

Optimization of supervised self-organizing maps with genetic algorithms for classification of urinary calculi

Igor Kuzmanovski^{a,*}, Mira Trpkovska^a, Bojan Šoptrajanov^{a,b}

^aInstitut za hemija, PMF, Univerzitet 'Sv. Kiril i Metodij', P.O. Box 162, 1001 Skopje, Macedonia

^bMakedonska akademija na naukite i umetnostite, 1000 Skopje, Macedonia

Received 20 December 2004; accepted 25 January 2005

Available online 8 March 2005

Abstract

Supervised self-organizing maps were used for classification of 160 infrared spectra of urinary calculi composed of calcium oxalates (whewellite and weddellite), pure or in binary or ternary mixtures with carbonate apatite, struvite or uric acid. The study was focused on such calculi since more than 80% of the samples analyzed contained some or all of the above-mentioned constituents. The classification was done on the basis of the infrared spectra in the 1450–450 cm⁻¹ region. Two procedures were used in order to find the most suitable size and for optimizing the self-organizing map of which that using the genetic algorithms gave better results. Using this procedure several sets of solutions with zero misclassifications were obtained. Thus, the self-organizing maps may be considered as a promising tool for qualitative analysis of urinary calculi.

© 2005 Elsevier B.V. All rights reserved.

Keywords: Urinary calculi; Classification; Self-organizing maps; Supervised self-organizing maps; Genetic algorithms; Infrared spectroscopy

1. Introduction

Urolithiasis (occurrence of urinary calculi) affects from 4 to 20% of the population according to country [1]. Thus the determination of the urinary calculi composition is important in clinical laboratories because it can provide information about the development of the calculi, about further treatments of the patients and about means (e.g. suitable diet) by which further reoccurrence of the urolithiasis could be prevented.

Since the infrared and Raman spectra are characteristic for a given compound [1,2], vibrational spectroscopy is one of the few instrumental methods suitable for analysis of urinary calculi by providing information about the exact chemical individuality of the constituents. Due to the importance of determining the composition of the calculi, many computerized methods have been developed [3–12]. Most of the methods found in the literature are based on

comparison of the sample spectra with the library of the spectra of urinary calculi or algorithm schemes where the calculi are classified according to the presence, absence or position of the band maxima, while other are based on principal component analysis [9], target factor analysis [10] and back-propagation neural networks [11,12].

In the last decade self-organizing maps (SOMs) have become a valuable tool for chemometricians [13–22], most often for unsupervised classification purposes [13–20], for process/reaction monitoring [21,22] and as a tool for variable selection [23]. The theoretical background for the self-organizing maps and their application in chemistry is in details described in the literature [24–26].

In this paper, an attempt was made to apply supervised self-organizing maps [26,27] for classification of 160 infrared spectra of four types of urinary calculi consisting of calcium oxalates (whewellite and weddellite) and/or their mixtures as well as of these two substances in mixtures with carbonate apatite, struvite or uric acid. The calculi used in this work have been analyzed in our laboratory since 1996 and these five substances were found in 80% of all analyzed calculi either as single constituents or in binary or ternary mixtures [28].

* Corresponding author. Tel.: +389 2311 7055; fax: +389 2322 6865.
E-mail address: shigor@iunona.pmf.ukim.edu.mk (I. Kuzmanovski).

2. Experimental

The infrared spectra of the samples were recorded (in the 1450–450 cm^{-1} region) on a Perkin–Elmer System 2000 Fourier-transform infrared spectrometer with a resolution of 4 cm^{-1} and sampling interval of 1 cm^{-1} . The samples were prepared as KBr pellets using 2 mg of homogenized sample and 250 mg spectroscopy-grade KBr. If the maximum value of 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.

For training of the supervised self-organizing maps (SOMs) prepared mixtures of whewellite, weddellite, carbonate apatite, struvite and uric acid were used. Whewellite, weddellite, carbonate apatite and struvite were synthesized according to procedures found in the literature [29–31] while anhydrous uric acid was a Merck product. The infrared spectra of these substances were compared with those in the database of infrared spectra from Dao and Daudon [1]. The comparison showed that the desired constituents have indeed been prepared and that the infrared spectra are of quality comparable to that in the database.

Some of the samples utilized for training were used in our previous works [11,12] but additional 69 mixtures of whewellite and weddellite as well as of whewellite, weddellite and struvite were prepared amounting to a total number of 179 mixtures. The infrared spectra of these mixtures were recorded and used for training of the self-organizing maps.

The infrared spectra of the samples of urinary calculi were recorded using the same procedure as that described above. Whenever possible (depending of the size of the calculi) infrared spectra were recorded from different calculi layers and then target factor analysis was used for determination of their qualitative composition [10]. In cases where the sample size did not permit application of the target factor analysis, the composition of the calculi was determined by comparing different regions of the sample spectra with the database by Dao and Daudon [1].

All together, 160 infrared spectra of urinary calculi were recorded and used for evaluating the performances of the optimized SOMs. Among these samples, 47 belonged to the whewellite-weddellite type of calculi, 20 samples to whewellite and weddellite in presence of uric acid, 11 samples consisted of oxalates and struvite, and 82 samples of oxalates and carbonate apatite.

3. Data analysis

3.1. Preprocessing

Prior to training the SOMs, the collected data were preprocessed. The infrared spectra were normalized to unit length and were stored in a single data matrix (D).

In order to make further calculations faster, the obtained data were reduced from 1000 to 100 absorbance values according to the following equation:

$$d'_{i,m} = \frac{\sum_{j=(m-1)10+1}^{m10} d_{i,j}}{10} \quad (1)$$

where $d_{i,j}$ represents the data from the preprocessed matrix, i is the sample number, j represent the absorbance values at different wavenumbers, while $d'_{i,m}$ is data from the i -th sample in the reduced data matrix.

Then the variables in the reduced data matrix were autoscaled. In order to extract as much as possible information in as less as possible data points and to make the training process faster, principal component analysis (PCA) was applied.

3.2. Genetic algorithms

In order to obtain as good results as possible, genetic algorithms were applied for the wavenumber selection as well as for selecting the most suitable training parameters and map size. It should be pointed out that the genetic algorithms have been proven to be an effective optimization tool [32–34] allowing relatively fast convergence without the need of running every permutation of variables. In the chemistry literature, the theory and use of genetic algorithms as a variable selection tool has been reported several times [35–41] so that only the procedure used in this work is explained here.

An initial population of eighty chromosomes was randomly generated. Each chromosome was represented using a binary vector with length of 126 genes. The first 100 bits in the binary vectors represented absorbances at different wavenumber values while the presence of the corresponding wavenumber interval was coded with 1 and its absence with 0.

Other genes were used for:

- selection of the most suitable number of principal components (PCs) used for training of the SOMs—four genes (from 1 up to 16 PCs);
- determining of the map size—eight genes were used (four genes for length and four genes for width); these parameters were changed in the interval from 4 to 19;
- determining the optimal number of iteration cycles for the ‘rough’ training phase (six genes); this parameter was searched in the interval between 1 and 64;
- finding the optimal number of iteration cycles for the ‘fine’ training phase; eight genes were used and the number of training cycles in this phase was changed in the interval between 1 and 256 increased by twice the number of training cycles in the ‘rough’ training phase.

The number of misclassified samples obtained by the SOM trained with parameters determined by each chromosome was used as a measure for its fitness. After calculating

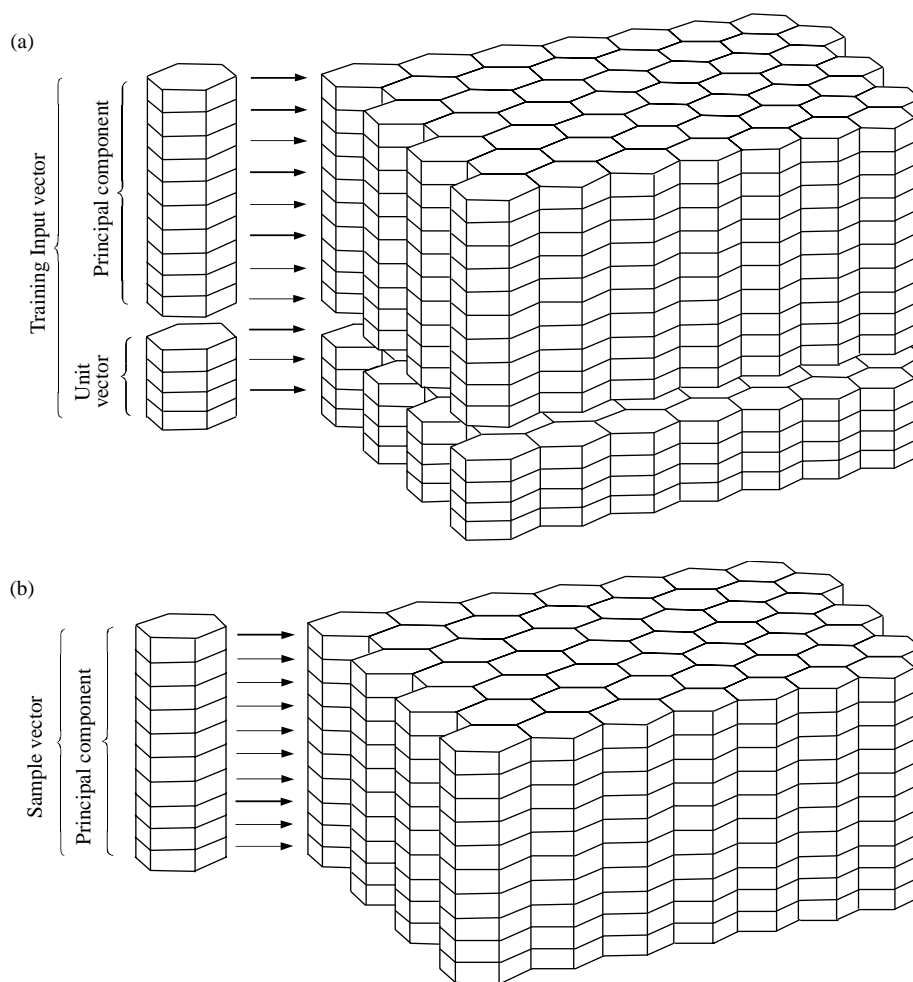


Fig. 1. Illustration of training (a) and prediction phase (b) for supervised self-organizing maps.

the fitness of the whole population, sixteen chromosomes (20% of the total population) with the best performances were selected (in what follows, these chromosomes are referred to as *studs*). The studs were kept unchanged in each successive generation until a different chromosome(s) produced better performances. In such a case, the better chromosome(s) would replace the stud(s). New offspring chromosomes were then created by the two-point crossover technique, which means that two random values between 1 and 126 were chosen. The values in the parent chromosomes between these two values were exchanged to form new chromosomes. After that, the chromosomes were mutated in order to prevent the genetic algorithm from converging too fast in the search space.

All the calculations were done in a Matlab environment [42] using the Self-Organizing Maps Toolbox by Vesanto [43] and Genetic Algorithm Toolbox [44].

3.3. Supervised self-organizing maps

Self-organizing maps were initially developed as an algorithm for unsupervised learning. But in the cases where poor class separation is obtained, applying slight

modifications of the algorithm could transform SOMs as a tool for supervised classification [27]. In order to make SOMs supervised, the input vectors for the samples in the training set \mathbf{d}_s (in our case—the principal components of the corresponding samples), were augmented by a unit vector \mathbf{d}_u (Fig. 1(a)) with its components assigned into one of the four classes of urinary calculi. In the present study each ‘1’ in the unit vector was multiplied by the maximal value in the data matrix consisting of PCs extracted from the training set. During the phase of prediction the part of the weight vectors of SOMs that correspond to unit vector is excluded (Fig. 1(b)). In other words, for each sample in the training set \mathbf{d}_s the corresponding \mathbf{d}_u must be used during the training while during the recognition of an unknown sample \mathbf{x} only the \mathbf{x}_s part is compared with the corresponding part of the weight vectors of the trained SOM.

4. Results and discussion

According to the data found in the literature it is recommended that the number of neurons in the map should be nearly equal to the number of samples in the

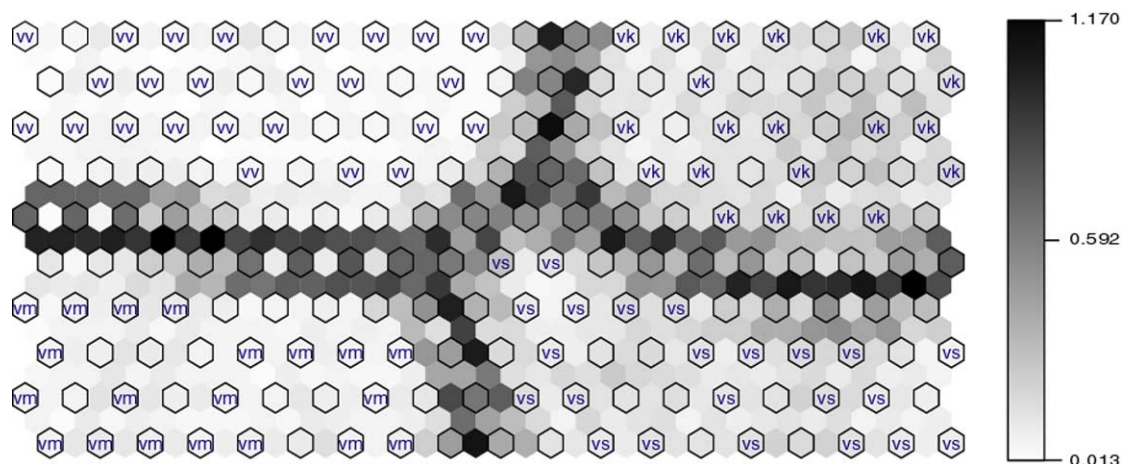


Fig. 2. Unified distance matrix for map trained by use of principal components obtained from the full spectrum.

training set and the length and width of the map should be proportional to the magnitude of the first eigenvalues obtained by the decomposition of the training set [26]. The ratio calculated by the first two calculated eigenvalues in this case is 1.96. Having that in mind (and also the recommendations [25,26] that the number of map neurons should be similar to the number of samples in the training set) we started the search for the optimal size of the map. After several trials, we chose a map with a size 19×10 which was trained using the first six principal components obtained from the mean-centered data matrix. The used map had plain boundary

conditions, a hexagonal grid, Gaussian neighborhood function, and linearly decreasing learning rate. The weight vectors were initialized along the first two principal components obtained by decomposition of the data matrix [26].

The SOMs were trained using the batch training algorithm [45] in two phases [26]: (1) rough training phase which lasted 50 iterations with an initial neighborhood radius equal to five, a final neighborhood radius equal to one, a learning rate of 0.5, and (2) fine training phase which lasted 500 iteration cycles, an initial and final neighborhood radius equal to one and a learning rate of 0.1.

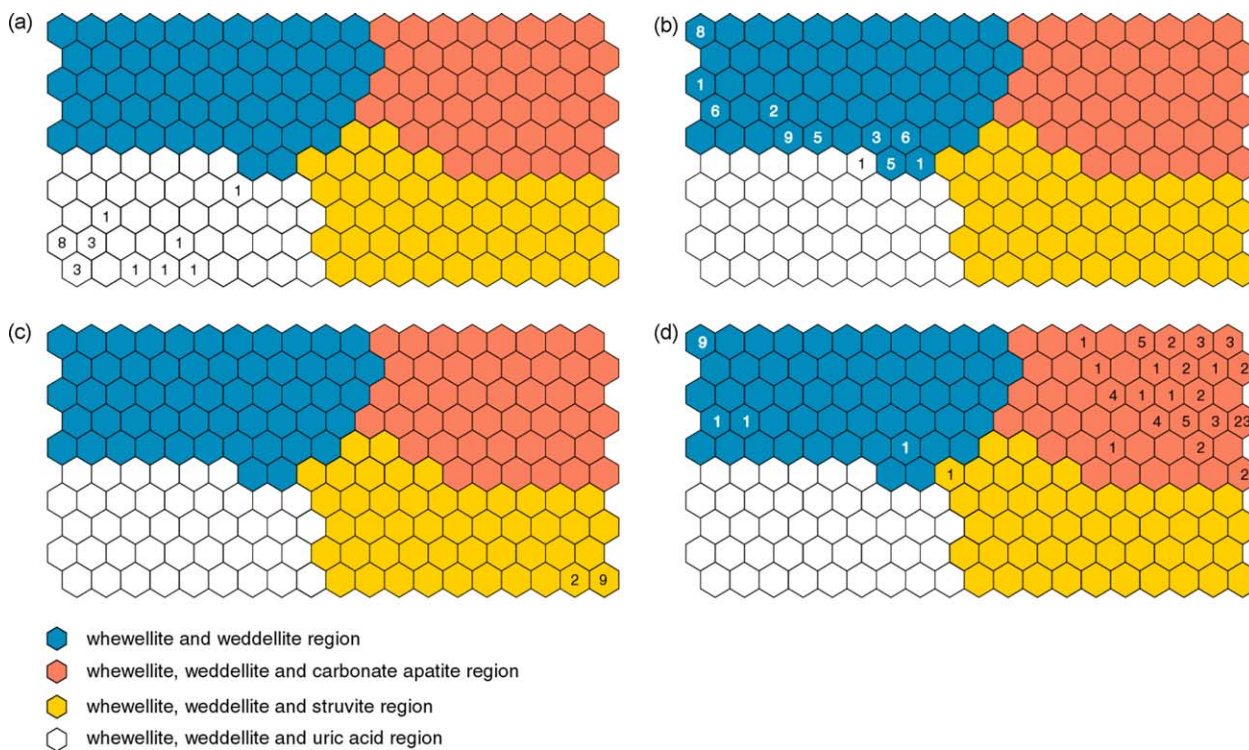


Fig. 3. The distribution of the samples from all four types of calculi on the trained map (a—whewellite, weddellite and uric acid samples; b—whewellite and weddellite samples; c—whewellite, weddellite and struvite samples; d—whewellite, weddellite and carbonate apatite samples).

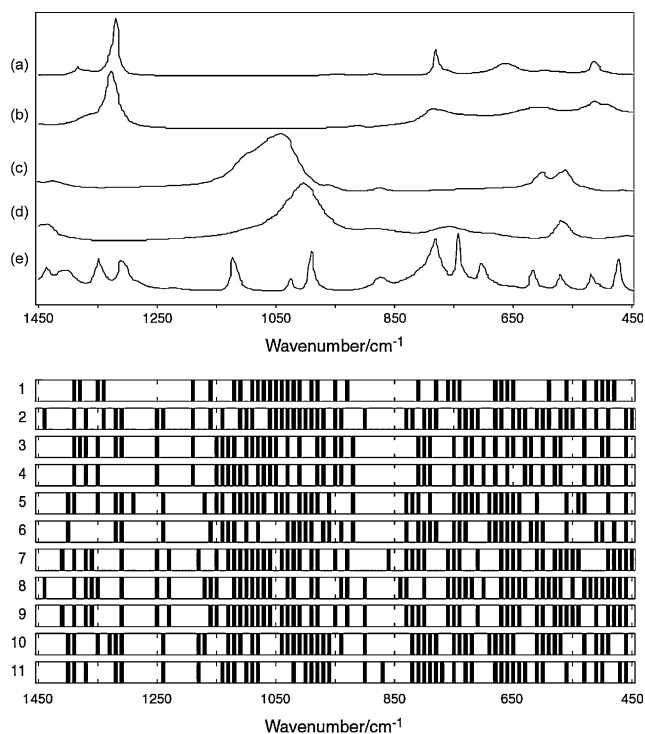


Fig. 4. Selected wavenumber regions for some of the best solutions obtained by use of genetic algorithms.

After the training was finished, the prediction abilities of the SOMs were examined using the data set consisting of suitably preprocessed infrared spectra of real samples.

The regions of the self-organizing map obtained using the principal components calculated from the full spectrum produced good separation of the samples in the training set which can be seen from the unified distance matrix presented in Fig. 2. However, using this map 14 samples were misclassified: one calculus consisting of oxalates was classified as an oxalates-uric acid concrement, further 12 calculi consisting of oxalates and carbonate apatite were classified as belonging to the oxalate type of calculi and one calculus consisting of oxalates and carbonate apatite was

Table 1
Map sizes and training parameters for some of the obtained solutions using the genetic algorithms

No.	Principal components	Length	Width	Rough training phase	Fine training phase
1	6	18	18	29	113
2	6	18	13	60	220
3	6	16	16	50	200
4	6	16	16	50	328
5	7	18	13	62	190
6	7	18	13	63	164
7	9	18	13	58	216
8	9	18	13	58	149
9	9	18	13	58	152
10	15	18	13	63	224
11	16	18	13	63	130

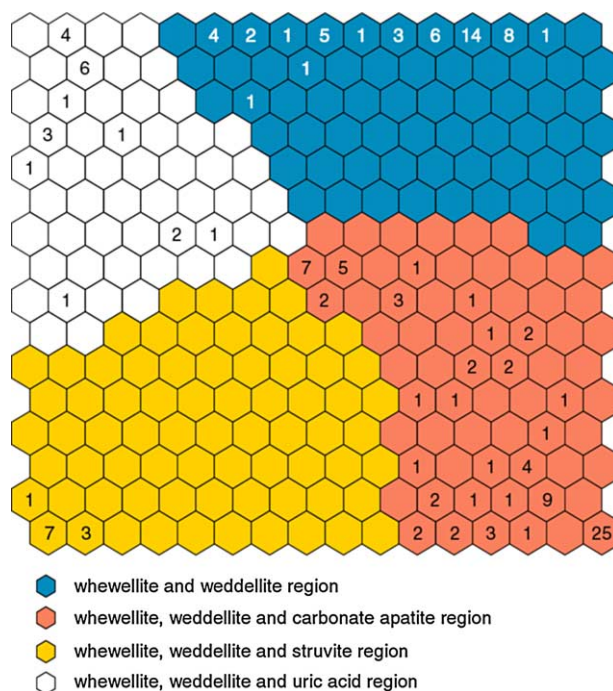


Fig. 5. Trained map for solution no. 3 (presented in Table 1) together with the distribution of all 160 samples of urinary calculi.

classified as belonging to the oxalates-struvite type. The distribution of the samples from all four types of calculi on the trained map, together with the misclassified ones is presented in Fig. 3.

The relatively high number of misclassified samples was the reason why we decided to use (prior to the extraction of principal components) genetic algorithms for the variable selection as well as for finding the most suitable map size and training parameters.

The procedure for variable selection using genetic algorithms was repeated several times for six hundred generations with an initial mutation rate of 0.10 in the initial population and linearly decreasing down to 0.05 until generation 300. After that the mutation rate was kept at the 0.05 value.

After a few repetitions of the optimization process, several solutions without misclassifications were obtained. The wavenumber regions for some of these chromosomes, together with the infrared spectra of the pure substances, are presented in Fig. 4. The map sizes and the training parameters for these same chromosomes are presented in Table 1. The self-organizing map for solution no. 3 (presented in Table 1) together with the distribution of all 160 samples in it are presented in Fig. 5.

5. Conclusions

Genetic algorithms can be successfully used for optimization of supervised self-organizing maps and for selection

of the most suitable wavenumber regions for classification of human urinary calculi. It has to be emphasized that genetic algorithms could help in the selection of suitable wavenumber regions without a need of any previous spectroscopic knowledge. Furthermore, the results of the optimization could be easily implemented into suitable graphical user interface which could then be of real help for the use of the results presented here in the clinical laboratories, a task on which we are presently working in our laboratory.

Acknowledgements

The financial support by the Ministry of Education and Science of Republic of Macedonia is gratefully acknowledged.

References

- [1] N.Q. Dao, M. Daudon, *Infrared and Raman Spectra of Calculi*, Elsevier, Paris, 1997.
- [2] M. Daudon, R.J. Reveillaud, *Presse Med.* 16 (1987) 627.
- [3] S.H. Kandil, T.A. Abou El Azm, A.M. Gad, M.M. Abdou, *Comput. Enhanced. Spectrosc.* 3 (1986) 171.
- [4] A. Hesse, M. Gergeleit, P. Schüller, K. Möller, *J. Clin. Chem. Clin. Biochem.* 27 (1989) 639.
- [5] M. Berthelot, G. Cornu, M. Daudon, *Clin. Chem.* 33 (1987) 2070.
- [6] C.A. Lehmann, G.L. McClure, I. Smolens, *Clin. Chim. Acta* 173 (1988) 107.
- [7] G. Rebentisch, M. Doll, J. Mueche, *Lab. Med.* 16 (1992) 224.
- [8] E. Peuchant, X. Heches, D. Sess, M. Clerc, *Clin. Chim. Acta* 205 (1992) 19.
- [9] H. Hobert, K. Meyer, *Fresenius' J. Anal. Chem.* 334 (1992) 178.
- [10] I. Kuzmanovski, M. Trpkovska, B. Šoptrajanov, V. Stefov, *Vib. Spectrosc.* 19 (1999) 249.
- [11] I. Kuzmanovski, Z. Zografski, M. Trpkovska, B. Šoptrajanov, V. Stefov, *Fresenius' J. Anal. Chem.* 370 (2001) 919.
- [12] I. Kuzmanovski, M. Trpkovska, B. Šoptrajanov, V. Stefov, *Anal. Chim. Acta* 491 (2003) 211.
- [13] P.K. Hopke, X.H. Song, *Anal. Chim. Acta* 348 (1997) 375.
- [14] D. Wienke, Y. Xie, P.K. Hopke, *Anal. Chim. Acta* 310 (1995) 1.
- [15] R. Goodacre, J. Pygall, D.B. Kell, *Chemometr. Intell. Lab. Syst.* 38 (1997) 1.
- [16] J. Zupan, M. Novič, *Anal. Chim. Acta* 192 (1994) 219.
- [17] H. Yang, I.R. Lewis, P.R. Griffiths, *Spectrochim. Acta* 55 (1999) 2783.
- [18] Y.V. Heyden, P. Vankeerberghen, M. Novic, J. Zupan, D.L. Massart, *Talanta* 51 (2000) 455.
- [19] I.V. Pletnev, V.V. Zernov, *Anal. Chim. Acta* 455 (2002) 131.
- [20] F. Vanderestraeten, C. Wojciechowski, N. Dupuy, J.-P. Huvenne, *Analisis* 26 (1998) 57.
- [21] M. Kolehmainen, P. Rönkkö, O. Raatikainen, *Anal. Chim. Acta* 484 (2003) 93.
- [22] C. Ruckebusch, L. Duponchel, J.-P. Huvenne, *Anal. Chim. Acta* 446 (2001) 257.
- [23] R. Todeschini, D. Galvagni, J.L. Vílchez, M. del Olmo, N. Navas, *Trends Anal. Chem.* 18 (1999) 93.
- [24] J. Zupan, M. Novič, I. Ruisánchez, *Chemometr. Intell. Lab. Syst.* 38 (1997) 1.
- [25] J. Zupan, J. Gasteiger, *Neural Networks in Chemistry and Drug Design*, WCH, Weinheim, 1999.
- [26] T. Kohonen, *Self-organizing maps*, third ed., Springer, Berlin, 2001.
- [27] T. Kohonen, *Computer* 21 (1988) 11.
- [28] I. Kuzmanovski, M. Trpkovska, B. Šoptrajanov, *Maked. Med. Pregled* 53 (1999) 251.
- [29] P. Brown, D. Ackermann, B. Finlayson, *J. Cryst. Growth* 98 (1989) 285.
- [30] D. Ackermann, P. Brown, B. Finlayson, *Urol. Res.* 16 (1988) 219.
- [31] M. Santos, P.F. González-Díaz, *Inorg. Chem.* 16 (1977) 2131.
- [32] J. Holland, *J. Comput. Machinery* 3 (1962) 297.
- [33] B. Kermani, S. Schiffman, H.T. Nagle, *IEEE Trans. Biomed. Eng.* 46 (1999) 429.
- [34] C. Henderson, W. Potter, R. McClendon, G. Hoogenboom, *Appl. Intell.* 12 (2000) 183.
- [35] D. Jouan-Rimbaud, D.L. Massart, R. Leardi, O.E. de Noord, *Anal. Chem.* 67 (1995) 4295.
- [36] R. Leardi, A. Lupiáñez, Gonzáles, *Chemometr. Intell. Lab. Syst.* 41 (1998) 195.
- [37] K. Hasegawa, Y. Miyashita, K. Funatsu, *J. Chem. Inf. Comput. Sci.* 37 (1997) 306.
- [38] B.M. Smith, P.J. Gemperline, *Anal. Chim. Acta* 423 (2000) 167.
- [39] H. Handels, T. Roß, J. Kreuzsch, H.H. Wolff, S.J. Pöppel, *Artif. Intell. Med.* 16 (1999) 283.
- [40] S.S. So, M. Karplus, *J. Med. Chem.* 39 (1996) 5246.
- [41] H. Yoshida, R. Leardi, K. Funatsu, K. Varmuza, *Anal. Chim. Acta* 446 (2001) 485.
- [42] MATLAB 5.2, 1984–1998 Mathworks.
- [43] J. Vesanto, *Intell. Data Anal.* 6 (1999) 111.
- [44] A. Chipperfield, P. Fleming, H. Pohlheim, C. Fonseca, *Genetic algorithm toolbox user's guide*, University of Sheffield, Sheffield, 1994.
- [45] W.-P. Tai, in: F. Fogelman-Soulié, P. Gallinari (Eds.), *Proc. ICANN'95, Int. Conf. Artif. Neural Networks vol. II*, EC2, Nanterre, France, 1995, p. II-33.