UTM Computing Proceedings
Innovations in Computing Technology and Applications
Volume 2 | Year: 2017 | ISBN: 978-967-0194-95-0

Classification of DNA Sequences Using Convolutional Neural Network Approach

Nurul Amerah Kassim[1], and Dr Afnizanfaizal Abdullah[2]

Faculty of Computing, Universiti Teknologi Malaysia (UTM), Malaysia

[1]amerahkassim@gmail.com, [2]afnizanfaizal@utm.my

Abstract. Extraction of meaningful information from the DNA is a key elements in bioinformatics research and DNA sequence classification has a wide range of presentations such as genomic analysis, and biomedical data analysis. Nowadays, deep learning approach has become an attention to many researcher. This models contains multiple of non-linear transforming layers which practice to represent a data at successively high-level abstractions. With many hidden layer, this innovative model are expected to be able to elucidate any complex problems. Thus, Convolutional Neural Network approach is proposed to classify the whole genomic sequences of an organisms. As the purpose of this research is to evaluate the performance of the proposed model by implementing convolutional neural network approach, the research framework is focused to identify genetic marker for liver cancer from Hepatitis B Virus DNA sequences using deep learning principle. The results show that convolutional neural network have more than 90 percent accuracy in training the data sets. Moreover, this research also analysed different size of sequence length to observe the performance of the proposed model. The overall outcome in this research achieve within the expected results.

Keywords: Deep Learning, DNA classification, convolutional neural network, hepatits B virus

1 Introduction

In recent years, a new branch techniques of machine learning models called deep learning was presented [1]. This models contains multiple of non-linear transforming layers which practice to represent a data at successively high-level abstractions. With many hidden layer, this innovative model are expected to be able to elucidate any complex problems. In this past years, there are several researchers who try to implement the DNA sequence classification to deep learning algorithm.

Eichholt and Cheng [2] have applied deep networks to predict protein disorder. In their research, they had a several sequences based features as an inputs of the predictor and they gain 0.82 of an accuracy which quiet high compare to others. Next in 2014, Leung et al. [3], proposed to train a sequences using deep neural network model to predict splicing patterns in human tissues. During the experiment, they extract 1393 of genomic features as a training data and achieved ominously improved performance in comparison with previous researches' models. Notice that, the previous conducted research still need to expertized features to symbolise sequences. Regarding to that situation, some of crucial and needed information about locus of each nucleotide in the sequence may and was lost. These matter will led to the lessening of models' performance.

Raw nucleotide sequences do not have explicit features and commonly techniques depictions introduce the key problem of the high dimensionality. Besides, machine learning method to supervised classification problems are intensely reliant on the feature extraction step. Somehow, some of important information about locus of each nucleotide in the sequence may and was lost. These matter will led to the lessening of classifier models' performance.

The goal of this research is to implement deep neural networks principles in DNA sequence classification. The specific objectives of this research are to analyse and study the machine learning approaches in classify the DNA sequences, to propose deep neural networks principles which is convolutional neural networks in DNA sequences classification, and to evaluate the performance of proposed models (convolutional neural network) in classify the DNA sequence.

The research will focus on two different class of DNA which are Hepatitis B Virus (HBV) Genotype B nucleotide sequence and Non-HBV nucleotide sequences used by Leung et al., [4]. The source of data information of the selected chassis can be taken from databases sources: Genebank is the database of the nucleotide sequences as describe on reference paper. Next, the research focus on using Python as programming language with help with Tensorflow as a framework which provide various of deep learning library. Deep learning principles that had been proposed is Convolutional Neural Network (CNN).

2 Methodology

The study contains two main phase which is pre-processing phase and post-processing phase. In pre-processing phase, the workflow focus on data pre-processing steps while in post-processing, the workflow can be broken down into two sub-steps which is model learning and model evaluation. Figure 2.1 shows an overview of research study representation.

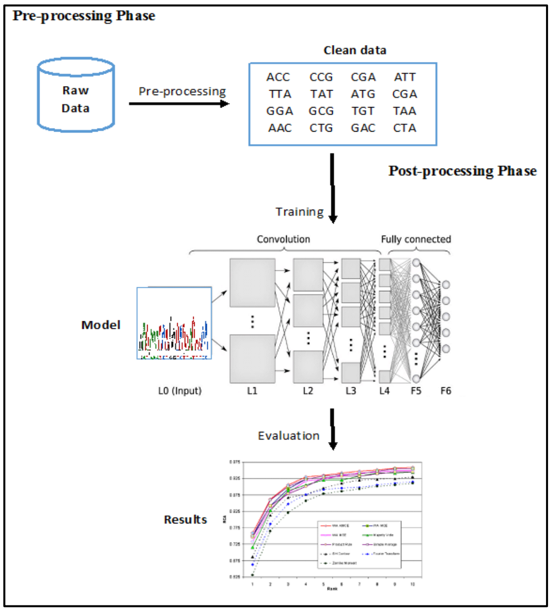

Figure 2.1: Representation of experimental setup

FASTA format is a text-based format, however, contrasting a text data, the DNA sequences in FASTA is lines of sequential letters without spacing. As being discussed in chapter 2, the convolutional neural network also had been applied in text data problems. Hence, the DNA sequences need to be translate as sequence to sequences of words in order to implement the same representation method of the proposed approach. This steps is very crucial as to avoid any missing data and neglect any losing locus information of each nucleotide in sequences.

Generally, the proposed model includes four steps in total regarding to different layer embedded. The model contains one embedding layer which will encoded the sequences and one convolutional layer followed by a max-pooling layer which extract features from representation matrixes of sequences. Then, all the extracted features is merged into one big feature vector using fully connected layer. Finally, the accuracy of the tested model is calculate and will be analysed as a performance result.

3 Results

İn this section, the proposed model is compared with others classifier which is Support Vector Machine (SVM), Neural Network (NN), Decision Tree (DT), Naïve Bayes (NB), Nonlinear Integral Classifier (NIC), Rule Learning (RL), to validate the improvement of classification.

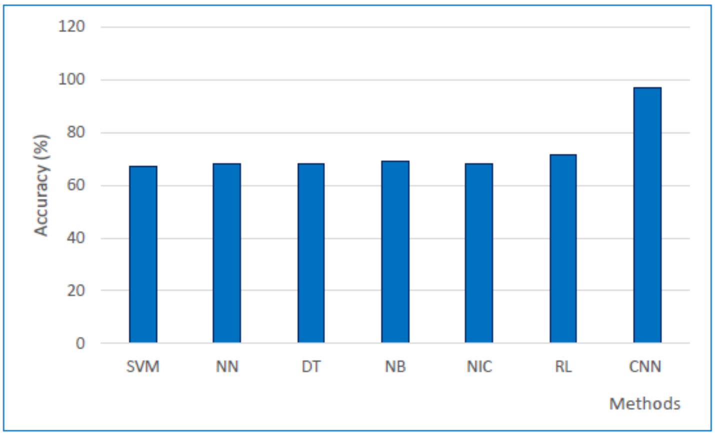

Figure 3.1: Comparison of all methods

Figure 3.1 shows a comparison of accuracy of all compared methods including the proposed methods. This experiment performed and run with whole nucleotide sequences (3215 bp for training set and 1675 bp for testing set). It is clearly shows that CNN has the highest value of accuracy with 96.83% followed by RL, NB, DT, NIC, NN and SVM with 71.6%, 68.9%, 68.2%, 68.2%, 68.1%, and 67.4% respectively.

Meanwhile, figure 3.2 shows a comparison of accuracy of four different sequences length run using proposed model. This set up is to analyse the performance of the proposed model on different size of data. From the graph, the outcome of the training are, with 500bp length to 2000bp length is steadily increase with 93.3%, 93.8%, 94.2%, and 95.6%.

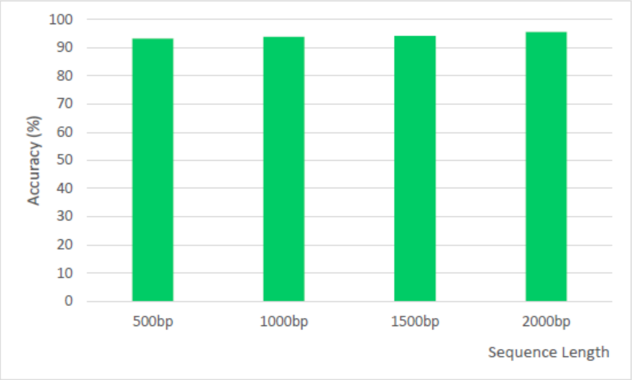

Figure 3.2: Comparison of four different sequences length

4 Discussion

In average the improvement of the proposed model compared to previous model is 28.43%. These improvements are quite high in comparison with other approaches such as finding good representations for sequences or feature selection which were applied before. It showed that features extracted by convolutional layers of the convolutional neural network are very useful for the classifier to classify sequences into true categories. This also means that this method can have higher prediction power in identify the genetic marker in liver cancer (HCC) nucleotide sequence. In this research, an improvements is achieved by using convolutional neural network, a deep learning model with the high power of representing complicated problems. One-hot vectors to represent sequences is applied, so the model could preserve specific position information of each individual nucleotide in sequences.

Besides, classification scores considering shorter to longer sequences showed very interesting results. Using CNN approach, clearly outperforms all the other classifiers in terms of accuracy. From the table, the results of different length of sequences obtained are nearly in the same scores within 90%. It showed consistency of the classifier accuracy even with different size of input. The performance of the model is increase as the size is increase. This phenomena is due to CNN model is working at the best with the larger size of input as it had been applied to process a thousand of image to identify an object.

The significant in recorded the loss of the model is to observed the learning rate perform by the proposed model. First, TensorFlow provide the smoothing function to visualize the graph in more clear shape. Figure 4.1 and 4.2 shows the comparison between validation graph and experimental result graph.

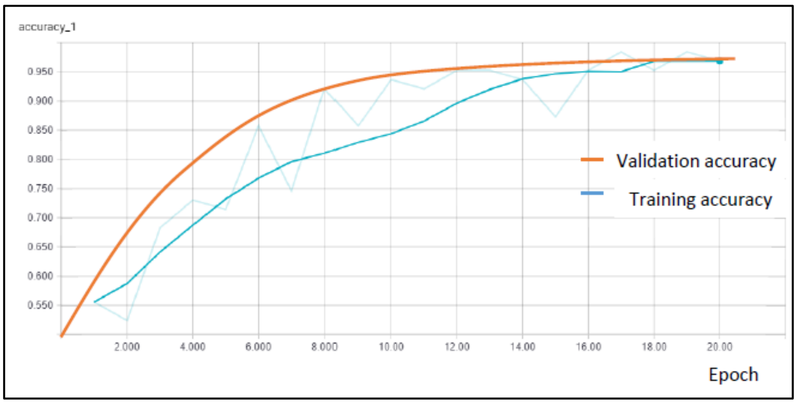

Figure 4.1: Comparison of validation accuracy and training accuracy

Based on the figure 4.1, the observed gap between the training and validation accuracy indicates the amount of overfitting. The smaller the distance between the two line graph (validation and training) the smaller the overfitting may occur. Thus, this research model able to minimize the amount of the overfitting with help of dropout parameter.

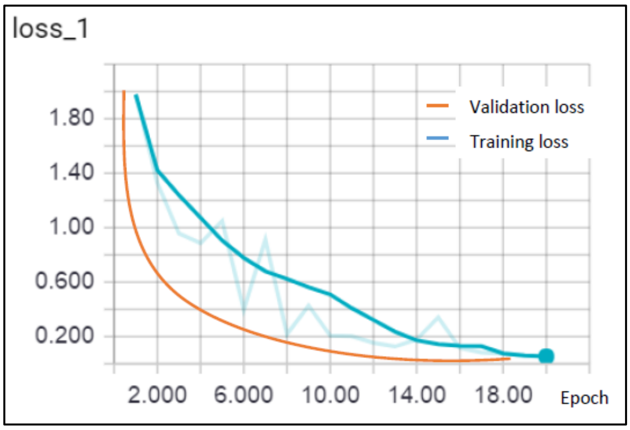

Figure 4.2: Comparison of validation loss and training loss

The loss depicting the effects of different learning rates. Low learning rates the curves will be linear while with high learning rates they will start to look more exponential. From figure 4.2, the curve of a typical loss function over time of a training loss looks reasonable. The curve almost exponential like the validation loss which indicates good learning rate. This explained that the model quickly learn the features maps from time to time during forward backward propagation in the CNN layer.

5 Future Outlook

Database of DNA sequences is still evolving. This is due to many research has been conduct, many new viruses, bacteria or genomics sequences have been found and this matter contribute in many unsolved problems involving the DNA sequence of the organisms. Deep learning also being evolve by years. Thus, for the future, deep learning is looking even more promising. It is to be expect something new invention or discovery will be found such a new technique of a hybrid techniques between two methods. A more elegant future, when deep learning is discover and can design and integrate a system that can evolve and adapt to any of environmental and contextual differences.

6 Conclusion

This study was conducted to analyse the performance of the proposed model in classify DNA sequence. The study also compared the result with the previous best performance to evaluate an improvement of the proposed model. Moreover, this research also analysed different size of sequence length to observe the performance of the proposed model. The overall outcome in this research achieve within the expected results.

Acknowledgements. Express the sincere gratitude to the parties involved such as sponsors, departments, individuals, companies, proprietors of premises and so on in conducting research/ project. Special thanks to my supervisor, reviewers and the bioinformatics members of UTM batch 2013/14 for insightful comments. This study was supported and guide by the Final Year Project supervisor, Dr Afnizanfaizal Abdullah.

References

41. Nguyen, N.G., Tran, V.A., Ngo, D.L., Phan, D., Lumbanraja, F.R., Faisal, M.R., Abapihi, B., Kubo, M. and Satou, K. DNA Sequence Classification by Convolutional Neural Network. J. Biomedical Science and Engineering, 9, 280-286. (2016)
42. Eickholt, Jesse, and Jianlin Cheng. "DNdisorder: predicting protein disorder using boosting and deep networks." BMC bioinformatics 14.1 (2013): 88.
43. Lee, Honglak, et al. "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations." Proceedings of the 26th annual international conference on machine learning. ACM, 2009.
44. Leung, KwongSak, et al. "Data mining on dna sequences of hepatitis b virus." IEEE/ACM transactions on computational biology and bioinformatics 8.2 (2011): 428-440.