

# Identification of dental bacteria using statistical and neural approaches

**Chaw Koh Yong, Choo Min Lim, Mark D. Plumbley, David Beighton, and Ross Davidson**

Proceedings of the 9th International Conference on Neural Information Processing (ICONIP'02), 18-22 Nov. 2002. Vol.2, Pages 606-610.

## ***Abstract***

This paper is devoted to enhancing rapid decision-making and identification of lactobacilli from dental plaque using statistical and neural network methods. Current techniques of identification such as clustering and principal component analysis are discussed with respect to the field of bacterial taxonomy. Decision-making using multilayer perceptron neural network and Kohonen self-organizing feature map is highlighted. Simulation work and corresponding results are presented with main emphasis on neural network convergence and identification capability using resubstitution, leave-one-out and cross validation techniques. Rapid analyses on two separate sets of bacterial data from dental plaque revealed accuracy of more than 90% in the identification process. The risk of misdiagnosis was estimated at 14% worst case. Test with unknown strains yields close correlation to cluster dendograms. The use of the AXEON VindAX simulator indicated close correlations of the results. The paper concludes that artificial neural networks are suitable for use in the rapid identification of dental bacteria.

©2002 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.



## IDENTIFICATION OF DENTAL BACTERIA USING STATISTICAL AND NEURAL APPROACHES

Chaw Koh Yong and Choo Min Lim  
Ngee Ann Polytechnic

School of Engineering, Electronic and Computer Engineering Division  
535 Clementi Road, Singapore 599489.

Mark Plumbley and David Beighton  
King's College, University of London  
The Strand, London WC2R 2LS  
United Kingdom.

Ross Davidson  
University of Aberdeen, Department of Engineering  
Fraser Noble Building, King's College  
Aberdeen AB24 3UF  
United Kingdom.

### ABSTRACT

This paper is devoted to enhancing rapid decision-making and identification of lactobacilli from dental plaque using statistical and neural network methods. Current techniques of identification such as clustering and principal component analysis are discussed with respect to the field of bacterial taxonomy. Decision-making using multilayer perceptron neural network and Kohonen self-organizing feature map is highlighted. Simulation work and corresponding results are presented with main emphasis on neural network convergence and identification capability using resubstitution, leave-one-out and cross validation techniques. Rapid analyses on 2 separate sets of bacterial data from dental plaque revealed accuracy of more than 90% in the identification process. The risk of misdiagnosis was estimated at 14% worst case. Test with unknown strains yields close correlation to cluster dendograms. The use of the AXEON VindAX simulator indicated close correlations of the results. The paper concludes that artificial neural networks are suitable for use in the rapid identification of dental bacteria.

### 1. INTRODUCTION

For decades, dental researchers have been waging an all-out war on microbes that live in the human mouth. These microbes are made up of hundreds of species of fungi, protozoa, viruses, intracellular parasites and above all, bacteria. There are about 200 to 500 species of dental bacteria i.e. dental plaque, of which only a minority have been named [1]. Conventional diagnostics microbiology and "fingerprinting" [2] of bacterial systems are inherently slow due to the need to grow the organism and carry out diagnostic tests. A desirable objective is to perform this characterization of bacteria as quickly and accurately as possible. Such diagnosis and rapid decision making will lead to improved hospital treatment and reduction in costs.

New methodologies such as computerized databases, statistical procedures and neural networks have rendered

bacterial identification quick and decisive. Goodfellow [2] has introduced the concepts of bacterial systematics, which is the study of the kinds and diversity of organisms and their inter-relationship among them. Magee [2] has reported the use of whole-cell pyrolysis for organism "fingerprinting". Pyrolysis is the thermal degradation of complex material into pyrolysate in an inert atmosphere or a vacuum. Goodacre [3 - 7] and Freeman [8 - 9] have carried out detailed analysis and discrimination of closely related microbial strains using pyrolysis method and neural networks. Electrophoresis method of bacteria identification has been studied by Vohradsky [12]. This technique is used to detect and identify proteins and enzymes in extracts of micro-organisms. The proteins extracted from a bacteria strain are separated by virtue of their dissimilar migration rates in an electric field producing a banding pattern [13]. Bacterial strains grown under identical conditions produce constant banding patterns that are used as "fingerprints" of the strains being investigated.

The importance of using artificial neural networks for rapid analysis and identification of bacteria has gained widespread attention. This work aims to verify and enhanced some of the research work done and to show that identification of dental bacteria using neural networks can be done quickly and accurately. The studies focus on the implementation of a robust clinical diagnostic system based on the investigation of the multilayer perceptron network using error back-propagation algorithm, unsupervised Kohonen self-organizing map, principal component analysis technique and hierarchical clustering method.

### 2. BACTERIAL TAXONOMY

Taxonomy is regarded as the theory of classification in the life sciences. The principal aim of classification is to provide a scheme to identify unknown organisms into groups. Classification of micro-organisms usually requires knowledge of their characteristics. This knowledge is acquired using experimental and observational techniques. Identification deals with the process of allocating a new specimen to a current and previously defined group. It involves determining the relevant features of the unknown strain and the matching of these with

an appropriate database containing information on known species. The determination of taxonomic structure can be done using clustering and principal component analyses.

Cluster analysis [14-17] is a technique used for combining observations into groups or clusters. Each cluster is homogenous with respect to certain characteristics. The first step in cluster analysis is to select a measure of similarity. The most common one is the Euclidean distance. A hierarchical or non-hierarchical technique of clustering maybe used.

Principal components analysis [15] is a statistical technique that linearly transforms an original set of variables into substantially smaller set of uncorrelated variables that represents most of the information in the original set of variables. A small set of uncorrelated variables is much easier to understand and use in further analyses than a larger set of correlated variables. In algebraic terms, the first principal component,  $y_1$  is a linear combination of  $x_1, x_2, \dots, x_n$  such that the variance  $y_1$  is maximized given the constraint that the sum of the squared weights is equal to one. The analysis finds the optimal weight vector and the associated variance  $\lambda_1$ . The second principal component  $y_2$ , involves finding a second vector such that the variance is maximized subject to the constraints that it is uncorrelated with the first component. The variances of the principal components get smaller as successive components are extracted.

### 3. ARTIFICIAL NEURAL NETWORKS

An artificial neural network is a structure composed of a number of interconnected units called artificial neurons. Each neuron has an input-output characteristic and implements a local computation or function. The output of any neuron is determined by its input-output characteristics, its interconnection to other neurons and possibly external inputs. The network usually develops an overall functionality through one or more forms of training.

Rosenblatt (1958) defines a perceptron to be a machine that learns using examples to assign input vectors to different classes. The multilayer perceptron network is composed of a hierarchy of processing units organized in a series of two or more mutually exclusive sets of neurons or layers. Multilayer perceptron networks [18 - 21] are able to solve many difficult and diverse problems by training them in a supervised manner with the error back-propagation algorithm. The training strategy employed is the gradient descent in the form of the delta rule or the generalized delta rule. The learning signal  $\delta$  is given as

$$\delta = (d - o) f'(u) \quad (3.1)$$

and the network weight correction at time  $t+1$  is given by

$$w_{ij}(t+1) = w_{ij}(t) + \epsilon d_i o_j \quad (3.2)$$

where  $d$  is the desired target,  $o = f(u)$  is the actual response,  $f'(u)$  is the derivative of the activation function  $f$  computed for  $u = w_{ij}x_j$  and  $\epsilon$  is the learning rate. This rule can be readily derived from condition of least squared error, by calculating the gradient vector with respect to the connection weight [19]. This rule can be generalized for the hidden layers of the multilayer perceptron network.

The Kohonen self-organizing map combines a competitive learning principle [22] with topological structuring of nodes such that adjacent node tends to have similar weight vector around a "winner" node. The learning algorithm ensures that the winning neuron as well as its neighbor moves towards a sample input vector. The first layer of a self-organizing map is the input layer. Each node in the second layer is the winner for

all input vectors in a region of input space. It also contains output nodes with intra-layer connections set up according to a predetermined topology such as a grid in which each node has four neighbors. Each  $k$ 'th output node has connections from all inputs, with connection strengths given by the  $n$ -dimensional vector  $w_k = \{w_{k1}, w_{k2}, \dots, w_{kn}\}$ . Their value changes during the learning process with each weight vector moving towards the centroid of some subset of input patterns.

The training algorithm updates the winner node and also nodes in its topological vicinity when an input pattern  $x$ , is presented. The neighborhood  $N_k(t)$  contains nodes that are within a topological distance  $D(t)$  from the node  $k$  at time  $t$  where  $D(t)$  decreases with time. At each training iteration, a distance measure  $d_m = x \cdot w_m(t)$  in the network is computed. A winner is selected for a minimum  $d_m$ . Weights in the neighborhood  $N_k$  containing this winner is change at time  $t$  at a rate  $\gamma(t)$  which decrease with time. The weight change rule [22] for the self-organizing map is given as

$$w_m(t+1) = w_m(t) + \gamma(t) (x - w_m(t)) \quad \text{if } m \in N_k(t) \quad (3.3)$$

$$= w_m(t) \quad \text{if } m \notin N_k(t)$$

### 4. DATA PREPARATION AND ANALYSES

Samples of bacteria cells were washed and suspended in special buffers. A heating process at 100°C is used to disrupt the cells. High-speed centrifugation was used to separate the cell proteins from the cell walls and membranes. Electrophoresis separation of these proteins in an electric field was applied to obtain the electrophoresis patterns of the bacteria strains. This method allows the separation of these proteins on the basis of molecular weight. This information is arranged in gel strips (Figure 4.1) each displaying a particular pattern for a typical strain. The positions of the bands in the strip differ with the molecular weight of the extracted proteins of the bacteria from electrophoresis. High molecular weight bands are usually better stacked [13] resulting in sharp, thin and dark bands. Low molecular weight bands have migrated a longer distance, causing broad and weak bands. The information on the gel strips was then scanned into GelCompar software [13] for numerical processing via a frame grabber in a machine vision system.

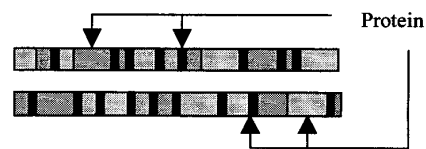
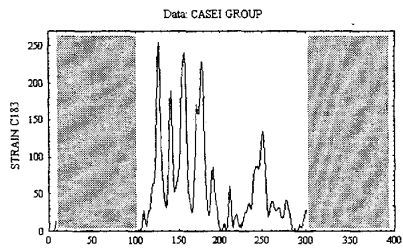
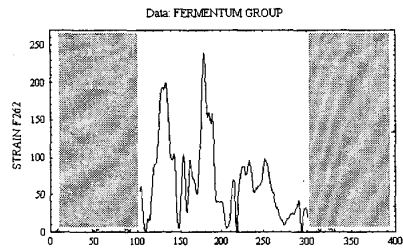


Figure 4.1 Gel strips showing two patterns

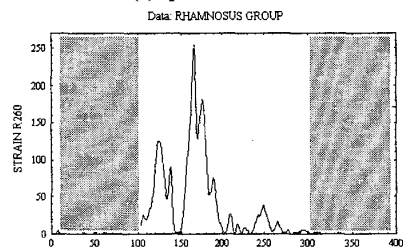
The intensity information on the gel strips was extracted into a code having a resolution of 256 grey-levels. Each code represents a bacteria strain of 400 points. Some of these codes are shown graphically in Figure 4.2. Only points from 101 to 300 were used as these contain the most essential information regarding the biological make-up of the bacteria strains. Bacterial taxonomy using hierarchical clustering via unweighted pair-group average method was performed yielding the dendrogram results shown in Figure 4.3.



(b) species: casei



(c) species: fermentum



(d) species: rhamnosus

Figure 4.2 Lactobacilli information

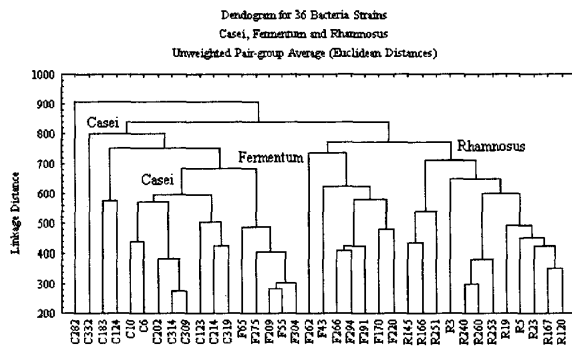


Figure 4.3 Dendrogram for bacteria strain samples.

## 5. RESULTS AND DISCUSSIONS

The primary objective of this work is to implement a robust clinical diagnostic system for enhancing rapid decision-making and identification of 3 species of lactobacilli. Two different groups of bacteria strains were used in this work. Group A contained 36 strains while group B contained 132 strains. Each group comprised approximately the same number of strains of casei, fermentum and rhamnosus species.

## 5.1 Multilayer Perceptron Network

The initial work began by selecting suitable network architecture and network parameters for the neural network. To identify the 3 species of lactobacilli, the neural network was designed with three outputs. The binary codes for classifying each bacterium into its own group are  $\{([1\ 0\ 0] \in \text{casei}), ([0\ 1\ 0] \in \text{fermentum}), ([0\ 0\ 1] \in \text{rhamnosus})\}$ .

### 5.1.1 Group A: Set of 36 Strains

A single hidden layer with sigmoid activation function was used. The number of hidden neurons and learning rate were varied to check how long training would take. Two query techniques were used namely resubstitution and leave-one-out method. The resubstitution method gave good results confirming similar works by Goodacre [6 - 7] and Chun [10 - 11]. In the classification process, any outputs greater than 0.5 is considered a "1" and less than 0.5 a "0".

The percentage of misclassifications or error rates was investigated for both the resubstitution and leave-one-out method. The resubstitution method gave zero misclassification error. The leave-one-out method revealed on average 15% misclassification error (Table 5.1). The results clearly indicated that resubstitution method must be used with caution as it gives an over-optimistic error rate result of misclassification. This is expected as the network is tested with the same input patterns that it was trained on. On the other hand, the leave-one-out technique reveals that a good point to stop training is at an error of 0.03 or 0.01 and 10 hidden neurons for the network is a good starting point. Cross-validation [19] is a popular technique used for checking the generalization capability of the network. Simulations were performed using 30 samples for the training set and 6 samples for the validation set. Test error is observed to decrease alongside with the training error indicating good generalization.

Table 5.1 Error rates using various network architecture with learning rate = 0.2 and momentum = 0.9.

Stopping at Training Error of:	MISCLASSIFICATION	
	Resubstitution	Leave-one-out
0.1	0 %	8/36 = 22 %
0.05	0 %	6/36 = 17 %
0.03	0 %	4/36 = 11 %
0.01	0 %	4/36 = 11 %
<b>Misclassified</b>	0 %	15 %
Number of Hidden Neurons	MISCLASSIFICATION	
	Resubstitution	Leave-one-out
20	0 %	5/36 = 14 %
10	0 %	4/36 = 11 %
5	0 %	7/36 = 19 %
<b>Misclassified</b>	0 %	15 %

Further analysis was done by extracting the first 10 principal components from the bacteria data. The plot of the first two components scores (Figure 5.1) revealed three distinct clusters. Simulation was performed with different number of principal components for the input to the neural network. The lowest test errors were recorded for 5 principal components at 5000 epochs using two validation sets.



A similar 6 fold cross-validation with query was used to test the effectiveness of the trained Kohonen map. Learning was performed using a training set of 108 samples. Numerous validation sets of 24 different bacteria strains were used for testing the trained map. The overall assessment yields an accuracy of 94%.

For a robust clinical system it must give a low risk of misdiagnosis. The system should give a definite yes to correct identification of test strains. It should prompt for expert advice to test strains located at cluster boundaries and those that are unsure of. The average risk factor computed from the above results is 14% (Table 5.4). This is a worst-case condition where only strains located within a cluster is identified as good.

Table 5.4 Misdiagnosis risk factor.

RUN	SYSTEM RESPONSE		RISK OF MISDIAGNOSIS
	YES: Pass	NO: Seek Advice	
A	19/24	5/24	21 %
B	24/24	0/24	0 %
C	23/24	1/24	4 %
D	18/24	6/24	25 %
E	19/24	5/24	21 %
F	21/24	3/24	13 %
Average =			14 %

## 6. AXEON VINDAX LEARNING PROCESSOR

The VindAX Learning Processor manufactured by AXEON Technology [23] is a neural network comprising of a parallel array of processing elements that take the form of a 64, 256 or 1024 interconnected RISC type processor. The processor uses a learning algorithm that is similar to the Kohonen's learning rule. The solutions that have been obtained using the simulator can be developed further into hardware using the VindAX Learning Processor card.

During training, the simulator created topological maps that showed clusters of processing elements that had formed. Labeling of clusters was then performed with color codes added in. The results obtained were similar to those obtained earlier showing good correlations. Hence, the results from the simulator could be further developed and converted into a specific hardware platform supported by the AXEON card. This means that the identification of dental bacteria could be performed more rapidly and easily with a dedicated workstation. Further work will involve the development of the actual application using the hardware platform for the identification of bacteria as a whole.

## 7. CONCLUSION

The current work on the rapid decision-making and identification of bacteria from dental plaque using neural networks has been successfully carried out using the prescribed methodologies and procedures. Practical understanding of the underlying concepts as applied to the results was achieved. Further work on bacteria identification will be carried out using neural hardware processor board to achieve even higher accuracy.

## 8. REFERENCES

[1] G. Hamilton, "Open wide, we're going to explore", *New Scientist*, 2125, p33-37, 1998.

[2] M. Goodfellow, *Handbook of New Bacterial System*, Academic Press, 1993.

[3] R. Goodacre and D. Kell, Pyrolysis mass spectroscopy and its application in biotechnology, *Current Opinion in Biotechnology*, 7, p20-28, 1996.

[4] R. Goodacre, Characterisation and quantification of microbial systems using pyrolysis mass spectrometry: Introducing neural network to analytical pyrolysis, *Microbiology Europe* 2(2), p16-22, 1994.

[5] R. Goodacre, M. Neal and D. Kell, Quantitative analysis of multivariate data using artificial neural network: A tutorial review and applications to the deconvolution of pyrolysis mass spectra, *Zentralblatt fur Bacteriologie*, 1995.

[6] R. Goodacre, S. Hiom, S. Cheeseman and D. Murdoch, Identification and discrimination of oral asaccharolytic Eubacterium spp. by pyrolysis mass spectrometry and artificial neural network, *Current Microbiology*, 32, p77-84, 1996.

[7] G. Goodacre, M. Neal and D. Kell, Rapid identification using pyrolysis mass spectrometry and artificial neural network of propionibacterium acnes isolated from dogs, *Journal of Applied Bacteriology* 76, p124-134, 1994.

[8] G. Freeman and R. Goodacre, Rapid identification of species within the mycobacterium tuberculosis complex by artificial neural network analysis of pyrolysis mass spectra, *Journal of Medical Microbiology*, 40, p170-173, 1994.

[9] G. Freeman and R. Goodacre, Pyrolysis mass spectrometry for the rapid epidemiological typing of clinically significant bacterial pathogens, *Journal of Medical Microbiology*, 32, p283, 1990.

[10] J. Chun and M. Goodfellow, Rapid identification of streptomycetes by artificial neural network analysis of pyrolysis mass spectra, *FEMs Microbiology Letters*, 114, p115-2120, 1993.

[11] J. Chun and M. Goodfellow, Artificial neural network analysis of pyrolysis mass spectrometric data in the identification of streptomyces strains, *FEMs Microbiology Letters*, 107, p321-326, 1993.

[12] J. Vohradsky, Adaptive classification of two-dimensional gel electrophoresis spot patterns by neural networks and cluster analysis, *Electrophoresis*, v18, 15, p2749-2754, 1997.

[13] Applied Maths BVBA, *GelCompar: Comparative Analysis of Electrophoresis Patterns*, manual, 1996.

[14] B. Austin and F. Riest, *Modern Bacterial Taxonomy*, Van Nostrand Reinhold (UK), 1986.

[15] S. Sharma, *Applied Multivariate Techniques*, John Wiley and Sons Inc., 1996.

[16] P.H.A. Sneath, *Classification of Micro-organisms*, John Wiley and Sons, Inc., 1978.

[17] B. Everitt, *Cluster Analysis*, Heinemann Educational Books, 1974

[18] R.J. Schalkoff, *Artificial Neural Networks*, McGraw-Hill Book Co., 1997.

[19] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2<sup>nd</sup> Edition, Prentice Hall Inc., 1999.

[20] J. Hertz, A. Krogh, and R.G. Palmer, *Introduction to the Theory of Neural Computation*, Addison-Wesley Publishing Company, 1991.

[21] C.K. Yong and P.I. Rockett, *A Modular High Order Neural Network for Vision Inspection*, IEEE International Conference on Automation, Robotics and Computer Vision (Singapore), 1994.

[22] T. Kohonen, *Self-organising Maps*, Springer-Verlag, 1997.

[23] Axion Technology, VindAX System Development Manual, 2001.