Journal of Theoretical and Applied Information Technology
15th June 2018. Vol.96. No 11
© 2005 – ongoing JATIT & LLS

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN ENHANCE CNN-RNN MODEL FOR PREDICTING FUNCTIONAL NON-CODING VARIANTS

[1]JALILAH ARIJAH MOHD KAMARUDIN, [1]NUR AFIFAH AHMAD AHYAD
[1]AFNIZANFAIZAL ABDULLAH, [2]ROSELINA SALLEHUDDIN

[1]Synthetic Biology Research Group,

[2]Soft Computing Research Group,

Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Bharu, Johor, Malaysia

E-mail: [1]jalilaharijah@gmail.com, [1]afnizanfaizal@utm.my, [2]roselina@utm.my

ABSTRACT

In the era of big data, deep learning has advanced rapidly particularly in the field of computational biology and bioinformatics. In comparison to conventional analysis strategies, deep learning method performs accurate structure prediction because it can handle high coverage biological data such as DNA sequence and RNA measurement using high-level features. However, predicting functions of non-coding DNA sequence using deep learning method have not been widely used and require further study. The purpose of this study is to develop a new algorithm to predict the function of non-coding DNA sequence using deep learning approach. We propose an enhanced CNN-RNN model to predict the function of non-coding DNA sequence. In this model, we train an algorithm to automatically find the optimal initial weight and hyper-parameter to increase prediction accuracy which outperforms other prediction models.

Keywords: *Functional Non-coding Variant, Machine Learning, Deep Learning, Convolutional Neural Network, Recurrent Neural Network*

1. INTRODUCTION

DNA strands consist of coding sequence which encodes for proteins and much of the DNA does not encode for proteins which is known as non-coding DNA or junk DNA. The coding DNA are transcribed into messenger RNA (mRNA) which is then used to produce proteins such as ribosomal. Meanwhile, the non-coding DNA sequence often referred as the conserved non-coding sequence (CNS) [1] produces introns, non-coding functional RNA (rRNA and tRNA), cis- and trans-regulatory elements, repeat sequence telomeres, scaffold attachment regions (SARs) and many others. However, the functions of most non-coding sequence have not been determined yet. Additionally, in the process of gene transcription and replication

Additionally, in the process of gene transcription and replication, CNS provides the binding site or regulatory site for proteins such as the promoter region and origin. Mutation in non-coding variant is also believed to be associated with the development of certain diseases.

Genome-wide association study (GWAS) focuses on non-coding regions and potentially to non-coding variants. Some of the genetic disorder has been identifies based on disease-associated mutations by sequencing the exome or coding region [2,3]. However, many of cases still remain undetermined because of the failure in analysis due to limitation of exome sequencing. This problem gives the strong reason that some of causative variants actually occur outside of coding region and inside regulatory. For example, mutation of the functional regulatory elements such as enhancers and insulators may results in cancer, diabetes, heart disease, obesity [4,5] and a rare disease called Hirschsprung's disease [6].

The whole genome sequencing (WGS) approach could characterize non-coding functional variant. Besides, WGS also enable to predict a structural variant (SV) includes copy number variants (CNVs) and copy number neutral SV. However, new methods are required to perform functional prediction of a large number of non-coding variant within single a human genome [7]. To overcome this limitation, computer algorithm was developed in order to deal with high-dimension of biological data and large numbers of datasets. The development and application of computer algorithms is related to the field of machine learning that aims to improve with experience. Machine learning is a supervised learning and used to classify objects in images, match news items, translate speech into text, products or posts with interest of users and select significant results of a search. Generally, all of these applications of machine learning are using a class of approaches known as deep learning [8].

Deep learning has clear potential to make analysis in high-dimensional data by training complex networks with various layers that represent their internal structure. This approach can be applied in image recognition [9-12], speech recognition [13-15], predicting potential drug [16], reconstructing brain circuits and many others. It has a major advance in solving problem including function prediction of non-coding variants. However, the main issue is difficulty and lacking of powerful computational approach in predicting the functional non-coding variants in human genome that potentially leads to detect disease-associated variants.

In order to achieve the aim of this study, we propose an enhanced CNN-RNN model to predict the function of non-coding sequence. This paper provides details explanation on the proposed model specifically on the algorithm used. The comparison between the proposed model and the existing models will also be discussed.

2. LITERATURE REVIEW

In this study, we highlight two deep learning method commonly used includes convolutional neural network (CNN) and recurrent neural network (RNN).

2.1 CNN

The input data of CNN available in the form of multidimensional arrays for example, one-dimensional genomic sequence or two-dimensional images. Large dimensional data becomes a challenge for a neural network because it requires a high number of parameters compared to the number of training data that fit in a model. CNN makes change to the network's structure by reducing the number of parameters during learning stage to overcome the challenge.

CNN consists of three layers of network known as convolutional layers, nonlinear layers and pooling layers. The convolutional layer in CNN composeds of various maps of neurons known as feature maps or filters [17]. The size of feature maps is equivalent to the dimension of the input image. To reduce the number of model parameters, CNN uses the concept of parameter sharing and local connectivity where by each neuron in the feature maps are only connected with a local patch of neurons in the previous layer or receptive field. Next, all neurons in the feature maps share the same parameters and scan for the same features as in the previous layer but at different locations. The weighted sum of input neurons and activation function is computed using discrete convolution to obtain the activity of neuron [18].

Usually, the frequency and exact position of features are irrelevant for final prediction for example, to identify objects in an image. By that, pooling layer summarizes adjacent neurons by calculating the average or maximum activities to represent features activities. This pooling operation effectively reduce the number of model parameters and down-sampled the input image [18]. Mostly, CNN includes several convolutional and pooling layers to learn various numbers of abstract features from small edges to whole objects. The number of convolutional layers, the size of receptive fields and the number of features maps dependently follows the application. Besides, the validation data set need to be strictly selected.

2.2 RNN

Recurrent Neural Network (RNN) was designed with a cyclic connection in the basic structure and used sequential information. Since the cyclic connection exists, RNN was performed in the hidden unit and the input data was processed sequentially. Thus, past information is stored in the hidden units which are known as state vectors [18].

Although RNN is different from CNNs based on their number of layers but it can be considered as deeper structure if unrolled in time. However, researchers is facing up disappear gradient problem and learning long-term dependency of data during training RNNs. This can be solved by replacing the simple perceptron hidden units with complex units including gated recurrent unit (GRU) and long short term memory (LSTM) that running as memory cells. RNNs still provide effective analysis methods particularly in sequential information even it has been less explored compared to DNNs and CNNs. In addition, this technique is a promising method to perform functional prediction of fixed size sequence and for mapping a variable-length input sequence [18].

3. CNN-RNN MODEL

We propose a CNN-RNN model, a framework that combines CNN and RNN with additional swarm optimization. This enhanced model uses the same features and framework as the DeepSEA [19] and DanQ [20] models.

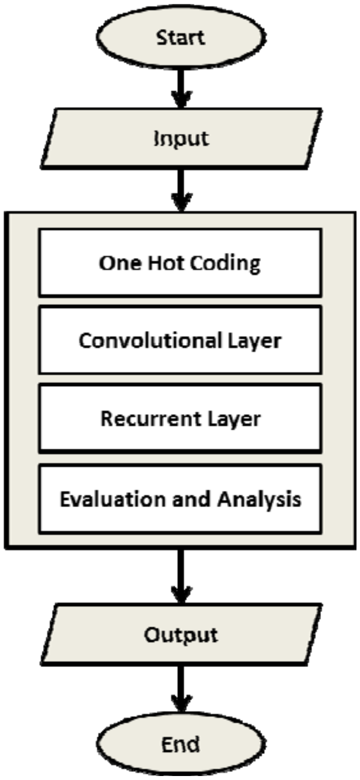


Figure 1: Framework of an enhance CNN-RNN

Firstly, an input sequence is encoded into a 4-row bit matrix. Next, the convolutional layer act as a motif scanner and max pooling to reduce the size of the output matrix. Recurrent layer that uses bidirectional long short term memory (BLSTM) considers the orientations and spatial distances among the motifs. Lastly, the dense layers of rectified linear units (ReLU) and a multi-task sigmoid output are used to evaluate the prediction.

3.1 One Hot Coding

We used the raw nucleotide bases A, C, G and T as input and each of the bases is converted into one-hot coding. A binary vectors with matching characters either A, C, G or T were encoded as 1 and the rest as 0s.

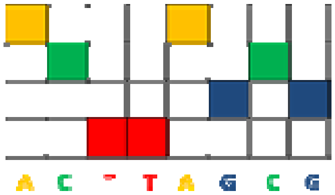


Figure 2: Basic structure of one hot encoding

This encoding matrix becomes the input for CNN-RNN algorithm and produce outputs of a vector of fixed dimension.

3.2 Convolutional Neural Network

It is reported that convolutional neural network (CNN) is suitable for predicting function from sequences by DeepSEA and DanQ models. To capture local patterns in the sequences, CNN uses a weight-sharing strategy. This weight-sharing strategy is useful for studying DNA composing convolution filters that able to capture sequence motifs.

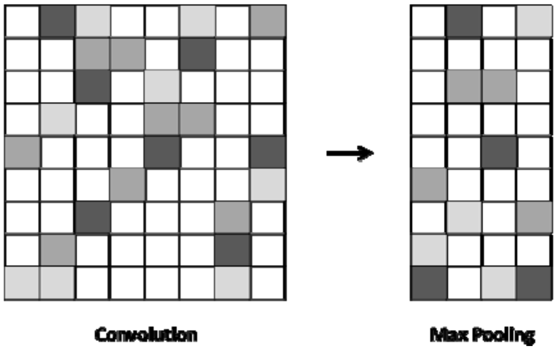


Figure 3: Basic structure of convolutional layer

The CNN is trained to predict large-scale chromatin-profiling data such as transcription factors (TF) binding, DNase I sensitivity and histone-mark profiles within various cell types. This method also predicts the regulatory function and the effect of regulatory variation. The CNN consists of one convolution layer and one max pooling layer.

3.2.1 Convolution layer

Convolution layer acts as motif scanner within the input matrix to produce output matrix by using rectifier activation. The output is produced with a row for every convolution kernel and a column for every position in the output. A convolution layer calculates ReLU using function above:

$$convolutional(X)_{ik} = ReLU(\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} W^{k}_{mn} X_{i+m,n}) \quad (1)$$

where *X* is the input, *i* is the index of the output position and k represents the index of kernels. Each convolutional kernel W^k is *M* × *N* (-weight matrix) which *M* is the window size and *N* is the number of output channel which equals to 4.

3.2.2 Max Pooling layer

Max pooling layer act to reduce the size of the output matrix with the spatial axis by preserving the number of channels. It calculates with a step size equivalent to the size of pooling window. The size of the output is reduced and allows learning sequence feature. The pooling operation is expressed as

$$pooling(X)_{ik} = \max(\{X_{iM,k}, X_{(iM+1),k}, \ldots, X_{(iM+M-1),k}\}) \quad (2)$$

where *X* is the input, *i* is the index of output position, *k* is the index for kernels and *M* represents pooling window size.

3.3 Recurrent Neural Network

Another layer in CNN-RNN model is RNN that form a directed cycle of connections between units. A variant of RNN which bidirectional long short term memory network (BLSTM) combines the outputs of two RNNs. In other words, one is processing the sequence from left to right and the other one processing the sequence from right to left.

Figure 4 illustrates the basic structure of BLSTM. The two RNNs that contain LSTM blocks are smart network units that can remember a value for an arbitrary length of time.

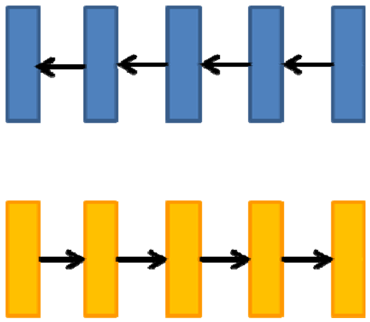


Figure 4: Basic structure of bidirectional long short term memory (BLSTM)

BLSTMs effectively capture long term dependencies and powerful for machine learning application [20].

3.4 Evaluation and Analysis CNN-RNN

A dense layer is one of the basic layers of the deep convolutional network. It calculates output using one-dimensional convolutional operation with a specific number of kernel or weight matrices. ReLU represents the rectified linear function

$$ReLU(x) = \begin{cases} x & if\ x \geq 0 \\ 0 & if\ x < 0 \end{cases} \quad (3)$$

The fully connected layer calculates ReLU (WX), where X represents the input and W is the weight matrix for a fully connected layer.

The final layer known as sigmoid output layer is used in this framework to makes prediction for 919 chromatin features such as 125 DNase features, 104 histone features and 690 TF features. The scale for predictions range from the 0 to 1 using the function of sigmoid

$$Sigmoid(x) = \frac{1}{1+e^{-x}} \quad (4)$$

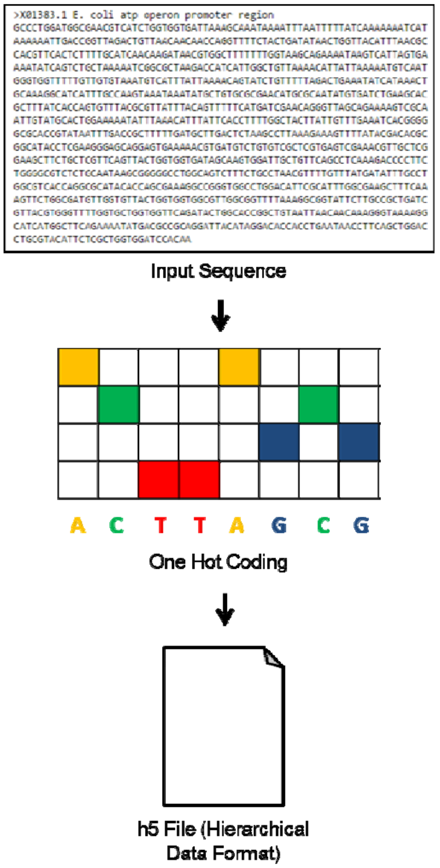
where X represents the input and W is the weight matrix for the sigmoid of output layer which calculates for Sigmoid (WX).

4. INITIAL RESULT

The initial result for this study focuses on the output of one coding algorithm that converts an input sequences into a binary vector.

The dataset of *E.coli* at operon promoter region retrieved from National Center for Biotechnology Information (NCBI) was used. This dataset consists of 1,150-bp of linear DNA and it is running on one hot coding algorithm as shown in Figure 5 above.

Precisely, the raw nucleotides which are A, C, G and T bases are converted into (1 0 0 0), (0 0 1 0), (0 1 0 0) and (0 0 0 1) respectively. The output at this stage was used as an input for the CNN-RNN model in the h5 file format.



```
>X01383.1 E. coli atp operon promoter region
GCCCTGGATGGCGAACGTCATCTGGTGGTGATTAAAGCAAATAAAATTTAATTTTTATCAAAAAAATCAT
AAAAAATTGACCGGTTAGACTGTTAACAACAACCAGGTTTTCTACTGATATAACTGGTTACATTTAACGC
CACGTTCACTCTTTTGCATCAACAAGATAACGTGGCTTTTTTTGGTAAGCAGAAAATAAGTCATTAGTGA
AAATATCAGTCTGCTAAAAATCGGCGCTAAGACCATCATTGGCTGTTAAAACATTATTAAAAATGTCAAT
GGGTGGTTTTTGTTGTGTAAATGTCATTTATTAAAACAGTATCTGTTTTTAGACTGAAATATCATAAACT
GCAAAGGCATCATTTGCCAAGTAAATAAATATGCTGTGCGCGAACATGCGCAATATGTGATCTGAAGCAC
GCTTTATCACCAGTGTTTACGCGTTATTTACAGTTTTTCATGATCGAACAGGGTTAGCAGAAAAGTCGCA
ATTGTATGCACTGGAAAAATATTTAAACATTTATTCACCTTTTTGGCTACTTATTGTTTGAAATCACGGGG
GCGCACCGTATAATTTGACCGCTTTTTGATGCTTGACTCTAAGCCTTAAAGAAAGTTTTATACGACACGC
GGCATACCTCGAAGGGAGCAGGAGTGAAAAACGTGATGTCTGTGTCGCTCGTGAGTCGAAACGTTGCTCG
GAAGCTTCTGCTCGTTCAGTTACTGGTGGTGATAGCAAGTGGATTGCTGTTCAGCCTCAAAGACCCCTTC
TGGGGCGTCTCTGCAATAAGCGGGGGCCTGGCAGTCTTTCTGCCTAACGTTTTGTTTATGATATTTGCCT
GGCGTCACCAGGCGCATACACCAGCGAAAGGCCGGGTGGCCTGGACATTCGCATTTGGCGAAGCTTTCAA
AGTTCTGGCGATGTTGGTGTTACTGGTGGTGGCGTTGGCGGTTTTAAAGGCGGTATTCTTGCCGCTGATC
GTTACGTGGGTTTTGGTGCTGGTGGTTCAGATACTGGCACCGGCTGTAATTAACAACAAAGGGTAAAAGG
CATCATGGCTTCAGAAAATATGACGCCGCAGGATTACATAGGACACCACCTGAATAACCTTCAGCTGGAC
CTGCGTACATTCTCGCTGGTGGATCCACAA
```

Figure 5: The flow of the input, one hot encoding algorithm and the output

5. DISCUSSION

This study presents an enhanced CNN-RNN model as a suitable method for functional prediction of non-coding variants. In comparison to DeepSEA and DanQ models, we strongly believe that this proposed model can outperform other prediction models and effectively predict the function directly from the sequence.

Table 1: Comparison between propose model and the existing prediction models.

Model	Purpose	Remarks
DanQ	Predicting the functional DNA sequences	Manually selects model parameters that can influence performance
DeepSEA	Predicting the effect of non-coding variants	Non-coding genomic regions and potential functions of complex disease or trait-associated SNPs are currently poorly understood
Enhance CNN-RNN	Predicting the functional non-coding variants	Turning algorithm to automatically find the optimal initial weight and hyper-parameters and train algorithm with 1024 convolutional kernel and use human genome data set

DanQ model as a predicting model for the function of DNA sequences uses a combination of convolutional and recurrent neural network as the framework. This prediction model consists of the convolutional layer to capture regulatory motifs and recurrent layer to capture long-term dependencies between motifs. In other words, the recurrent layer is used to learn a regulatory grammar in order to improve predictions [20]. DanQ outperforms others model including DeepSEA model against some metrics. This model achieved 50% improvement in area under the precision-recall curve (PR AUC) metric compared to other models [20]. However, DanQ model manually selects model parameters that can influence the performance caused by weight initialization. For an enhance CNN-RNN model, we currently tuning the algorithm to automatically find the optimal initial weight and hyper-parameter.

Slightly different with DeepSEA model, it only uses a convolutional neural network in their framework. As it is used for predicting the effect of non-coding variants, the framework of DeepSEA consists of three convolution layers and two max pooling layers in order to learn motifs [20]. In this model, the non-coding genomic regions and potential functions of complex disease or trait-associated SNPs are still poorly understood and warrant further studies [19]. This is the main reason the convolutional neural network and recurrent neural network were used in our propose model which is to increase the prediction accuracy. Besides, we also uses human genomic sequences as training dataset to provides better training data and insights of functional variants.

Since this study is still in its infancy, more investigations will be carried out to improve CNN-RNN model. Optimization method are proved to be helpful based in its implementation in solving biological problem [21-23], so based on this regard, we plan to implement optimization algorithm to this proposed model to increase the prediction accuracy. We hope that this proposed model will be utilized in medical applications to identify disease-associated variants, early diagnosis and interventions for patients in future.

REFRENCES:

[1] Woolfe, A., Goodson, M., Goode, D. K., Snell, P., McEwen, G. K., Vavouri, T., ... & Walter, K. (2004). Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol*, *3*(1), e7.

[2] Yang, Y., Muzny, D. M., Reid, J. G., Bainbridge, M. N., Willis, A., Ward, P. A., ... & Hardison, M. (2013). Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *New England Journal of Medicine*, *369*(16), 1502-1511.

[3] Beaulieu, C. L., Majewski, J., Schwartzentruber, J., Samuels, M. E., Fernandez, B. A., Bernier, F. P., ... & Adam, S. (2014). FORGE Canada Consortium: outcomes of a 2-year national rare-disease gene-discovery project. *The American Journal of Human Genetics*, *94*(6), 809-817.

[4] Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., ... & Amin, V. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, *518*(7539), 317-330.

[5] Corradin, O., & Scacheri, P. C. (2014). Enhancer variants: evaluating functions in common disease. *Genome medicine*, *6*(10), 85.

[6] Emison, E. S., McCallion, A. S., Kashuk, C. S., Bush, R. T., Grice, E., Lin, S., ... & Chakravarti, A. (2005). A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. *Nature*, *434*(7035), 857-863.

[7] Lupski, J. R., Reid, J. G., Gonzaga-Jauregui, C., Rio Deiros, D., Chen, D. C., Nazareth, L., ... & McGuire, A. L. (2010). Whole-genome sequencing in a patient with Charcot–Marie–Tooth neuropathy. *New England Journal of Medicine*, *362*(13), 1181-1191.

[8] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444.

[9] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

[10]Farabet, C., Couprie, C., Najman, L., & LeCun, Y. (2013). Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1915-1929.

[11]Tompson, J. J., Jain, A., LeCun, Y., & Bregler, C. (2014). Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in neural information processing systems* (pp. 1799-1807).

[12]Szegedy, C. et al. (2014). Going deeper with convolutions. Preprint at http://arxiv.org/abs/1409.4842.

[13]Mikolov, T., Deoras, A., Povey, D., Burget, L., & Černocký, J. (2011, December). Strategies for training large scale neural network language models. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on* (pp. 196-201). IEEE.

[14]Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *29*(6), 82-97.

[15]Sainath, T. N., Mohamed, A. R., Kingsbury, B., & Ramabhadran, B. (2013, May). Deep convolutional neural networks for LVCSR. In *Acoustics, speech and signal processing (ICASSP), 2013 IEEE international conference on* (pp. 8614-8618). IEEE.

[16]Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E., & Svetnik, V. (2015). Deep neural nets as a method for quantitative structure–activity relationships. *Journal of chemical information and modeling*, *55*(2), 263-274.

[17]Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. *Molecular systems biology*, *12*(7), 878.

[18]Min, S., Lee, B., & Yoon, S. (2016). Deep learning in bioinformatics. *Briefings in Bioinformatics*, bbw068.

[19]Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nature methods*, *12*(10), 931-934.

[20]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

[21]Abdullah, A., Deris, S., Hashim, S. Z. M., Mohamad, M. S., & Arjunan, S. N. V. IEEE (2011, December). An improved local best searching in particle swarm optimization using

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

differential evolution. In Hybrid Intelligent Systems (HIS), 2011 11th International Conference on (pp. 115-120).

[22]Abdullah, A., Deris, S., Mohamad, M. S., & Anwar, S. (2013). An improved swarm optimization for parameter estimation and biological model selection. PloS one, 8(4), e61258.

[23]Ismail, M. A., Deris, S., Mohamad, M. S., & Abdullah, A. (2015) A Newton cooperative genetic algorithm method for in silico optimization of metabolic pathway production. PloS one, 10(5), e0126199.