Application of Artificial Neural Networks to the Detection of *Mycobacterium tuberculosis*, its Antibiotic Resistance and Prediction of Pathogenicity Amongst *Mycobacterium spp.* Based on Signature Lipid Biomarkers

Jonas S. Almeida[1,4], Anders Sonesson[1,2*], David B. Ringelberg[1,2] and David C. White[1,3]

[1]*Center for Environmental Biotechnology, University of Tennessee*
10515 Research Drive, Suite 300, Knoxville, TN 37932, USA
[2] *Microbial Insights, Inc., 201 Center Park Drive, Suite 1140, Knoxville, TN 37922-2105, USA*
[3]*Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA*
[4]*Dpt.Química, FCT/Universidade Nova de Lisboa, 2825 Mt.Caparica, Portugal*
[*]*Present Address, Astra Draco AB, Division of Bioanalytical Chemistry*
PO Box 34, S-221 00 Lund, Sweden

Received : 9 October 1995; revised 30 October 1995

The potential of artificial feed forward neural networks in assessing the detection of ***Mycobacterium tuberculosis****, its antibiotic resistance, and the pathogenicity of 19 Mycobacterial species was tested with a set of 67 strains based on signature lipid analysis. The lipid signature biomarkers were based on concentration patterns of 36 wax neutral lipid alcohols and fatty acids, which to date are found exclusively in* ***Mycobacteria****. The strains were assigned to species by independent clinical laboratories by their cultural properties and in some cases on the basis of DNA gene probes. The trained artificial neural network was able to identify correctly* ***M. tuberculosis*** *strains from their lipid signatures using one hidden node. The predictive accuracy was independently tested with 10 lipid profiles not used to train the artificial neural network. The analysis of predictive sensitivity showed that most of* ***M. tuberculosis*** *lipid signature is separable from other* ***Mycobacterium*** spp. *based on these data. Recognition of antibiotic resistance and pathogenicity was non-linear requiring multiple hidden nodes. When the same data set was used to train an artificial neural network with two hidden nodes to recognize different nutritional behaviours, pathogenic and saprotrophic* ***Mycobacterium*** spp. *were successfully recognized with 5% and 4% error, respectively. The successful recognition of strict pathogenicity was more complex requiring a hidden layer of 23 nodes. An association between the lipid signatures of 15 strains of* ***Mycobacterium tuberculosis*** *with resistance to isoniazid and streptomycin was achieved with a 7% error by a artificial neural network with 13 hidden nodes. These results suggest that the analysis of signature lipids by artificial neural network can be used for species detection/identification, pathogenicity, and drug resistance. Since signature lipid biomarker analysis does not require isolation and culture of microbes, has the potential for automation, rapid analysis, and exquisite sensitivity, the technique offers great promise in the detection and management of Mycobacterial disease.*

Introduction

Signature lipid biomarker patterns in cells reflect both genotype and phenotype: gene expression and metabolic activity in response to environmental parameters (Vestal & White, 1989, White, 1995). In the identification or classification of micro-organisms, the inferences from exogenous factors is lost with the routine analysis of isolates grown under standard culture conditions. DNA based identifications often cannot adequately define the phenotypic expressions of important traits (White, 1994). Since the expression of phenotypic traits such as drug resistance is crucial to the management of the infections, an analysis that provides both genotypic and phenotypic properties is desirable. We tested the hypothesis that utilizing artificial neural network (ANN) analysis of neutral lipid signature biomarkers. The possibility exists that classical definitions based on cultural properties of a particular species include biovars with larger phenotypic differences among themselves than with other close species.

The association between signature lipid composition and the type of microbial activity is particularly complicated due to the intrinsic variability between different species. In this case linear associations offer very little help, as exemplified below in identifying the pathogenicity of a number of *Mycobacterium* spp.

The accuracy of ANN is necessarily limited by the data set used to train it. However, a properly trained ANN nally, as ANN learns from experience (Hinton, 1992), its accuracy will increase as new data becomes available and is used to retraining.

Methods

Signature Lipids

The analysis utilized fatty acids and alcohols which appear thus far to be unique to *Mycobacteria* and are derived from the phthiocerol and phenophthiocerol wax neutral lipids, the phenolic glycolipids, glycopeptidolipids, and trehalose-containing lipooligosaccharides.

Isolation of the Mycobacterial phthiocerol waxes, phenophthiocerol waxes, and mycolate secondary alcohols

Lyophilized samples were extracted after suspension in methanol 0.3% NaCl (10:1, v/v) with sonication. Hexane was then added and the upper layer recovered. The lower phase of the methanol-salt solution was extracted twice with additional hexane. The combined hexane extractants were evaporated in a stream of nitrogen, dissolved in toluene and hydrolyzed with 30% methanolic KOH at 100°C. After cooling, the mixture was acidified and extracted with diethyl ether. The fatty acids were then methylated and suspended in hexane for analysis by GC/MS. The diethyl ether extracts were also analysed for secondary alcohols (Dobson *et al.*, 1985; Minnikin *et al.*, 1987). The secondary alcohols were analyzed as TMS ethers using methyl tricosanoate as an internal standard.

Isolation of the 3-hydroxy fatty acids was accomplished by subjecting the lyophilized bacteria to 4 N HCl in methanol at 85°C for 18 hours. Methyl esters of hydroxy fatty acids were purified on silicic acid columns (Jantzen *et al.*, 1993) and after the formation of the trimethylsilyl esters (TMS) analyzed by GC/MS. GC/MS analysis was by electron impact with positive ion detection of molecular ions and fragment ions specific to the components of interest. The Mycobacteria strains analyzed and their sources are listed in Table 1. The straight chain, saturated 3-0H-fatty acids of between 12 and 26 carbons were not utilized in the analyses as they were present in low (and variable) amounts and are also found in the related *Actinomycetes*, *Nocardia*, *Rhodococcus* and *Corynebacterium* genera (Table 2).

Table 1 Mycobacterium species and strains analysed listed with the corresponding reference codes and sources. The reference code consists of a species code followed by a strain number.

Species	Code	Strain	Provider
M. africanum	MAFR1	ATCC 35711	ATCC, Maryland
	MAFR2	ATCC 25420	ATCC, Maryland
M. avium	MAVI1	716/93	NPHI, Turku, Finland
	MAVI2	760/93	NPHI, Turku, Finland
	MAVI3	640/93	NPHI, Turku, Finland
	MAVI4	ATCC 35714	ATCC, Maryland
M. bovis BCG	MBOV1	BCG CO	Colorado State University
	MBOV2	TMC 1011 BCG	Dept Health, Nashville TN
M. chelonae	MCHE1	ATCC 51130	CDC, Atlanta, GA
	MCHE2	ATCC 51131	CDC, Atlanta, GA
M. gastri	MGAS1	ATCC 15754	ATCC, Maryland
	MGAS2	ATCC 25157	ATCC, Maryland
M. gordonae	MGOR1	563/93	NPHI, Turku, Finland
	MGOR2	755/93	NPHI, Turku, Finland
	MGOR3	833/93	NPHI, Turku, Finland
	MGOR4	TMC 1318	Dept Health, Nashville, TN
M. intracellulare	MINT1	603/93	NPHI, Turku, Finland
	MINT2	737/93	NPHI, Turku, Finland
	MINT3	816/93	NPHI, Turku, Finland
	MINT4	ATCC 35761	Dept Health, Nashville TN
M. kansasii	MKAN1	TMC 1201	Dept Health, Nashville, TN
	MKAN2	139R	Dept Health, Nashville TN
	MKAN3	7943	Dept Health, Nashville TN
	MKAN4	8246	Dept Health, Nashville TN
M. leprae	MLEP	CO	Colorado State University
M. malmoense	MMAL1	596/93	NPHI, Turku, Finland
	MMAL2	597/93	NPHI, Turku, Finland
	MMAL3	619/93	NPHI, Turku, Finland
M. marianum	MMAR	464/92	NPHI, Turku, Finland
M. microti	MMIC1	ATCC 35781	ATCC, Maryland
	MMIC2	116/93	NPHI, Turku, Finland
	MMIC3	ATCC 19422	ATCC, Maryland
M. phleii	MPHL	TMC 1516	Dept Health, Nashville TN
M. scrofulaceum	MSCR1	966 R	Dept Health, Nashville TN
	MSCR2	ATCC 19981	Dept Health, Nashville TN
	MSCR3	ATCC 35787	Dept Health, Nashville TN

Table 1 Continued on facing page

Table 2 3-0H-Fatty Acids in *Actinomycetes* Strains (in pmoles/mg dry weight).

Species	Source	Number of carbons:number of double bonds									
		14:0	16:0	18:0	20:0	21:0	22:0	23:0	24:0	25:0	26:0
Rhodococcus equi *	ATCC 6939	80	120								40
Rhodococcus rhodochrous **	ATCC 21197			110	1300	160	1050	2500	300	150	60
Nocardia asteroides **	ATCC 3308			90	70						
Corynebacterium pseudotuberculosis *	ATCC 19410	40	70	320	20						

Data Analysis

The lipid profile of 67 *Mycobacterium* spp .isolates were processed by cluster analysis and ANN. Lipid concentrations were normalized by dividing the observed value by the maximum value found for that particular lipid in any strain. The cluster analysis was performed by unweighted pair-group average (UPGA) of Euclidean distances (Sneath & Sokal, 1973). Fully connected feed forward artificial neural networks (ANN) were implemented in *Braincel*™ 2.5. The connection weight optimization was performed by the Back Percolation algorithm (Jurik Research & Consulting Lic., 1991). In all cases, only one hidden layer was considered and the number of hidden nodes was set as the minimum that still allows cross validation with a data subset not used to train the ANN (option "best net search" within *Braincel*™ 2.5). Cross validation consisted of comparing the predictive error for the training data set with the predictive error for a data set not used to train the ANN (test data set), and stop the training process when the they diverge. A positive identification of a particular trait was quantified as '1' and a negative identification as '0'. The prediction error was defined as the standard deviation between the predicted and observed values. Therefore, the prediction error describes primarily the ANN goodness of fit and not the predictive success: if the value 0.5 was set as the threshold for a positive identification, the predictions of streptomycin resistance in Table 11 would be all correct, yet the predictive error was 15%. Tables 3, 4, 8 and 9 also include the standard deviation, defined with reference to the average output value.

The predictive sensitivity to each lipid was calculated by the maximum variation induced in the prediction by varying the given lipid concentration. The resulting values were plotted as cumulative percentages—the sum of sensitivities to all lipids was 100%.

In order to further understand the recognition process, the correlation between each lipid concentration (*l*) and a positive identification was evaluated by a similarity (*SiI*) and dissimilarity (*DiI*) index. These indexes were calculated as the standard deviations among the positive identifications and between the positive and negative identifications, respectively, using both the training and test data sets (Equation 1). The association between a difference in lipid concentration and the target trait would be reflected by a difference between the similarity and dissimilarity indexes for that lipid. The opposite situation where the individual lipid is not correlated with the trait would be reflected by approximately equal values of *SiI* and *DiI*. However, a low *SiI* and high *DiI* does not necessarily indicate that the target trait can be recognized solely by that lipid concentration—the difference would have to be consistent for all lipid concentrations. On the other hand, lipids with high values for both indexes can still be crucial in the recognition of the target trait when considered together with other lipids. In this case, the recognition is non-linear and requires ANN with multiple hidden nodes.

Table 1 Continued from facing page

Species	Code	Strain	Provider
M. smegmatis	MSME	TMC 1515	Dept Health, Nashville, TN
M. szulgai	MSZU1	525R	Dept Health, Nashville, TN
	MSZU2	579R	Dept Health, Nashville TN
	MSZU3	719R	Dept Health, Nashville TN
M. tuberculosis	MTUB1	H_{37} Rv CO	Colorado State University
	MTUB10	H_{37} Rv Lund	Lund University Hospital
	MTUB11	H_{37} Rv Nashville	Dept Health, Nashville, TN
	MTUB12	TMC 201	Dept Health, Nashville, TN
	MTUB13	Clin isol Lund	Lund University Hospital
	MTUB14	3503468	Clinical isolate, The Toronto Hospital
	MTUB15	3403080	Clinical isolate, The Toronto Hospital
	MTUB16	3491069	Clinical isolate, The Toronto Hospital
	MTUB17	2605879	Clinical isolate, The Toronto Hospital
	MTUB18	126614	Clinical isolate, The Toronto Hospital
	MTUB19	2161727	Clinical isolate, The Toronto Hospital
	MTUB2	2750789	Clinical isolate, The Toronto Hospital
	MTUB20	228712070	Clinical isolate, The Toronto Hospital
	MTUB3	91684	Clinical isolate, The Toronto Hospital
	MTUB4	3899823	Clinical isolate, The Toronto Hospital
	MTUB5	3954352	Clinical isolate, The Toronto Hospital
	MTUB6	3716681	Clinical isolate, The Toronto Hospital
	MTUB7	3718423	Clinical isolate, The Toronto Hospital
	MTUB8	3446253	Clinical isolate, The Toronto Hospital
	MTUB9	5376722	Clinical isolate, The Toronto Hospital
M. ulcerans	MULC1	ATCC 19423	ATCC, Maryland
	MULC2	ATCC 25896	ATCC, Maryland
M. xenopii	MXEN1	7735	Dept Health, Nashville, TN
	MXEN2	148/92	NPHI, Turku, Finland
	MXEN3	239R	Dept Health, Nashville, TN
	MXEN4	1469R	Dept Health, Nashville, TN
	MXEN5	TMC 1470	Dept Health, Nashville, TN

$$SiI_l = \frac{1}{(SA)^2}\sqrt{\sum_i^{SA}\sum_j^{SA}(A_{l,j} - A_{l,j})^2}$$

$$DiI_l = \frac{1}{(SA \cdot SB)}\sqrt{\sum_i^{SA}\sum_j^{SB}(A_{l,j} - B_{l,j})^2} \qquad (1)$$

$$\sum l = T$$

A = data subset with positive identifications consisting of *T* lipid concentrations x *SA* strains.

B = data subset with negative identifications consisting of

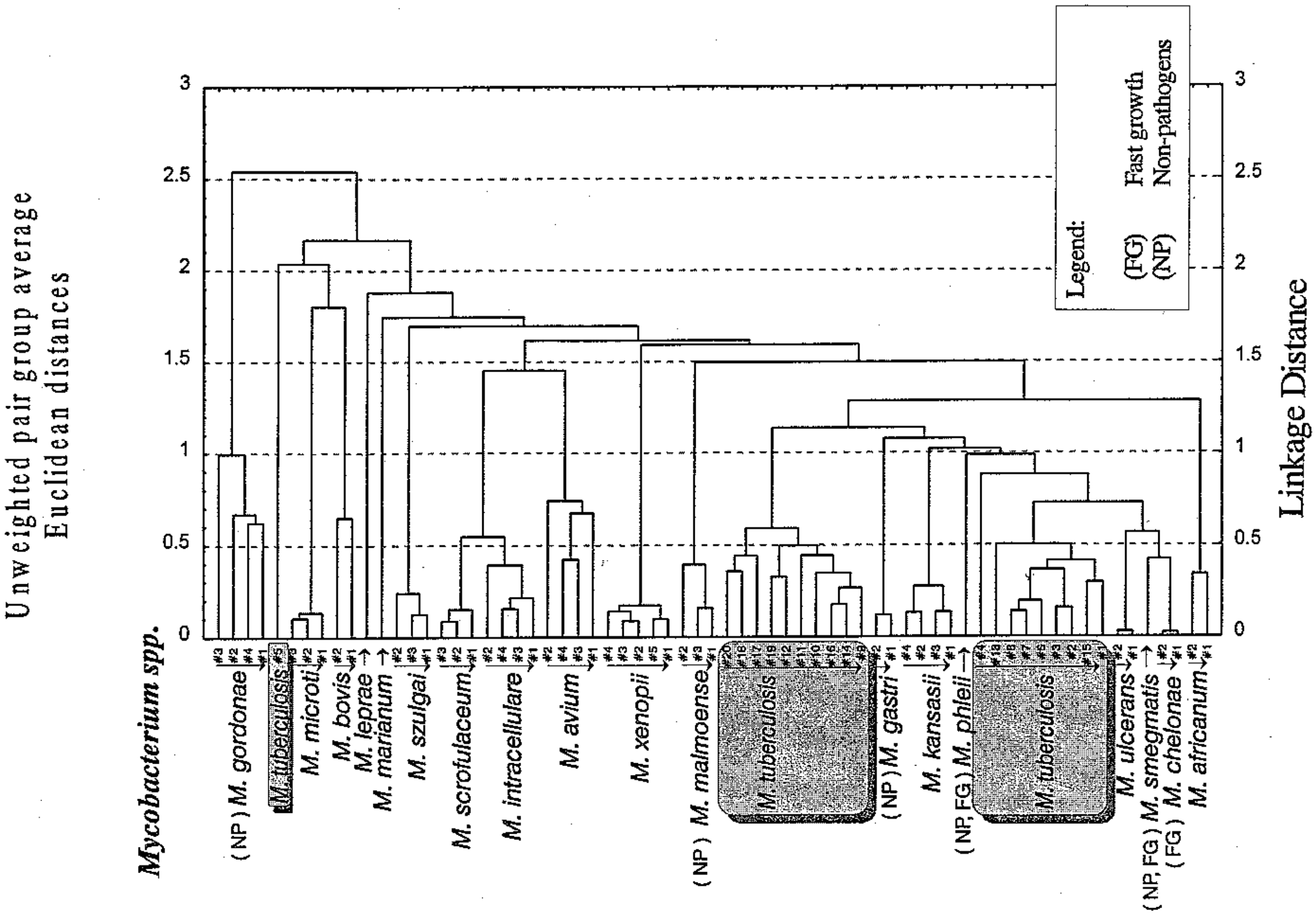

Figure 1 Cluster analysis of normalised lipid profiles. The *M. tuberculosis* strains are framed, the fast growing species of *Mycobacterium* spp. are marked 'FG' and the non-pathogenic 'NP'.

The ANN were trained to recognize the following traits based on the lipid concentration profile (lipid signature):

1. **Identification of *M. tuberculosis***
   A set of 67 lipid profiles corresponding to identified *Mycobacterium* spp strains was used to train an ANN to identify *M. tuberculosis*.
2. **Prediction of Mycobacterial pathogenicity**
   The same set of data was used to try to infer the pathogenicity and other nutritional behaviours of the corresponding species. The same set of 67 *Mycobacterium* spp lipid profiles clustered in Figure 1 was used to infer the pathogenicity and other nutritional behaviours of the corresponding species. Instead of using the raw data, the lipid profile for each species (therefore each cluster) was averaged, except for *M. tuberculosis* where each of the three clusters was averaged individually. This procedure ensures that the test data set will not include lipid signatures very similar to the ones already present in the training
3. **Prediction of resistance of *M. tuberculosis* strains to isoniazid and streptomycin**
   The lipid profiles of 15 *M. tuberculosis* strains isolated at the Toronto Hospital were used to train a ANN to associate it with resistance to Isoniazid and streptomycin.

Results

1. Identification of ***M. tuberculosis***

The set of 67 lipid profiles corresponding to identified *Mycobacterium spp* strains was analyzed by hierarchical clustering for its ability to discriminate the different species (Figure 1).

The lipid profiles of the different *Mycobacterium* spp. cluster by species with the exception of *M. tuberculosis* which is distributed between two clusters with one differentiated strain (MTUB5). Because cluster analysis failed to uniquely discriminate *M. tuberculosis*, a ANN was trained to accomplish the task. This goal was reached using only one hidden node. The summary statistics of predictive accuracy for the training set are presented in

The accuracy of predictions and the simplicity of the ANN suggest that the different *M. tuberculosis* lipid profiles are clearly separable. The predictive accuracy was independently tested with 10 lipid profiles not used to train the ANN (Table 4).

It should be noted that the *M. tuberculosis* strain #5 (MTUB5) was included in the test data set. The fact that its lipid profile clusters apart from the ones used to train the ANN (Figure 1) did not hinder its correct identification (Table 4). The importance of each individual lipid in the recognition of *M. tuberculosis* can be ascertained by plotting the corresponding sensitivity, similarity and dissimilarity indexes (Figure 2).

2. Prediction of Mycobacterial Pathogenicity

The averaged profiles (see Methods section) were associated with the species nutritional behaviour (Table 5, 6 and 7). If the lipid profile in some way depends on the nutritional behaviour, a ANN might recognize it. As previously, the data set was divided into two subsets, a larger one to train the ANN (Table 6) and the remaining one to independently test its predictive accuracy (Table 7).

The data presented in Table 6 does not include 4 species which were randomly selected to independently evaluate the ANN predictive accuracy (Table 7).

Table 3 Identification of *M. tuberculosis* from its lipid signature using training data set. The ANN configuration consisted of one hidden layer and one hidden node.

Target output	Average output	Standard deviation	Max-Min	Number of signatures
1	1.00	0.01	0.05	17
0	0.00	0.01	0.04	40

Table 4 Actual and predicted *M. tuberculosis* identification for lipid profiles used to test the ANN.

Target output	Average output	Standard deviation	Max-Min	Number of signatures
1	1.00	0.00	0.00	3
0	0.00	0.00	0.01	7

Table 5 Key for nutritional behavioural code used in Tables 7 and 8 based on the classifications of Good (1985).

Code	Nutritional behaviour
A	Pathogen
B	Saprophyte
C	Fast Growth
D	Photochromogen
E	Scotochromogen
F	Nonphotochromogen
G	Strict Pathogen

Table 6 Data used to train the ANN to recognize nutritional behaviour. The actual and the predicted outputs are displayed. The ANN configuration consisted of one hidden layer and 2 hidden nodes. The overall prediction error for this data was 7%. The prediction error for pathogenicity only (code A) was 0.1%.

	Actual nutrition behaviour							Predicted nutrition behaviour						
Cluster (Figure 1)	A	B	C	D	E	F	G	A	B	C	D	E	F	G
M. africanum	1	0	0	0	0	0	1	1.01	-0.01	0.04	0.01	-0.01	0.02	0.99
M. avium	1	0	0	0	0	1	0	1.01	-0.01	-0.01	0.01	-0.01	0.85	0.22
M. bovis	1	0	0	0	0	0	1	1.01	-0.01	0.02	0.01	-0.01	0.02	1.01
M. chelonae	1	0	1	0	0	0	0	0.97	0.02	0.33	0.02	-0.01	0.01	0.37
M. gastri	0	1	0	0	0	1	0	0.04	0.81	0.63	0.03	0.17	0.01	-0.01
M. gorodnae	0	1	0	0	1	0	0	0.00	0.92	0.15	0.04	1.01	0.04	-0.01
M. intercellulare	1	0	0	0	0	1	0	1.01	-0.01	-0.01	0.01	-0.01	0.65	0.17
M. kansasii	1	0	0	1	0	0	0	1.00	0.00	0.08	0.02	0.01	0.03	0.03
M. malmoense	0	0	0	0	0	1	0	0.69	0.16	0.19	0.02	0.14	0.02	-0.01
M. microti	1	0	0	0	0	0	1	1.01	-0.01	0.04	0.01	-0.01	0.02	1.00
M. phleii	0	1	1	0	0	0	0	0.03	0.85	0.88	0.02	0.00	0.00	0.08
M. scrofulaceum	1	0	0	0	1	0	0	1.00	0.00	0.00	0.03	0.73	0.08	-0.01
M. szulgai	1	0	0	0	1	0	0	1.00	-0.01	-0.01	0.03	0.86	0.09	-0.01
M. tuberculosis 5	1	0	0	0	0	0	1	1.01	-0.01	0.19	0.01	-0.01	0.01	0.97
M. tuberculosis A	1	0	0	0	0	0	1	1.01	-0.01	0.04	0.01	-0.01	0.02	1.00
M. tuberculosis B	1	0	0	0	0	0	1	1.01	-0.01	0.12	0.01	-0.01	0.02	0.96
M. xenopi	1	0	0	0	1	0	0	1.00	0.00	-0.01	0.03	0.84	0.09	-0.01

Table 7 Actual and predicted *M. tuberculosis* identification for lipid profiles not used to train the ANN. The overall prediction error for this data was 9%.

	Actual nutrition behaviour							Predicted nutrition behaviour						
Cluster (Figure 1)	A	B	C	D	E	F	G	A	B	C	D	E	F	G
M. leprae	1	0	0	0	0	0	1	1.00	0.00	0.18	0.02	0.00	0.02	0.33
M. marinum	1	0	0	1	0	0	0	1.01	0.00	0.05	0.01	0.00	0.02	0.83

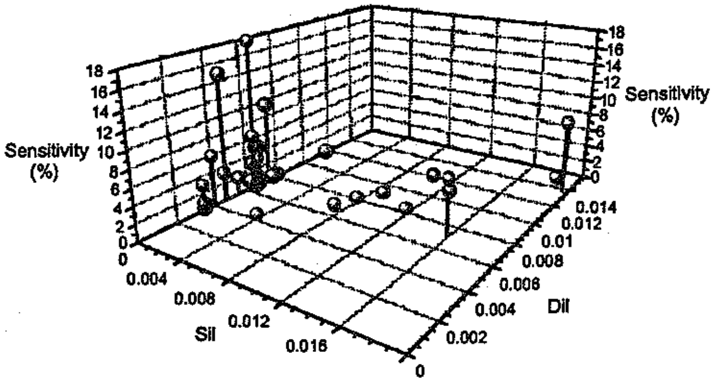

Figure 2 Lipid signature predictive sensitivity, similarity (*SiI*) and dissimilarity (*DiI*) indexes for the recognition of *M. tuberculosis*. The lower the similarity index, the higher the homogeneity for that lipid content within *M. tuberculosis* strains. The higher the dissimilarity index the higher the differences between *M. tuberculosis* strains and other strains.

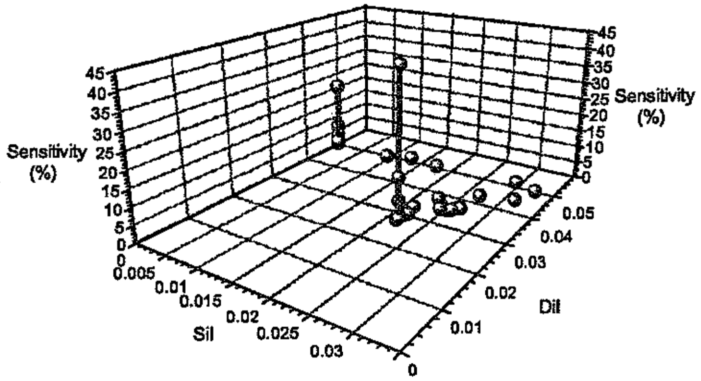

Figure 3 Lipid signature predictive sensitivity, similarity (*SiI*) and dissimilarity (*DiI*) indexes for the recognition of Pathogenicity (nutritional behaviour A) among *Mycobacterium* spp. The ANN used was trained to recognize other nutritional behaviours along with

The results presented in Tables 6 and 7 suggest that pathogenicity (code A) and some extent Saprotrophic (B) and fast growers (C), can be identified by analysis of lipid profiles. However, the nutritional behaviours other than pathogenicity include too few positive results to risk a conclusion. The sensitivity plot for pathogenicity recognition is presented in Figure 3. On the other hand this ANN failed to predict strict pathogenicity (code G) for the independent data set (Table 7). Another ANN was trained solely to recognize strict pathogenicity (Table 8) yielding a better prediction accuracy (Table 9).

The use of a separate ANN to solely recognize strict pathogenicity (code G) significantly improved the prediction accuracy. A separate ANN was also trained for pathogenicity (code A) but the prediction accuracy was much poorer than the one that considers the 7 nutritional behaviours simultaneously (results not shown). It could be speculated that the recognition of a pathogen is improved if information is available on alternative behaviours. On the contrary, a strict pathogen may not show other behaviours and therefore information on alternative behaviours only interferes with strict pathogenicity recognition.

The relatively small size of the data set for nutritional behaviours cautions against further elaboration on the information obtained.

3. Prediction of resistance of **M. tuberculosis** *Strains to Isoniazid and Streptomycin*

The 15 *M. tuberculosis* strains isolated at the Toronto Hospital were characterized for resistance to the antibiotics izoniazil and streptomycin (Tables 10 and 11) and sent to Tennessee (unidentified) for signature lipid analysis. The lipid signatures were simplified by excluding those lipids present in only 3 or less *M. tuberculosis* strains. The resulting profiles consisted of 9 of the original 36 signature lipid biomarkers. An ANN was optimized to associate the lipid signature with the antibiotic resistance data

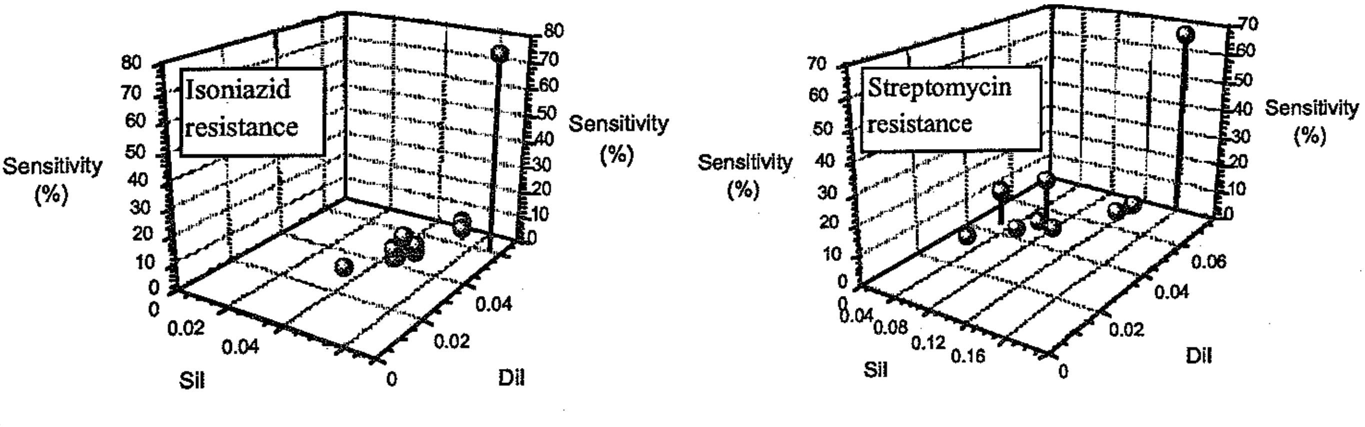

Figure 4 Lipid signature predictive sensitivity, similarity (*SiI*) and dissimilarity (*DiI*) indexes for the recognition of resistance to streptomycin and izoniazil.

(Table 10). Three strains were not used to train the ANN in order to test its predictive accuracy (Table 11). The sensitivity plots for Izoniazil and streptomycin resistance are presented in Figure 4.

The relative large number of hidden nodes and the sensitivity plots assert the non-linearity of antibiotic resistance prediction. Moreover, the significantly larger predictive errors observed with the test data (Table 11) suggest a weaker generalization than that observed for *M. tuberculosis* identification and pathogenicity/strict pathogenicity predictions.

Generalisation of ANN predictions

In order to evaluate if the independent data subsets were within the range of training data, the dendrogram in Figure 2 was used: an outlyer profile would not have as close neighbours any of the signatures used to train the ANN. The only independent data subset observed to include a clear outlyer (MTUB5) was presented in Table 4. Therefore, in the other cases, a poorer prediction for the test than for the training data subset can be interpreted as a measure of the ANN inability to generalize its predictions within the training range, as was indicated with the antibiotic resistance analyses presented in Tables 9 and 10.

Discussion

The use of ANN for microbial identifications is a well established practice (Boddy & Morris, 1993; Goodacre *et al.*, 1994; Kennedy & Takur, 1993; Schindler *et al.*, 1994) and has been used before to identify species within the *Mycobacterium tuberculosis* complex (Freeman *et al.*, 1994). The results presented hereby for *M. tuberculosis* identifications strongly support this practice. The additional inference of nutritional behaviour and antibiotic resistance from the lipid profile is more complex as reflected by the higher number of hidden nodes required to make the association. Although the small data set analyzed advises caution, the results obtained show great potential for the use of ANN to infer behaviour of possible clinical importance.

Plotting predictive sensitivity with similarity and dissimilarity indexes was used to analyze the recognition process. The non linearity associated with multiple hidden nodes was reflected in predictive sensitivity of lipids which did not correlate with the target trait (i.e. similar *SiI* and *DiI* values).

To sum up, the ANN may negotiate the complex interactions that yield a particular lipid profile and recognize its primary endogenous and exogenous variables. Furthermore, even if ANN have mostly been used as black box filters of complex inputs, its ability to put into evidence underlying dependencies should not be neglected.

The signature lipid biomarker analysis reported herein has the potential to be made into a rapid, sensitive, potentially automatable, quantitative detection/ identification system that can be used to predict potential

Table 8 Recognition of strict pathogenicity alone. The ANN configuration consisted of one hidden layer and 23 hidden nodes.

Target output	Average output	Standard deviation	Max-Min	Number of signatures
1	0.99	0.03	0.08	6
0	0.02	0.04	0.12	11

Table 9 Actual and predicted strict pathogenicity recognition from lipid profiles not used to train the ANN (Table 8).

Target output	Average output	Standard deviation	Max-Min	Number of signatures
1	0.90	0.07	0.1	2
0	0.01	0.02	0.03	2

Table 10 Antibiotic resistance data and ANN predictions for the training subset. 13 hidden nodes were used and the predictive error was 1% for Izoniazid and 2% for streptomycin resistance.

	Antibiotic resistance		ANN predictions	
Code	ISO	STR	ISO	STR
MTUB2	1	1	0.95	0.91
MTUB4	1	0	0.99	0.00
MTUB5	1	0	1.01	-0.01
MTUB7	0	0	0.01	-0.01
MTUB8	0	0	0.00	-0.01
MTUB9	0	0	-0.01	-0.01
MTUB14	0	0	0.00	0.00
MTUB15	1	1	0.97	0.95
MTUB16	0	0	-0.01	0.00
MTUB17	1	1	0.99	1.00
MTUB18	1	0	0.99	0.00
MTUB19	1	0	1.01	0.00

Table 11 Antibiotic resistance data and ANN predictions for the lipid signatures not used to train it. The predictive error was 7% for Izoniazid and 15% for streptomycin resistance.

	Antibiotic resistance		ANN predictions	
Code	ISO	STR	ISO	STR
MTUB3	1	0	0.96	0.19
MTUB6	0	0	-0.01	-0.01
MTUB20	1	1	0.85	0.75

Acknowledgements

This work was partially supported by SBIR grant CDC-WO93-0681 of the Center for Disease Control, by JNICT (Junta Nacional de Investigação Científica, Portugal) grant 1/B/94/PO, by FLAD (Fundação Luso-Americana para o Desenvolvimento, Portugal) grant 3B/A12/579/94, and funds from Microbial Insights, Inc. The authors would like to thank the laboratories who supplied the cultures for analysis, including C. Sterling of the Tennessee State Public Health Laboratory, M. Viljanan, of the Mycobacterial Reference Laboratory in Finland and M.W. Mittelman, Centre for Infection and Biomaterials Research, The Toronto Genearal

References

BODDY, L. & MORRIS, C.W. (1993). Analysis of flow cytometry data - a neural network approach. *Binary* **5**, 17-22.

DOBSON, G., MINNIKIN,D.E., MINNIKIN, S.M., PARLETT, J.H., GOODFELLOW, M., RIDELL, M. & MAGNUSSON, M. (1985). Systematic analysis of complex *Mycobacterial* lipids. In *Chemical Methods in Bacterial Systematics*, pp. 237-265. Edited by M. Goodfellow & D.E. Minnikin. New York, NY: Academic Press.

FREEMAN, R., GOODACRE, R., MAGEE, P.R., WARD, A.C. & LIGHTFOOT, N.F. (1994). Rapid identification of species within the *Mycobacterium tuberculosis* complex by artificial neural networks analysis of pyrolysis mass spectra. *Journal of Medical Microbiology* **40**, 170-173.

GOOD, R.C. (1985). Opportunistic pathogens in the genus *Mycobacterium. Annual Reviews in Microbiology* **39**, 347-369.

GOODACRE, R., NEAL, M.R., KELL, D.B., GREENHAM, L.W., NOBLE, W.C. & HARVEY, W.G. (1994). Rapid identification using pyrolysis mass spectrophotometry and artificial neural networks of *Propionibacterium acnes* isolated from dogs. *Journal of Applied Bacteriology* **76**, 124-134.

HINTON, G.E. (1992). How neural networks learn from experience. *Scientific American* **267**, 144-151.

HAYKIN, S. (1994). *Neural networks, a Comprehensive Foundation*, pp. 22-23. NY: Macmillan Pub.

JANTZEN, E., SONESSON, A., TANGEN, T. & ENG, J. (1993). Hydroxy-fatty acid profiles of *Legionella* species: diagnostic usefulness assesse by principal component analysis. *Journal of Clinical Microbiology* **31**, 1413-1419.

KENNEDY, M.J. & TAKUR, M.S. (1993). The use of neural networks to aid in microorganism identification: a case study of *Haemophilus* species identification. *Antonie van Leeuvenboek* **63**, 35-38.

MINNIKIN, D.E., BOLTON, R.C., DOBSON, G. & MALLET, A.I. (1987). Mass spectrometric analysis of multimethyl branched fatty acids and phthiocerols from clinically-significant *Mycobacteria. Proceedings Japanese Society Medical Mass Spectrometry* **12**, 23-32.

SCHINDLER, J., PARYZEK, P. & FARMER, J. (1994). Identification of bacteria by artificial neural networks. *Binary* **6**, 191-196

SNEATH, P.H.A. & SOKAL, R.R. (1973). *Numerical Taxonomy*. San Francisco, CA: Freeman.

VESTAL, J.R. & WHITE, D.C. (1989). Lipid Analysis in Microbial Ecology. *BioScience* **39**, 535-541.

WHITE, D.C. (1994). Is there anything else you need to understand about the microbiota that cannot be derived from analysis of nucleic acids? *Microbial Ecology* **28**, 163-166.

WHITE, D.C. (1995). Chemical ecology: Possible linkage between macro-and microbial ecology. *Oikos* in press.