

APPLICATION OF ARTIFICIAL NEURAL NETWORK IN BACTERIA IDENTIFICATION BASED ON BERGEY'S MANUAL : *HYDROGENOPHILACEAE* FAMILY

Amirul Hafiz Ruhaimi¹, Normadyzah Ahmad^{2*}, Hazlina Husin³, Syazana Mohamad Pauzi⁴

Faculty of Chemical Engineering, Universiti Teknologi MARA, 40450 Shah Alam , Selangor, Malaysia

¹hafiiz.ruhaimi@gmail.com

²madyzah@hotmail.com

³hazlina858@salam.uitm.edu.my

⁴syazana.pauzi@gmail.com

*Corresponding author

Abstract: Due to limitations and disadvantages of current identification process of unknown bacteria, artificial neural network can be employed as an alternative technique in bacteria identification at low cost and less time consuming. Artificial Neural Network (ANN) is a developed biological neurons principle based system in MATrix LABoratory (MATLAB) computer software that connects known input value to desired output or target value of the system and with the help of Bergey's manual as data sources of chemical and physical characteristic of selected bacteria. Therefore, unknown bacteria can be successfully determined. Analysing and data extraction from Bergey's manual require high understanding of selected microorganism in order to prevent any error or inaccurate result generated from ANN. Therefore, this study was conducted on Gram-Negative Bacillus shape bacteria under *Betaproteobacteria* Class and Order *Hydrogenophilales*. Selected bacteria under *Hydrogenophilales* order was Bacteria family of *Hydrogenophilaceae*. Levenberg Marquardt algorithm based Feed-forward backpropagation with Multilayer perceptrons type of ANN was used in the training and learning sessions of the ANN development in order to obtain high accuracy simulation results. within short period time.

Keywords: Artificial neural network, Bacteria identification, Bergey's manual , Feed-forward backpropagation

1. INTRODUCTION

Identification and classification of the microorganism can be done through several testing such as morphological characteristics, differential staining, biochemical testing, serological test, phage typing, fatty acid profile, flow cytometry, Deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) based method including Polymerase Chain Reaction (PCR), ribosomal ribonucleic acid (rRNA) sequencing, % guanine-cytosine content (G+C) comparisons[1], and DNA fingerprinting[2]. These identification methods was used by previous researchers in determining unknown bacteria and bacteria data will be extracted from Bergey's manual that was initiated by American Bacteriologists with David H. Bergey[3]. The classification and identification of the microorganism in Bergey's manual was based on morphology, biochemical characteristic and composition, physiological, ecological and genetic characteristic. Complexity of the manual will be a challenge to the first-time user in finding desired data[4].

Conventional method such as PCR, BIOLOG, Fatty Acid Methyl Esters (FAME) and others are widely used in identification of unknown bacteria nowadays. However, there are several drawbacks of these method such as long procedural time[5]; expensive due to bio-chemical material usage [6] and hazardous due to handling infectious disease caused bacteria[7]. Furthermore, results generated by conventional methods are sometimes not consistent and confusing[8].

Collaboration of the Bergey's manual as a data sources of the bacteria and the neural network system will eliminate the drawback of the conventional method in bacteria identification. Artificial neural network (ANN) was developed by using MATLAB (MATrix LABoratory) computer programme that is influenced by several parameters such as type of neural network used, algorithm used, number of hidden layers and number of training[9]. Every parameter selected is vital in obtaining successful result with high accuracy. In recent study conducted by Dzuiba & Nalepa (2012), combination of Fourier Transform Infra Red (FTIR) with ANN is the most advanced chemometric method for identifying bacteria. In this study, FTIR was used in revealing the characteristic featuring all component such as membrane protein and nucleic acid. This showed that application of the ANN can

fit in various field of study including Engineering, Technology[10], Chemical[11], Biology[12] and Mathematic[13].

In applying biological neurons principle, ANNs have recently been found in solving complex real-world problem and useful alternative in modeling technique on applying variety of data sets representing non-linear relationships[14]. ANN has been used in variety of application including classification, modeling, identification, pattern recognition, and multivariate data analysis[15].

An application of ANN in solving, classifying, and identification of complex data in various field of research have been proved by Ismail, N & Abdullah (2016) in applying ANN in prediction of electricity demand[10] , Barko',G and Hlavay, J.Z (1997) using ANN in identification of volatile organic compound[17] and by Beresford, R and Kustrin, S.A (2000) in applying ANN on pharmaceutical research[18].

This study focused on, Gram-Negative Bacillus shape bacteria under Betaproteobacteria Class and Order Hydrogenophilales. Selected bacteria under Hydrogenophilales order was Bacteria family of Hydrogenophilaceae as shown in Figure 1 below.

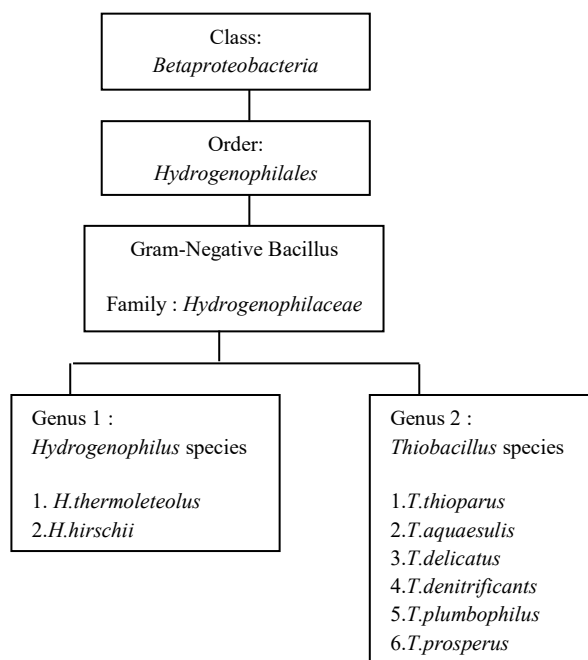


Figure 1. Taxonomical arrangement of bacterial order, family, genus and species of *Hydrogenophilales* Order

2. METHODOLOGY

A. Organism identification data

Input data of the bacteria identification was obtained and extracted from Bergey's manual and was used as a training data set in neural network. There are four sets of data which are for Family (*Hydrogenophilaceae*), Genus (*Hydrogenophilus* and *Thiobacillus*), Species in genus 1 (*Hydrngenophilus*) and species in genus 2 (*Thiobacillus*). Each of data set used in different neural network. Therefore, there are 4 ANN involved in this bacterial identification system development. Table 1 shows four (4) training data set for Family, Genus, Genus 1 (Species) and Genus 2 (Species). Coding system for the bacterial properties for *Hydrogenophilaceae* family is shown in Table 2.

Table 1. Training data set for Bacterial species in family *Hydrogenophilaceae*

Family <i>Hydrogenophilaceae</i>		Genus (<i>Hydrogenophilus</i> and <i>Thiobacillus</i>)	
Training Data Set	Target Value	Training Data Set	Target Value
[1.1, 2.1, 3.1, 4.1], [1.1, 2.2, 3.1, 4.1], [1.1, 2.1, 3.1, 4.2], [1.1, 2.2, 3.1, 4.2],	1 - <i>Hydrogenophilaceae</i>	[1.1, 2.1, 3.1]	1 - <i>Hydrogenophilus</i>
		[1.2, 2.2, 3.1]	2 - <i>Thiobacillus</i>
[1.2, 2.1, 3.1, 4.1] , [1.2, 2.2, 3.1, 4.1], [1.1, 2.3, 3.1, 4.1], [1.2, 2.3, 3.1, 4.1], [1.1, 2.1, 3.2, 4.1], [1.2, 2.1, 3.2, 4.1], etc....	0 - None	[1.2, 2.1, 3.1], [1.3, 2.1, 3.1], [1.1, 2.2, 3.1], [1.3, 2.2, 3.1], [1.1, 2.3, 3.1], [1.2, 2.3, 3.1], [1.3, 2.3, 3.1], [1.1, 2.1, 3.2], etc....	0 - None
Genus I (<i>Hydrogenophilus</i>)		Genus II (<i>Thiobacillus</i>)	
Training Data Set	Target Value	Training Data Set	Target Value
[1.1, 2.1, 3.1, 4.1]	1 – <i>H.thermoleteolus</i>	[1.1, 2.1, 3.1, 4.1, 5.1]	1 - <i>T.thioparus</i>
[1.2, 2.2, 3.2, 4.2]	2 - <i>H.hirschii</i>	[1.2, 2.1, 3.2, 4.2, 5.2]	2 - <i>T.aquaesulis</i>
[1.2, 2.1, 3.1, 4.1], [1.3, 2.1, 3.1, 4.1], [1.1, 2.2, 3.1, 4.1], [1.2, 2.2, 3.1, 4.1], [1.3, 2.2, 3.1, 4.1], [1.1, 2.3, 3.1, 4.1], [1.2, 2.3, 3.1, 4.1], [1.3, 2.3, 3.1, 4.1], etc	0 - None	[1.3, 2.1, 3.3, 4.3, 5.1]	3 - <i>T.delicatus</i>
		[1.4, 2.2, 3.4, 4.4, 5.1]	4 - <i>T.denitrificants</i>
		[1.5, 2.2, 3.5, 4.5, 5.3]	5 - <i>T.plumbophilus</i>
		[1.6, 2.2, 3.6, 4.6, 5.3]	6 - <i>T.prosperus</i>
		[1.2, 2.1, 3.1, 4.1, 5.1], [1.3, 2.1, 3.1, 4.1, 5.1], [1.4, 2.1, 3.1, 4.1, 5.1], [1.5, 2.1, 3.1, 4.1, 5.1], ...etc	0 - None

Table 2. Coding system for the bacterial properties for *Hydrogenophilaceae* family

Category	Family	Genus	Genus I	Genus II
Prefix Code No.	Bacterial Properties			
1	Gram Type	Size	Size	Size
2	Metabolism Type	Optimum Growth Temperature	Optimum Growth Temperature	Colonies Diameter
3	Shape	Colonies Color	Respiration Type	Optimum Growth Temperature
4	Motility	-	Fatty Acid Composition	Optimum Growth pH
5	-	-	-	Nitrogen Sources

B. Neural Network Software

MATLAB Version 7.11 (R2010b) (Math Work, U.S) software was used in developing Neural Network. Four separate NN model was constructed regarding to bacteria family, genus, genus 1 and genus 2 containing 4 sets of input data respectively with 1-2 numbers of hidden layer. Each hidden layer consist of various number of neurons and single output nodes was set for every neural network model.

C. Developing neural network

Several parameters were used in developing proper artificial neural network model; number of neurons in input layer, output layer and hidden layer, activation function used in each layer ,number of hidden layer , number of epoch, performance goal set (mean square error), and weights and biases of hidden layer. Several types of activation function are available in constructing NN model; linear, sigmoid, sigmoid stepwise, sigmoid symmetric, sigmoid symmetric stepwise, gaussian, gaussian symmetric and elliot[16].

Four sets of training data were used for separated NN model in training and learning sessions. Different input data lead to various number of hidden layers with various number of neurons in each layer.

As the ANN was based on biological neurons information processing, the input data was inserted and transferred into input neuron as a raw data and linked to the output neuron on output layer. The network was connected through hidden layer with a hidden neuron that lead to the generation of appropriate target value. Input and output data was inserted in form of matrices. The connection of input and hidden layer were varied during training sessions and the study showed that network performance were influenced by number of connections; less amount of hidden neurons will produce a difficulty that will slow the learning process. Nevertheless, large amount of hidden neurons will decrease the prediction abilities through overtraining and become excessively time-consuming. In optimizing the number of hidden layer and hidden neurons, there are several rules of thumb that are available in literature and the most popular approach is by trial and error [15] and etc. In order to produce desired output value, weight values are necessary to be inserted in the neural system. Weights are used as analytical tools in prediction of output value from data inserted in input layer. The generating of the weight occurs during training process; produced in order to obtain minimum error produced and achieve desired accuracy of the output value that is closed to the target value.

In this research study scope, the input neurons, as shown in Figure 2 and Tables 1 and 2, represeting bacterias such as the colonies color, fatty acid composition, nitrogen sources, etc. Meanwhile, various number of neurons in hidden layer and 1-2 number of hidden layer were specified and conducted in order to gain output value that close to the target value with smallest number of error as shown in Table 3. Mean square error (MSE) is one of the indicators in evaluating NN performance. MSE was described as:

$$mse = \frac{1}{N} \sum_{i=1}^N (e_i)^2 = \frac{1}{N} \sum_{i=1}^N (t_i - a_i)^2 \quad (1)$$

Where e is the difference between t , targeted value and simulated value, a ; N is the number of input neurons.

Suitable algorithm used in this NN development is the feed-forward backpropagation with Levenberg-Marquardt backpropagation training type. Selection of suitable type of activation function depends on layer number, input output relationship[9]. Selected activation functions are tangent sigmoidal (*tansig*) and log sigmoid (*logsig*) for hidden layer whereas pure linear (*purelin*) activation function was set for output layer.

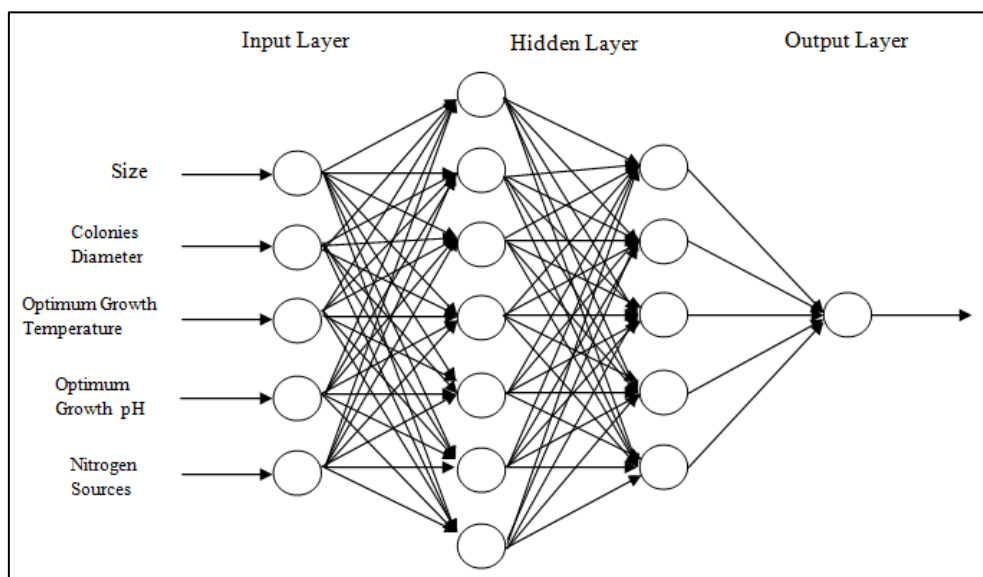


Figure 2. Neural Network architecture for identification of bacterial species under Genus II

D. Developing Bacterial Identification Programme.

Final stage of bacterial identification process via ANN is programme development. Weights and biases generated from the training session will be used in order to complete the programming and to successfully generate accurate results in identification process. This programme developed was to simplify the identification process when conducting biochemical testing without referring Bergey's manual[4].

Table 3. Specification of the neural network architecture for the training set.

Category Training Set	Family	Genus	Genus I	Genus II
Network Structure:				
Input Layer	4	3	4	5
1 st Hidden Layer	2-20	2-20	2-20	2-20
2 nd Hidden Layer	-	-	2-20	2-20
Output Layer	1	1	1	1
Training method /no. Of Epoch	LM/1000	LM/1000	LM/1000	LM/1000
Performance goal met / Performance goal set	$5.89 \times 10^{-8} / 1.00 \times 10^{-5}$	$8.72 \times 10^{-7} / 1.00 \times 10^{-6}$	$8.47 \times 10^{-7} / 1.00 \times 10^{-6}$	$6.37 \times 10^{-6} / 1.00 \times 10^{-5}$

3. RESULT AND DISCUSSIONS

Factors that influence the identification programme development are the selection of the suitable number of hidden layers and number of hidden neurons in each hidden layer. Increasing the number of epochs in training session, the complexity of the data generated and the type of activation functions used and the effects of the high number of hidden layers used. In obtaining accurate output value, several factors need to keep track are the goals (error or mse) and gradient regression. These two factors used are indicators of the output value accuracy where goal was set as a setpoint as low as 1.0×10^{-6} . Therefore, the most accurate result is the output value generated with the smallest mean square error. Similarly, selection number of neurons is vital in generating accurate result and it also influences the time consumed in training session; higher number of neurons in hidden layers, longer time needed in complete each training and besides that, it influenced the accuracy of the programming result regarding to the weights and biases generated.

Performance of training and regression of training also important in identifying accurate results. As shown in Figure 3, NN performance for this study is evaluated by referring to mean square error and number of epochs generated after training session where 74/200 number of epochs (iteration) needed in accomplishing NN goal of 5.8851×10^{-8} from the training sessions of *Hydrogenophilaceae* family.

Besides that, activation function selected also influenced in result generated. For example, *tansig* and *logsig* activation function; *logsig* most suitable for multilayer NN compare to *tansig*. For single hidden layer of complex input data, *logsig* is suggested and more accurate in terms of regression gradient. As shown in Figure 4 (a), (b), (c) and (d) for simple input data with same number of hidden layers in range of 2-20 neuron in hidden layer; both activation functions generated same maximum gradient result of training session. This shows that both *tansig* and *logsig* activation function are suitable in generating good output value in training session for simple data. However, in developing identification programme, *tansig* are the most suitable in generating accurate identification results in terms of weights and biases inserted in producing accurate data.

The weights and biases obtained from training depending on number of neurons and number of layer in NN and it were important and were used in developing identification programme. Table 4 shows the comparison of target value with the simulated value from developed ANN. The simulated output values of ANN (Table 4) showed that error (the difference between the target value and simulated outputs) of less than 2% is obtained for each category. The identification ability of ANN is very good despite the relatively large amounts of data required to develop the network. Consequently, the developed bacterial identification programme has successfully provided a means of capturing the data in a form that is easily accessible as shown in Figure 5.

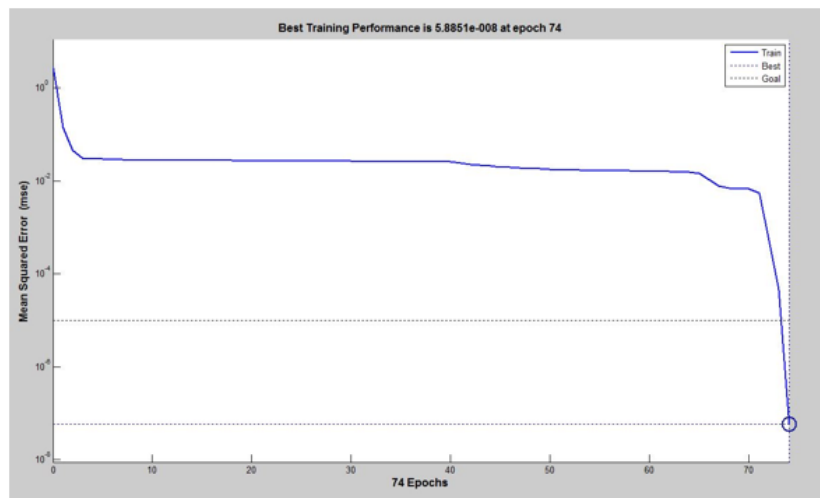


Figure 3. Training performance of family *Hydrogenophilaceae*

Figure 4. Plots of output and target values for Species in Genus II.

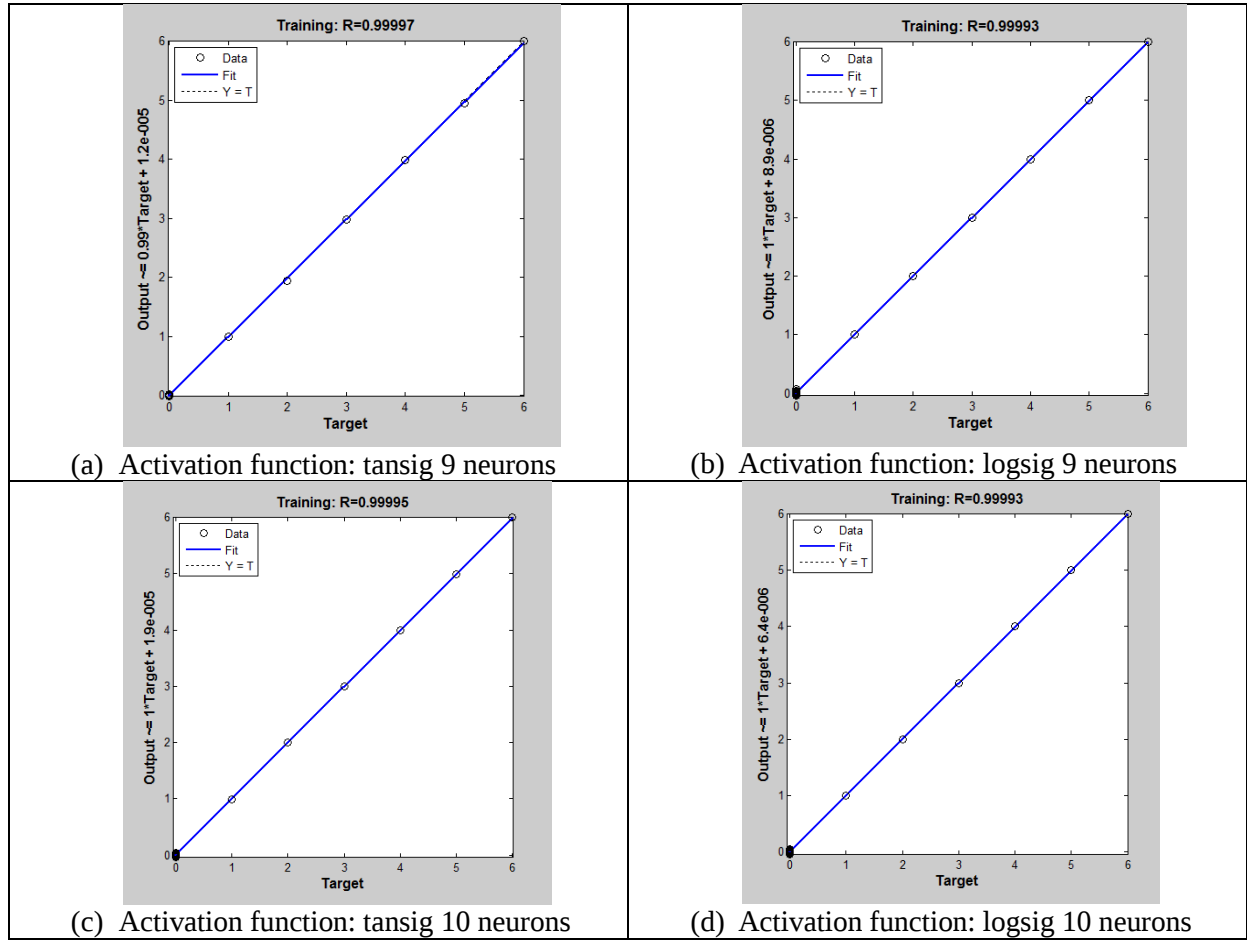


Table 4. Comparison of target values with the simulated values from the developed ANN

Name	Target Value	Output Value	MSE
Family : <i>Hydrogenophilaceae</i>	1	0.9998	5.89×10^{-8}
	1	0.9996	
	1	1.0002	
	1	0.9997	
Genus: <i>Hydrogenophilus</i>	1	1.0000	8.72×10^{-7}
<i>Thiobacillus</i>	2	1.9999	
Genus I: <i>H.themoluteolus</i>	1	0.9991	8.47×10^{-7}
<i>H.hirschii</i>	2	1.9947	
Genus II: <i>T.thioparus</i>	1	1.0010	6.37×10^{-6}
<i>T.aquaesulis</i>	2	1.9965	
<i>T. delicatus</i>	3	2.9658	
<i>T.denitrificants</i>	4	3.9847	
<i>T.plumbophilus</i>	5	5.0016	
<i>T.prosperus</i>	6	5.9239	

4. CONCLUSION

To conclude, as results presented in this study suggested that artificial neural network can advantageously employed for bacteria identification. Based on literature, basic ANN development used whether for identification, classification, prediction or other , 3 layers are required which are input layer, hidden layer and output layer. In this research scope, one of the crucial part is using Bergey's manual. As a first user, using manual can be quite confusing and time consuming due to the weakness of the manual in obtaining accurate data. By using data from the manual and using ANN as a identification and classification system, bacteria identification can be performed without any limitation that has been faced by conventional method. Eventhough there are no embeded physical or mathematical equation involved in NN modeling, the model can still relating the input to the output data[12]. Number of hidden layers used depend on the complexity of the input data. As in genus 2 NN , higher complexity of the data lead to the uses of the 2 hidden layer in obtaing accurate result. However, using 2 hidden layer affect the diffulty in developing identification programme for first time user. On the other hand, higher number of hidden layer may lead to the increasing the complexity of the generated data. Proper selection of parameters are important in obtaining accurate result.

~~~~~  
 Identification For Bacteria In Genus I (Hydrogenophilus)  
 ~~~~~

Code	Bacteria Properties
	<u>Size:</u>
1.1	0.5-0.6 x 2.0-3.0 µm
1.2	0.6-0.8 x 1.0-2.0 µm
1.3	Out of range
	<u>Optimum Growth Temperature:</u>
2.1	50-52°C
2.2	63-68 °C
2.3	Out of range
	<u>Respiration Type:</u>
3.1	Aerobic
3.2	Microaerophilic
3.3	Other than that
	<u>Fatty Acid Composition:</u>
4.1	C _{16:0} , C _{18:0}
4.2	C _{16:0} , C _{17:0} Cyclo , C _{19:0} Cyclo
4.3	Other than that

Type the unknown bacterial species coding by refering to the given table:

Size	:	1.2
Optimum Growth Temperature	:	2.2
Respiration Type	:	3.2
Fatty Acid	:	4.2

Bacteria Identified: *Hydrogenophilus hirschii*

Figure 5. Display of the Bacterial Identification Programme (*Hydrogenophilus*)

ACKNOWLEDGEMENT

The authors gratefully acknowledge the support from the Universiti Teknologi MARA (UiTM). The authors wish to thank the referee for making this a better paper.

REFERENCES

- [1] S. Mann and Y.-P. P. Chen, "Bacterial genomic G + C composition-eliciting environmental adaptation," *Genomics*, pp. 7-15, 2010.
- [2] A. Hayatdavoudi, N. Chegenizadeh, A. Christoserdov, R. Bajpai, and F. Boukadi, "Application of bacterial DNA fingerprinting in crude oil for evaluating the reservoir-Part 1," *Search and Discovery Article*, pp. 1-14, 2013.
- [3] N.A., "Bergey's Manual trust," in *History of Bergey's Manual and the Trust*, ed, 2014.
- [4] N. Ahmad, S. R. S.Abdullah and N. Anuar, "Bacterial identification using artificial neural network : A case study of *Peptococcaceae* family identification," *SIRIM Bhd*, pp. 149-154, 2013.
- [5] J. Y. Jo, J. H. Jung, M. H. Kim, H. J. Park, S. J. Kang, and Y. K. Park, "Label-free identification of individual bacteria using Fourier transform light scattering," *Optics Express*, 2015.
- [6] P. H. Rodriguez and G. A. Ramirez, "Polymerase chain reaction: Types, utilities and limitations," *InTech Open Science*, pp. 155-172, 2012.
- [7] S. Q. v. Veen, E. C. J. Claas, and E. J. Kuijper, "High-throughput identification of bacteria and yeast by matrix-assisted laser desorption ionization–time of flight mass spectrometry in conventional medical microbiology laboratories," *Journal Of Clinical Microbiology*, pp. 900-907, 2010.
- [8] S. Salsabil, "Bacterial identification based on Bergey's Manual via neural network," *National Universiti of Malaysia*, 2005.
- [9] P. A. Chandak, A. R. Lende, and J. P. Modak, "A Literature Review on Methodology & Fundamentals of Development of Mathematical Model through Simulation of Artificial Neural Network," *International Journal of Computer Application*, pp. 63-75, 2014.
- [10] N. Ismail and S. Abdullah, "Principal component regression with artificial neural network to improve prediction of electricity demand," *The International Arab Journal of Information Technology* pp. 196-202, 2016.
- [11] R. M. Balabin, R. Z. Safieva, and E. I. Lomakina, "Comparison of linear and nonlinear calibration models based on near infrared (NIR) spectroscopy data for gasoline properties prediction," *Chemometrics and Intelligent Laboratory Systems*, pp. 183-188, 2007.
- [12] N. Ahmad, S. R. S.Abdullah, N. Anuar, and H. Husin, "Bacterial identification of *Peptococcaceae* Family Based on Bergey's Manual Via Artificial Neural Network Approach," in *International Conference on Electronic Design*, Penang, 2008, p. DOI: 10.1109/ICED.2008.4786682.
- [13] B. Dziuba and B. Nalepa, "Identification of lactic acid bacteria and propionic acid bacteria using FTIR Spectroscopy and Artificial Neural Networks," *Food Technol. Biotechnol*, pp. 399-405, 2012.
- [14] J. Bourquin, H. Schmidli, P. v. Hoogevest, and H. Leuenberger, "Advantages of Artificial Neural Networks (ANNs) as alternative modelling technique for data sets showing non-linear relationships using data from a galenical study on a solid dosage form," *European Journal of Pharmaceutical Sciences*, pp. 5-16, 1998.
- [15] I. A. Basheer and M. Hajmeer, "Artificial neural networks: fundamentals, computing, design, and application," *Journal of Microbiological Methods*, pp. 3-31, 2000.

- [16] P. Sibi, S. A. Jones, and P. Siddarth, "Analysis of different activation functions using back propagation neural networks," *Journal of Theoretical and Applied Information Technology*, pp. 1264-1268, 2013.
- [17] Kustrin, S. A., & Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of Pharmaceutical and Biomedical Analysis* , 717-727.
- [18] Barko', G., & Hlavay, J. z. (1997). Application of an artificial neural network (ANN) and piezoelectric chemical sensor array for identification of volatile organic compounds. *Talanta* 44 , 2237-2245.