, a hybrid model based on CNN and LSTM was prepared (Gunasekaran et al., 2021).

When the literature is examined, although different studies have been carried out in different DNA studies, no deep learning study has been observed in TF DNAs, especially in TF DNAs of plants belonging to the Brassica species. The code dictionary preprocessed model we have prepared classifies the DNA sequences that make up the TF proteins in Brassica plants and provides a solution to the problem of detecting and classifying the source DNAs of the proteins to be formed.

3. Material and Method

TF proteins play a very important role in the regulation of vital functions in the life of Brassica plants. The expression of these TF proteins also depends on the information in the DNA of these TFs. For these reasons, knowing the biological and bioinformatic structure and classes of Brassica plants enables them to dominate many features of the plant. Looking at the literature, no deep learning studies were found about the TF DNA of Brassica plants. This situation reveals the importance and necessity of the classification of these DNAs. To create a data set to be used in this study, the data were downloaded and processed in scattered form from PlantTFDB, one of the large databases related to TFs.

PlantTFDB is a comprehensive and public database designed by a team of researchers to provide communication with the plant genome, TFs in gene families, and additional information about these TFs. The PlantTFDB website contains individual DNA and protein sequences and individual TF listings for each family and transcription factor (Jin et al., 2017).

3.1. Structure of the DNA Sequences

DNA sequences are genetic material that consists of a cascade of nucleotides. There are different symbols in the literature denoting adenine (A), guanine (G), cytosine (C) and thymine (T) nucleotides and other states other than these defined nucleotides ("Nomenclature for Incompletely Specified Bases in Nucleic Acid Sequences. Recommendations 1984. Nomenclature Committee of the International Union of Biochemistry (NC-IUB).," 1986). Symbols other than these 4 basic nucleotides are very few and have been ignored in the prepared model. Just as letters form words and words form sentences, so the ordering of the symbols of nucleotides creates DNA sequences. The resulting DNAs are also divided into classes according to their functions.

3.2. Data Preprocessing

The data downloaded from PlantTFDB, together with the sequences and their families, are collected in a single file and the data set is created. DNA sequences with a structure such as "AATGCAATTT...", expressed in characters, must be digitized to be given to the deep learning model. Numbers from 1 to 4 are assigned to 4 nucleotides in 14384 DNA sequences. For other negligible cases, the number 0 is assigned. Figure 2 shows the frequency with which nucleotides are found in sequences.

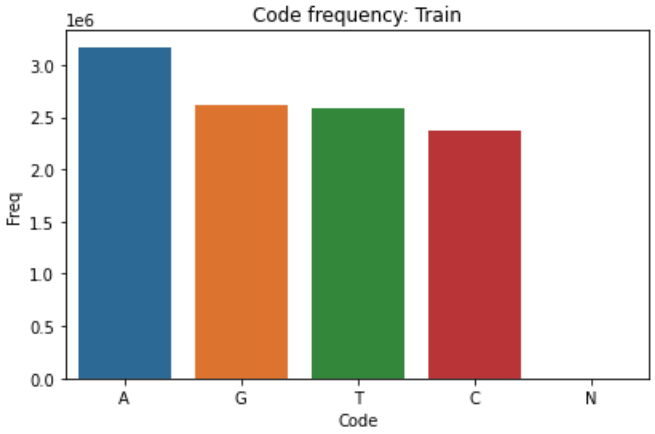

Fig. 2 Nucleotide frequency

In this way, the sequences are digitized. The length of the digitized sequences should be equalized for the deep learning model to have a healthy training process. For this reason, it was decided to calculate the average length of the sequences and determine the sequence size as 1070. The graph showing the average lengths of the sequences is presented in Figure 3.

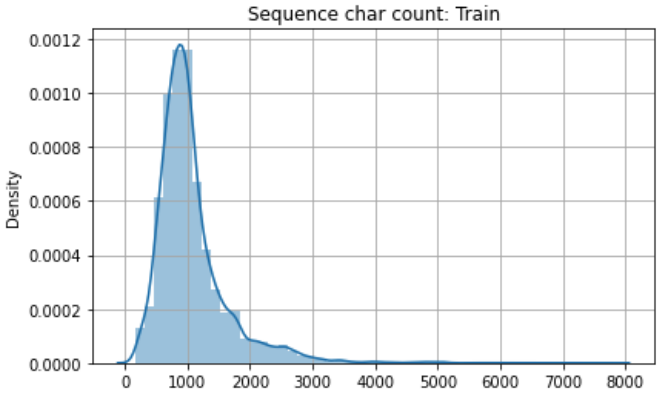

Fig. 3 Average lengths of the sequences

Among the sequences digitized in this way, the longer ones were shortened to 1070, and the shorter ones were extended by filling in 0's at the end (Bileschi et al., 2022). Then, the classes of the sequences, that is, the 58 families to which the DNAs belong, were coded with one-hot encoding, just as stated in the study of Yang et al., and all the data were prepared (Yang et al., 2018). The data prepared and separated as 80% train, 10% validation and 10% test are ready to be given to the deep learning model.

3.2. Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU)

LSTM and GRU networks use sequential information in the data and make predictions by storing historical states and long-short-term dependencies, which are structurally similar and provide almost the same results. In LSTM, there are input, output and forget gates (Greff et al., 2017). On the other hand, in GRU, forget function is done with a key, not a gate (Gao & Glowacka, 2016). GRU has a relatively faster training and working time compared to LSTM since there is a missing gate compared to LSTM. These models can process long texts and data and successfully detect long-term dependencies and which information will be forgotten, thanks to the gates they contain and the calculations they make (Şeker et al., 2017).

In evaluating the prepared model, accuracy, precision, recall, f-score criteria (Luque et al., 2019) and a 5-fold cross-validation

method were used (Xiong et al., 2020). In addition, the train and validation accuracy and loss graphs of the model and the ROC curve were also used (KILIC, 2013).

The embedding layer of Keras was used in the prepared model, and the vector size was determined as 256. A 128-unit layer Bidirectional LSTM and a 256-unit layer Bidirectional GRU are used in the model. Since GRU is relatively fast compared to LSTM, more units are preferred in GRU than LSTM. Table 1 shows the design details of the model.

Table 1. Example of a table

Layer	Output Shape	Param #
Embedding	(*None*, 1070, 256)	1280
Bidirectional LSTM	(*None*, 1070, 256)	394240
Bidirectional GRU	(*None*, 512)	789504
Dropout (0.25)	(*None*, 512)	0
Flatten	(*None*, 512)	0
Dense	(*None*, 256)	131328
Dropout (0.25)	(*None*, 256)	0
Dense (Classification)	(*None*, 58)	14906

3. Results and Discussion

The data set used consists of 14384 sequences. In order to allocate more data for training, the data set is divided into three parts, 80% train, 10% validation and 10% test, with the Python Scikit-Learn library. A value of 0.001 was used in all layers as the learning rate in the model. The batch size value is set to 256 because long arrays are used. ADAM is used as the optimization function. Since a 58-class data set was used, the categorical cross-entropy function was chosen as the loss function. The model completed its excellent training in 28 epochs. The evaluation results of the model are as follows:

- Accuracy: 90.40%
- Precision: 90.37%
- Recall: 89.34%
- F-score: 88.77%
- 5-fold cross-validation: 86.75%
- Train time: 16.85 min.

Figure 4 shows the accuracy and loss graphs of the model, and Figure 5 shows the ROC curve.

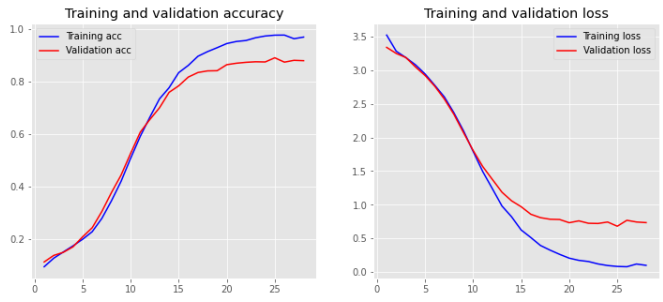

Fig. 4 Accuracy and loss graphs for deep learning model

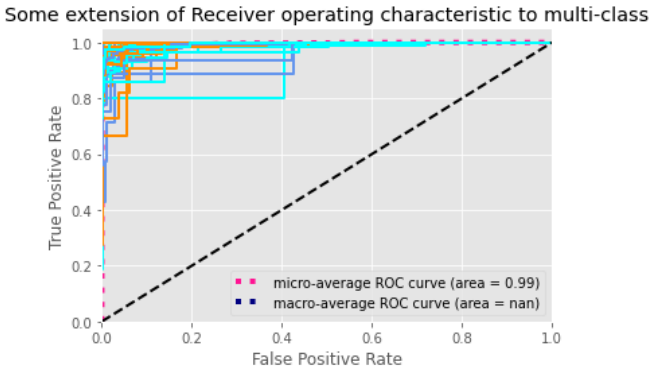

Fig. 5 ROC curve for deep learning model

The results in Table 2 and the graphs in Figure 4 and 5 showed that the model successfully classified TF DNAs in plants belonging to the Brassica species and made an important contribution to the literature. The proximity of all curves to the upper left corner in Figure 5 shows that the model has successfully classified all 58 classes.

When we look at the literature, although there are studies such as the N6-methyladenosine region determination study of Tang et al. (Tang et al., 2022), the study of Sakr et al. including DNA and amino acid approaches (Sakr et al., 2022), and the detection of DNA of viruses such as COVID and SARS, it is possible to determine the DNA of transcription factors. No study could be seen on the transcription factor DNAs of plants of the genus and Brassica species. In this study, which we have done on this gap in the literature, we have succeeded in classifying the DNA of transcription factors in Brassica species. The model I prepared includes 1 layer of Bidirectional LSTM and 1 layer of Bidirectional GRU. Thanks to these layers, the structures and motifs of long DNA sequences can be captured, and various long- and short-term dependencies can be detected, resulting in a successful classification. In this way, we were able to have preliminary research information and classification information of the proteins to be synthesized from these DNAs. In addition, with this model, we filled a gap in the literature.

4. Conclusions and Recommendations

By preparing this study, we classified the DNAs that synthesize the transcription factors that greatly affect the life cycles and functions of plants belonging to the Brassica plant species according to the transcription factor type. Thanks to this classification and model, the time to be spent with biological experiments, human-induced errors, and high costs have been tried to be avoided. In addition, thanks to this speed and success, more work can be done. This prepared model stands out as an important innovation that has been added to the literature regarding Brassica plants.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Baldi, P., & Brunak, S. (2001). *Bioinformatics, Second Edition: The Machine Learning Approach*. MIT Press.

Bileschi, M. L., Belanger, D., Bryant, D. H., Sanderson, T., Carter, B., Sculley, D., Bateman, A., DePristo, M. A., & Colwell, L. J. (2022). Using deep learning to annotate the protein universe. *Nature Biotechnology*, *40*(6), 932–937. https://doi.org/10.1038/s41587-021-01179-w

Du, X., Cai, Y., Wang, S., & Zhang, L. (2016). Overview of deep learning. *2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, 159–164. https://doi.org/10.1109/YAC.2016.7804882

Eddy, S. R. (1996). Hidden Markov models. *Current Opinion in Structural Biology*, *6*(3), 361–365. https://doi.org/10.1016/S0959-440X(96)80056-X

Fang, G., Zeng, F., Li, X., & Yao, L. (2021). Word2vec based deep learning network for DNA N4-methylcytosine sites identification. *Procedia Computer Science*, *187*, 270–277. https://doi.org/10.1016/j.procs.2021.04.062

Gao, Y., & Glowacka, D. (2016). Deep Gate Recurrent Neural Network. In R. J. Durrant & K.-E. Kim (Eds.), *Proceedings of The 8th Asian Conference on Machine Learning* (Vol. 63, pp. 350–365). PMLR. https://proceedings.mlr.press/v63/gao30.html

Greff, K., Srivastava, R. K., Koutnik, J., Steunebrink, B. R., & Schmidhuber, J. (2017). LSTM: A Search Space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(10), 2222–2232. https://doi.org/10.1109/TNNLS.2016.2582924

Gromiha, M. M. (2010). Protein Sequence Analysis. *Protein Bioinformatics*, 29–62. https://doi.org/10.1016/B978-8-1312-2297-3.50002-3

Gunasekaran, H., Ramalakshmi, K., Rex Macedo Arokiaraj, A., Deepa Kanmani, S., Venkatesan, C., & Suresh Gnana Dhas, C. (2021). Analysis of DNA Sequence Classification Using CNN and Hybrid Models. *Computational and Mathematical Methods in Medicine*, *2021*, 1–12. https://doi.org/10.1155/2021/1835056

Huerta, M., Haseltine, F., Liu, Y., Downing, G., & Seto, B. (2000). *NIH working definition of bioinformatics and computational biology*.

Jin, J., Tian, F., Yang, D.-C., Meng, Y.-Q., Kong, L., Luo, J., & Gao, G. (2017). PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Research*, *45*(D1), D1040–D1045. https://doi.org/10.1093/nar/gkw982

Jin, J., Yu, Y., & Wei, L. (2022). Mouse4mC-BGRU: Deep learning for predicting DNA N4-methylcytosine sites in mouse genome. *Methods*, *204*, 258–262. https://doi.org/10.1016/j.ymeth.2022.01.009

Karin, M. (1990). Too many transcription factors: positive and negative interactions. *The New Biologist*, *2*(2), 126–131.

KILIC, S. (2013). ROC Analysis in Clinical Decision Making. *Journal of Mood Disorders*, *3*(3), 135. https://doi.org/10.5455/jmood.20130830051624

Latchman, D. S. (1993). Transcription factors: an overview Function of transcription factors. *Int. J. Exp. Path*, *74*, 417–422.

Luque, A., Carrasco, A., Martín, A., & de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, *91*, 216–231. https://doi.org/10.1016/J.PATCOG.2019.02.023

Narayana, N., Ginell, S. L., Russu, I. M., & Berman, H. M. (1991). Crystal and molecular structure of a DNA fragment: d(CGTGAATTCACG). *Biochemistry*, *30*(18), 4449–4455. https://doi.org/10.1021/bi00232a011

Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984. Nomenclature Committee of the International Union of Biochemistry (NC-IUB). (1986). *Proceedings of the National Academy of Sciences*, *83*(1), 4–8. https://doi.org/10.1073/pnas.83.1.4

Piecyk, R. S., Schlegel, L., & Johannes, F. (2022). Predicting 3D chromatin interactions from DNA sequence using Deep Learning. *Computational and Structural Biotechnology Journal*, *20*, 3439–3448. https://doi.org/10.1016/j.csbj.2022.06.047

Price, M. N., Wetmore, K. M., Waters, R. J., Callaghan, M., Ray, J., Liu, H., Kuehl, J. v, Melnyk, R. A., Lamson, J. S., Suh, Y., Carlson, H. K., Esquivel, Z., Sadeeshkumar, H., Chakraborty, R., Zane, G. M., Rubin, B. E., Wall, J. D., Visel, A., Bristow, J., … Deutschbauer, A. M. (2018). Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature*, *557*(7706), 503—509. https://doi.org/10.1038/s41586-018-0124-0

Riaño-Pachón, D. M., Ruzicic, S., Dreyer, I., & Mueller-Roeber, B. (2007). PlnTFDB: an integrative plant transcription factor database. *BMC Bioinformatics*, *8*(1), 42. https://doi.org/10.1186/1471-2105-8-42

Sakr, A. S., Pławiak, P., Tadeusiewicz, R., & Hammad, M. (2022). Cancelable ECG biometric based on combination of deep transfer learning with DNA and amino acid approaches for human authentication. *Information Sciences*, *585*, 127–143. https://doi.org/10.1016/j.ins.2021.11.066

Şeker, A., Diri, B., & Balık, H. H. (2017). Derin Öğrenme Yöntemleri ve Uygulamaları Hakkında Bir İnceleme. *Gazi Mühendislik Bilimleri Dergisi*, *3*(3), 47–64.

Shu, J. J. (2017). A new integrated symmetrical table for genetic codes. *Biosystems*, *151*, 21–26. https://doi.org/10.1016/J.BIOSYSTEMS.2016.11.004

Strodthoff, N., Wagner, P., Wenzel, M., & Samek, W. (2020). UDSMProt: universal deep sequence models for protein classification. *Bioinformatics*, *36*(8), 2401–2409. https://doi.org/10.1093/bioinformatics/btaa003

Tang, X., Zheng, P., Li, X., Wu, H., Wei, D.-Q., Liu, Y., & Huang, G. (2022). Deep6mAPred: A CNN and Bi-LSTM-based deep learning method for predicting DNA N6-methyladenosine

sites across plant species. *Methods*, *204*, 142–150. https://doi.org/10.1016/j.ymeth.2022.04.011

WATSON, J. D., & CRICK, F. H. C. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, *171*(4356), 737–738. https://doi.org/10.1038/171737a0

Xiong, Z., Cui, Y., Liu, Z., Zhao, Y., Hu, M., & Hu, J. (2020). Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation. *Computational Materials Science*, *171*, 109203. https://doi.org/10.1016/j.commatsci.2019.109203

Yang, K. K., Wu, Z., Bedbrook, C. N., & Arnold, F. H. (2018). Learned protein embeddings for machine learning. *Bioinformatics*, *34*(15), 2642–2648. https://doi.org/10.1093/bioinformatics/bty178