Available online at www.sciencedirect.com
SCIENCE DIRECT®
Talanta 67 (2005) 360–367
Talanta
www.elsevier.com/locate/talanta

Chemical profiling and classification of illicit heroin by principal component analysis, calculation of inter sample correlation and artificial neural networks

Pierre Esseiva [a], Frederic Anglada [a], Laurence Dujourdy [a,1], Franco Taroni [a], Pierre Margot [a], Eric Du Pasquier [b,2], Michael Dawson [b], Claude Roux [b], Philip Doble [b,*]

[a] *Institut de Police Scientifique, University of Lausanne, Switzerland*
[b] *Centre for Forensic Science, Faculty of Science, University of Technology Sydney, P.O. Box 123, Broadway, NSW 2007, Australia*

Received 20 October 2004; received in revised form 4 March 2005; accepted 8 March 2005

Abstract

Artificial neural networks (ANNs) were utilised to validate illicit drug classification in the profiling method used at "Institut de Police Scientifique" of the University of Lausanne (IPS). This method established links between samples using a combination of principal component analysis (PCA) and calculation of a correlation value between samples.

Heroin seizures sent to the IPS laboratory were analysed using gas chromatography (GC) to separate the major alkaloids present in illicit heroin. Statistical analysis was then performed on 3371 samples. Initially, PCA was performed as a preliminary screen to identify samples of a similar chemical profile. A correlation value was then calculated for each sample previously identified with PCA. This correlation value was used to determine links between drug samples. These links were then recorded in an Ibase® database. From this database the notion of "chemical class" arises, where samples with similar chemical profiles are grouped together. Currently, about 20 "chemical classes" have been identified.

The normalised peak areas of six target compounds were then used to train an ANN to classify each sample into its appropriate class. Four hundred and sixty-eight samples were used as a training data set. Sixty samples were treated as blinds and 370 as non-linked samples. The results show that in 96% of cases the neural network attributed the seizure to the right "chemical class".

The application of a neural network was found to be a useful tool to validate the classification of new drug seizures in existing chemical classes. This tool should be increasingly used in such situations involving profile comparisons and classifications.

Keywords: Drug intelligence; Heroin profiling; Neural networks; Principal component analysis

1. Introduction

A variety of analytical techniques have been described in the literature for the analysis of heroin samples, particularly gas chromatography, which gives high resolution of target compounds as well as good sensitivity and reproducibility [1–6].

A simplified method was developed at the "Institut de Police Scientifique" (IPS) of the University of Lausanne that utilises GC–FID for the analysis of heroin samples seized in Switzerland in order to obtain rapid intelligence information. This method separates the major alkaloids, and also allows identification of the principal cutting agent [7–10]. This method requires simple sample preparation when compared to the more common approach of GC analysis of the minor acidic and neutral impurities [11].

* Corresponding author. Tel.: +61 2 9514 1792; fax: +61 295141460.
E-mail address: philip.doble@uts.edu.au (P. Doble).
[1] Present address: Laboratoire de Police Scientifique, Lyon, France.
[2] Present address: Etablissement Cantonal d'Assurance, ECA, Pully, Switzerland.

0039-9140/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2005.03.041

In our study, a preliminary determination of chemical similarity between samples was established via principal component analysis (PCA). Correlation between samples was then calculated by converting the sample data into vector representations, and calculating the square of the cosine of the angle between the vectors. This correlation value gave a measure of similarity between samples. When this correlation value exceeded a certain threshold limit, defined in previous research [7,8], the samples were considered to have a similar chemical profile. Multiple seizures with similar chemical profiles were then considered to belong to a distinct chemical class. Membership of these classes infers a link between samples and is useful for intelligence purposes.

This paper also demonstrates the use of a simulated artificial neural network (ANN) to recognise and validate links between seized heroin samples, which were previously classified by the combination of PCA and calculation of the correlation value. The ANN was trained using six target compounds as input variables and the chemical class as the output variable.

2. Data collection

Since 1992, the IPS has analysed samples of illicit drugs for strategic and operational purposes to control the flow of illicit drugs into Switzerland. The heroin samples were street samples seized by different State Police. Each sample was analysed using the GC–FID technique described in [7]. Each of the six target compounds' normalised area was extracted using specific macros written in Visual Basic® in Excel® software. This chemical data was then transferred to a File Maker Pro 6® database, along with other relevant investigation and chemical information including seizure details, arresting officers' details, number of samples in the seizure, purity, cutting agents, the PCA scores for PC1 and PC2 (allowing the possibility to perform a preliminary selection) and the chemical class. Fig. 1 presents a screen capture of this database.

To date, this database comprises of 8000 heroin samples. Three thousand three hundred and seventy-one samples from the past 3 years were selected from this database for PCA and similarity determination using the correlation value.

3. Results and discussion

3.1. Principal component analysis

PCA allows reduction of a data set by sequential linear transformation of the data where often the first few principal components (PC) retain much of the variability of the original dataset [13,14]. PCA was performed on the peak areas of the six target alkaloids: meconine, acetylcodeine, acetylthebaol, 6-monoacetylmorphine, papaverine and noscapine. The peak areas of each of these compounds were normalised to diacetylmorphine. The data was further normalised to zero mean and unit variance prior to PCA. PCA was performed with Unscrambler® 9.1 from CAMO. Computation of the

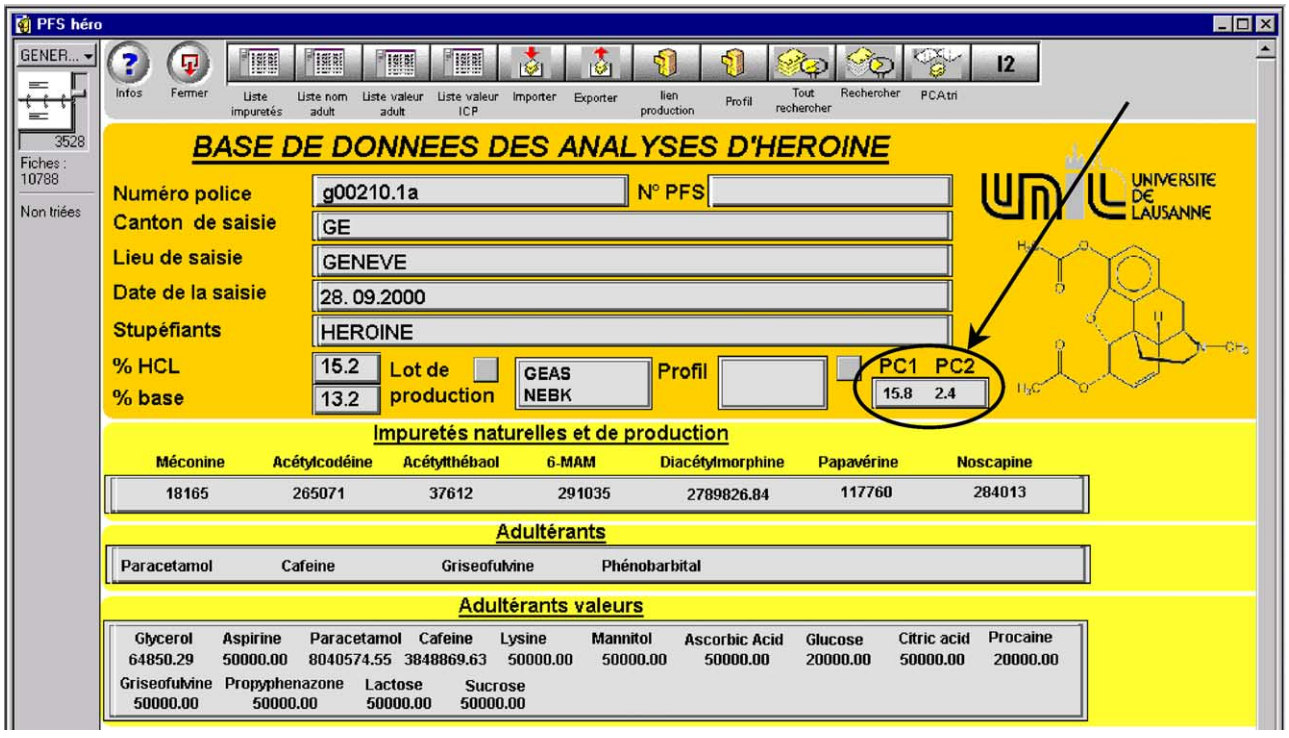

Fig. 1. Screenshot of the File Maker Pro 6® database of a heroin sample. The arrows illustrate the PC1 and PC2 information.

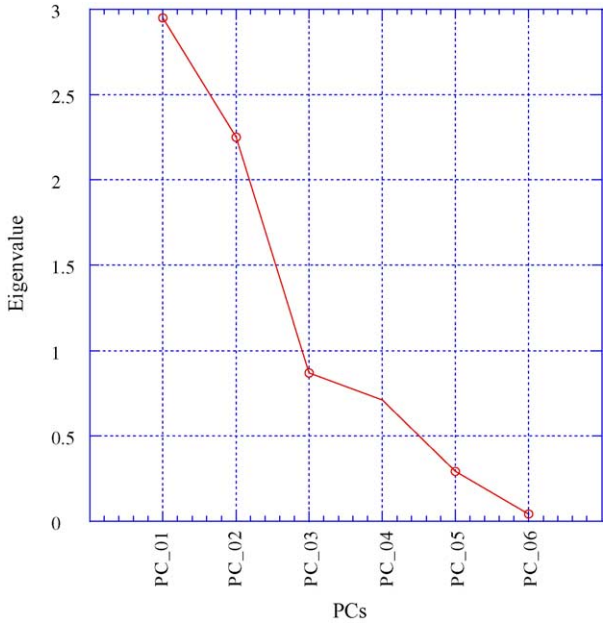

Fig. 2. Scree plot of the eigenvalue vs. principal components.

PCs resulted in the first and second principal components describing 42.7 and 31.7% of the variability in the original observations, respectively, while both principal components account for 74.4% of the total variance. Thus, the first two PCs reduced the six-dimensional data set to a two-dimensional data set, with an average of 25.6% loss of detail (Fig. 2).

Fig. 3(a) shows a scatter plot of PC2 versus PC1 after PCA of the data showing only the samples that belong to distinct chemical classes [7]. Fig. 3(b) is a zoomed view of the data cluster in lower quarter of the plot in Fig. 3(a). Visual inspection of these plots clearly shows clusters, each corresponding to a chemical class. A 98.9% confidence envelope is constructed around each of these clusters based on three times the standard deviation of the PC scores within each cluster. These envelopes are then used to determine the chemical class of new seizures via computation of the PC scores. Seizures falling within these envelopes are then extracted and compared sample by sample with the correlation function as described below and in references [6,7].

It is necessary to perform this initial chemical class selection to remove any unrelated samples before further refinement of link assignment using the correlation measurement, which takes into consideration the complete variance of the data set. Failure to do this results in unnecessary complication of interpretation of the data and is also time-consuming as the calculation of the correlation value is a sample-by-sample comparison.

3.2. Calculation of inter-sample correlation

Consider two vectors as shown in Fig. 4.

The scalar product of the two vectors is:

$$\vec{a} \cdot \vec{b} = ||\vec{a}||\,||\vec{b}|| \cos\theta \tag{1}$$

If the expression of the scalar product according to the vector components compared to an orthonormal base in space is:

$$\vec{a} = \begin{pmatrix} a_1 \\ a_2 \\ \cdots \\ a_n \end{pmatrix} \quad \text{and} \quad \vec{b} = \begin{pmatrix} b_1 \\ b_2 \\ \cdots \\ b_n \end{pmatrix}, \quad \text{then}$$

$$\vec{a} \cdot \vec{b} = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n \tag{2}$$

And the vector norm according to its components in space is:

$$||\vec{a}|| = \sqrt{a_1^2 + a_2^2 + \cdots + a_n^2} \tag{3}$$

Then the square of the cosine of the angle between the two vectors is:

$$\cos^2\theta = \frac{(\vec{a} \cdot \vec{b})^2}{||\vec{a}||^2 ||\vec{b}||^2} \Rightarrow \cos^2\theta = \frac{(a_1 b_1 + a_2 b_2 + \cdots + a_n b_n)^2}{(a_1^2 + a_2^2 + \cdots + a_n^2)(b_1^2 + b_2^2 + \cdots + b_n^2)} \tag{4}$$

Therefore, the correlation value, C, between the two vectors is given by:

$$C = 100 \left[\frac{(a_1 b_1 + a_2 b_2 + \cdots + a_n b_n)^2}{(a_1^2 + a_2^2 + \cdots + a_n^2)(b_1^2 + b_2^2 + \cdots + b_n^2)} \right] \tag{5}$$

Where $a_1, a_2, \ldots, a_n$ represent the values of the variables 1–n for the matrix **a**, respectively, and $b_1, b_2, \ldots, b_n$ represent the values of the variables 1–n for the matrix **b**, respectively.

The samples were considered to be linked, i.e. the same, when a correlation value greater than or equal to 99.8 was obtained. This threshold was determined in a previous study [7].

3.3. Calculation of the discriminating power

It is essential to estimate the capability of the method to discriminate the samples. This is particularly important for the assessment the usefulness of the calculation of the correlation value and its applicability for intelligence. This estimate was evaluated by calculation of the discriminating power by selection of samples analysed from the past 3 years (3371 samples).

Twenty chemical classes were identified in all of the 3371 samples. These classes contained more than one seizure and accounted for approximately 20% of the total samples. For the other 80% of the samples, 238 unique chemical profiles were identified, but no chemical classes were established because the samples of these specific profiles belonged to a single seizure. Therefore, 258 unique chemical profiles were identified in the whole sample set.

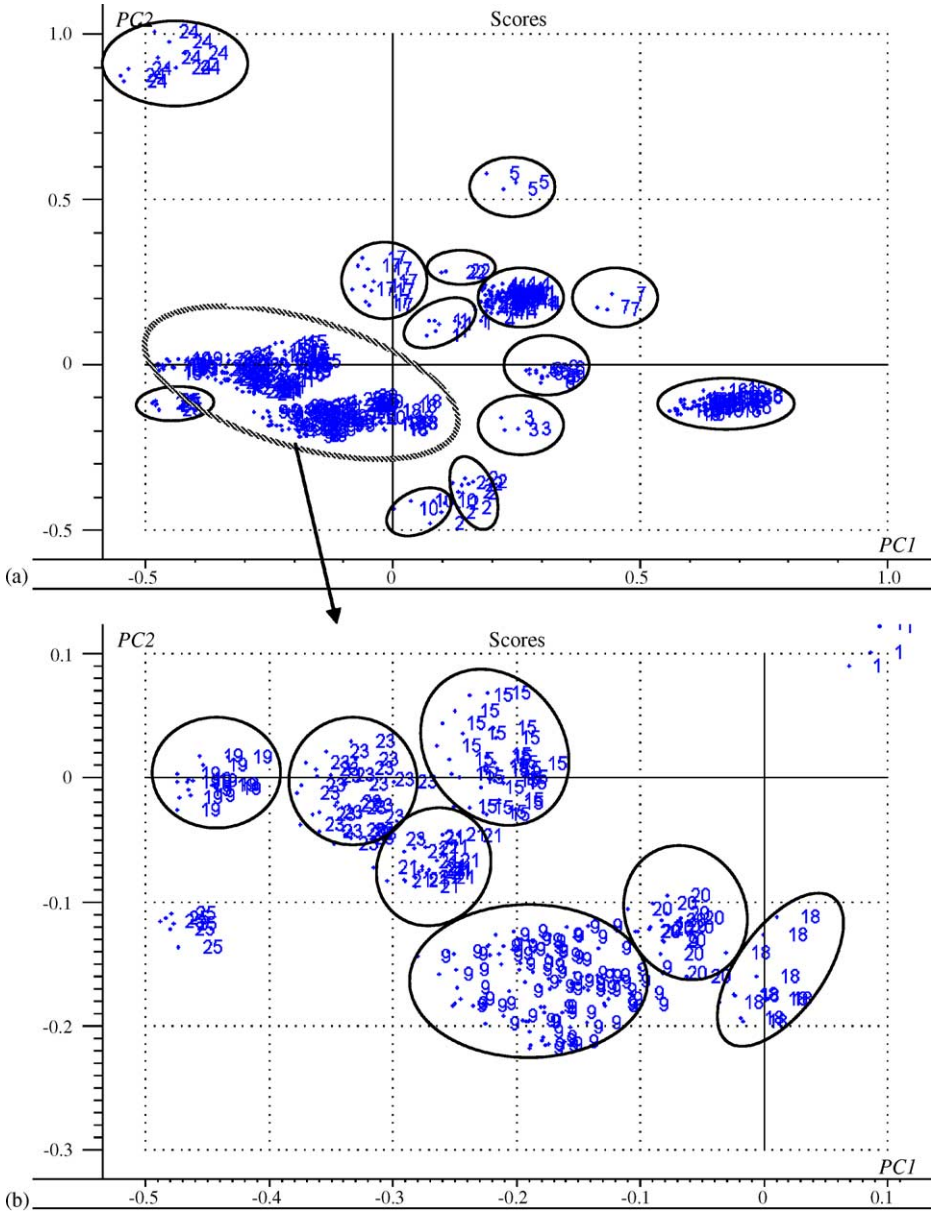

Fig. 3. PC score scatter plot of the samples making up the 20 chemical classes [7].

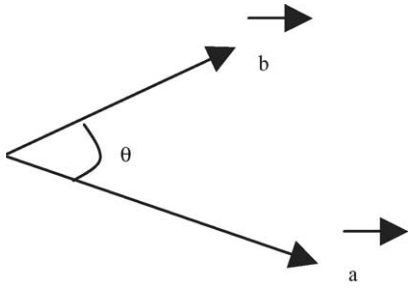

Fig. 4. Diagram of the angle between two vectors.

Assuming that the heroin samples are distributed among the $C_1, C_2, \ldots, C_{258}$ chemical profiles, the proportion in each profile is given by:

$$p_1 = \frac{c_1}{n}$$
$$p_2 = \frac{c_2}{n}$$
$$\vdots$$
$$p_{258} = \frac{c_{258}}{n}$$

where $c_1, c_2, \ldots, c_{258}$, are the number of samples belongings to class $C_1, C_2, \ldots, C_{258}$, respectively, and n is the total number of samples (in this case 3371).

If the population is assumed to be infinite, the overall proportion of combinations ($\hat{Q}$), which results in a match, is given by [15–17]:

$$\hat{Q} = \sum_{j=1}^{n} p_j^2 = 0.00377 \quad (6)$$

If the population is assumed to be finite, as in our case, we use this formula:

$$\hat{Q} = \frac{\sum_{j=1}^{k} \hat{p}_j^2 - \frac{1}{n}}{1 - \frac{1}{n}} \quad (7)$$

And the result is:

$$\hat{Q} = 0.00351$$

Therefore, the probability of randomly finding two samples (belonging to the same profile selected within the database population) using this method is 0.35%. Accordingly, the probability of discrimination between two samples with the selected method is 99.65%. So, a discriminating power equal to 0.9965 means that on average greater than 99% of samples will be discriminated into its unique chemical class.

3.4. Management of chemical classes

Determination of chemical classes is particularly useful when combined with existing information gathered by the police. Recording this information in a meaningful and useful way is crucial for the data to be utilised from a holistic intelligence perspective.

Accordingly, each link was recorded in an Ibase® database as well as additional information including seizure details, such as location, date, arresting officers, quantity, and sample details, such as cutting agents and purities. From an intelligence perspective, Ibase® combined with Analyst Notebook®, provides an easy way to interpret visual representation of the chemical classes, allowing a clickable drill down hierarchical interface as shown in Fig. 5.

Without Ibase® and the Analyst Notebook® it would be difficult to visualise and manage these links. This is particularly helpful to share this information with the police forces.

3.5. Artificial neural networks

Multi-layer perceptron (MLP) and radial basis functions (RBF) neural networks are supervised learning techniques that infer a relationship between input values and output values, which in this case are the normalised peak areas of the heroin target compounds obtained from the GC–FID analysis and the chemical classes, respectively. Therefore, it is necessary to have previously determined the existence of chemical classes to train these networks, i.e. the classification performed by PCA and subsequent calculation of sample similarity. A properly trained network is able to model the function that relates the input variables to the output variables, and can be used to make predictions where the output is not known.

Artificial neural networks use a set of processing elements (or nodes) loosely analogous to neurons in the brain. These nodes are interconnected in a network in such a way that allows patterns to be recognised in the data as the data is introduced to it.

Neural networks have been dealt with extensively in many publications [18–27] and a detailed description is beyond the scope of this paper.

In brief, an ANN consists of neurons arranged in a layered topology containing an input layer, a hidden layer and an output layer which are all interconnected. When the ANN is executed it attempts to identify patterns in the structure of the data by a feed forward iterative process that continually adjusts the weights of each of the neurons in the hidden and output layer to minimise the error of the response surface. The training of an ANN to find a suitable architecture to model the data in question is performed via a heuristic process.

The following demonstrates the use of ANNs to identify and classify heroin samples into their respective chemical classes and/or chemical profiles based on the six compounds normalised peak areas as inputs and the previously determined classes as outputs. Two types of network configurations were tested: multi-layer perceptron and radial basis functions.

The MLP is one of the most popular network architectures (Fig. 6) [18].

MLP networks have either threshold or sigmoidal activations functions. Its greatest strength is the use of non-linear solutions to solve problems. Two training algorithms are generally used, back propagation and conjugate gradient descent. Back propagation involves the calculation of the gradient vector of the error surface that points along the direction of steepest descent. Moving along the vector a short distance will decrease the error. Repeating this process and moving along the vector in shorter distances will eventually find a minimum. Conjugate gradient is a more sophisticated training technique in which the line of descent directions of the error response surface are selected to maintain the second derivative of the error surface at zero. Conjugate gradient typically requires fewer epochs than back propagation, and usually converges to a lower minimum.

Radial basis function neural networks have an input layer of branching nodes, a hidden layer of radial units, each modeling a Gaussian response surface, and an output layer (Fig. 7).

The activation of a hidden unit is determined by the distance between the input vector and the prototype vector. This network uses a two-stage training procedure. In the first stage, the parameters governing the basis functions (corresponding to the hidden units) are determined using relatively fast, unsupervised training methods. The second stage of training then involves the determination of the final-layer weights, which requires the solution of a linear problem, and therefore is also fast [19].

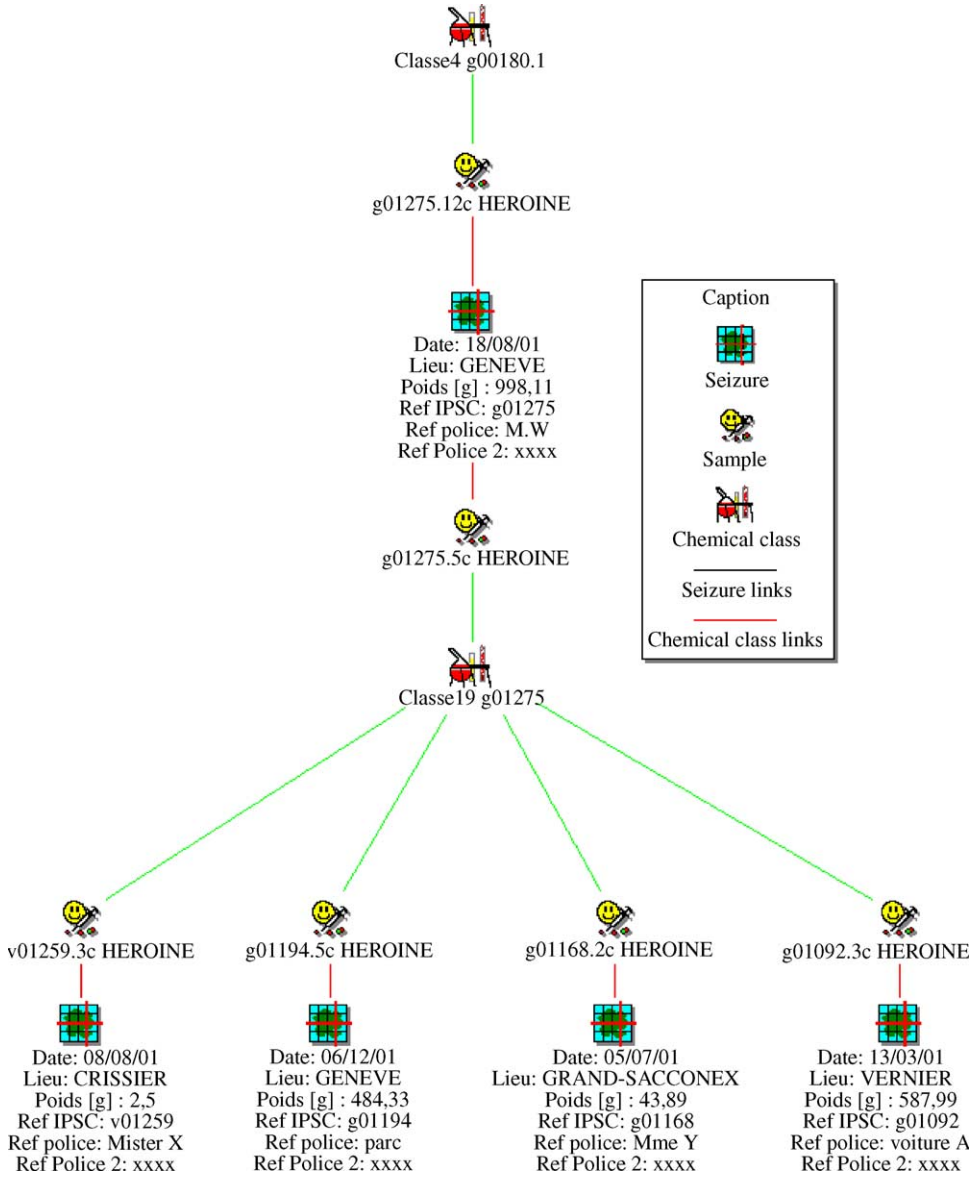

Fig. 5. Ibase® and Analyst Notebook® screenshots detailing a linked heroin seizure.

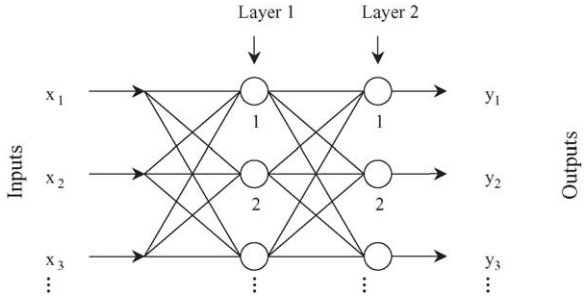

Fig. 6. Architecture of a multilayer perceptron network. The network consists of nodes and interconnecting arcs that form signal paths from left to right through the network.

The neural network software used in this research was Trajan Neural Networks, Version 6.0®. This software has an algorithm designed to mimic the heuristic process. The algorithm searches through the possible ANN architectures and combinations by sequentially changing the number of nodes in the hidden layer for both MLP and RBF networks. Some network configurations are trained a number of times because each training run starts from a random selection of weights on the nodes of the neural network. Approximately 2000 artificial network configurations were tested to find the most efficient network. Each test consisted of different training algorithms and/or network.

The data were broken into a number of groups to facilitate the training of the networks, and to validate the outputs. These were a training set and verification set, which consisted of

Table 1
Best eight networks

Hidden	Training correct (%)	Training wrong (%)	Training unknown (%)	Test correct (%)	Test wrong (%)	Test unknown (%)	False positive (%)
MLP 6:35:20	98.68	0.00	1.32	88.14	3.39	8.47	1.88
MLP 6:23:20	98.95	0.26	0.79	91.53	1.69	6.78	51.08
MLP 6:22:20	97.63	0.53	1.84	88.14	3.39	8.47	35.48
MLP 6:20:20	97.63	0.79	1.58	72.88	3.39	23.73	19.35
RBF 6:120:20	97.38	0.40	2.20	83.05	16.95	0.00	5.01
RBF 6:116:20	96.78	0.40	2.81	84.74	15.26	0.00	5.30
RBF 6:126:20	96.58	0.40	3.01	83.05	1.69	15.25	5.60
RBF 6:121:20	97.38	0.60	2.00	96.61	1.69	1.69	4.12

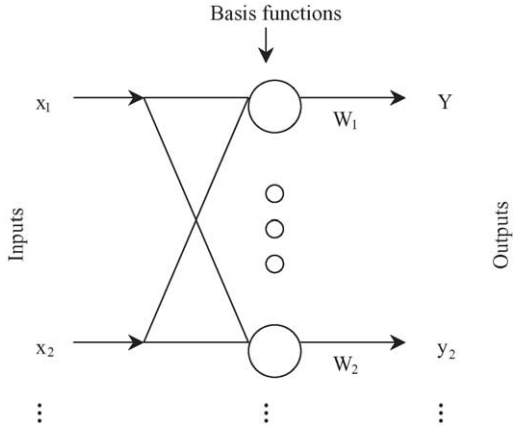

Fig. 7. The traditional radial basis function network. Each of n components of the input vector x_i feeds forward to basis functions whose outputs are linearly combined with weights (W_i) into the network output.

Table 2
ANN topology and training methods for the best eight networks

Type	Hidden	Training[a]
MLP	35	BP50, CG50, CG113b
MLP	23	BP50, CG50, CG96b
MLP	22	BP50, CG74b
MLP	20	BP50, CG75b
RBF	120	KM, KN, PI
RBF	116	KM, KN, PI
RBF	126	KM, KN, PI
RBF	121	SS, EX, PI

KM, K-means, center assignment; KN, K-nearest neighbour, deviation assignment; PI, pseudo-invert, linear least squares optimisation; SS, sub-sample; EX, explicit deviation assignment are training algorithms used by the RBF network.

[a] Indicates how the network was trained, e.g. BP50, CG50, CG113b means the network was trained initially with back propagation for 50 epochs, then conjugated gradient for 50 epochs and then a conjugated gradient for 113 epochs.

330 and 168 samples, respectively, known to be members of the 20 classes. The verification points ensured that the network did not suffer from overlearning [28,29]. A further 60 samples known to be members from 1 of the 20 classes were treated as test samples to ensure that the predictions of class membership made by the ANN were accurate. In addition, 370 known non-linked samples (samples not belonging to the 20 chemical classes) were used to ensure the ANN did not produce false positives.

Table 1 shows the performance of the best eight networks and Table 2 summarizes the architecture of these networks.

As shown in Table 1, the neural network, which gave the best performance, was an RBF consisting in 6 inputs (6 target

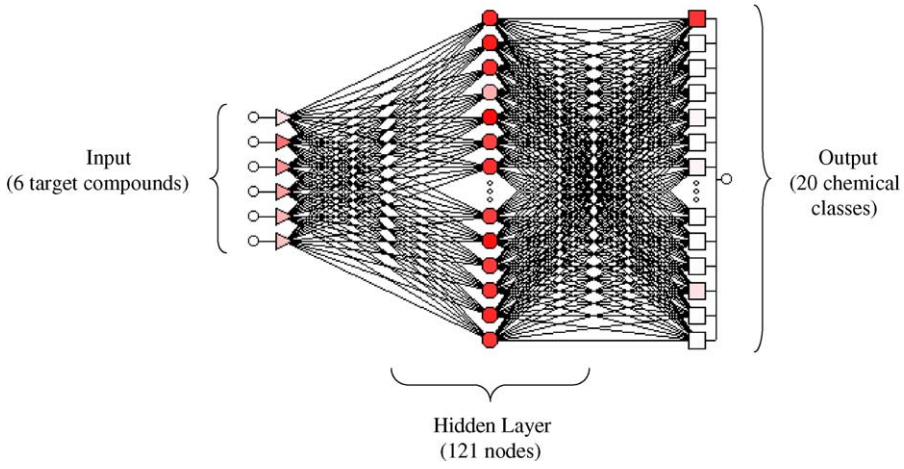

Fig. 8. Schematic of best performing network.

compounds), 1 hidden layer (121 nodes) and 20 outputs (20 chemical classes). The schematic of this neural network is presented in Fig. 8.

This network correctly classified 97.4%, misclassified 0.6% and could not classify 2% of the training and verification samples; 96.6% of the test samples were correctly classified, while 1.7% were misclassified and 1.7% were unknown. Most importantly, this network produced a false positive rate of less than 4%. The utilisation of the ANN model is acceptable in an operational perspective where the information is initially dedicated to the police forces (supporting the inquiry) and not for court purposes. The applicability of the ANN can be extended to new chemical classes via retraining.

4. Conclusions

PCA filtering followed by the calculation of the sample correlation is an efficient and accurate method to identify links between heroin samples. It overcame many of the difficulties associated with the variability of a heroin signature, which can vary slightly from one sample to the next by repeated sampling from the same batch. Of the 1000 heroin samples analysed by GC–FID and selected for the validation of the neural network architecture, 498 were found to group into one of 20 classes. The rest of the samples all had a unique chemical profile and were not linked.

An efficient system to manage the links of the chemical signatures using Ibase®, and visualisation of these links using Analyst Notebook can aid the law enforcement agencies in better interpretation of links between seizures and thus improve intelligence.

Chemical class determination can be achieved using an ANN trained on data in which chemical classes have been previously identified. The ANN can then be used to rapidly validate the previous classification and to classify future seizures.

This tool should be increasingly used in such situations involving profile comparisons as well as profile classifications.

References

[1] B.A. Perillo, R.F.X. Klein, E.S. Franzosa, Forensic Sci. Int. 69 (1994) 1–6.
[2] H. Neuman, Forensic Sci. Int. 69 (1994) 7–16.
[3] F. Besacier, H. Chaudron-Thozet, M. Rousseau-Tsangaris, J. Girard, A. Lamotte, Forensic Sci. Int. 85 (1997) 113–125.
[4] L. Stromberg, L. Lundberg, H. Neumann, B. Bobon, H. Huizer, N.W. Van Der Stelt, Forensic Sci. Int. 114 (2000) 67–88.
[5] M. Chiarotti, N. Fucci, J. Chromatogr. B 773 (1999) 127–136.
[6] H. Huizer, Analytical Studies on Illicit Heroin, Doctorate, Rijswijk, Netherlands, 1988.
[7] P. Esseiva, L. Dujourdy, F. Anglada, F. Taroni, P. Margot, Forensic Sci. Int. 132 (2003) 139–152.
[8] L. Dujourdy, G. Barbati, F. Taroni, O. Guéniat, P. Esseiva, F. Anglada, P. Margot, Forensic Sci. Int. 131 (2003) 171–183.
[9] O. Guéniat, Le profilage de l'heroine et de la cocaine—les methods d'analyse, la modelisation du concept du profilage, la gestion et l'exploitation des liens, Doctorate, Universite de Lausanne, Institut de Police Scientifique et de Criminologie, Switzerland, 2001.
[10] P. Esseiva, L. Dujourdy, F. Anglada, F. Taroni, O. Guéniat, P. Margot, Utilite et gestion de l'information RICPT 55 (2002) 104–111.
[11] R.B. Myors, P.T. Crisp, S.V. Skopec, R.J. Wells, Analyst 126 (2001) 679–689.
[13] P. Legendre, L. Legendre, Numerical Ecology. Developments in Environmental Modelling, vol. 20, Elsevier, 1998.
[14] The Unscrambler® 9.1 CAMO PROCESS AS Nedre Vollgate 8, N-0158 OSLO, Norway.
[15] K.W. Smalldon, A.C. Moffat, J. Forensic Sci. Soc. 13 (1973) 291–295.
[16] D.A. Jones, J. Forensic Sci. Soc. 12 (1972) 355–359.
[17] J.B. Parker, J. Forensic Sci. Soc. 7 (1967) 134–144.
[18] C. Kingston, J. Forensic Sci. 37 (1992) 252–264.
[19] J.F. Casale, J.W. Watterson, J. Forensic Sci. 38 (1993) 292–301.
[20] N.J. Paulsson, F. Winquist, Forensic Sci. Int. 105 (1999) 95–114.
[21] C.S. Tong, K.C. Cheng, Chemometrics Intell. Lab. Syst. 49 (1999) 135–150.
[22] P. Sinha, Forensic Sci. Int. 98 (1998) 67–89.
[23] G. Zeno, J. Keijzer, Forensic Sci. Int. 82 (1996) 21–31.
[24] P. Castellano, S.A. Sridharan, Speech Commun. 18 (1996) 139–149.
[25] Y. Tominaga, Chemometrics Intell. Lab. Syst. 49 (1999) 105–115.
[26] C. Wilson, G. Candela, C. Watson, J. Artif. Neural Network 1 (1992) 1–26.
[27] D. Maio, D. Maltoni, Neural network based minutiae filtering in fingerprints, in: 14th International Conference on Pattern Recognition (ICPR), Brisbane, Australia, 1998, pp. 1654–1658.
[28] C. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, Oxford, 1995.
[29] L. Fausset, Fundamentals of Neural Networks, Architectures, Algorithms and Applications, Prentice Hall, New Jersey, 1994.