# New approaches to identification of bacterial pathogens by surface enhanced laser desorption/ionization time of flight mass spectrometry in concert with artificial neural networks, with special reference to *Neisseria gonorrhoeae*

Oliver Schmid,[1]† Graham Ball,[2]† Lee Lancashire,[2] Renata Culak[1] and Haroun Shah[1]

[1]Molecular Identification Services Unit, Centre for Infections, Health Protection Agency, London, UK

[2]The Nottingham Trent University, School of Biomedical and Natural Sciences, Nottingham, UK

Correspondence
Oliver Schmid
oliver.schmid@hpa.org.uk

Received 1 July 2005
Accepted 16 August 2005

Surface enhanced laser desorption/ionization-time of flight mass spectrometry (SELDI-TOF MS) has been applied in large numbers of oncological studies but the microbiological field has not been extensively explored to date. This paper describes the application of SELDI-TOF MS in concert with a multi-layer perceptron artificial neural network (ANN) with a back propagation algorithm for the identification of *Neisseria gonorrhoeae*. *N. gonorrhoeae*, the aetiological agent of gonorrhoea, is the second most common sexually transmitted disease in the UK and USA. Analysis of over 350 strains of *N. gonorrhoeae* and closely related species by SELDI-TOF MS facilitated the design of an ANN model and revealed 20 ion peak descriptors of positive, negative and secondary nature that were paramount for the identification of the pathogen. The model performed with over 96 % efficiency when based on these 20 ion peak descriptors and exhibited a sensitivity of 95·7 % and a specificity of 97·1 %, with an area under the curve value of 0·996. The technology has the potential to link several ANN models for a comprehensive rapid identification platform for clinically important pathogens.

## INTRODUCTION

Gonococcal infection is the second most common bacterial sexually transmitted infection in the UK, with more than 24 000 infections diagnosed in 2003. The number of confirmed cases has risen steadily since 1995. Young people are most commonly infected, with males aged 20–24 years and females aged 16–19 years showing the highest rates of infection, especially in urban areas (http://www.hpa.org.uk/infections/topics_az/hiv_and_sti/sti-gonorrhoea/gonorrhoea.htm; Gerbase *et al.*, 1998).

The causative organism, *Neisseria gonorrhoeae*, is currently confirmed in diagnostic laboratories using traditional biochemical tests and more recent 16S rDNA analysis. Conventional diagnostic methods for *N. gonorrhoeae* include direct microscopy, selective culturing, immunological tests, enzyme reaction tests and nucleic acid amplification tests, among others (Johnson *et al.*, 2002; Knapp, 1988). The population structure of *N. gonorrhoeae* is of an unstructured random mating or panmictic nature (Smith *et al.*, 1993) causing frequent horizontal genetic exchange, which in addition to natural mutation causes the high levels of variability that enable bacterial adaptation and immune system evasion (Fredlund *et al.*, 2004).

In this pilot study, to explore more specific and rapid methods of diagnosis, we assessed the potential application of surface enhanced laser desorption/ionization-time of flight mass spectrometry (SELDI-TOF MS) in concert with artificial neural networks (ANNs) for bacterial identification. The former is a modified version of matrix-assisted laser desorption/ionization-TOF MS (MALDI-TOF MS), in that it utilizes ProteinChip arrays for selective capture of chemically or biochemically distinct proteins from a mixed population (Fung & Enderwick, 2002). The protein arrays in SELDI-TOF MS are available with different types of receptors bound to their surface. In this study we used reverse phase H50 ProteinChip arrays. Their active surface contains 16 methylene groups that bind proteins through reverse phase

†These authors contributed equally to this paper.

Abbreviations: ANN, artificial neural network; MLP, multi-layer perceptron; PCA, principle component analysis; SELDI, surface enhanced laser desorption/ionization; TOF, time of flight.

chemistry. The binding occurs with proteins rich in alanine, valine, leucine, isoleucine, phenylalanine, tryptophan and tyrosine.

Prior to analysis, the extracted proteins are bound to the array, covered with a matrix solution to initiate crystallization and subsequently bombarded with a nitrogen laser to create an ion cloud. Depending on their size, the ions move, via the TOF tube, towards the detector at different speeds and are thus separated according to their mass/charge ratio. The resulting spectra can be used to represent different bacterial species or in some instances subspecies or serogroups (unpublished data). However, the resulting output from the instrument is complex, and the accompanying data analysis software is designed for searching for biomarkers among limited datasets. Consequently alternative approaches for data analysis were investigated in several preliminary studies. Among them, ANNs appeared the most promising.

ANNs are a form of artificial intelligence that allows for the modelling of complex systems, such as those found in the physical sciences. They are able to identify biomarkers that are capable of differentiating between two states while also providing a measure of certainty for the prediction (Lancashire *et al.*, 2005; Mian *et al.*, 2003). They are non-linear and hence able to process data containing complex interactions, whilst also being able to model for data that may be noisy, fuzzy or even incomplete. ANNs have successfully been used in the analysis of single nucleotide polymorphism (SNP) data (Tomita *et al.*, 2004) and microarray data (Khan *et al.*, 2001). They store their knowledge by means of inter-neuron connection weights, which are determined by the training process. After training, ANNs do not require information about the data source with which they are challenged (Agatonovic-Kustrin & Beresford, 2000).

There are various learning algorithms that may be applied to ANNs. This study utilized a multi-layer perceptron (MLP) ANN, together with a back-propagation algorithm, because of its wide-application capabilities and ability to manage data with high levels of background noise (Wei *et al.*, 1998; Ball *et al.*, 2002).

In this study we analysed over 350 strains of *N. gonorrhoeae*, other neisseriae and closely related species such as *Kingella denitrificans* and *Moraxella osloensis* isolated from around the UK. All strains were analysed on hydrophobic H50 ProteinChip arrays. Comparative 16S rDNA sequence analysis and standard biochemical tests were used to establish the identity of the strains prior to SELDI-TOF MS analysis.

## METHODS

**Strains.** The strains analysed were as follows: *N. gonorrhoeae*, 305 (126 training; 42 testing; 42 validation; 95 extra); *Neisseria meningitidis*, 48; *Neisseria animalis*, 1; *Neisseria caviae*, 1; *Neisseria cinerea*, 4; *Neisseria cuniculi*, 1; *Neisseria denitrificans*, 1; *Neisseria elongata*, 1; *Neisseria flava*, 1; *Mycoplasma hominis*, 2; *Neisseria lactamica*, 1; *Neisseria mucosa*, 1; *Neisseria ovis*, 1; *Neisseria siccus*, 1; *Neisseria polysaccharea*, 2; *Neisseria subflava*, 2; *Neisseria weaveri*, 1; *M. osloensis*, 1; *K. denitrificans*, 11. In addition the model was confronted with a number of unrelated microorganisms such as salmonellae and staphylococci.

**Identification of *N. gonorrhoeae* by 16S rDNA sequencing.** Isolates were grown on chocolate agar (Medical Department, HPA) for 24 h at 37 °C in 10 % $CO_2$. Cells were suspended in sterile water and heated for 10 min at 95 °C. The suspension was cooled on ice and centrifuged for 20 s at 10 000 ***g***, and 1 µl of the supernatant was used for the PCR. The primers (MWG) used were 27, 5′-AGAGTTTGATCMTGGCTC-3′, and 1522, 5′-GGAGGTGATCCANCCRCA-3′ (product size, 1495 bp). PCR cycle conditions using PCR ReadyMix (Sigma) were as follows: 95 °C for 2 min, followed by 35 cycles of 95 °C for 45 s, 56 °C for 45 s and 72 °C for 60 s. Final extension was carried out at 72 °C for 5 min. Products were cleaned using the PCR purification kit (Qiagen). The sequencing primers used were 3, 5′-GTTGCGCTCGTTGCGGGACT-3′, and ANT1, 5′-AGAGTTTGATCCTGGCTCAG-3′. Sequences were analysed on a Positive Beckman Capillary sequencer (CEQ 8000).

The bacteria detected by 16S rDNA PCR were identified by sequence comparison to the GenBank database using BLAST (http://www.ncbi.nlm.nih.gov) and to an in-house database of over 400 16S rDNA sequences of *N. gonorrhoeae.*

**SELDI-TOF MS.** Strains were grown on chocolate agar (Medical Department, HPA) at 37 °C for 48 h in a 10 % $CO_2$ atmosphere and resuspended in cooled distilled water containing glass beads (< 105 µm, Sigma). Cells were disrupted for four successive periods of 1 min using a Mickle cell disintegrator (Mickle) with 1 min intervals on ice in between, and afterwards were left on ice for 10 min before being pulsed down. The supernatant was transferred into fresh tubes and spun at 21 000 ***g*** for 40 min at 4 °C.

All SELDI-TOF MS ProteinChip arrays used were H50 hydrophobic arrays (Ciphergen Biosystems). All samples were run in duplicate. Chip arrays were bulk washed in 50 % methanol (BDH) for 5 min. Following 30 min of air-drying, each spot was pre-wetted twice with 5 µl 0·1 % trifluoroacetic acid (Sigma) for 5 min. Samples were diluted 1 : 1 with ammonium acetate in 25 % acetonitrile at pH 4. Four microlitres of sample was applied to array spots and incubated in a humidity chamber for 60 min. Subsequently, array spots were washed twice with 5 µl 50 % methanol for 2 min and left to air-dry. One microlitre of matrix was applied and left to dry for 10 min. The matrix consisted of 1 % trifluoroacetic acid (Sigma) in 100 % acetonitrile (Sigma) (1 : 1) saturated with 35 mg sinapic acid $ml^{-1}$ (Fluka Chemie AG). The ProteinChip arrays were analysed using the SELDI-TOF mass spectrometer (PBS II Protein Chip Array Reader, Ciphergen Biosystems), applying a focus mass of 26·5 kDa and a mean laser intensity of 180 eV. Spectra were collected using ProteinChip Software v3.1 (Ciphergen Biosystems). All spectra were normalized for total ion current and mass calibrated applying a pre-recorded spectrum of Ciphergens all-in-one protein mix TLF 15kDa [calibrants: cytochrome C (equine cardiac) (12360·2+1H), myoglobin (equine cardiac) (16951·5+1H) and albumin (bovine serum) (66433+1H)]. Preliminary principle component analysis (PCA) was performed using Ciphergen Express Data Manager v2.0 (Ciphergen Biosystems).

### ANN architecture and training

**Data collection for analysis for ANN.** A constrained MLP network was used containing two nodes in the hidden layer using algorithms implemented in Neuroshell (WardSystems, www.wardsystems.com). The momentum was set to a value of 0·5 and the learning rate to 0·1 as these values had produced encouraging results previously. During training, the samples were split so that 60 % were used for training, 20 % were used as a test set and the remaining 20 % were set aside as a production (validation) dataset. Training continued until the model reached convergence, which was established by a failure of the model to improve the minimum mean squared error on the test dataset for 20 000 training events, so as to avoid over-training.

**Model parameterization.** This allows for the identification of the inputs that have the most influence on correct sample classification.

Data in the mass range of 3–6 kDa were trained over 50 random training/test/production subsets (bootstrapping/random sample cross validation). Relative importance values of each input were calculated based on the weight of the ANN model (Ball *et al.*, 2002), allowing ranking of inputs in order of importance according to their influence on sample classification. A rolling approach was used to shift the data along 1 kDa so that the input range spanned the 4–7 kDa mass range and then trained as previously. The process was repeated, producing a proteomic profile displaying the relative importance of ions over a data range of 3–30 kDa.

Inputs with the greatest influence in the model were selected for further training to reduce model complexity, yet increase classification performance. The top 1000 inputs of greatest relative importance and training over 50 random training/test/production subsets were selected and relative importance values were recorded to deduce the top 100 inputs from these 1000. Training was repeated to determine the top 30 inputs and finally the top 20 inputs from an initial cohort of approximately 13 000 inputs within the 3–30 kDa mass range.

## RESULTS AND DISCUSSION

The proteome of the cell represents a wide range of hitherto untapped intracellular and membrane-bound molecules that may serve as biomarkers for microbial classification and identification, and consequently studies of microbial infections. The absence of high-throughput technologies and appropriate data analysis software is largely responsible for this deficiency. Proteins represent the functional molecules of the cell and in tandem with genomics provide a new perspective on the characterization of microbes. SELDI-TOF MS is the first of these new technologies that is amenable to high throughput analysis which, in many instances, is sensitive down to femtomole or even attomole concentrations of the analyte (Grus *et al.*, 2003). Its salient property is its accompanying ProteinChip assays, which have the capacity to selectively capture even low abundance proteins in the complex biological milieu of the cell with low levels of signal to background ratios. The resulting mass spectrum varies considerably among bacterial species (Fig. 1) and, using the method described here, we have identified biomarkers up to 180 kDa among species (unpublished).

Previous experiments on a range of micro-organisms comparing ProteinChip arrays such as the hydrophobic- (H50), anion exchange- (SAX2) as well as cation exchange- (WCX) surface characteristics have shown the H50 ProteinChip arrays yield the broadest spectrum of mass ions and mass intensities (data not shown). Therefore these were selected for all further work. Using the H50 array, the SELDI-TOF mass spectrum generated results in excess of 34 000 data points in the analysed mass range alone (3–30 kDa), which makes successful and exhaustive data mining a formidable task.

Preliminary investigation of *N. gonorrhoeae* strains from a relatively isolated group of individuals showed subtle variation in mass spectral profiles, suggesting that SELDI-TOF MS is capable of detecting small variations in differential protein expression among strains. However, the accompanying Biomarker Wizard software was found to be inadequate as a tool for microbial characterization. A single mutation, down-regulation or modification of a particular biomarker (e.g. post-translational modification) may result in the loss of a key biomarker and consequently incorrect assignment of a profile to a given taxon. More recent developments in data analysis software that utilizes principle component analysis (PCA) have considerably improved the capacity of SELDI-TOF MS to delineate closely related species to a degree. However, when the SELDI-TOF MS mass ions of the two closely related species *N. gonorrhoeae* and *N. meningitidis* were analysed, the two data vectors merged to a point of overlap that contained several intermediary samples (Fig. 2).

The problem most likely lies within the nature of PCA

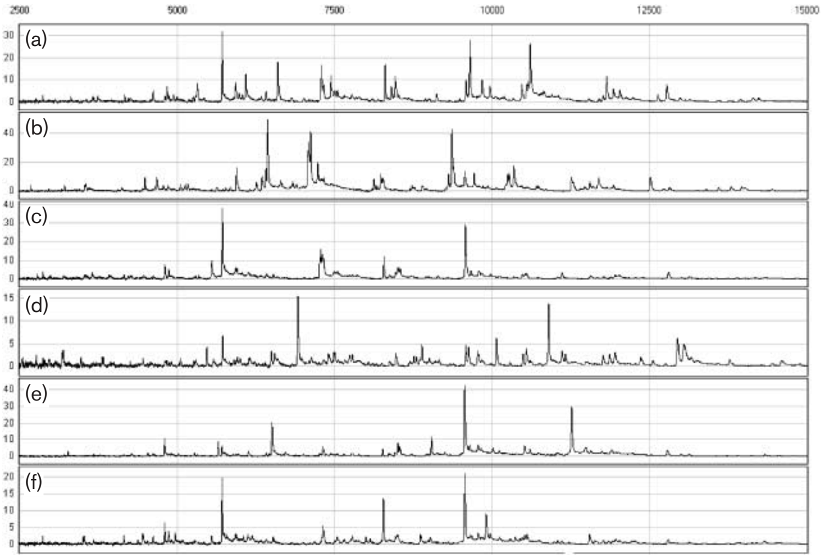

**Fig. 1.** Surface enhanced laser desorption/ionization time of flight mass spectra (SELDI-TOF MS) comparison of *N. gonorrhoeae* (a) and closely related species *N. meningitidis* (b), *N. cinerea* (c), *K. denitrificans* (d), *N. mucosa* (e) and *M. osloensis* (f). The mass spectral profiles display the ability of the technology to delineate species that are difficult to separate by conventional means.

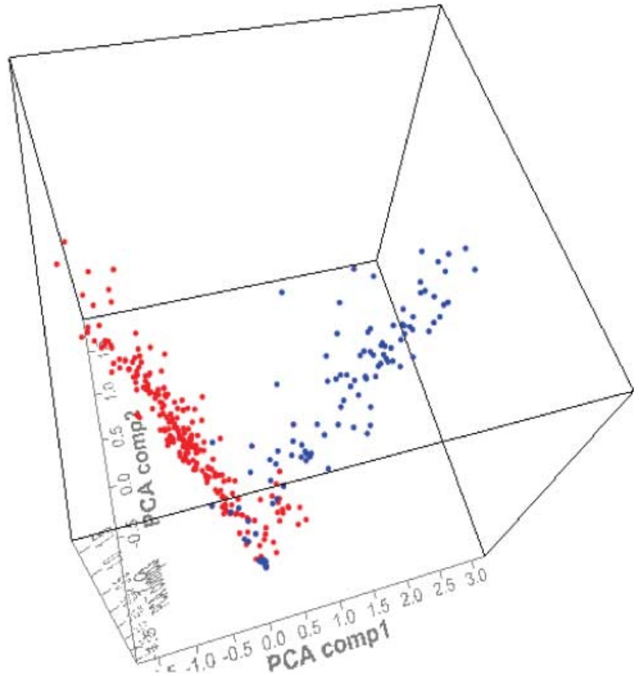

**Fig. 2.** Principle component analysis (PCA) of SELDI-TOF MS data of *N. gonorrhoeae* (red) and *N. meningitidis* (blue). The inability to distinguish some isolates of these closely related micro-organisms is shown in these datasets. Two distinct vectors merging to a point and the overlap of several intermediary samples are shown.

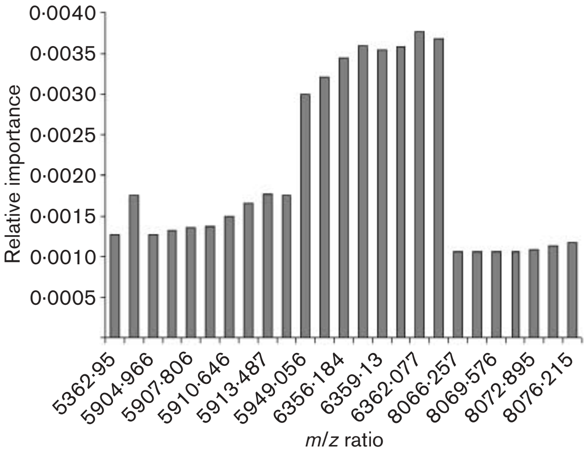

**Fig. 3.** Ion channel mapping of putative biomarkers for *N. gonorrhoeae*. The graph represents the relative importance of eight ion channels, charting the spread of the potential biomarker throughout the production/validation population. Compared to more conventional analysis applications, which detect a peak at a certain *m/z* value, the ANN charts the peak spread on the basis of the raw ion channel data.

analysis. PCA is a form of data compression that reduces data dimensionality to the most descriptive data separation vector. The algorithms utilized to compress data dimensionality in PCA are based on a straight decision boundary concept and work in a limited ($k$) dimensional space defined by the number of experimental observations per sample. This could explain some of the convergence of the data vector, especially when comparing organisms that are closely related. In addition, strains with trans-species characteristics have been reported in which biochemical tests defined the organism as *N. meningitidis* and serological tests defined it as *N. gonorrhoeae* (Vazquez *et al.*, 1995), further blurring the species boundaries.

Consequently alternative data analysis involving non-linear decision boundary algorithms such as ANNs were investigated. In contrast to PCA, ANNs have the ability to utilize a curved decision boundary as they operate in unlimited ($n$) dimensional space and do not employ data compression to reduce data dimensionality. The ion peak predictors utilized in the ANN approach are not just arbitrary peaks detected by the software algorithm but correlate to the end result as the identity of the training/validation population is known in advance. Furthermore, the ANN model not only considers the presence or absence of an ion peak at the detected highest signal-to-noise ratio but maps out the potential drift of a peak throughout the entire training sample population (Fig. 3) based on the raw ion channel data of the mass spectrometer. Throughout a sample population, an ion peak can be affected by a slight mass drift across a ProteinChip array and also can exhibit a small amount of mass drift affected by the resolution of a solely linear mode TOF mass spectrometer.

For verification of the necessary 'teacher signal' we performed 16S rDNA sequence analysis and phenotypic (biochemical and Phadebact immunological) tests on all strains used for building, validation and subsequent blind data testing of the ANN model. Previous experience in our laboratory indicates that comparative sequence analysis based upon 1 kb contigs (i.e. sets of overlapping segments of DNA) provides an accurate means of confirmation of *N. gonorrhoeae* samples (unpublished database of 400 samples).

Training of the ANN started by defining the 1000 most important peaks within the 3–30 kDa range as this was found to be the most descriptive mass range for all micro-organisms analysed in our lab. Subsequently, the number of ion inputs was reduced to minimize noise interference and increase the speed of the model by determining the relative importance values for individual ion peaks and selecting the peaks with the most predictive power. The process of selecting ion peaks with superior descriptive performance was repeated and the model re-trained until it did not exhibit any significant improvement in predictive power or reduction in error value. The final model was based on the 20 most predictive ion peaks (Table 1), at which point the model exhibited the lowest mean squared error value (Fig. 4).

The output of the ANN was set to a value between 1 and 2 with a cut-off point of 1·5, where all samples with a prediction value below 1·5 are considered to be *N. gonor-*

**Table 1.** Mass-to-charge values of ion channels and nature of 20 SELDI-TOF MS ion peak predictors for identification of *N. gonorrhoeae*

| Mass-to-charge value | Response descriptor |
|---|---|
| 3219·92 | Negative |
| 3220·97 | Negative |
| 4505·98 | Negative |
| 5223·17 | Negative |
| 5224·51 | Secondary |
| 5232·52 | Negative |
| 5949·56 | Secondary |
| 5950·48 | Negative |
| 6356·18 | Secondary |
| 6357·66 | Negative |
| 6359·13 | Secondary |
| 8121·1 | Negative |
| 9389·33 | Positive |
| 9390·82 | Positive |
| 10165·22 | Positive |
| 10167·78 | Positive |
| 10279·15 | Secondary |
| 10410·68 | Positive |
| 10412·56 | Positive |
| 12506·63 | Negative |

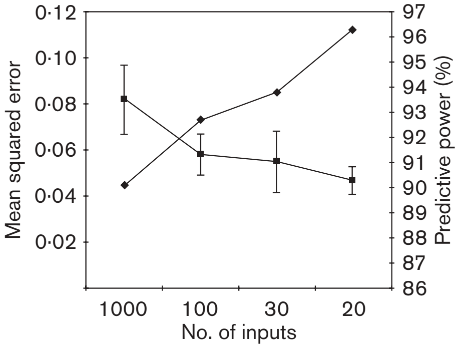

**Fig. 4.** Prediction capability of the *N. gonorrhoeae* ANN model. The number of ion peak inputs plotted against correct prediction percentage are represented by ◆. The first test model based on 1000 ion inputs displayed a 90 % correct prediction rate. Reducing this number to the 20 ion peaks with the highest relative importance values minimized noise interference and increased the predictive power to >96 %. The mean squared error rate for the individual models, represented by ■, consistently decreased to the point where the model verified 20 ion peak markers per strain.

*rhoeae* and all samples with a value over 1·5 are considered to be of other origin. The prediction values are arbitrarily assigned to function as result descriptors. The correct prediction rate of the model for the production/validation dataset of *N. gonorrhoeae* and closely related taxa, such as other *Neisseria*, *Kingella* and *Moraxella* species, was >96 %. The sensitivity of the model was determined to be 95·7 %, which indicates that out of a population of 100 positive samples, 96 would be classified as positives and four samples would be wrongly labelled as negative for gonorrhoeal infection. Conversely, the specificity of the test was determined to be 97·1 %, which corresponds to an identification of 97 true negatives for a population of 100 gonorrhoea-negative samples and leaving three samples categorized as *N. gonorrhoeae* positive when in fact they are not. The positive predictive value was determined to be 97·78 %, with a negative predictive value of 94·37 %. The area under the curve was calculated to be 0·991 (Fig. 5).

ANNs are considered to be a black box because they do not provide information on how a particular output is reached (Schwarzer *et al.*, 2000; Geeraerd *et al.*, 2004). However, information can be derived by analysing the nature of the input layer as well as the information contained in the output layer. Response plots of the 20 ion peaks (three of which are shown in Fig. 6) demonstrated that the model computes six positive-, nine negative- and five secondary-predictors. Positive predictors are either ion peaks only present in *N. gonorrhoeae* or ion peaks displaying a distinct intensity differential when compared to the mean intensity for that peak. Negative predictors may be absent in *N. gonorrhoeae* but present in the non-gonorrhoeae sample population or they may exhibit a discrete intensity disparity to the *N. gonorrhoeae* sample population. Secondary predictors do not affect the output in a discriminate way but rather fine-tune the output value. The slope of the curve contains further information and acts as an indicator for the discriminatory power of the ion peak; the steeper the slope, the more discriminatory the ion peak.

Unfortunately, molecular mass comparison of the positive

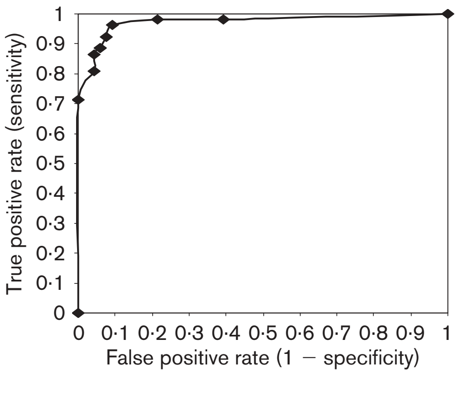

**Fig. 5.** Receiver operating characteristic (ROC) curve demonstrating the high specificity and sensitivity of the ANN model for blind data. In an ROC curve, the number of true positives is plotted against the number of false positives at different prediction thresholds. The greater the area under the curve, the more accurate the model. Conversely, the closer it comes to the 45° diagonal, the less accurate the test.

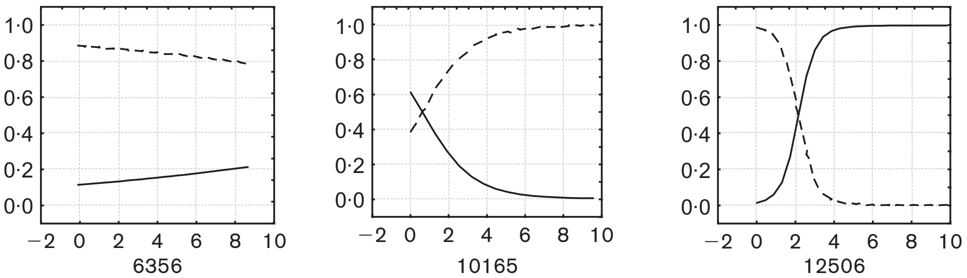

**Fig. 6.** ANN response plots for three of the 20 ion peaks on which the ANN model is based. The graphs plot relative intensity (*x* axis) versus predictive output (*y* axis) of a given ion peak. The dotted lines indicate predictive response of the model to the ion peak for *N. gonorrhoeae* and the solid lines indicate the predictive response to the ion peak for non-gonorrhoeae. A dotted line graph plotted from a high predictive output value with low relative intensity towards a low predictive output value with a high relative intensity (e.g. 12506) indicates a negative predictor, i.e. an ion peak that is present (or at least more intense compared to the mean) in non-gonorrhoeae samples but not in *N. gonorrhoeae* samples. The opposite graph indicates a positive predictor (e.g. 10165), which is an ion peak present in *N. gonorrhoeae* but not in non-gonorrhoeae samples. The steeper the curve, the more descriptive the predictor, as the changeover from *N. gonorrhoeae* to non-gonorrhoeae occurs over a small change in relative intensity. Graphs where the two lines do not cross represent secondary descriptors (e.g. 6356), which do not directly affect the output but rather fine-tune the output result. The range of intensity values in the data are indicated by the lengths of the lines on the *x* axis.

predictors, which are the only predictors species-specific status could be assigned, did not give conclusive results. This is most likely to be due to the preparatory method causing a decrease in protein integrity. Maintaining protein integrity through the extraction process is very involved (Fedarko, 1994) and did not present itself as a necessity to the study. The molecular masses of all the predictor ions are relatively small already, indicating that these are not whole proteins.

In addition to the analysis of the predictors, further information regarding the relationships of the strains can be achieved by comparing the output values. It is likely that strains displaying similar output values are potentially more related to each other than strains that demonstrate a high differential in output values. This aspect of the ANN analysis will be investigated further in future studies.

## Concluding remarks

SELDI-TOF MS has presented itself as a capable technology for microbial characterization and identification, and in synergy with ANNs the potential of the technology can be amplified. The present method for generation, training and validation of an ANN model for a bacterial organism can be achieved in real-time (~3 h), after which the model can be challenged with new unknown data at the click of a button. The model therefore has the potential for high-throughput screening of micro-organisms. Several ANN models for a number of human pathogens, among them *N. meningitidis* (Lancashire *et al.*, 2005) and *Staphylococcus aureus*, have been built and trained with the aim of aligning them for the detection of important human pathogens. This would open up the possibility of devising a single and rapid platform, based on SELDI-TOF MS, for simultaneous identification of a wide range of microbial pathogens.

## REFERENCES

**Agatonovic-Kustrin, S. & Beresford, R. (2000).** Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *J Pharm Biomed Anal* **22**, 717–727.

**Ball, G., Mian, S., Holding, F. & 8 other authors (2002).** An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers. *Bioinformatics* **18**, 395–404.

**Fedarko, N. S. (1994).** Isolation and purification of proteoglycans. *EXS* **70**, 9–35.

**Fredlund, H., Falk, L., Jurstrand, M. & Unemo, M. (2004).** Molecular genetic methods for diagnosis and characterisation of *Chlamydia trachomatis* and *Neisseria gonorrhoeae*: impact on epidemiological surveillance and interventions. *APMIS* **112**, 771–784.

**Fung, E. T. & Enderwick, C. (2002).** ProteinChip clinical proteomics: computational challenges and solutions. *Biotechniques* **32 Suppl.**, S34–S41.

**Geeraerd, A. H., Valdramidis, V. P., Devlieghere, F., Bernaert, H., Debevere, J. & Van Impe, J. F. (2004).** Development of a novel approach for secondary modelling in predictive microbiology: incorporation of microbiological knowledge in black box polynomial modelling. *Int J Food Microbiol* **91**, 229–244.

**Gerbase, A. C., Rowley, J. T., Heymann, D. H., Berkley, S. F. & Piot, P. (1998).** Global prevalence and incidence estimates of selected curable STDs. *Sex Transm Infect* **74 Suppl 1**, S12–S16.

**Grus, F. H., Joachim, S. C. & Pfeiffer, N. (2003).** Analysis of complex autoantibody repertoires by surface-enhanced laser desorption/ionization-time of flight mass spectrometry. *Proteomics* **3**, 957–961.

**Johnson, R. E., Newhall, W. J., Papp, J. R. & 12 other authors (2002).** Screening tests to detect *Chlamydia trachomatis* and *Neisseria gonorrhoeae* infections – 2002. *MMWR Recomm Rep* **51**, 1–38.

**Khan, J., Wei, J. S., Ringner, M. & 8 other authors (2001).** Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med* **7**, 673–679.

**Knapp, J. S. (1988).** Historical perspectives and identification of *Neisseria* and related species. *Clin Microbiol Rev* **1**, 415–431.

**Lancashire, L., Schmid, O., Shah, H. & Ball, G. (2005).** Classification of bacterial species from proteomic data using combinatorial approaches incorporating artificial neural networks, cluster analysis and principal components analysis. *Bioinformatics* **21**, 2191–2199.

**Mian, S., Ball, G., Hornbuckle, J. & 8 other authors (2003).** A prototype methodology combining surface-enhanced laser desorption/ionization protein chip technology and artificial neural network algorithms to predict the chemoresponsiveness of breast cancer cell lines exposed to Paclitaxel and Doxorubicin under *in vitro* conditions. *Proteomics* **3**, 1725–1737.

**Schwarzer, G., Vach, W. & Schumacher, M. (2000).** On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. *Stat Med* **19**, 541–561.

**Smith, J. M., Smith, N. H., O'Rourke, M. & Spratt, B. G. (1993).** How clonal are bacteria? *Proc Natl Acad Sci U S A* **90**, 4384–4388.

**Tomita, Y., Tomida, S., Hasegawa, Y., Suzuki, Y., Shirakawa, T., Kobayashi, T. & Honda, H. (2004).** Artificial neural network approach for selection of susceptible single nucleotide polymorphisms and construction of prediction model on childhood allergic asthma. *BMC Bioinformatics* **5**, 120.

**Vazquez, J. A., Berron, S., O'Rourke, M., Carpenter, G., Feil, E., Smith, N. H. & Spratt, B. G. (1995).** Interspecies recombination in nature: a meningococcus that has acquired a gonococcal PIB porin. *Mol Microbiol* **15**, 1001–1007.

**Wei, J. T., Zhang, Z., Barnhill, S. D., Madyastha, K. R., Zhang, H. & Oesterling, J. E. (1998).** Understanding artificial neural networks and exploring their potential applications for the practicing urologist. *Urology* **52**, 161–172.