

# Identification of Dental Bacteria using Multilayer Perceptron Neural Networks

C. K. Yong<sup>1</sup>, M. D. Plumbley<sup>2\*</sup> and D. Beighton<sup>3</sup>

<sup>1</sup> - Department of Mathematics, King's College London, Strand, London WC2R 2LS, UK

<sup>2</sup> - Department of Electronic Engineering, King's College London, Strand, London WC2R 2LS, UK

<sup>3</sup> - Department of Oral Microbiology, GKT Dental Institute, Caldecot Road, London SE5 9RW, UK

\* - Corresponding author

## Summary

Strains of dental bacteria (Lactobacilli) were classified into three groups (casei, fermentun and rhamnosus) using gel electrophoresis 'fingerprints' and multilayer perceptrons (MLPs). Comparison with traditional statistical visualisation methods such as hierarchical cluster dendrograms and principal components analysis shows that there is some correlation between extreme strains on these diagrams and those that are difficult to classify by the MLP. We discuss the types of failure and suggest that for this problem a 'safety-margin' method gives a good trade-off of undetected misclassifications against unclassified strains.

## Introduction

Bacteria, for example in the human mouth [11], are a major cause of disease. However, not all of these cause serious problems, and reducing the use of antibiotics when this is not necessary may help to slow the development of so-called 'superbugs' which are resistant to antibiotics.

Bacteria can be identified by using 'fingerprinting' [13]. One technique is *electrophoresis*, where the different proteins extracted from a strain are separated into a banding pattern, since they migrate at different rates according to their mass in a gel placed in an electric field [1]. Another technique is pyrolysis mass spectrometry, which produces a different pattern based on the mass/charge ratio of thermally degraded molecules. These techniques are traditionally analysed by statistical techniques such as hierarchical cluster analysis [4].

Neural networks are starting to be used to identify various strains of neural networks, mostly from pyrolysis mass spectra [2,3,5,7,10]. Goodacre and colleagues [6,15]

have extended this to quantitative analysis of mixtures of strains. See e.g. [8,9] for a review.

Recently, Vhoradsky [16] used 2-dimensional gel electrophoresis patterns as input to a neural network for classification of *Streptomyces*, and Wang et. al. [17] used a complex multiple-expert approach to recognise *Escherichia coli* 0157:H7, which has caused serious food poisoning outbreaks in various countries over the last few years.

Here we will use 1-dimensional sodium dodecyl sulphate poly acrylamide gel electrophoresis (SDS-PAGE) patterns and simple multi-layer perceptron (MLP) neural networks (see e.g. [12]). These will be applied to the classification of Lactobacilli into three strain groups: casei, fermentun and rhamnosus.

## Data preparation

Gel electrophoresis strips for each of 132 strains of Lactobacilli were prepared and scanned in to the GelCompar software [1]. This resulted in a 400-element vector per strain with a resolution of 256 grey levels for each element. The mid-range values (elements 101-300) were kept, resulting in a 200-element vector for further processing. The resulting vectors were imported into Statistica version 5 (StatSoft) to produce a clustering dendrogram (Figure 1) and principal component analysis (Figure 2) to visualise the data.

The 200-element data was imported into NeuFrame 3 (Neural Computer Sciences) to simulate the MLPs. This has a user-friendly interface, making it accessible to non-expert clinical users. The software was run under Windows 95 on a 200MHz Pentium MMX PC with 32Mb RAM.

## Learning on 36 strains

For initial investigations a subset of 36 strains was selected from the 132 available strains, with 12 chosen from each of the 3 well separated groups. A multilayer perceptron with 200 inputs and 3 outputs was used with a 1-of-N output representation: 100 for casei, 010 for fermentun, and 001 for rhamnosus. A 200-10-3 architecture was chosen following experiments measuring training time and misclassification rates on a range of single-hidden-layer architectures with between 5 and 35 hidden units.

To confirm that the network could represent the data, it was initially trained on all 36 strains, using standard Error Back-Propagation with a learning rate of 0.2 and momentum of 0.9 down to a final training error of 0.3. This gave reasonably fast training, and allowed us to analyse the types of errors made by the network when not trained too 'hard'. As expected (with a network with many more weights than training samples) the network correctly classified all of the training set examples, to within 0.08 of 0 or 1 as required.

To investigate the generalisation ability, we trained the network 36 times, each time leaving out one of the training set to check the results (the leave-one-out method). A correct classification was determined to be made when only the correct output was above 0.5. Four of the 36 strains failed the leave-one-out test: C309, C282, F262 and R145. These failed either because no output reached 0.5 (C309 and C282) or more than one output exceeded 0.5 (F262 and R145), so none of the four would have been an undetected misclassification of one group for another under this criterion.

These results are qualitatively similar to those obtained from pyrolysis mass spectra of with different species of bacteria by Chun et al [2,3] and Goodacre et al [7,10], although the latter use a smaller training error and get somewhat tighter classification results.

On the 36-strain dendrogram (Figure 1) C282 (left) and F262 (middle) both have a high linkage distance to their clusters, indicating they are a long way from the other samples. However, the other failures (C309, mid-left, and R145, mid-right) seem well embedded within their species group clusters, indicating that the dendrogram cannot reliably predict which strains will be classified successfully.

Turning to the principal component analysis (Figure 2), the first 2 principal components shown represented 72% of the total variance of the dataset. Three of the casei strains (C123, C282 and C332) are a long way from the rest of the strains: one of these (C282) is a 'fail' case, while the other two were 'difficult' cases which had one or more output further than 0.2 away from the ideal 0 or 1 on the leave-one-out test above. The failed fermentun strain (F262) is the most extreme in the direction of lowest first principal component, but other apparently extreme cases (e.g. F209) did not cause problems. For rhamnosus, there were three difficult/fail cases (R251, R166 and R145), which are all in the upper region close to the border with casei, but another strain in that area (R240) was classified correctly with no difficulty.

It therefore seems that strains that cause problems are more likely to be found in the extremes or internal borders of the principal component plot. However, this is not a definitive indicator of problems. There is also a very tight bunching around the fermentun/rhamnosus border without problems, indicating that separation here may be more evident with a higher dimensionality.

As well as the leave-one-out method, we also used cross-validation with simulations performed keeping 6 of the 36 samples aside during training. On our original 200-10-3 network, we found that the performance on the validation set can be very variable (Figure 3). Validation set 1 performed as expected, with the validation error slightly more than the training error. Validation set 2 performed very poorly: investigation showed that this contained all four strains that failed on the leave-one-out test, illustrating the danger of an unlucky choice of validation set. For validation set 3 the validation error was *less* than the training error, perhaps indicating that this validation set was particularly representative and that the remaining samples in the training set were causing more classification problems than those in this validation set.

In addition to using principal components analysis (PCA) for visualisation, we can also use it to reduce the dimensionality of the data before it is given to the network. For example, using 10 components contributes about 94% of the data variance. We trained the network with up to 10 principal components (Figure 4), with the training and validation error

measured after 5000 epochs. Preliminary results with 4 hidden units indicated that the lowest validation errors occurred for 6 and 5 principal components. It would be interesting to see if this number increases with more hidden units.

### Learning on 132-strain dataset

On the full 132-strain dataset we used the same 200-10-3 architecture, as used for the 36-strain set, but this time stopping at an error of 0.01. The training times (in real time) were a little longer, but comparable to the 36-strain dataset at the same error level: although the training set was much larger, the number of epochs required tended to be proportionally smaller.

Again testing on the training set gave 100% accuracy, although we know this will be an overestimate of the performance on unseen data. To estimate this generalisation performance, we used 6-fold cross-validation, with 24 samples (8 from each group) kept aside for testing and 108 for training. Some fermentun and rhamnosus strains were used for validation more than once, since these groups have slightly fewer strains (40 and 44) than casei (48).

Similar numbers of epochs (about 1200±100) were needed to reach the required training error. Out of 144 tests, 13 (9%) failed to be classified correctly according to the same criteria used as for the 36 strain set (Table 1).

Three of these 13 (C282, C309 and R145) were also 'fail' cases for the 36 strain set. This indicates that some samples are inherently difficult to classify, rather than failure simply being a random effect changing from network to network. It could also indicate that these are outliers, needing further investigation, although we should take care not to call these 'outliers' simply because our networks find it difficult to classify them!

It is possible to improve the misclassification performance using different accept/reject criteria for the data. Varying either the 'closeness' to 1 or 0 required or the 'safety margin' between best and others required, indicated that a safety margin gives a better trade-off between undetected misclassifications and unclassified samples, with a 20% margin being a good figure here. A similar result was obtained on the 36-strain dataset (not shown).

### Other strains

There are other small strain groups which cluster near to the three strain groups used here (Figure 6). Using these as test data on the network trained on the 132-strain dataset yields results which are generally consistent with the position on the dendrogram (Figure 7).

### Conclusion

We have used a multilayer perceptron neural network to classify Lactobacilli strains into their strain groups. While using a simple neural network simulator accessible to non-experts, we gained an accuracy of 91% correct classification on 132 strains validation set.

There is some correlation between strains that are difficult to classify correctly, and their position in hierarchical cluster dendrograms or principal component plots, although not reliable enough for these to predict all problem cases.

To reduce undetected misclassifications, the 'safety-margin' criterion seems to offer a better trade-off against unclassified samples than the 'closeness to 1 or 0' criterion.

### Acknowledgements

We would like to thank Sadie Marchant for her help in generating the dental bacterial data in a readable format and Doug Clark for his Excel macro program for data conversion.

### References

- [1] Applied Maths BVBA (1996), GelCompar: Comparative Analysis of Electrophoresis Patterns, manual v4.0.
- [2] Chun, J., Atalan, E., Kim, S.-B., Kim, H.-J., Mamid, M.E., Trujillo, M.E., Magee, J.G., Manifo, G.P., Ward, A.C., and Goodfellow, M. (1993), Rapid identification of *Streptomyces* by artificial neural network analysis of pyrolysis mass spectra, *FEMS Microbiology Letters* **114**, 115–120.
- [3] Chun, J., Atalan, E., Ward, A.C. and Goodfellow, M. (1993), Artificial neural network analysis of pyrolysis mass spectrometric data in the identification of *Streptomyces* strains, *FEMS Microbiology Letters* **107**, 321–325.
- [4] Everitt, B. (1974), Cluster Analysis, Heinemann Educational Books.

- [5] Freeman, R., Goodacre, R., Sisson P.R., Magee J.G., Ward A.C. and Lightfoot N.F. (1994), Rapid identification of species within the *Mycobacterium tuberculosis* complex by artificial neural network analysis of pyrolysis mass spectra, *Journal of Medical Microbiology* **40**, 170–173.
- [6] Goodacre, R. (1994), Characterisation and quantification of microbial systems using pyrolysis mass spectrometry: Introducing neural networks to analytical pyrolysis, *Microbiology Europe* **2**(2), 16–22.
- [7] Goodacre, R., Hiom, S.J., Cheeseman, S.L., Murdoch, D., Weightman A.J., and Wade W.G. (1996), Identification and discrimination of oral asaccharolytic *Eubacterium* spp. using pyrolysis mass spectrometry and artificial neural networks, *Current Microbiology* **32**, 77–84.
- [8] Goodacre, R. and Kell, D.B. (1996), Pyrolysis mass spectroscopy and its application in biotechnology. *Current Opinion in Biotechnology* **7**, 20–28.
- [9] Goodacre, R., Neal, M.J. and Kell, D.B. (1996), Quantitative analysis of multivariate data using artificial neural network: a tutorial review and applications to the deconvolution of pyrolysis mass spectra. *Zentralblatt für Bacteriologie* **284**, 516–539.
- [10] Goodacre, R., Neal, M.J., Kell, D.B., Greenham, L.W., Noble, W.C. and Harvey, R.G. (1994), Rapid identification using pyrolysis mass spectrometry and artificial neural networks of *Propionibacterium acnes* isolated from dogs. *Journal of Applied Bacteriology* **76**, 124–134.
- [11] Hamilton, G. (1998), Open wide, we're going to explore. *New Scientist* **2125**, 33–37.
- [12] Haykin, S. (1994), *Neural Networks: A Comprehensive Foundation*, Prentice Hall.
- [13] Magee, J.T. (1993), Whole-organism fingerprinting. In Goodfellow, M. and O'Donnell, A.G., Eds, *Handbook of New Bacterial Systematics*, pp. 383–427. Academic Press, London.
- [14] Mehrotra, K., Mohan, C. K. and Ranka, S. (1997), *Elements of Artificial Neural Networks*, Bradford Book, MIT Press.
- [15] Timmins, É.M. and Goodacre, R. (1997). Rapid quantitative analysis of binary mixtures of *Escherichia coli* strains using pyrolysis mass spectrometry with multivariate calibration and artificial neural networks. *Journal of Applied Microbiology* **83**, 208–218.
- [16] Vohradsky, J. (1997), Adaptive classification of two-dimensional gel electrophoretic spot patterns by neural networks and cluster analysis, *Electrophoresis* **18**, 2749–2754.
- [17] Wang, D.Y., Keller, J.M., Carson, C.A., McAdoo Edwards, K.K. and Bailey, C.W. (1998), Use of fuzzy-logic-inspired features to improve bacterial recognition through classifier fusion. *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics* **28**, 583–591.



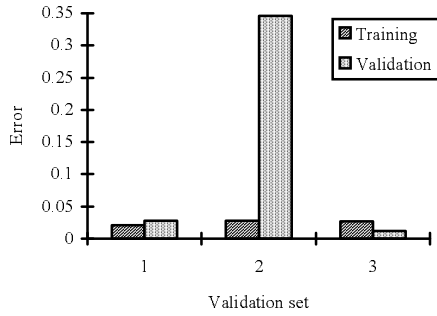


Figure 3 Training and validation error after 1000 epochs for 200-10-3 network with learning rate = 0.2 and momentum = 0.9.

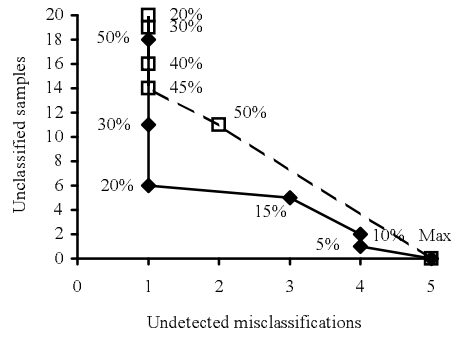


Figure 5 Unclassified samples obtained against undetected misclassifications remaining for various levels of (a) closeness to 0 or 1 required to accept a classification (dashed curve), or (b) margin required between maximum and second-best output (solid curve). "Max" indicates the criterion of simply taking the maximum output to be the classification result.

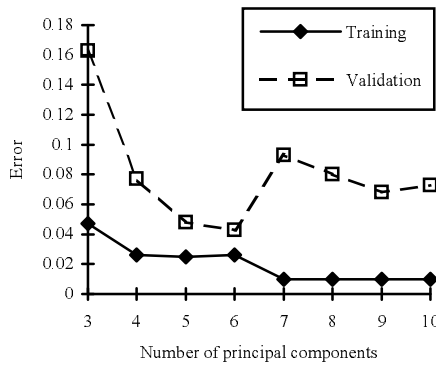


Figure 4 Effect of principal component feature extraction on training and validation error. Learning rate = 0.2, momentum = 0.9, with 4 hidden units. Training was stopped once the training error reached 0.01 (which occurred for 7, 8, 9 and 10 hidden units).

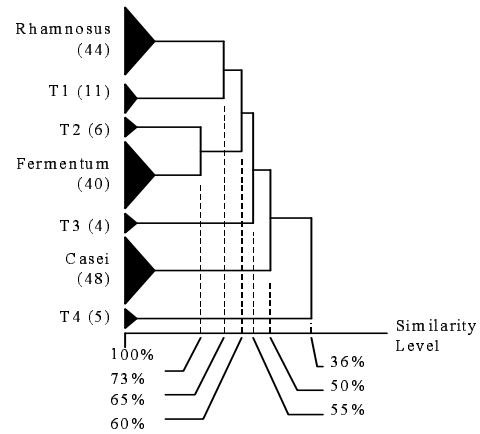


Figure 6 Dendrogram showing the relative clustering of the new test strain groups (T1-T4) with the training groups, under UPGMA clustering. The number in parentheses indicates the number of strains within each group. (Not to scale).

	Strain	C	F	R
Casei	C213	(0.206)	0.029	0.001
	C282	(0.268)	0.010	0.015
	C302	(0.481)	0.055	0.003
	C309	(0.228)	0.006	(0.329)
	C316	(0.451)	<u>0.620</u>	0.005
	C245	0.839	0.000	(0.560)
Fermentum	F34	0.039	(0.131)	0.002
	F43	(0.336)	0.999	0.000
	F62	0.153	0.026	(0.310)
	F86	0.000	0.892	(0.427)
	F148	0.060	(0.302)	0.002
	F161	<u>0.855</u>	0.012	0.025
	F291	0.123	(0.394)	0.041
	F295*	0.002	(0.599)	(0.200)
F295*	0.002	(0.544)	(0.268)	
Rhamnosus	R3	0.058	0.023	(0.372)
	R16	0.001	(0.376)	0.852
	R25	(0.201)	0.000	0.942
	R141	(0.475)	0.000	0.812
	R142	0.178	0.000	(0.614)
	R145	(0.346)	0.000	(0.344)

Table 1 Network output values for cross-validation results on misclassified or borderline cases. Outputs significantly different from 0 or 1 are in parenthesis, and those leading to incorrect classifications (using the criterion of one node above 0.5) are underlined. Strain F295 (indicated with an asterisk) appeared in two validation sets.

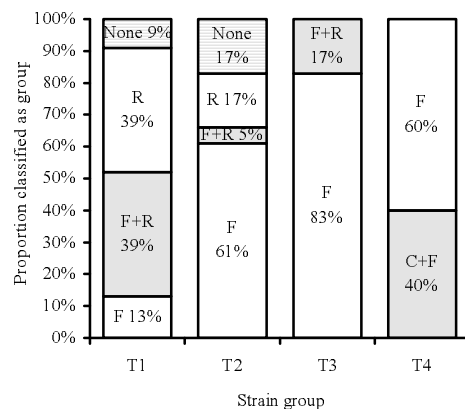


Figure 7 Proportion of new strains T1-T4, not in the training set, classified as each of Casei (C), Fermentum (F), Rhamnosus (R) or combinations of these (C+F) and (F+R), averaged over three runs with different initial weight sets. A strain was classified as "None" if no output reached 0.5.