I. Kuzmanovski, K. Zdravkova and M. Trpkovska,
S. Afr. J. Chem., 2006, **59**, 12–16,
<http://journals.sabinet.co.za/sajchem/>.

Classification of Urinary Calculi using Feed-Forward Neural Networks

Igor Kuzmanovski,[a]* Katerina Zdravkova[b] and Mira Trpkovska[a]

[a]*Institut za hemija, PMF, Univerzitet 'Sv. Kiril i Metodij', P.O. Box 162, 1001 Skopje, Macedonia.*

[b]*Institut za informatika, PMF, Univerzitet 'Sv. Kiril i Metodij', P.O. Box 162, 1001 Skopje, Macedonia.*

Received 20 July 2005; revised 9 January 2006; accepted 10 January 2006.

ABSTRACT

Recent studies have shown that more than 80% of the analysed samples of urinary calculi in our laboratory were mainly composed of four types of calculi, consisting of the following substances: (1) whewellite and weddellite, (2) whewellite, weddellite and uric acid, (3) whewellite, weddellite and struvite and (4) whewellite, weddellite and carbonate apatite. In this work the results of classification of these types of calculi (using their infrared spectra in the region 1450–450 cm^{-1}) by feed-forward neural networks are presented. Genetic algorithms were used for optimization of neural networks and for selection of the spectral regions most suitable for classification purposes. The generalization abilities of the neural networks were controlled by an early stopping procedure. The best network architecture and the most suitable spectral regions were chosen using twentyfold cross-validation. The cross-validation error for the real samples varies from 5.3% to 5.9% misclassifications, which makes the proposed method a promising tool for the identification of these types of calculi.

KEY WORDS

Urinary calculi, infrared spectroscopy, classification, neural networks, variable selection, genetic algorithms.

1. Introduction

Owing to the considerable advances in instrumentation used in clinical laboratories today, it is easy and relatively inexpensive to obtain great quantities of data. However, the main problem with these data is their suitable evaluation and interpretation. Since the beginning of the 1980s with the introduction of computerized instrumentation in clinical laboratories many computerized methods for the determination of the composition of urinary calculi have been developed.[1–11] Most of them are suitable for the determination of the qualitative composition of the calculi comparing the collected data with the data stored in the database.

Determination of the composition of the calculi is important due to the fact that it could help in finding the reasons for the occurrence of the calculi and could help determine suitable diets in order to prevent their further recurrence. One of the most suitable instrumental techniques for the analysis of urinary calculi is infrared spectroscopy.[12] This method, coupled with chemometric techniques such as factor-based methods,[5,7,9,13] as well as those based on artificial neural networks (ANN)[8–11] have been previously used for quantitative and qualitative analysis of urinary calculi.

The specific problem in this study was to identify the type of the calculi according to their infrared spectra. Hence we are dealing with the classification problem. A recent study[14] revealed that more than 80% of the samples were composed of mainly four types of calculi: (1) whewellite and weddellite, (2) whewellite, weddellite and uric acid, (3) whewellite, weddellite and struvite and (4) whewellite, weddellite and carbonate apatite. Because of the low frequency of appearance at this stage we do not have considerable numbers of other types of calculi so we will focus our study on the above noted types.

* To whom correspondence should be addressed.
E–mail: shigor@iunona.pmf.ukim.edu.mk;
tel. ++389-2-3117055; fax: ++389-2-3226865

In this work the results are presented of our attempt to develop a method for the classification of urinary calculi using artificial neural networks. Since ANN and their application in chemistry are well documented[15,16] here only the procedure for their optimization will be explained in detail.

The network architecture which shows the best performance, as well as the selection of the most suitable wavenumber intervals, was determined using genetic algorithms (GA). The genetic algorithms are explained in more detail below.

Genetic algorithms are optimisation tools[17–20] based on a stochastic search of the optimal parameters guided by the principles of natural evolution and genetics. They are especially suitable for optimization of discrete functions. GA are valuable optimization tools allowing a search for the optimal parameters to be achieved without the need to run every single parameter permutation. When GA are used the parameters are represented by binary strings called genes. The genes from different parameters are combined into chromosomes. The algorithm searches, for the chromosome(s) with best performance(s) among the group of chromosomes so called population. Using natural selection and genetic operations (crossover and mutation), chromosomes with better performances are selected. Natural selection retains the genetic material from chromosomes with best performances for the reproduction procedure. During crossover new chromosomes are produced by swapping genetic material between selected parent chromosomes. This operation, in general, produces individuals with better performances and improves the algorithm convergence. Different crossover techniques can be found in the literature.[20] One-point crossover (Fig. 1a) consists of randomly partitioning the binary strings that represent the parent chromosomes into two sections. New chromosomes are formed by swapping the genetic material. They contain genetic material from both parent chromosomes. Two-point cross-over (Fig. 1b) is similar to the previously explained procedure;

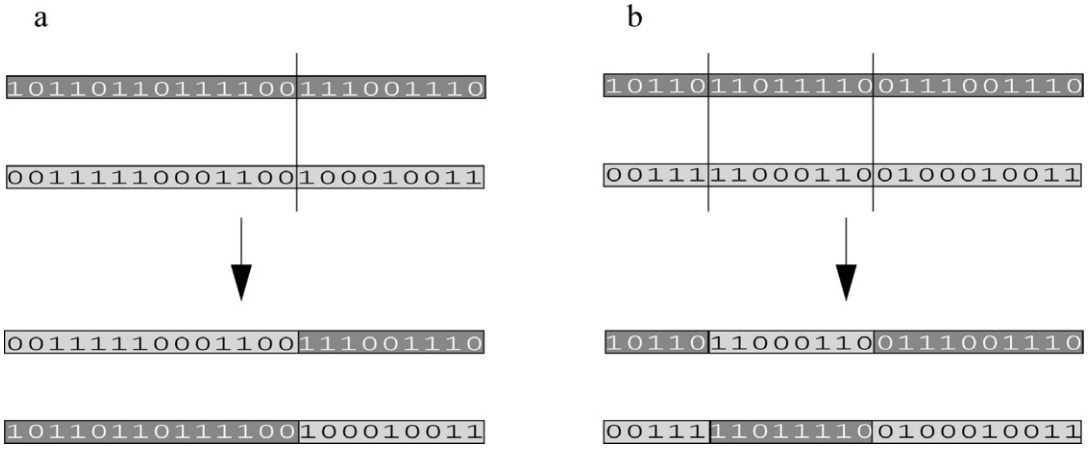

Figure 1 Crossover techniques: (**a**) single point crossover; (**b**) double point crossover.

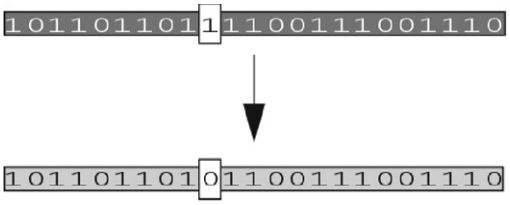

Figure 2 Mutation of the gene in the chromosome.

however, in this case the parent chromosomes are split into three parts, and newly formed chromosomes contain two parts from one parent and one part from the other parent chromosome. Mutation is the procedure of random alternation of one or more genes in the chromosomes (Fig. 2). This operation prevents the GA from converging too quickly in a small area in the parameter search space. The probability at which mutation occurs in the whole population is defined by the user.

2. Experimental

A total number of 179 mixtures was used for the training of ANN. Some of the mixtures (whewellite, weddellite and uric acid; whewellite, weddellite and carbonate apatite) were used in our previous works,[9,10] while additional mixtures consisting of whewellite and weddellite as well as of whewellite, weddellite and struvite were prepared. The mixture design for the samples composed of oxalates and uric acid are given in Fig. 3a. The mixture designs for the samples consisting of whewellite, weddellite and carbonate apatite as well as of whewellite, weddellite and struvite are presented in Fig. 3b. Twelve additional samples of whewellite and weddellite were prepared with the mass fraction of the constituents varying from 0 to 1.

Whewellite, weddellite, carbonate apatite and struvite were synthesized according to procedures found in the literature,[21–22] while uric acid was a Merck product (Darmstadt, 99.9%). The spectra of these compounds were compared using a digital database of infrared spectra of constituents of urinary calculi from Dao and Daudon.[23]

In this study we used 160 infrared spectra of the urinary calculi collected in the period between 1996 and 2003 in our laboratory.[14] The qualitative composition of these samples (depending on the size of the calculi) was determined if possible by target factor analysis.[5,13] In all other cases the composition was determined by comparing different spectral regions of the sample with the database of infrared spectra from Dao and Daudon.[23] The data set consisted of 47 samples of whewellite and weddellite type of calculi, 20 samples of whewellite, weddellite and uric acid type of calculi, 11 samples consisted of whewellite, weddellite and struvite, while 82 belonged to whewellite, weddellite and carbonate apatite type of calculi.

Prior to recording the infrared spectra, up to 2 mg of the sample was homogenized with 250 mg spectroscopy grade KBr. Pellets were prepared after application of pressure of 10 tonne per square centimetre on homogenized mixtures. The infrared spectra were recorded at room temperature in the region of 1450–450 cm^{-1}, using a Perkin–Elmer System 2000 FTIR instrument (with resolution of 4 cm^{-1} and sampling interval of 1 cm^{-1}), with 32 scans for the samples and 32 scans for the background (atmosphere and 250 mg KBr pellet). If the maximum value of

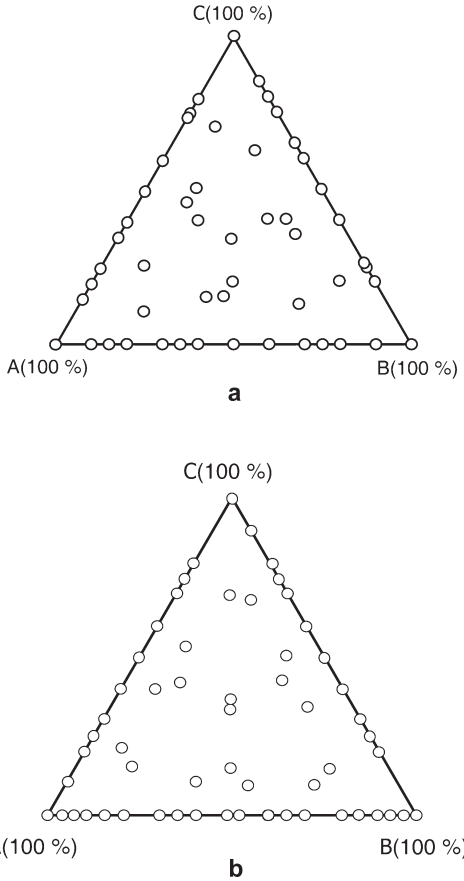

Figure 3 Mixture designs for the training samples prepared from binary and ternary mixtures of (**a**) whewellite, weddellite, carbonate apatite; (**b**) whewellite, weddellite and carbonate apatite or struvite. (A – whewellite; B – weddellite; C – uric acid, carbonate apatite or struvite).

the absorbance in the recorded spectrum exceeded 1, the mass of the sample in the pellet was proportionally reduced in order to achieve the desired maximum value of absorbance.

3. Data Analysis

3.1. Preprocessing

Before the optimization of ANN started, proper preprocessing of the data was applied. The spectra of the prepared mixtures for the training of the networks as well as the spectra of the urinary calculi were stored in a single data matrix (D). The rows in the data matrix correspond to infrared spectra of different samples, and the columns correspond to different wavenumbers. The spectra were offset corrected and normalized to unit area. In order to make further calculation faster the columns in the matrix D were reduced according to the equation:

$$d^r_{i,m} = \frac{\sum_{j=(m-1)\cdot 10+1}^{m\,10} d_{i,j}}{10}. \quad (1)$$

$d_{i,j}$ in Equation (1) represents the data from the normalized matrix, i is the sample number, j represents absorbance values at different wavenumbers, while $d^r_{i,m}$ are data from the i-th sample in the reduced data matrix. Then the variables in the reduced data matrix were autoscaled. The autoscaling inflates the error in the baseline region; however, since we used genetic algorithms not only for finding the optimal ANN architecture but also for variable selection, here it helps for better selection of the spectral regions which carry information suitable for classification purposes.

In order to extract as much information as possible within the fewest possible data points and to make the training process faster, principal component analysis (PCA) was applied. Calculated principal components were used for the training of the ANN.

The output data were stored in another data matrix. The composition of the samples was expressed using a unit vector with length four.

3.2. Artificial Neural Networks

Throughout this work a three-layered feed-forward neural network with sigmoid transfer function in the hidden layer and linear transfer function in the output layer was used. The network architecture with the best performance was searched, changing the number of input neurons (principal components) and number of hidden neurons. The number of output neurons was fixed at four. Each output neuron serves as an indicator for different types of calculi.

The weights and the biases of the networks were initialized according to the Nguyen-Widrow algorithm[24] to force the active regions of the layer's neurons to be distributed roughly evenly over the input space. The weights and biases of the networks were adjusted using the Levenberg-Marquardt algorithm[25] for back-propagation of error. In order to avoid overtraining, which could produce networks with poor generalization abilities, an early stopping procedure[26] was applied. The procedure requires division of the data into three sets: training, validation and test sets. When early stopping is used the weights and biases of the networks are adjusted using the training set. The validation set serves to monitor the performances during the training process. In the beginning of the training process the error in the validation set decreases together with the error in the training set. However, when the network starts to overfit the data the error in the validation set starts increasing. This finding was used to control the generalization abilities of the networks. The network training was stopped when all the samples were correctly classified or if in ten consecutive iteration cycles the number of misclassified samples in the validation set increased. In the latter case, the weights and biases that correspond to a minimum in the validation error are restored. The test set serves to check the performances of the trained network.

Table 1 Percentage variance captured by first 16 principal components obtained by decomposition of the data matrix.

Number of PCs	Percentage variance	Cumulative percentage variance
1	43.6	43.6
2	22.2	65.8
3	13.0	78.8
4	10.6	89.4
5	6.0	95.4
6	1.4	96.7
7	1.1	97.9
8	0.6	98.4
9	0.4	98.9
10	0.3	99.1
11	0.2	99.3
12	0.2	99.5
13	0.2	99.7
14	0.1	99.9
15	0.0	100.0
16	0.0	100.0

4. Results and Discussion

As previously stated, the genetic algorithm was used for variable selection as well as for a search of the optimal network architecture. The percentage variances captured by each principal component obtained by decomposition of the reduced data matrix are presented in Table 1. It can be seen that 15 principal components carry 100% of the variance in the data matrix. These principal components are usually used to determine the optimal number of input neurons.[10] However, in our case while using genetic algorithms where the genes are represented using binary strings the number of possible combinations is 2^n, where n is a positive integer, so we decided to change the number of input neurons in the interval from 1 to 16. The optimal number of hidden neurons of the networks was searched in the same range. Together with the genes representing the absorbances at 100 different wavenumber values, the chromosomes consisted of a total number of 108 genes (four genes were used to determine the optimal number of hidden neurons and the other four to determine the optimal number of input neurons). The initial population of 80 chromosomes was randomly generated.

Prior to the training of different network architectures, variable selection was performed, followed by decomposition of the data matrix (consisting of selected wavenumber regions) by PCA. The principal components obtained were used for training of ANN. The performances of each chromosome were obtained as an average number of misclassified samples after twentyfold cross-validation. The weights and biases were adjusted using a training set consisting of the 179 prepared mixtures, while the 160 real samples were randomised and divided into five subsets. Four of the subsets (80% of the samples) formed the validation set, while the fifth was used as a test set. During five consecutive optimizations of ANN (using network architecture as well as the wavenumber intervals determined by the same chromosome) each of the five subsets was used once as a test set. After five repetitions of the cross-validation the order of the samples in the data matrix was randomly changed.

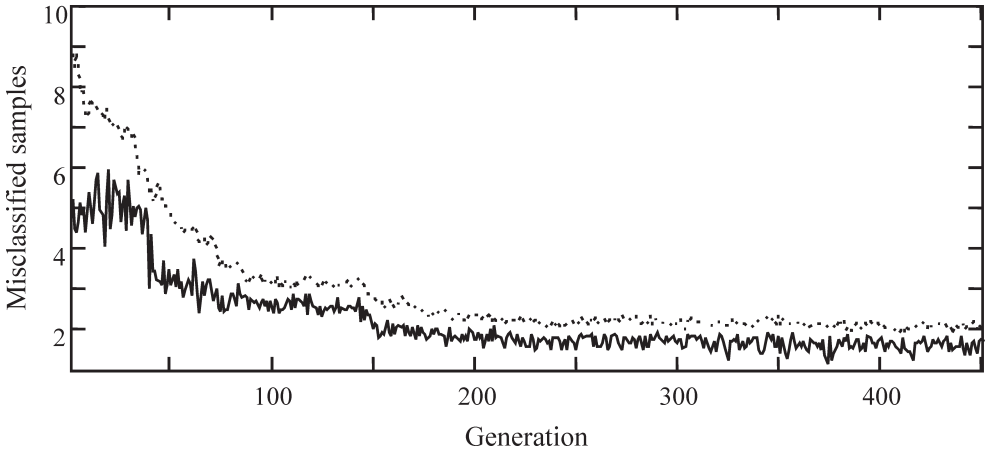

Figure 4 Average number of misclassified samples in the chromosomes used as parents (- - - -) and the average performances of the best chromosome in the population in each generation (——).

The genetic algorithm was repeated within 450 generations. The population consisted of 80 chromosomes. During each generation the 16 chromosomes (20% of the whole population) with best performances were kept unchanged. The mating pairs were formed with selection of chromosomes according to the 'roulette wheel selection' rule.[20] When this selection rule is used all chromosomes are placed on the 'wheel'. The fraction of the wheel which belongs to a different chromosome is proportional to its performance (better performances – larger fraction). The use of this technique allows a better propagation of genetic material from chromosomes with better performances to be favourable compared with the rest of the chromosomes.

The genetic material among the chromosomes was changed using the two-point crossover technique (Fig. 1b). The initial mutation rate was 0.10 and it decreased linearly to 0.05 in generation 150. After that the mutation rate was kept at 0.05.

Because the weights and biases were randomly initialized and also because the validation and test sets were randomly generated before each optimization the chromosomes which showed best performance, even if the optimization was repeated twenty times, did not always show the same performances in the successive generations. This may also serve as an explanation why the chromosomes with best performances presented in the Fig. 4 did not gradually approach a minimal average classification error for the test set. That is why the intermediate results for the chromosomes, and the chromosomes themselves which showed the best results, were saved.

The absorbance intervals from four chromosomes with best performances are presented in Fig. 5. The average percentage of misclassified samples for these solutions as well as the optimal number of input and hidden neurons are presented in Table 2. The average percentage of misclassified samples was calculated

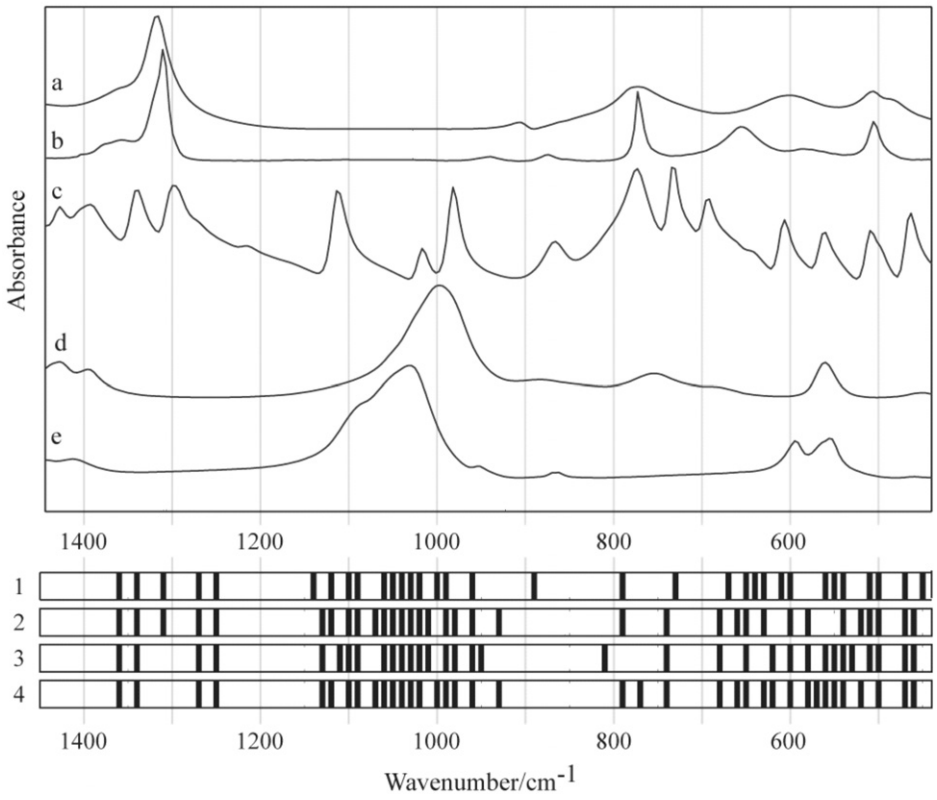

Figure 5 Infrared spectra of five substances used in this study: (**a**) whewellite; (**b**) weddellite; (**c**) uric acid; (**d**) struvite; (**e**) carbonate apatite, and the four best solutions with wavenumber intervals selected by genetic algorithms.

Table 2 Average percentage of misclassified samples in the test set determined by two-hundredfold cross-validation (ww = whewellite and weddellite; wwu = whewellite, weddellite and uric acid; wwc = whewellite, weddellite and carbonate apatite; wws = whewellite, weddellite and struvite).

Solution	Average misclassification error					Network architecture	
	Overall	ww	wwu	wws	wwc	Input neurons	Hidden neurons
1	5.9	7.1	1.6	2.0	6.8	3	12
2	5.8	4.3	1.1	2.1	8.3	3	14
3	5.3	3.8	0.8	2.7	7.6	3	14
4	5.7	4.5	0.9	4.3	7.7	3	16

using the test set after two-hundredfold cross-validation. Note that the overall performance varies from 5.3 to 5.9%. Examining the classification performances for different types of calculi we note that the best accuracy is obtained for whewellite, weddellite and uric acid (0.8–1.6%) and whewellite, weddellite and struvite (2.0–4.3%). The average percentages of misclassified samples for whewellite and weddellite type of calculi vary from 3.8 to 7.1%, while the percentage of misclassified samples of calculi composed of whewellite, weddellite and carbonate apatite is in the interval between 6.8 and 7.7% and is the highest compared with the other types of calculi.

5. Conclusion

Here we have shown that feed-forward neural networks, coupled with genetic algorithms for finding the optimal network architecture as well as for variable selection, have been successfully used for the classification of the infrared spectra of urinary calculi. The low overall percentage of misclassified samples (varying from 5.3% to 5.9%) makes this procedure a promising tool for the identification of these types of calculi. Although, due to the limited number of samples from other (less frequent) types of calculi, this study focused on the classification of the four most frequent types,[14] the procedure described here could be easily extended for classification of other types of urinary calculi.

References

1 A. Hesse, M. Gergeleit, P. Schüller and K. Möller, *J. Clin. Chem. Clin. Biochem.*, 1989, **27**, 639–642.

2 G. Rebentisch, B. Vetter and J. Muche, *Laboratoriumsmedizin*, 1991, **15**, 598–602.

3 G. Rebentisch, M. Doll and J. Muche, *Laboratoriumsmedizin,* 1992, **16**, 224–228.

4 E. Peuchant, X. Heches, D. Sess and M. Clerc, *Clin. Chim. Acta,* 1992, **205**, 19–30.

5 I. Kuzmanovski, M. Trpkovska, B. Šoptrajanov and V. Stefov, *Vib. Spectrosc.,* 1999, **19**, 249–253.

6 H. Hobert and K. Meyer., *Fresenius' J. Anal. Chem.,* 1992, **334**, 178–185.

7 M. Volmer, A. Block, B.G. Wolthers, A.J. de Ruiter, D.A. Doornbos and W. van der Slik, *Clin. Chem.,* 1993, **39**, 948–954.

8 M. Volmer, A. Block, H.J. Metting, T.H.Y. de Haan, P.M.J. Coenegracht and W. van der Slik, *Clin. Chem.,* 1994, **40**, 1692–1697.

9 I. Kuzmanovski, Z. Zografski, M. Trpkovska, B. Šoptrajanov and V. Stefov, *Fresenius' J. Anal. Chem.,* 2001, **370**, 919–923.

10 I. Kuzmanovski, M. Trpkovska, B. Šoptrajanov and V. Stefov, *Anal. Chim. Acta,* 2003, **491**, 211–218.

11 M. Volmer, J.C. de Vries and H.M. Goldschmidt, *Clin. Chem.,* 2001, **47**, 1287–1296.

12 M. Daudon and R.J. Reveillaud, *Presse Med.,* 1987, **16**, 627–631.

13 E.R. Malinowski, *Factor Analysis in Chemistry,* 2nd edn., John Wiley, New York, NY, USA, 1991, p. 32.

14 I. Kuzmanovski, M. Trpkovska and B. Šoptrajanov, *Maked. Med. Pregled,* 1999, **53**, 251–255.

15 A. Bos, M. Bos and W.E. van der Linden, *Anal. Chim. Acta,* 1992, **256**, 133–144.

16 J. Zupan and J. Gasteiger, *Neural Networks in Chemistry and Drug Design,* VCH, Weinheim, Germany, 1999.

17 C.B. Lucasius and G. Kateman., *Chemom. Intell. Lab. Syst.,* 1993, **19**, 1–33.

18 C.B. Lucasius and G. Kateman, *Trends Anal. Chem.,* 1991, **10**, 254–261.

19 C.B. Lucasius, M.L.M. Becker and G. Kateman, *Anal. Chim. Acta,* 1994, **286**, 135–153.

20 K.E. Kinnear, *Advances in Genetic Programming,* MIT Press, Cambridge, MA, USA, 1994, p. 87.

21 P. Brown, D. Ackermann and B. Finlayson, *J. Cryst. Growth.,* 1989, **98**, 285–292.

22 M. Santos and P.F. González-Diaz, *Inorg. Chem.,* 1977, **16**, 2131–2133.

23 N.Q. Dao and M. Daudon, *Infrared and Raman Spectra of Calculi,* Elsevier, Paris, France, 1997.

24 D. Nguyen and B. Widrow, *Proc. Int. Joint Conf. Neural Networks,* 1990, **3**, 21–26.

25 M.T. Hagan and M. Menhaj, *IEEE Trans. Neural Networks,* 1994, **5**, 989.

26 H. Demuth and M. Beale, *Neural Network Toolbox,* Mathworks, Natick, MA, USA, 1997, pp. 5–39.