Introduction

In an Escherichia Coli (E. Coli) bacteria that investigated by “in silico” (simulation) method which is important role nowadays has been discovered an immune system named CRISPR (clustered regularly interspaced short palindromic repeats)/Cas9 [1]. According to this system an E. Coli bacteria which is infected by any virus, add the virus DNA its memory and recognize when any other virus attack accrues. This defined virus DNA cuts this virus DNA from its own DNA through the CAS9 enzyme. Thus, DNA repair will occur. According to recent research the leading factor in gene regulation has been observed the “microRNA (miRNA)” targeted by the “single-guide RNA (sgRNA)” s. MiRNAs are small and non-coding RNA molecules [2]. In diseases, for example, different miRNAs involved in all stages of cancer make cancer diagnosis and treatment available [3]. CRISPR/CAS9 is used to destroy targeted miRNAs in cells [4]. It uses “guideRNA (gRNA)” as a guide to target the CRISPR/CAS9 nucleus to the DNA sequence and trigger the double-strand break at the desired location. Breaking and repair of these threads can cause random addition and alteration of DNA [5].

MiRNA activities are different in each cell type [6]. So, the miRNA activities of a human cell and a bacterial cell will not be the same. From this point, it is important to find a dataset about the genomes to be studied in this respect. Various miRNA target data sets used in “in silico” studies will be used [7]. The aim of this study, it is made mirRNA target estimation with machine learning algorithms. The result of wrong targeted miRNA may be cause of undesirable gene mutations [8]. It is aimed to minimize the Type 1 and Type 2 errors of miRNAs by applying machine learning and deep learning method of Type 1 and Type 2 errors as output. Thus, the wrong target estimate will be minimized. In this way, it will be possible to reliably repair gene damage by correctly targeting mis-targeted miRNAs. Genetic diseases will be eliminated by repair of gene damage.

The purpose of the “in silico” technique is to increase the accuracy of disease prevention by single pointing with the help of mechanization. The large base readings provided by mechanization also contribute greatly. Studies for miRNA targeting, support vector machines (Support Vector Machine -SVM) [9], deep learning, constrained logic programming (Constraint Logic Programming) [10] and a class classification (One Class Classification-OCC) [11] methods were used. Besides, in the future big data may be decreased to its most valuable parts with digital data forgetting concepts. It makes computations faster and the machines will store less data in disks [37].
Materials and Methods

CRISPR

CRISPR/CAS9 system is a strong genome editing mechanism used in many biotechnology applications [12]. MiRNAs are RNA molecules that play an important role in gene regulation in animals and plants [13]. Damaged miRNAs are targeted by sgRNAs. However, the effectiveness of sgRNAs have not been defined firmly to target area [14]. In the systems of CRISPR/CAS9 (clustered regularly interspaced short palindromic repeats), Cas9 nuclease creates short array (about 20 nucleotide) with RNA guideline two stranded in the determined region at DNA. CRISPR/CAS9 has been a unique technology that allows genetic and medical researchers to add, subtract or modify DNA sequences in various parts of the genome. In contrast to ZFN (Zinc-Finger Nuclease) and TALENs (Transcription Activator-Like Effector Nuclease), CRISPR/CAS9 is not man-made; currently the system is part of the bacterial immune system, which helps protect from invasive phages. Originality in CRISPR is achieved using an RNA molecule that is complementary to the gene of interest. After binding, this RNA molecule (also called guide RNA) traps a CAS9 nuclease, a double-strand break that causes a frame shift when repaired by NHEJ (Non-Homologous End Joining). CRISPR is the most effective process to date in gene repair and editing tools. Efficiency varies depending on the organism and the target site. Since it is the simplest, versatile and sensitive method of genetic manipulation that is currently available, it attracts great attention in the world of science. In summary, CRISPR/CAS9 differs from ZFN and TALENs in an important aspect that makes them superior to genomic regulation applications: The ZFN and TALENs bind to DNA through a direct protein-DNA interaction that requires redesigning the protein for each new target.

DATA Set

Applied “in silico” research and review, BLAST (Basic Local Alignment Search Tool) was used for CRISPR [15]. BLAST is a search tool that analyzes the amino acids and DNA sequences of proteins and finds similarities between them. Besides BLAST, the data set resources have been used such as National Human Genome Research Institute (NHGRI) [16], miRBase [17], Genome Crispr [18], CrisprInc [19], ENSEMBL [20], ENCODE [21], CRISPRz [22], CRISPOR [23], CRISPR Local [24]. In the algorithm studies performed with these data sets, estimation tools such as mirWalk [25], TargetScan (ID2 PPI analysis network) [26], miranda [27], mirbase [28], mirTarget [29], TarBase [30] have been developed.

In this study, CRISPR Local data set has been used for the analysis [31]. The source of CRISPR Local dataset is ENSEMBL Plants. There are approximately 854,610 lines of CRISPR data in the original. In this study, we used 34,200 lines of data because of technical issues. There are 11 column features in this dataset which are examples of “Cyanidioschyzon merolae” alga. This features; the gene in which the sgRNA, on target estimated chromosome and its coordinate, sgRNA sequence with 23'nt., on target prediction score, off target prediction gene which has the greatest CFD score, the chromosome on target prediction, its coordinate and beginning position, off-target prediction sequence, the number of sgRNA and mismatch on the off-target sequence, axon name, axon start position, all off-target and sgRNA having the highest CFD score.

In this study, CNN, MLP and BLSTM models were constituted and compared.

Experimental Studies

In this study, CNN, MLP and BLSTM models were constituted and compared.

Multilayer Perceptron -MLP

![Figure 2: MLP Model.](image_url)
A 4-layer MLP model has been developed according to the MLP algorithm run using the Google Colab GPU (https://colab.research.google.com). In the MLP model a fully connected structure of dense layers was formed. Information on the model used is shown in Figure 2. The rates of logistic regression and accuracy according to the MLP model are shown in Figure 3 and Figure 4. Accuracy obtained 80.38% according to MLP model.

Convolutional Neural Network - CNN

Convolutional Neural Networks are a model of artificial neural network which is used successfully in image processing, bioinformatics, robotics, data mining, finance and many other areas. However, except for image analysis, surprisingly high accuracy ratio was obtained in emotion analysis, text classification and question answering applications. According to this model, it is applied to nxn matrix with nxn filtering method (dot product), with acceptation of n>m. Thus, it allows the identification and classification of properties. In Figure 5, 3x3 matrix as a result of the intrinsic product of a 5x5 matrix and filtering was obtained. In this study, data set has been divided two group one of training 70 per cent, the other test 30 per cent. The model has been fixed up to non-linear by using the tangent and sigmoid activation functions. Convolution network is used to clarify the properties. The convolution network helps to create a new matrix with the results of the multiplication of the matrices. In order to prevent over fitting, maxpooling layer was used. It selects the elements with the maximum value from the matrixpool of the specified size in the maxpooling layer. Accordingly, the information obtained when the CNN model is generated can be seen in Figure 6. The rate of accuracy according to the CNN model can be seen from Figure 7.
Bidirectional Long-Short Term Memory - BSLTM

Bidirectional LSTM is different from other feed forward models in neural networks feedback system. Accordingly, the information obtained when the bidirectional LSTM model is generated can be seen in Figure 8. The rate of accuracy according to the bidirectional LSTM model can be seen from Figure 9. Accuracy has obtained 80.88% with bidirectional LSTM model. Accuracy was 96.7% compared to the CNN model. Stochastic Gradient Descent (SGD) optimization method was used. Learning rate was set to 0.0005. Binary cross entropy logarithmic regression function was used. The following table (Table 1) shows the MLP, CNN and bidirectional LSTM model accuracy rates. Table 2 shows the MLP, CNN and bidirectional LSTM model accuracy rates, precision-recall, f-measure values. In another study [35] authors used the same data set. Source of Crispr Local which is Ensembl Plants [36] data contains on a set of plant and animal cell data. Support Vector Machine (SVM) algorithm was applied. Located accuracy of ratio is 87 %. Table 3 shows a comparison of our algorithm and last studies accuracy rate.

Table 1: MLP, CNN and Bidirectional LSTM accuracy.

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convolutional Neural Network-CNN</td>
<td>96.7%</td>
</tr>
<tr>
<td>Multilayer Perceptron-MLP</td>
<td>80.38%</td>
</tr>
<tr>
<td>Bidirectional Long Short-Term - BLSTM</td>
<td>80.88%</td>
</tr>
</tbody>
</table>

Table 2: MLP, CNN and BLSTM results.

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Accuracy</th>
<th>Loss</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN</td>
<td>96.79%</td>
<td>0.05</td>
<td>1.00%</td>
<td>96.79%</td>
<td>98.36%</td>
</tr>
<tr>
<td>MLP</td>
<td>80.38%</td>
<td>0.24</td>
<td>1.00%</td>
<td>80.38%</td>
<td>89.12%</td>
</tr>
<tr>
<td>BLSTM</td>
<td>87.62%</td>
<td>0.69</td>
<td>1.00%</td>
<td>87.61%</td>
<td>93.39%</td>
</tr>
</tbody>
</table>

Table 3: Comparing with other studies.

<table>
<thead>
<tr>
<th>Study</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhu &amp; Liang (2018) (SVM)</td>
<td>87.0%</td>
</tr>
<tr>
<td>Our research (CNN)</td>
<td>96.7%</td>
</tr>
</tbody>
</table>

Conclusion

In this study, the algorithms of Multilayer Perceptron-MLP, Convolutional Neural Networks-CNN and Bidirectional Long Short-Term Memory-BLSTM have been compared with use of CRISPR data set. As a result, according to this data set, the accuracy rate in the MLP model was 80.38% and Bidirectional LSTM model was 80.88% whereas for CNN this result was found 96.7%. According to the results, a highest accuracy rate was obtained with the CNN model than MLP and BLSTM. In the CNN model, revised CRISPR has reached up to 7 layers according to the ENSEMBL Plants dataset and 4 layers have been formed in MLP and 2 layers have been formed in BSLTM. Comparing other algorithms with our algorithm is better performance according to results. Research that used with SVM algorithm performed 87.0% accuracy result. However, our model achieved 96.7 %. This result is more reliable performance than the research used SVM algorithm. Any mistargeted position causes unwanted genome distortions. For this reason, the accuracy rate is urgent in sgRNA targeting.

References


10.26717/BJSTR.2019.21.003652


This work is licensed under Creative Commons Attribution 4.0 License

Submission Link: https://biomedres.us/submit-manuscript.php

https://biomedres.us/