Normadyzah Ahmad[1]*,
S. Rozaimah S. Abdullah[2],
Nurina Anuar[2]

*[1]System Design Centre, SIRIM Berhad,
40700 Shah Alam, Malaysia*

*[2]Department of Chemical and Process Engineering,
Faculty of Engineering and Built Environment, University
Kebangsaan Malaysia, 43600 UKM Bangi, Malaysia*

(madyzah@sirim.my)

BACTERIAL IDENTIFICATION OF PEPTOCOCCACEAE FAMILY BASED ON BERGEY'S MANUAL VIA ARTIFICIAL NEURAL NETWORK APPROACH

RINGKASAN: *Satu kaedah baru mengaplikasikan sistem rangkaian neural telah diperkenalkan dalam kajian ini untuk mengenalpasti sekumpulan spesies bakteria dari famili Peptococcaceae. Untuk mencapai tujuan tersebut, terdapat beberapa kaedah kajian yang perlu dijalankan termasuklah mengumpul segala data maklumat bakteria yang terkandung dalam Manual Bergey dan melatih data tersebut dengan menggunakan sistem rangkaian neural dalam perisian program MATLAB Versi 7.0. Pembinaan rangkaian neural dalam kajian ini melibatkan beberapa parameter penting seperti bilangan neuron dalam setiap lapisan rangkaian serta penggunaan fungsi pengaktifan dan algoritma pembelajaran dalam mod sesekumpul. Secara keseluruhannya, hasil simulasi rangkaian neural telah berjaya memberi keputusan yang memiliki kejituan yang tinggi iaitu sebanyak 92 %. Hasil simulasi juga membuktikan bahawa penggunaan bilangan neuron yang tinggi di dalam lapisan tersembunyi telah berjaya menghasilkan keputusan rangkaian neural dengan kejituan yang tinggi untuk saiz data latihan yang besar. Manakala untuk saiz data latihan yang kecil dan sederhana, bilangan neuron yang rendah adalah memadai. Kejituan keputusan hasil simulasi telah dibandingkan dengan kejituan keputusan ujikaji pengenalpastian tiga spesies bakteria Gram-Positif menggunakan Microstation BIOLOG (MicroLog3 4.20.05) yang turut dijalankan dalam kajian ini. Hasilnya, penggunaan sistem rangkaian neural yang telah dibangunkan lebih mudah digunakan dalam kadar masa yang singkat.*

ABSTRACT: A novel technique employing a neural network system was used as a tool for identifying bacterial species from Peptococcaceae family. The study was initiated by extracting the bacterial properties from the Bergey's Manual and the data was then trained using neural network tool in MATLAB programme Version 7.0 (Math Works, U.S.) software. Standard parameters were involved during the development of the neural network system such as neuron numbers, activation function used and learning algorithm in a batch training mode. The results indicated that the developed neural network for the identification programme worked successfully with an accuracy of 92 %. The application of large neuron numbers in the hidden layer for a large set of training data has improved the network prediction capabilities with high accuracy. Otherwise, small neuron numbers in the hidden layer were adequate for small and medium sets of training data. The accuracy of developed neural network system was compared with a conventional identification system of bacteriology tool, Microstation BIOLOG (MicroLog3 4.20.05), by conducting bacterial identification experiments on three Gram-Positive bacterial species'. The developed neural network system used much simpler steps and required much shorter time.

Keywords: Bacterial Identification, Bergey's Manual, *Peptococcaceae*, Neural Network

INTRODUCTION

Conventional bacterial identification is normally based on the Bergey's Manual after conducting a series of biochemical tests. These tests include morphological description, ability of various substrate hydrolysis and ability of various product formation, which are useful for differentiating amongst closely related species (Kennedy and Thakur, 1993). The Bergey's Manual was developed by Dr. David Bergey in 1923 and has been widely used by microbiologists (Buchanan and Gibbons, 1974). This manual is a collection of procedures intended to aid in the identification of over 1600 bacterial species based on its structural, biochemical and chemoorganotrophic, physiological, ecological and genetic characteristics (Booth, 1978). The major drawbacks of this manual are that it is difficult to use and comprehend, time consuming as well as confusing (Yabuuchi, 1980; Holder-Franklin *et al.*, 1992), particularly by infrequent or occasional users.

However, advances in science and technology have made the identification process less lengthy and simpler to handle. Currently, systems such as API 20E, Enterotube, Oxi/ferm, BIOLOG and Polymerase Chain Reaction (PCR) are the most commonly used methods to identify microorganisms (Ibrahim and Che Omar, 2004). The sample preparations are less tedious and results obtained are almost accurate and reliable without the fuss of referring to the Bergey's Manual. All these commercialised systems give results based on the utilisation of various carbon and nitrogen sources, enzyme pre-formation, growth of an isolated colony of the test organisms and DNA fingerprint (Ibrahim and Che Omar, 2004).

Although manufacturers of these systems continuously improve the accuracy of their systems, it is worth noting that their systems may sometimes yield false results (Herman and de Ridder, 1993; Candrian, 1995; O'Connell and Garland, 2002). This is because some species from different genera might have similar biochemical properties. Furthermore, their systems are specific for certain types of bacterial species for example, API 20E are meant for identification of rod-shaped

Gram-Negative and *Enterobacteriaceae* bacterial family. BIOLOG and PCR have their own libraries and databases which do not cover the whole range of bacterial species that exist across the globe (Ibrahim and Che Omar, 2004). O'Connell and Garland (2002) have reported the dissimilar microbial community responses exhibited in two Gram-Negative plates from BIOLOG. Herman and de Ridder (1993) encountered false negative PCR results in cases where they obtained 105 or less cells per ml after enrichment for the detection of *Listeria monocytogenes* in dairy products. This could be explained by the elimination of selective plating steps plus the biochemical and serological identification in PCR assay that reduced the total analysis time compared to conventional and long identification process (Candrian, 1995).

Due to the fact that only certain species can be identified using the available systems, microbiologists still refer to Bergey's Manual for definite answers despite the controversy and the complexity of using the manual. In view of this fact, numerous techniques were devised aimed at simplifying the bacterial identification process by comparing the characteristics of an unknown organism with those of identified organisms listed in the microorganism identification tables in the manual. Among the numerous techniques summarised by Cowan (1974) are the use of paper strips, card, plastic wood and metal and later assisted by the use of computer programmes. The computer programmes consist of a series of nested IF-THEN statements. For example, if the response to degradation of tyrosine is positive, and the response to growth at 65 °C is positive and **if.... then** the organism is... However, this manual approach was very rigid and unable to deal with biological variability from a large amount of data present in the Bergey's Manual.

Therefore, it is suggested that it is reasonable to adopt an appropriate computational method to well represent the enormous amount of information the manual possessed in a complete and informational data. In order to obtain consistent results from the manual in identifying the bacteria accurately, artificial neural network (ANN) system is suggested as an appropriate method.

ANN is a computer system developed to mimic the operations of the human brain by mathematically modeling its neuro-physiological structure. ANN constitutes a supervised procedure which is able to generalise, to cope with non-linear problems, and suited for the recognition of degraded, missing, or noisy input data (Haykin, 1999; Basheer and Hajmeer, 2000). These properties exhibit ANN as an interesting tool for learning, interpolation, redundancy, and non-linear function approximation (Arab-Alibeik and Setayashi, 2005). Furthermore, ANN can also compute large amounts of data and discern complex relationships without specific rules having to be programmed into the computer (Nelson and Illingworth, 1991). ANN has numerous applications in the field of bioinformatics (Dopazo *et al.*, 1997), which include protein structure prediction (Bohr *et al.*, 1990), prediction of mycobacterial promoter sequences (Kalate *et al.*, 2003), mosquito gene sequences classification and identification (Banerjee *et al.*, 2008), DNA sequence analysis and biological pattern recognition (Blinder *et al.*, 2005; Sabbatini, 1993; Simpson *et al.*, 1992). ANN has also been employed in the identification of microorganisms using a variety of data: fatty acid profiles (Bertone *et al.*, 1996; Giacomini *et al.*, 1997; Noble *et al*., 2000; Xu *et al*., 2003), pyrolysis/mass spectrometry (Chun *et al*., 1993; Freeman*et al*., 1994; Goodacre *et al.*, 1994), whole-cell protein data (Giacomini *et al.*, 2000) and FT-IR spectra (Goodacre *et al.*, 1996; Udelhoven *et al.*, 2000; Tintelnot *et al.*, 2000).

This paper describes the development of a bacterial identification system by employing ANN focusing on anaerobic bacterial species belonging to the Peptococcaceae family (Figure 1) which is well characterised in Bergey's Manual. The microorganism's data characteristic were collected from the manual and then trained using neural network system in MATLAB programme Version 7.0 (Math Works, U.S.) software. A computerised bacterial identification programme was also developed by MATLAB, which consists of a particular question-answer and series of nested IF-ELSE statements to aid in the task of identifying a bacteriological specimen. Experimental work was conducted to confirm this study in order to justify the identification process of three bacteria using conventional commercialised biochemical test.

MATERIALS AND METHODS

Bacterial identification data

Data on bacterial identification of various species of Peptococcaceae were extracted from Bergey's Manual (Buchanan and Gibbons, 1974). The data were used as training sets in the neural network system. In order to overcome the difficulty of using the bacterial hierarchical taxonomy outlined in the manual, an approach suggested by Mullin (1970) was adopted to construct the training sets. The approach was to first select the family, followed by genus and lastly the species. In this project, six sets of training data for Peptococcaceae family (Figure 1) were established describing the species within the genera and genera within the family. The data of bacterial characterisation were encoded in numerical terms. Table 1 shows the coding system for the bacterial properties for each category.

Developing ANN model

ANN model used to identify the Peptococcaceae bacterial species were developed using MATLAB Version 7.0 programme (Math Works, U.S.) software. The ANN model for each species consisted of three layers; one input layer, one hidden layer, and one output layer and was constructed individually. Six sets of training data were established and corresponded to the taxonomy of Gram-Positive cocci bacteria as shown in Figure 1. Therefore, six individual ANN models with distinguished architectures were also developed exclusively for the corresponding six sets of training data. All the six networks established have different sizes of training data to be used with different number of neurons in the hidden layer. Specifications of the particular training set used for identification of the bacterial species are presented in Table 2.

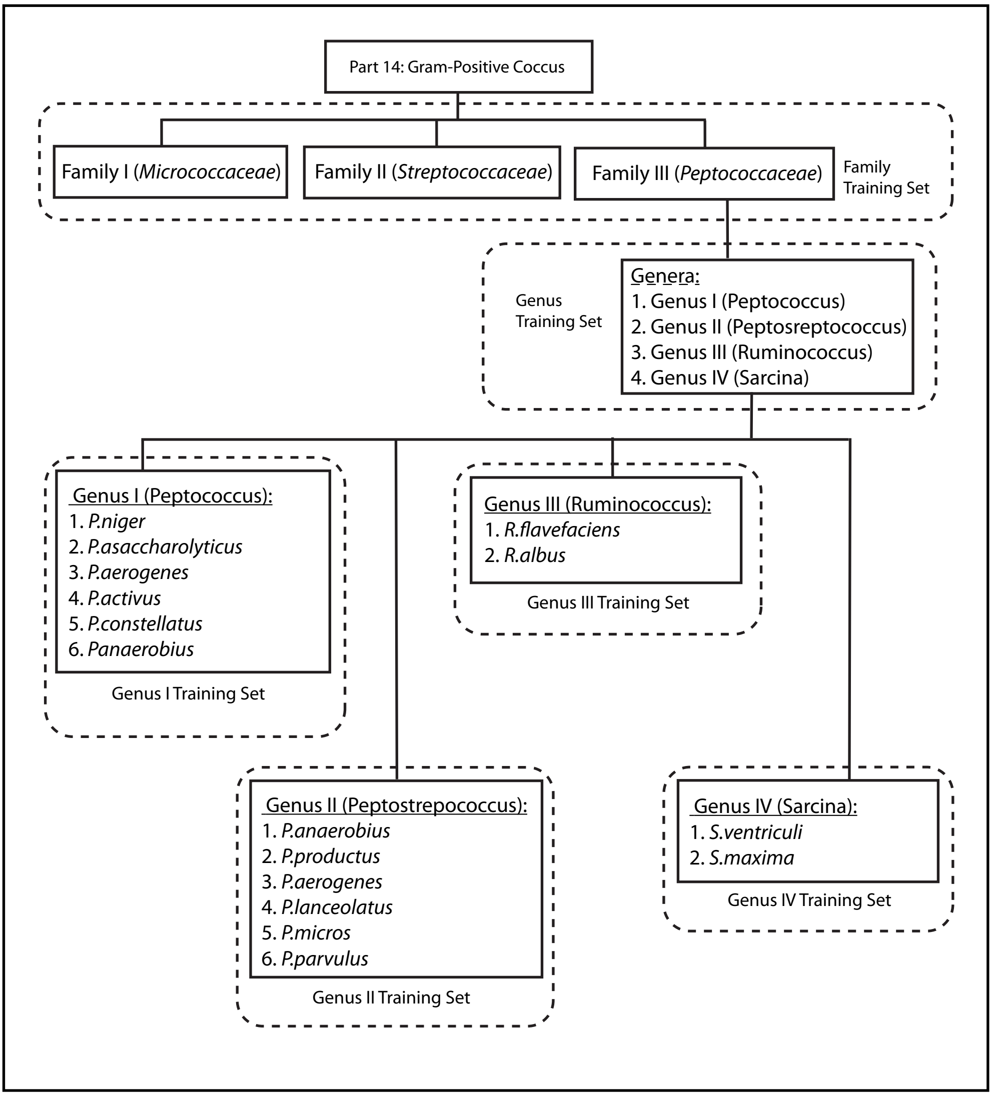


Figure 1. *Taxonomical arrangement of bacterial family, genus and species of Gram-Positive cocci group.*

Table 1. Coding system for the bacterial properties for Peptococcaceae family.

Category	**Family**	**Genus**	**Genus I**	**Genus II**	**Genus III**	**Genus IV**
Prefix Code No.	**Bacterial Properties (Input neurons)**					
1	Diameter	Growth pH	Sugar fermentation	Carbohydrate Fermentation	Cell chain numbers	Cellucose in outer layer
2	Arrangement	G + C of DNA	Visible gas	Fetid odour gas	Cellobioes fermentation	G + C of DNA
3	Motility	Carbohydrate Fermentation	Amino acids fermented	Acid produced	Iodophilic	Spore formation
4	Aerobic/ Anaerobic	Cellucose Digestion	Purines fermented	Propionate produced	G + C of DNA	-
5	-	Reaction with amino acid	Nitrate reduction	Coagulation of litmus milk acid	-	-
6	-	Peptone & amino acid as energy source	H25 produced	-	-	-

G + C – Guanine & Cytosine, DNA – Deoxyribonucleic Acid.

The input neurons, as shown in the Figure 2 and Table 2, represent the bacterial properties such as the cell chain numbers, Guanine & Cystosine content of DNA and cellobïose fermentation. The outputs are the bacterial species that corresponds to the bacterial properties outlined in the hidden layer. The output layer consists of only one neuron that sends out the result of ANN simulation.

Table 2. Specification of the neural network architecture for the training sets.

Category Training Set	**Family**	**Genus**	**Genus I**	**Genus II**	**Genus III**	**Genus IV**
Network Structure:	**Number of neurons**					
Input layer	4	6	8	5	4	3
Hidden layer	10-Feb	20-May	7-20	6-20	3-10	3-10
Output layer	1	1	1	1	1	1

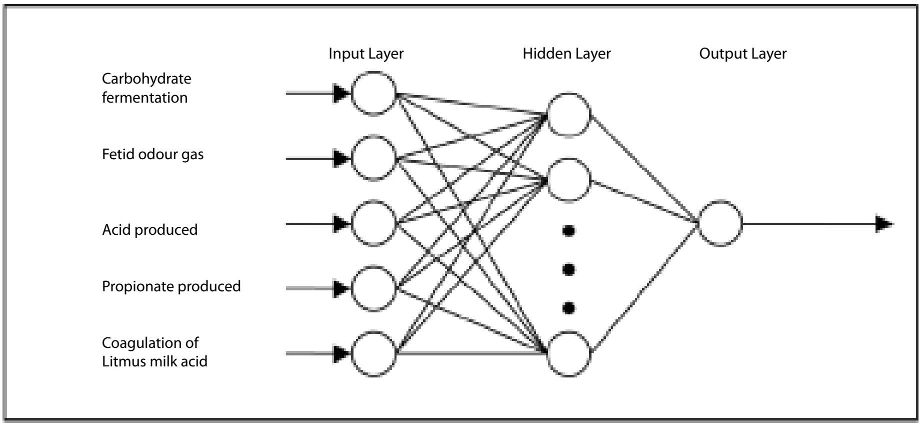


Figure 2. *Neural network architecture for identification of bacterial species in Genus II*

Developing bacterial identification programme

A computerised bacterial identification programme was developed in order to simplify the identification process should a user wish to identify the bacterial species after conducting conventional phenotypic tests. The programme was developed employing the values of connection weights and biases simulated from the successful training sessions of the developed neural networks. The programme was constructed by employing Mullin approach consisting of a question-answer system and a series of nested IF-ELSE statements. Prior to developing a complete bacterial identification programme for the whole species in Peptococcaceae family, six individual bacterial identification programmes were established corresponding to the six developed neural networks. All the six programmes were then combined and tested to ensure that the results are sufficiently comprehensive and exclusive only for the bacterial species that correspond to the programme.

Biochemical test for bacterial identification

Commercialised biochemical tests for identification of three Gram-Positive bacterial species; Staphylococcus aureus ATCC 25923, Micrococcus luteus ATCC 10240 and Aerococcus viridans ATCC 10400 were carried out using BIOLOG instrument (MicroLog3 4.20.05). The experiments were conducted to investigate the accuracy of the results that will be obtained. BIOLOG gives results based on the utilization of carbon sources in the 96 wells in a GP2 microplate kit which employs the patented reduction of tetrazolium. The pure bacterial strains were prepared according to the instruction manual which included cultivation of isolate on BIOLOG Universal Growth (BUG) Agar, preparation of a standardised liquid suspension based on turbidity and inoculation on GP2 Microplate per culture. After incubation at 37 °C for 24 hours, the Microplates were read on a BIOLOG plate reader and the strains were identified based on its database system.

RESULTS AND DISCUSSION

Simulation of developed ANNI

Coding system was established for each species and property data, for which these data were used as the input vectors for the training sets as shown in Table 1. The ANN developed for this process learned the input data given in the training

Table 3. Optimised neuron numbers in hidden layer

IDENTIFICATION OF GRAM -POSITIVE COCCI BACTERIAL FAMILY

Code	Bacteria Properties
	Diameter:
1.1	0.5 - 3.5□m
1.2	Out of range
	Arrangement:
2.1	Chains
2.2	Packets
	Motility:
3.1	Motile
3.2	Non-motile
	Growth:
4.1	Aerobic
4.2	Anaerobic

Key- in the bacterial property codes:

Diameter	: 1.1
Arrangement	: 2.1
Motility	: 3.2
Growth	: 4.2

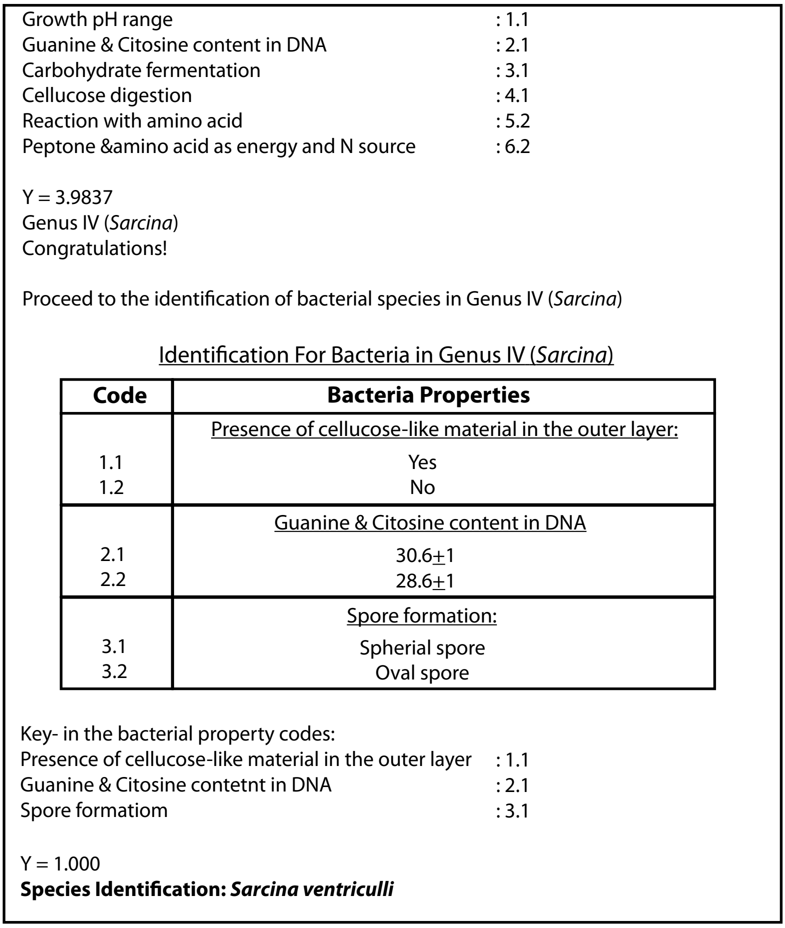
Y + 1.00035
Bacterial Family Identificatied: *Peptococcaceae*
Congratulations!

Proceed to the identification of bacterial genera in *Peptococcaceae* Family

IDENTIFICATION OF BACTERIAL GENUS IN PEPTOCOCCACEAE FAMILY

Code	Bacteria Properties
	Growth pH range:
1.1	2 - 2.5
1.2	6.0 - 8.0
	Guanine & Citosine content in DNA
2.1	28.6 - 30.6
2.2	33.5
2.3	35.7 - 36.7
2.4	39.8 - 45.4
	Carbohydrate fermentation:
3.1	Yes
3.2	90% only
	Cellucose Digestion:
4.1	Yes
4.2	No
	Reaction with amino acid:
5.1	Succinate
5.2	Ethanol
5.3	Volatile acid
	Peptone & amino acid as energy an N source:
6.1	Yes
6.2	No

Key- in the bacterial property codes:



```
Growth pH range                                   : 1.1
Guanine & Citosine content in DNA                 : 2.1
Carbohydrate fermentation                         : 3.1
Cellucose digestion                               : 4.1
Reaction with amino acid                          : 5.2
Peptone &amino acid as energy and N source        : 6.2
```

Y = 3.9837
Genus IV (*Sarcina*)
Congratulations!

Proceed to the identification of bacterial species in Genus IV (*Sarcina*)

Identification For Bacteria in Genus IV (*Sarcina*)

Code	Bacteria Properties
	Presence of cellucose-like material in the outer layer:
1.1	Yes
1.2	No
	Guanine & Citosine content in DNA
2.1	30.6±1
2.2	28.6±1
	Spore formation:
3.1	Spherial spore
3.2	Oval spore

```
Key- in the bacterial property codes:
Presence of cellucose-like material in the outer layer   : 1.1
Guanine & Citosine contetnt in DNA                        : 2.1
Spore formatiom                                           : 3.1
```

Y = 1.000
Species Identification: *Sarcina ventriculli*

Figure 4. *Simulated output from the developed bacterial identification programme.*

This is the main reason why microbiologists still need to revert and conduct the conventional phenotypic procedures with lengthy series of biochemical tests and finally refer to Bergey's Manual for resolving the mystery of unknown microorganism, especially when dealing with mixed isolates.

Table 4. Bacterial identification results using BIOLOG.

CONCLUSION

This study showed that ANN provided a very accurate identification of the bacterial species over an enormous amount of input data considered. ANN was able to use the data in Bergey's Manual to predict the identity of the microorganism. Conventionally, accurate bacterial identification based on the manual was relatively difficult to obtain which led to many speculations about the manual's weaknesses. The identification ability of the ANN is very good in spite of the relatively large amounts of data required to establish the network. The results indicated that the developed neural network worked successfully with high accuracy. Bacterial species from Peptococcaceae family is just a model used in this case study in which this group is well characterised in Bergey's Manual. In the future, the ANN model should be investigated to identify bacteria from other families in Bergey's Manual.

ACKNOWLEDGEMENTREFERENCESArab-Alibeiki, H and Setayash S. Adaptive control of a PWR core

The first author is grateful to SIRIM Berhad and Ministry of Science, Technology and Innovation of Malaysia for financial support through the given study grant.

entals, computing, design

sing & Chemical E

996). Aut
ty acid

B.
339 -

B.

.E., Mage
eural network analysis of

Giacomini, M., Ruggerio, C., Calegari, L., Bertone, S. (2000). Artificial neural network based identification of environmental bacteria by gas31 chromatographic and electrophoretic data. *J. Microbiol. Methods* **43**: pp 45 – 54.

Goodacre, R., Timmins, E.M., Rooney, P.J., Rowland, J.J., Kell, D.B. (1996). Rapid identification of Streptococcus and 1 Enterococcus species using diffuse reflectance-absorbance Fourier Transform Infra Red spectroscopy and artificial neural networks. *FEMS – Microbiol. Lett.* **140**: pp 233 – 239.

Goodacre, R., Neal, M.J., Kell, D.B., Greenham, L.W., Noble, W.C., Harvey, R.G. (1994). Rapid identification using pyrolysis mass spectrometry and artificial neural networks of Propionibacterium acnes isolated from dogs. *J. Appl. Bacteriol*. **76**: pp 124 – 134.

Haykin, S. (1999). Neural Networks: A Comprehensive Foundation. Prentice Hall International, London, UK.

Herman, L. and de Ridder, H. (1993). Comparison of different methods for detection of Listeria monocytogenes in dairy products. *Milchwissenschaft* **48**: pp 684 – 686.

Holder-Franklin, M.A., Thorpe, A., Wuest, L. (1992). Evaluation of tests employed in the numerical taxonomy of river bacteria. *J. Microbiological Methods* **15**: pp 263 – 277.

Kalate, R. N., Tambe, S. S., Kulkarni, B. D. (2003). Artificial neural networks for prediction of mycobacterial promoter sequences. *J. Computational Biology and Chemistry* **27**: pp 555 – 564.

Kennedy, M. J. and Thakur, M.S. (1993). The use of neural networks to aid in microorganism identification: a case study of Haemophilus species identification. *J. Antonie van Leeuwenhoek* **63**: pp 35 – 38.

Kreig, N.R. and Holt, J.B. (1984). Bergey's Manual of Systematic Bacteriology. Williams and Wilkins, Baltimore.

M.A. Kamaruzzaman, N. Anuar, S.R.S. Abdullah, N.H. Tan Kofli. Problem in determination of local bacterial community from wastewater treatment plants, Engineering Postgraduate Conference 2009 (EPC 2009), 20 – 21 Oct 2009, Kuala Lumpur

Mullin, J.K. (1970). COQAB: A computer optimized question asker for bacteriological specimen identification. *Mathematical Sciences* **6**: pp 55 – 66.

Negnevitsky, N. (2005). Artificial intelligence: A guide to intelligent system. 2nd Ed. Pearson Education Ltd., Essex. pp 165 – 217.

Nelson, M.M. and Illingworth, W.T. (1991). A Practical Guide to Neural Nets. Addison-Wesley Publishing Company Inc., Canada. pp 16

Noble, P.A., Almeida, J.S., Lovell, C.R. (2000). Application of neural computing methods for interpreting phospholipid fatty acid profiles of natural microbial communities. *J. Appl. Environ. Microbiol*. **66**: pp 694 – 699.

O'Connell, S.P. and Garland, J.L. (2002). Dissimilar response of microbial communities in Biolog GN and GN2 plates. *J. Soil and Biochemistry* **34**: pp 413 - 416.

Rumelhart, D.E., Hinton, G.E., Williams, R.J. (1986). Learning internal representation by error back propagation. In: D.E. Rumelhart, J.L. McCleland, (Eds.), Parallel Distributed Processing. M.I.T. Press, Cambridge, Massachussets. pp 318 -362.

Sabbatini, R.M.E. (1993). Neural networks for classification and pattern recognition of biological signals. In: Proceedings of the 15th Annual International Conference of the IEEE. pp 265 – 266.

Simpson, R.G., Williams, R., Ellis, R.E., Culverhouse, P.F. (1992). Biological pattern recognition of neural networks. *Mar. Ecol. Prog. Ser.* **79**: pp 303 – 308.

Tintelnot, K., Haasr, G., Seibold, M., Berhmann, F., Staemmler, M., Franz, T., Naumann, D. (2000). Evaluation of phenotypic markers for selection and identification of *Candida dubliniensis. J. Clin. Microbiol.* **38**: pp 1599 – 1608.

Trzaska, J. and Dobrzanski, L.A. (2007). Modelling of CCT diagrams for engineering and constructional steels. *J. Materials Processing Technology* **192 – 193**: pp 504 – 510.

Udelhoven, T., Naumann, D., Schmitt, J. (2000). Development of a hierarchical classification system with artificial neural networks and FT-IR spectra for the identification of bacteria. *Appl. Spectrosc.* **54**: pp 1471 – 1479.

Xu, M., Voorhees, K.J., Hadfield, T.L. (2003). Repeatability and pattern recognition of bacterial fatty acid profiles generated by direct mass spectrometric analysis of insitu thermal hyrolysis/methylation of whole cell, *Talanta*. **59**: pp 577 – 589.

Yabuuchi, E. (1980). Legitimacy of the names of subspecies of Campylobacter Fetus Proposed by Véron and Chatelain, Annales de l'Institut Pasteur. *Microbiologie* **134**: pp 3 – 8.