Indonesian Journal of Electrical Engineering and Computer Science
Vol. 5, No. 3, March 2017, pp. 691 ~ 696
DOI: 10.11591/ijeecs.v5.i3.pp691-696

Deep Neural Network Method for the Prediction of Xylitol Production

Siti Noorain Mohmad Yousoff*, 'Amirah Baharin, Afnizanfaizal Abdullah
Synthetic Biology Research Group, Faculty of Computing, Universiti Teknologi Malaysia,
81310 UTM, Johor, Malaysia
*Corresponding author, e-mail: ainyousoff@gmail.com

Abstract

Bio-based chemical products such as xylitol have achieved remarkable attentions both in pharmaceutical and food industries due to their several advantages such as sugar substitute that can help diabetic patients and help in preventing tooth decay problem. To produce xylitol, recently, microbial host such as E. Coli often used as it is predicted that E. Coli can produce high level of xylitol. Therefore, metabolic engineering need to be done towards E. Coli and powerful tools are needed to manipulate, simulate and analyse the E. Coli metabolic pathway. Artificial intelligence methods such as deep neural network offer an efficient and powerful approach to be used to analyse the xylitol production value and at the same time to predict which genes and pathway that give biggest effect in the process to produce xylitol in E. Coli. Results show that, with an absence of genes pgi, tkt and tala, xylitol production can be boosted up to the higher level.

Keywords*: Xylitol, Microbial host, Artificial Intelligence, Deep Leraning, Deep Neural Network*

1. Introduction

The range of bio-based chemical products in microbial host has shown an increment result [1] and has become one of the common interest among researchers. Recently, bio-based chemical products such as xylitol have been in the spotlight due to their several advantages especially both in pharmaceutical and also food industries. In the pharmaceutical industry, xylitol often used as a sugar substitute as it has a characteristic which is, it does not increase blood sugar level and insulin respond which make it highly recommended taken by diabetic patients. While in food industry, xylitol has been widely used in sugar free chewing gum and it is believed that it can help in preventing tooth decay problem. On the other hand, *E. Coli* usually has been used by researchers as a microbial hosts due to their flexible conditions which are easy to culture, grown and manipulate in a laboratory setting. Moreover, *E. Coli* is a versatile platform organism that always been used for the overproduction of non-native as well as native metabolites [2-8]. According to Cirino *et al.* [9], by engineers *E. Coli*, high level of xylitol can be produced from a mixture of xylose and glucose.

In accordance with this, a powerful tool needed to manipulate this microbial host (*E. Coli*) in order to find the best way to produce xylitol with higher value by considering all the effect that can affect the xylitol production in *E. Coli*. Therefore, advanced computational technology such as deep learning can be used as a tool to analyze and predict possible genes and pathway that can help *E. Coli* in producing more xylitol. Over these past few years, deep learning has been in the spotlight due to their advancement in technology. According to Le Chun *et al*., [10], deep learning has gained wide attention when it is already defeated other machine learning methods in the drug discovery and genomics fields.

Therefore, the aim of this paper is to manipulate, predict and analyse *E. Coli* metabolic pathway by using deep learning method. In this experiment, we used one of the deep learning methods called deep neural network (DNN) and we perform an analysis focusing on the certain pathway in *E. Coli* together with the gene deletion strategy in order to get high level of xylitol production. This paper is organized as follows. In the related work section, several related previous studies will be reviewed as a reference for this study. While in research method section, an overview of the deep learning methods and deep neural network will be discussed briefly. Besides, pathway that has been used in *E. Coli* will also be described briefly and the

Received December 3, 2016; Revised February 3, 2017; Accepted February 21, 2017

genes that involved in gene deletion strategy will be explained. In the results section, result will be discussed in three phases which are phase 1 is the result of simulation by using FBA, phase 2 is the result of analysis by using DNN and phase 3 is the overall results. In the next section which is a discussion section, reasons and explanation about the results gained in result section will be explained in more details. Lastly, conclusion section will discusses conclusion of this experiment and future work that can be done to improve this experiment.

2. Related Works

Over these past few years, deep learning has been used by many researchers in solving the prediction problems including prediction in bioinformatics field. Deep learning already helped biologist and computational biologist in solving many prediction problems from different cases such as protein sequence prediction, multiple sequence alignment prediction and many others.

Since neural network has successfully played a very important role in secondary structure predictions, Spencer *et al.,* [11] determine to develop the secondary structure predictor by using the position-specific scoring matrix generated by PSI-BLAST and deep neural network. Spencer *et al.,* [11] in their work called this proposed method as DNSS method. By the end of the experiment where they used fully independent test data set of 198 proteins to predict the secondary structure of protein, it shows that DNSS method gives 80.7% of prediction accuracy which can be categorized as high and almost accurate.

On different cases, Alipanahi *et al.,* [12] has developed an approach called DeepBind which is based on the deep learning method. It is believed that this new developed approach can discover new patterns, even when the locations of patterns within the sequences are unknown. Alipanahi *et al.* developed this approach after motivated by the fact that DNA and RNA binding proteins is important for developing models of the regulatory processes in biological systems. By using DepBind, result shows an increasing predictive power and they use its predictions to predict RNA editing and alternative splicing, discover regulatory motifs and to interpret genetic variants [13].

While for Di Lena *et al.,* [14], the importance of residue-residue contact prediction for protein structure prediction has become the greatest motivation for them to propose new novel machine learning for contact map prediction. They have implemented deep neural network architecture in their proposed new method to progressively refine and organize the prediction of contacts, integrating information over both time and space. The proposed method by Lena et al., have been shown a significant improvement to predict contact maps with an accuracy close up to 30%.

3. Research Method

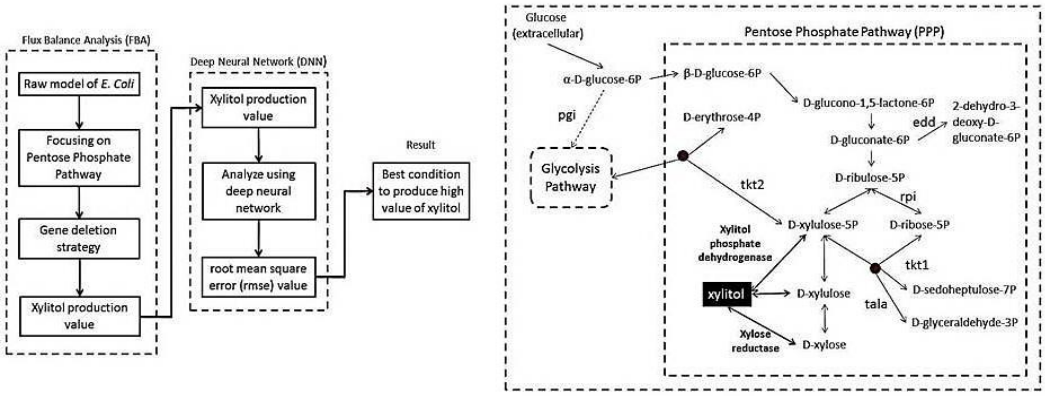
Deep learning method is one of the examples of representation learning which is a set of methods that allows automatic discovery of representations needed for classification or detection and allows a machine to detect raw data input [10]. One of the methods of deep learning that will be used to analyse dataset in this paper is called deep neural network (DNN). DNN is a feed forward, artificial neural network that contains more than one hidden layer between its inputs and its outputs [15]. In DNN, each hidden layer, *j*, typically used logistic function to map its total input from the layer below, p_j, to the scalar state, q_j that it is sent to the layer above.

$$q_j = logistic\left(p_j\right) = \frac{1}{1+e^{-p_j}}, \qquad p_j = b_j + \sum_i q_i n_{ij} \tag{1}$$

Where b_j is a bias unit for *j*, *i* is an index over units in the layer below, and $n_{ij}$is the weight on a connection to unit *j* from unit *i* in the layer below.

DNN can be discriminatively trained by utilizing backpropagation algorithm derivatives of a cost function that measures the error between predicted outputs and actual outputs produced for each training case [16]. One of the advantages of DNN is that it is flexible with a very large number of parameters when it has many hidden layers and many units per layer. With this characteristic, it makes them capable of modeling highly nonlinear and very complex

relationship between inputs and outputs. This paper will discuss on how deep neural network is used to analyse xylitol production value got from flux balance analysis (FBA) simulation.



(a) (b)

Figure 1. Flowchart work for this experiment. (a) work flow on how this experiment has been carried out starting from simulation by using FBA and then analysing by using DNN and lastly the overall result for this experiment. (b) details on pentose phosphate pathway with all the reactions and the genes that involved with gene deletion strategy

Figure 1 above shows work flow on how this experiment has been carried out. As shown in Figure 1(a) experiment has been carried out starting with the dataset has been encoded by using FBA and the simulation has been done by focusing only in the pentose phosphate pathway (PPP). While simulating the PPP, gene deletion strategy will be implemented to observe what is the effect of certain genes towards the production of xylitol in *E. Coli*. Genes that involved in this gene deletion strategy is as shown in Figure 1(b). There are glucose-6-phosphate isomerase (*pgi*), ribose-5-phosphate isomerase (*rpi*), 6-phosphogluconate dehydratase (*edd*), transketolase (*tkt*) and transaldolase (*tala*).

All these genes have been deleted in order to observe what is their effect towards the production of xylitol in *E. Coli*. These genes will be deleted by assuming that once these genes absences, all the carbon flux will go to the pathway that will produce xylitol and so that there will be less competition between the pathway to produce xylitol. This experiment has been set up with the gene deletion strategy in order to identify in which condition that *E. Coli* can produce the highest xylitol value.

4. Results and Analysis

Result for this experiment will be divided into three phases which are phase 1, result for simulation by using FBA, phase 2, result for analysing data by using DNN and lastly is phase 3, result for overall phases.

4.1. Simulation by using Flux Balance Analysis (FBA)

As mentioned before, gene deletion strategy will be implemented in this first phase. Objective of this phase is to observe in which conditions that *E. Coli* work the best to produce high level of xylitol. There are three main conditions that will be used which are (i) condition when *rpi* gene has been deleted, (ii) condition when *pgi*, *tkt* and *edd* genes have been deleted and lastly (iii) is the condition where *pgi*, *tkt* and *tala* genes have been deleted. Table 1 below will provide highest xylitol value that can be produced by each condition.

Table 1. Highest xylitol production value for each deletion condition

Deletion Condition	Xylitol Highest Value (mmol $gDW^{-1}hr^{-1}$)
RPI	153.8493023
PGIEDD	333.3333333
PGITALA	375.4333333

Result for this phase shows that with the condition where *pgi*, *tkt* and *tala* genes absence, xylitol can be produced with the highest value which is 375.4333333 mmol $gDW^{-1}hr^{-1}$ compared to the other two conditions. While condition where *rpi* gene absent shows that xylitol can only be produced up until 153.8493023 mmol $gDW^{-1}hr^{-1}$.

4.2. Analysis by using Deep Neural Network (DNN)

After finishing the simulation process in phase 1, it will give the xylitol production value as the output. By using this xylitol production value, it can be analysed by using deep neural network where we can verify and predict which condition is the best condition that can help *E.Coli* to produce high level of xylitol. By the end of this phase, deep neural network will give root mean square error value.

Root mean square error (RMSE) usually used to calculate error or differences between values predicted by a model and the values actually observed. Therefore, the small value of RMSE, the better it is as it is actually indicates that small values means less error. In this case, calculation of RMSE has been made between predicted and actual values. The predicted values comprised xylitol values when gene deletion strategy has been implemented. Whereas, the actual values are xylitol values with no modification or gene deletion strategy implemented in the model. It is also important to recall that RMSE usually has the same unit as the dependent variable which in this case the unit is mmol $gDW^{-1}hr^{-1}$. Result for this phase can be seen in the graph:

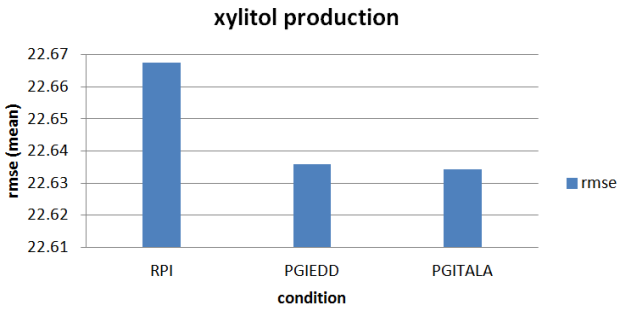


Figure 2. Result for analysing *E. Coli* metabolic pathway by using DNN

As shown in Figure 2 above, condition with the deletion of *pgi*, *tkt* and *tala* give the smallest RMSE value followed by condition with the deletion of *pgi*, *tkt* and *edd* with RMSE values 22.63435 and 22.63592 respectively. While condition with the deletion of *rpi* gene give the biggest RMSE value which is 22.66744 compared to the other two conditions.

4.3. Overall Results

In this phase, all the result obtained from phase 1 and phase 2 will be visualized in one graph so that comparison can be made. As shown in Figure 3 below, the red histogram indicates the phase 1 result while the blue histogram indicate the phase 2 result and all the values were shown in mini table below the histogram.

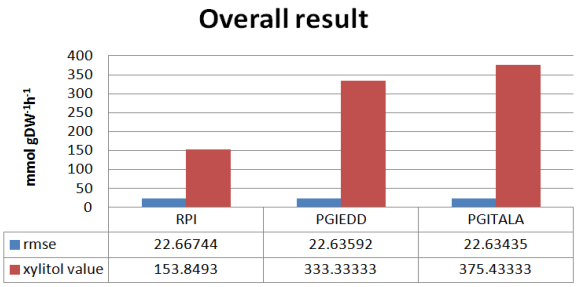


Figure 3. Overall result

As shown in Figure 3, it is clearly shows that as the xylitol production value obtained from FBA simulation phase higher, the RMSE value obtained from DNN will become lower. This is because the smaller the RMSE the better it is which means if certain condition has highest xylitol value, it will have lowest RMSE value.

5. Discussion

As mentioned before, for this experiment, simulation has been done focusing only in the Pentose Phosphate Pathway (PPP) in *E. Coli*. This is because, it is predicted that xylitol can be produced at the high level with the mixture of glucose and xylose. This is supported by Cirino *et al*., [17], where they engineered *E. Coli* to uptake xylose and at the same time they metabolizing glucose and the results shows that high levels of xylitol can be produced from a mixture of xylose and glucose. As xylose is one of the end products of PPP, which is why this experiment has been done focusing only in PPP pathway.

On the other hand, certain genes have been used and involved in the gene deletion strategy. It is a strategy where targeted genes will be deleted to observe what is their effect towards xylitol production in *E. Coli. Pgi* gene has been deleted so that all desired carbon flux that originally *E. Coli* share between glycolysis pathway and the pentose phosphate pathway can be straight go to the pentose phosphate pathway only. By doing so, competitiveness between glycolysis and the pentose phosphate pathway to get source from *pgi* can be reduced and the probability of xylitol production boost up to the high level is higher. While for the other genes such as *rpi*, *tkt*, *edd* and *tala*, the deletion has been made with the same reasons which are to reduce the competitiveness in the pathway and so that the desire carbon flux can go to the reactions that produce xylitol with an aim xylitol production can reach high value.

As shown in overall result, condition with the deletion of *pgi*, *tkt* and *tala* give the highest value of xylitol production. This is because; tala is an enzyme of the PPP where it catalyse the reversible interconversion of glyceraldehyde-3-phosphate (*g3p*), sedoheptulose-7-phosphate (*s7p*) and erythrose-4-phosphate (*e4p*). Therefore, by deleting *pgi*, *tkt* and *tala*, all reactions involved in glycolysis pathway can be cut off as well as the conversion of *g3p*, *s7p* and *e4p*. This will reduce many competitive pathways and reactions and all the desired carbon flux from all these deleted genes can go to the pathway that will produce xylitol and boost up the xylitol production value. On the other hand, condition with deletion of *rpi* can only give a small amount of xylitol because it only gets rid of one reaction in the pathway which is D-ribose-5P. Glycolysis pathway still in the pathway and it can increase the competitiveness to get carbon flux in the pathway which led to the small amount of xylitol that can be produced.

In this experiment, deep neural network has been used. Deep neural network is one of the methods from deep learning. Main reason deep learning is used in this experiment is because according to Park and Kellis [14], deep learning is well suited for genomics study. It is a powerful approach that can be used for learning complex patterns at multiple layers. This has been proved by Alipanahi *et al*., [15] when they used deep learning strategy to compute protein-nucleic acid interactions from various test datasets. They also developed new algorithm call DeepBind which is based on the deep learning method to predict binding affinity of a protein to a DNA or RNA sequence in two steps. This shows that deep learning is one of the advanced computational technologies that can be used in the bioinformatics field to make analysis, simulation and manipulation towards microbial organism. Besides, as modernization takes place, humankind tends to use technology to ease their burden and save time as well as cost-effective. By using deep neural network, desired output can be generated in a short time compared to when using wet laboratory experiment which always take longer time to give output.

6. Conclusion

In this paper, DNN is proposed to analyse xylitol production value that was gotten from the FBA simulation process. DNN will give the root mean square error which indicates the error that occurs between predicted xylitol value and actual xylitol value. The smallest value of root mean square error the better it is and by the end of the experiment we can see that condition with the deletion of *pgi*, *tkt* and *tala* genes in *E. Coli* show that xylitol can be boosted up to the high level.

For future work, it is even better if a deep neural network method can be combined with an optimization algorithm for better finding results. This is because deep neural network still suffers from several limitations that is believed can be solved by optimization algorithm such as a differential search algorithm (DSA). Moreover, according to Mohamad *et al.,* [18], hybrid methods are highly recommended compared to filter methods to produce better results.

Acknowledgements

We would like to express our appreciation to Malaysia Ministry of Higher Education for supporting this project under Fundamental Research Grant Scheme (Project Vot No. 4F481). We also would like to thank to Research Management Center, Universiti Teknologi Malaysia for managing this project.

References

[1] Ng CY, Khodayari A, Chowdhury A, Maranas CD. Advances in de novo strain design using integrated systems and synthetic biology tools. *Current opinion in chemical biology*. 2015; 28: 105-114.
[2] Akinterinwa O, Khankal R, Cirino PC. Metabolic engineering for bioproduction of sugar alcohols. *Curr Opin Biotechnol.* 2008; 19: 461-467.
[3] Boghigian BA, Pfeifer BA. Current status, strategies, and potential for the metabolic engineering of heterologous polyketides in Escherichia coli. *Biotechnol Lett*. 2008; 30: 1323-1330.
[4] Das A, Yoon SH, Lee SH, Kim JY, Oh DK, Kim SW. An update on microbial carotenoid production: application of recent metabolic engineering tools. *Appl Microbiol Biotechnol*. 2007; 77: 505-512.
[5] Jarboe LR, Grabar TB, Yomano LP, Shanmugan KT, Ingram LO. Development of ethanologenic bacteria. *Adv Biochem Eng Biotechnol*. 2007; 108: 237-261.
[6] Lee SY, Kim HU, Park JH, Park JM, Kim TY. Metabolic engineering of microorganisms: general strategies and drug production. *Drug Discov Today*. 2009; 14: 78-88.
[7] Neidhardt FC, Curtiss R. Escherichia coli and Salmonella: Cellular and Molecular Biology. 2nd edition. Washington, DC: ASM Press. 1996.
[8] Yan Y, Liao JC. Engineering metabolic systems for production of advanced fuels. *J Ind Microbiol Biotechnol.* 2009; 36: 471-479.
[9] Cirino PC, Chin JW, Ingram LO. Engineering Escherichia coli for xylitol production from glucose-xylose mixtures. *Biotechnol Bioeng*. 2006; 95: 1167-1176.
[10] LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015; 521: 436-444.
[11] Spencer M, Eickholt J, Cheng J. *A deep learning network approach to ab initio protein secondary structure prediction.* IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB). 2015; 12(1): 103-112.
[12] Alipanahi B, Delong A, Weirauch M, Frey B. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol Nature Biotechnology.* 2015; 33(8): 831-838.
[13] Park Y, Kellis M. Deep learning for regulatory genomics. *Nat Biotechnol Nature Biotechnology.* 2015; 33(8): 825-826.
[14] Di Lena P, Nagata K, Baldi P. Deep architectures for protein contact map prediction. *Bioinformatics.* 2012; 28(19): 2449-2457.
[15] Hinton G, Deng L, Yu D, Dahl GE, Mohamed AR, Jaitly N, Kingsbury B. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*. 2012; 29(6): 82-97.
[16] DE Rumelhart, GE Hinton, RJ Williams. Learning representations by back-propagating errors. *Nature*. 1986; 323(6088): 533-536.
[17] Cirino PC, Chin JW, Ingram LO. Engineering Escherichia coli for xylitol production from glucose-xylose mixtures. *Biotechnol Bioeng*. 2006; 95: 1167-1176.
[18] Mohamad MS, Omatu S, Deris S, Yoshioka M, Abdullah A, Ibrahim Z. An enhancement of binary particle swarm optimization for gene selection in classifying cancer classes. *Algorithms for Molecular Biology.* 2013; 8(1): 1.