ELSEVIER

Journal of Chromatography B, 780 (2002) 27–33

JOURNAL OF CHROMATOGRAPHY B

www.elsevier.com/locate/chromb

Artificial neural network classification based on high-performance liquid chromatography of urinary and serum nucleosides for the clinical diagnosis of cancer

Jun Yang, Guowang Xu*, Hongwei Kong, Yufang Zheng, Tao Pang, Qing Yang

National Chromatographic R.&A. Center, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian 116012, China

Received 4 February 2002; received in revised form 22 May 2002; accepted 11 June 2002

Abstract

Nucleosides in human urine and serum have frequently been studied as a possible biomedical marker for cancer, acquired immune deficiency syndrome (AIDS) and the whole-body turnover of RNAs. Fifteen normal and modified nucleosides were determined in 69 urine and 42 serum samples using high-performance liquid chromatography (HPLC). Artificial neural networks have been used as a powerful pattern recognition tool to distinguish cancer patients from healthy persons. The recognition rate for the training set reached 100%. In the validating set, 95.8 and 92.9% of people were correctly classified into cancer patients and healthy persons when urine and serum were used as the sample for measuring the nucleosides. The results show that the artificial neural network technique is better than principal component analysis for the classification of healthy persons and cancer patients based on nucleoside data.

Keywords: Artificial neural networks; Principal component analysis; Nucleosides

1. Introduction

Biochemical substances which may serve as tumor markers in body fluids have been the subject of several reviews [1]. Numerous molecules found in urine and serum have been studied as potential tumor markers [2]. It has been shown that nucleosides are excreted in abnormal amounts in the urine and serum of cancer patients [1–12]. Reversed-phase high-performance liquid chromatography (RP-HPLC) and immunoassays have been used as the main analytical methods for urinary and serum nucleosides [1,4–13].

The statistical techniques reported in these previous investigations included the use of threshold logic units [14], K-nearest neighbors [15] and Wilcoxon test correlations [16], principal component analysis, and stepwise discriminate analysis [3]. The artificial neural network (ANN) has recently become a main tool in this area due to its higher recognition rate [3,17].

The artificial neural network is a developing branch of chemometrics. Theoretical study has shown that three-layered feedforward networks (one hidden layer) can fit any continuous function at any precision. The ANN is particularly suitable for solving nonlinear multivariate problems such as modeling [18,19], optimization [20], pattern recognition [2,21–25], etc.

In this study, artificial neural networks were

*Corresponding author. Tel./fax: +86-411-369-3403.
E-mail address: dicp402@mail.dlptt.ln.cn (G. Xu).

1570-0232/02/$ – see front matter © 2002 Elsevier Science B.V. All rights reserved.
PII: S1570-0232(02)00408-7

trained by an algorithm called ‘‘gradient descent with momentum and adaptive learning rate back-propagation’’ to correlate normal cases or cancer cases with nucleoside levels in the corresponding urine and serum samples. The results showed that the ANN method employed is suitable for the classification of healthy persons and cancer patients based on urinary or serum nucleosides.

2. Experimental and ANN method

2.1. Analysis of urinary and serum nucleosides

The collected urine and serum samples were immediately frozen and stored at −20 °C. For the analysis of ribonucleosides, the samples were defrosted at room temperature. A phenyl boronate gel-affinity chromatographic method was used to isolate the nucleosides with a phenyl boronate column, and the extracted nucleosides were analyzed by RP-HPLC [13].

2.2. Data set

A data set of urinary nucleosides containing 69 patterns, representing 18 healthy persons and 51 cancer patients, was obtained. Another data set of serum nucleosides had 42 patterns consisting of 23 healthy persons and 19 cancer patients. Each pattern was described by 15 feature variables, which were the concentrations of 15 nucleosides: dihydrouridine (Dhu), pseudouridine (Pseu), cytidine (C), uridine (U), 1-methyladenosine (m1A), inosine (I), 5-methyluridine (m5U), guanosine (G), xanthosine (X), 3-methyluridine (m3U), 1-methylinosine (m1I), 1-methylguanosine (m1G), adenosine (A), 6-methyladenosine (m6A), and 5′-deoxy-5′-methylthioadenosine (MTA). The normalized variables of each pattern were taken as inputs to train the ANN.

To classify the patterns, binary values were used to represent the two groups of persons, 0 for a healthy person and 1 for a cancer patient.

2.3. ANN method

Different researchers have adequately introduced the theory behind the neural network [3,25]. Among the different training methods of neural networks, the back-propagation (BP) technique is the most popular and is often used in analytical applications. An artificial neural network consists of a number of ‘‘neurons’’ or ‘‘hidden units’’ that receive data from the outside, then process the data, and output a signal. A ‘‘neuron’’ is essentially a regression equation with a non-linear output. When more than one of these neurons is used, non-linear models can be fitted. The back-propagation network receives a set of inputs, which are multiplied by each neuron’s weight. These products are summed for each neuron and a non-linear transfer function is applied. The transformed sums are then multiplied by the output weights, summed at a final time, transformed and interpreted. Since a back-propagation network is a supervised method, the desired output must be known for each input vector, and an error can be calculated. This error is propagated backwards through the network, then the weights are adjusted to make them, at the next time, closer to the desired output. The patterns are repeated many times until the network learns the relationship.

In this study, Matlab was used to complete the work. An algorithm was employed to elevate the learning rate and algorithm reliability, which was called ‘‘gradient descent with momentum and adaptive learning rate back-propagation’’. Back-propagation is used to calculate derivatives of performance *perf* (usually the sum of squared errors) with respect to the weight and bias variables *X*. Each variable is adjusted according to gradient descent with momentum:

$$\Delta X = mc \cdot \Delta X^{(\mathrm{previous})} + lr \cdot (1 - mc) \cdot \left(\frac{\partial perf}{\partial X}\right)$$

where $\Delta X^{(\mathrm{prev}\upsilon\mathrm{ious})}$ is the previous change to the weight or bias, *mc* is the momentum term, and *lr* is the learning rate. The added momentum term helps to direct the search on the error hyperspace to the global minimum by allowing a portion of the previous updating (magnitude and direction) to be added to the current updating step.

3. Results and discussion

In order to remove systematic variation, the mean of the variables was first subtracted from the data set.

Because the PCA method is mostly used by researchers in this area, it was selected for comparison with the ANN in this work. The data were then subjected to PCA. Figs. 1 and 2 show the two-dimensional (the first two principal components) display for urine and serum patterns by PCA. The healthy person patterns were clustered, while the cancer patterns were rather scattered. Decision lines, which were determined by a perceptron algorithm, were found to separate two areas: healthy persons and cancer patients. The total consistency rates were 88.4 and 78.6% when urine and serum, respectively, were used as the sample to measure the nucleosides (Tables 1 and 2).

3.1. Neural network optimization

To evaluate the correct rate of classification, an artificial neural network was employed to implement the application. Similar to the literature [26], a randomly selected two-thirds of the data were treated as the training set, with the remaining being the validation set.

The performance of a neural network is affected by the following parameters: network architecture, initial weight value, learning rate, and momentum term. These parameters should be carefully defined to achieve a good result.

3.1.1. Number of hidden layers and nodes

It has been claimed that an arbitrary nonlinear mapping of an input domain to an output domain can be achieved by using three layers with one hidden layer in a neural network [3]. In this work we also employed artificial neural networks with three layers. Since there is no theoretical way of choosing the number of each layer node, the best configuration is to be found in practice. Networks were tested with 15 input nodes (equal to the input variables), one output node and hidden nodes ranging from 3 to 10.

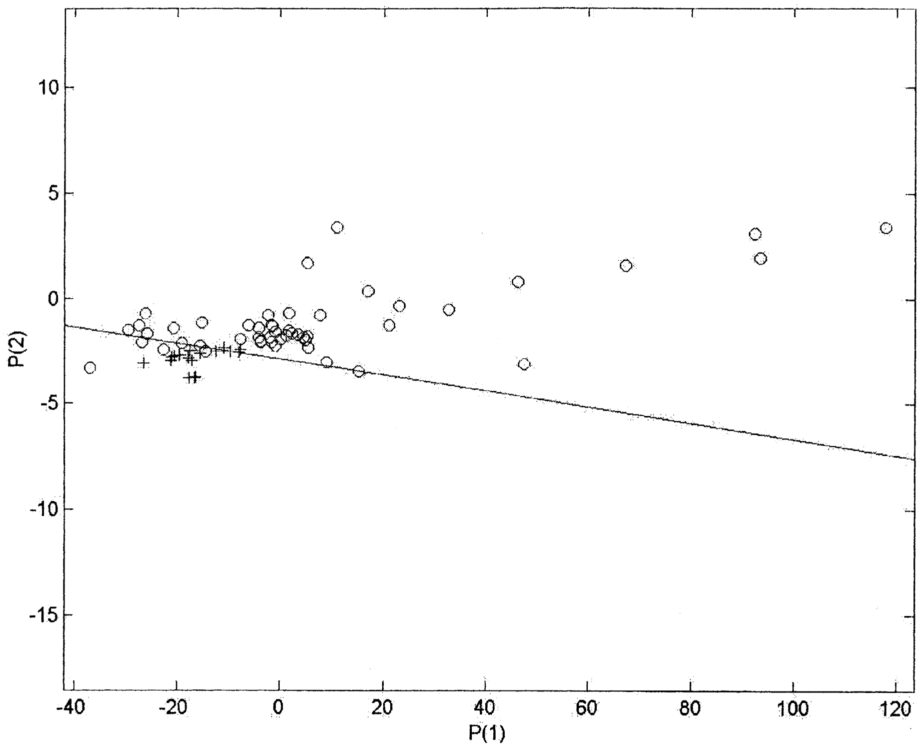

Fig. 1. PCA of the urine samples. (○) Cancer patients; (+) healthy persons (the most important variables are Dhu, Pseu, m1A, X, m1I, and m1G).

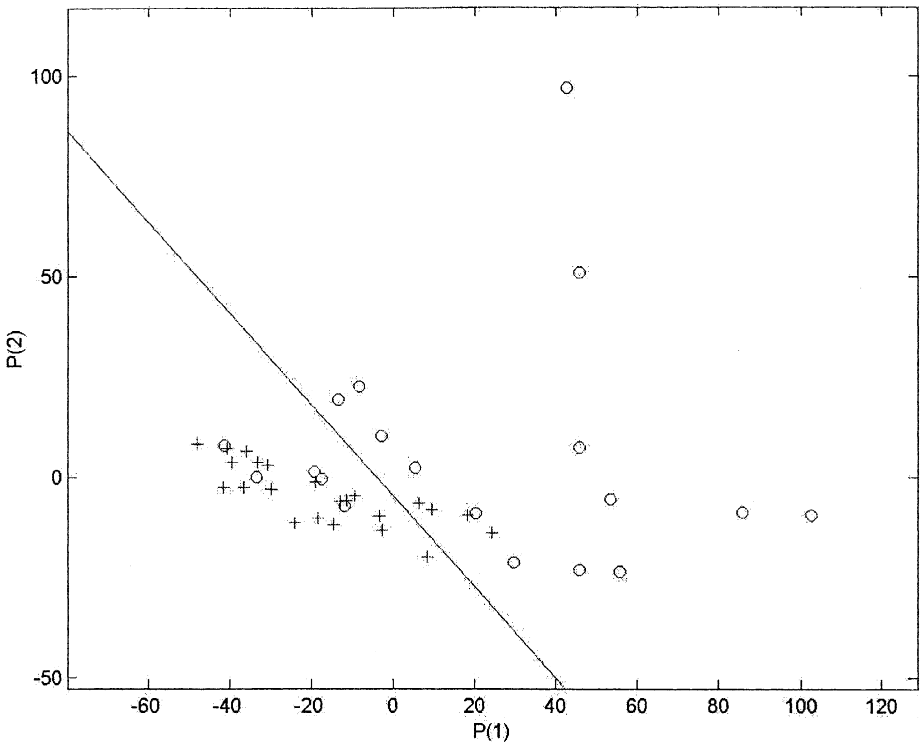

Fig. 2. PCA of the serum samples. (○) Cancer patients; (+) healthy persons (the most important variables are Dhu, Pseu, m1A, X, m1I, m1G, and mTA).

Table 1
Classification of the urine nucleoside data by the ANN and PCA methods[a]

	Training set			Validating set		
	Total No.	No. recognized as healthy persons	No. recognized as cancer patients	Total No.	No. recognized as healthy persons	No. recognized as cancer patients
ANN method						
Healthy persons	12	12	0	6	6	0
Cancer patients	34	0	34	17	1	16
Recognition rate (%)	Prediction rate (%)	Sensitivity (%)	Specificity (%)	Correct rate (%)		
100.0	95.7	98.0	94.7	98.6		
PCA method						
Healthy person	18	14	4			
Cancer patient	51	4	47			
Sensitivity (%)	Specificity (%)	Correct rate (%)				
92.2	77.8	88.4				

[a] Recognition rate is the rate of the correct classification of the training set. Prediction rate is the rate of the correct classification of the predicting set. Sensitivity is the number of true positives classified as positive. Specificity is the number of true negatives classified as negative [1].

Table 2
Classification of the serum nucleoside data by the ANN and PCA methods[a]

	Training set			Validating set		
	Total No.	No. recognized as healthy persons	No. recognized as cancer patients	Total No.	No. recognized as healthy persons	No.recognized as cancer patients
ANN method						
Healthy persons	15	15	0	8	8	0
Cancer patients	13	0	13	6	1	5
Recognition rate (%)	Prediction rate (%)	Sensitivity (%)	Specificity (%)	Correct rate (%)		
100.0	92.9	94.7	95.8	97.6		
PCA method						
Healthy persons	23	19	4			
Cancer patients	19	5	14			
Sensitivity (%)	Specificity (%)	Correct rate (%)				
73.7	79.2	78.6				

[a] As for Table 1.

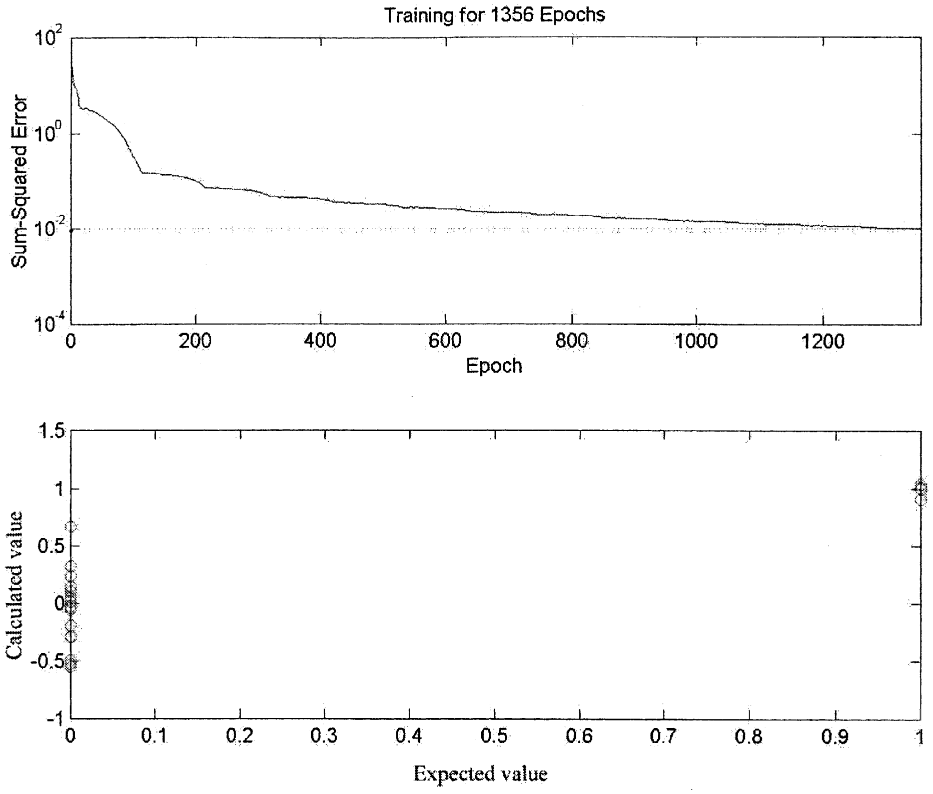

Fig. 3. Output value versus expected value of ANN for the classification of healthy persons and cancer patients based on the urinary nucleoside concentrations as the variables. Top panel: the figure of sum-squared error (the performance) versus epoch reveals the performance of the training. Bottom panel: the pattern points close to (0,0) and (1,1) are thought to be correctly classified.

3.1.2. Initial weight value

Early empirical studies concluded that the initial weight set should have small values that must not be approximately equal lest they fail to converge to a solution that has weights of differing values. Experience shows that the initial weights need not be small, but can be selected at random from a uniform distribution ranging from 0 to 1 [17].

It was found from our investigation that a network architecture consisting of 15 input nodes, one output node and five hidden nodes gave the best results. The suitable training iterations were about 10 000, while multi-pattern training was adopted and the learning rate was adjusted dynamically.

3.2. Classification of healthy humans and cancer patients based on the nucleoside concentrations in urine and serum

3.2.1. Nucleoside data from urine

The data set of 69 patterns of urinary nucleosides was randomly divided into two parts, a training set including 46 patterns (12 healthy persons, 34 cancer patients) was used to train the network model, and a validating set including 23 patterns (six healthy persons, 17 cancer patients) was used to evaluate the network model.

As the criterion for classification, an output value of >0.6 was considered as a cancer case and a value of <0.4 as a normal case. The classification results are shown in Fig. 3 and Table 1. From Table 1 it can be seen that a recognition rate of 100% for the training set and 92.8% for the validating set was obtained. The sensitivity and specificity were 98.0 and 94.7%, respectively, much higher than for PCA (Table 1).

3.2.2. Nucleoside data from serum

The data set of 42 patterns of serum nucleosides was randomly divided into two parts, a training set including 28 patterns (about two-thirds, 13 healthy

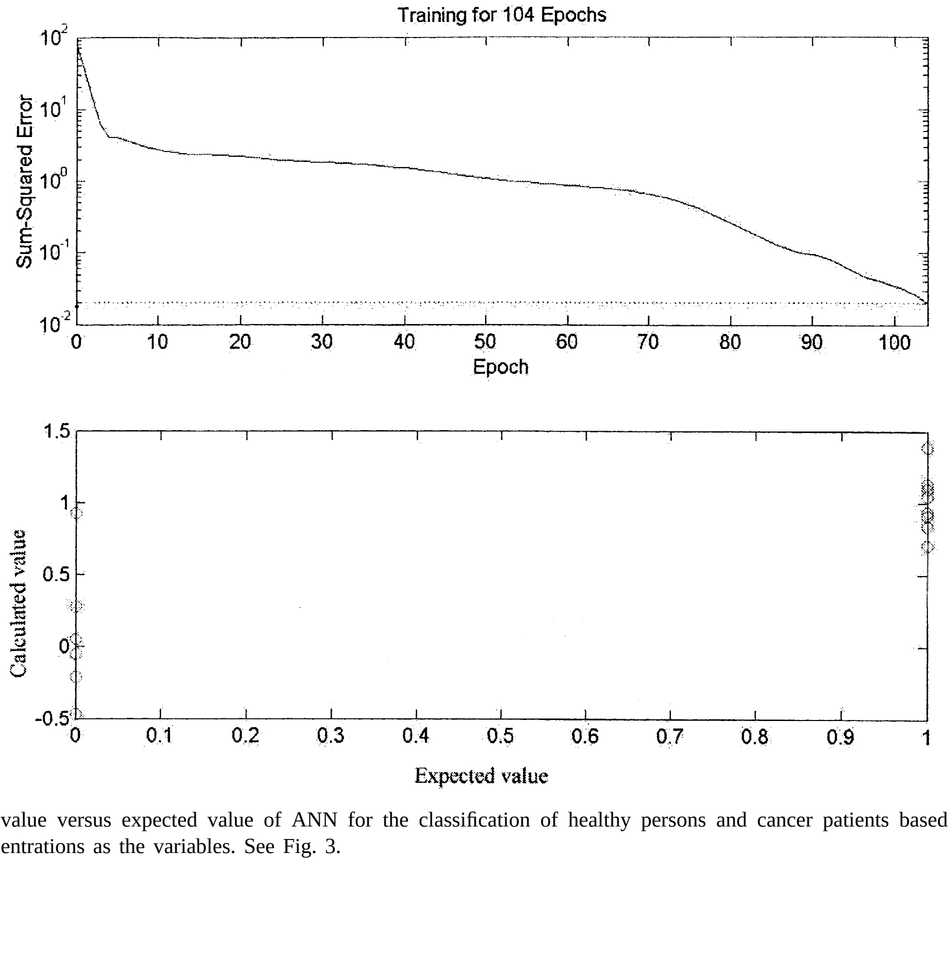

Fig. 4. Output value versus expected value of ANN for the classification of healthy persons and cancer patients based on the serum nucleoside concentrations as the variables. See Fig. 3.

persons, 15 cancer patients) was used to train the network model, and a validating set including 14 patterns (one-third, six healthy persons, eight cancer patients) was used to evaluate the network model of the serum nucleosides.

The classification results are shown in Fig. 4 and Table 2. It can be seen that the recognition rates of the training set and validating set were 100 and 92.9%, and the sensitivity and specificity were 94.7 and 95.8%, respectively, much higher than those from the PCA method.

It can be seen that the classification success when serum was the sample was a little worse than when urine was the sample. The reason for this is probably that the number in the training and predicting sets of the serum samples was smaller than that of the urine samples, and the network was not trained very well.

4. Conclusions

In this study, a rapid training ANN algorithm was employed to which a momentum term was added and the learning rate adjusted dynamically. It is concluded that the nucleosides in urine and serum analyzed by HPLC are possible candidates as markers for cancer. Compared with PCA methods, classification by artificial neural networks is more satisfactory, and it can hopefully be used as a powerful tool for tumor diagnosis.

References

[1] C.W. Gehrke, K.C. Kuo (Eds.), Chromatography and Modification of Nucleosides, Part C, Elsevier, Amsterdam, 1990.

[2] J.E. McEntire, K.C. Kuo, M.E. Smith, D.L. Stalling, Cancer Res. 49 (1989) 1057.

[3] R. Zhao, G. Xu, B. Yue, H.M. Liebich, Y. Zhang, J. Chromatogr. A 828 (1998) 489.

[4] T. Rasmuson, G.R. Bjork, Acta Oncol. 34 (1995) 61.

[5] A.J. Sasco, F. Rey, C. Reynaud, J. Bobin, M. Clavel, A. Niveleau, Cancer Lett. 108 (1996) 157.

[6] F. Pane, M. Savoia, G. Fortunato, A. Camera, B. Rotoli, F. Salvatore, L. Sacchetti, Clin. Biochem. 26 (1993) 513.

[7] E.P. Mitchell, L. Evans, P. Schultz, R. Madsen, J.W. Yarbro, C.W. Gehrke, K. Kuo, J. Chromatogr. 581 (1992) 31.

[8] C. Reynaud, C. Bruno, P. Boullanger, J. Grange, S. Barbesti, A. Niveleau, Cancer Lett. 61 (1992) 255.

[9] G. Xu, C. Di Stefano, H.M. Liebich, Y. Zhang, P. Lu, J. Chromatogr. B 732 (1999) 307.

[10] H.M. Liebich, C. Di Stefano, A. Wixforth, H.R. Schmid, J. Chromatogr. A 763 (1997) 193.

[11] G. Xu, H.R. Schmid, X. Lu, H.M. Liebich, P. Lu, Biomed. Chromatogr. 14 (2000) 459.

[12] G. Xu, H.R. Enderle, H.M. Liebich, P. Lu, Chromatographia 52 (2000) 152.

[13] G. Xu, H.M. Liebich, Am. Clin. Lab. (2001) 22.

[14] M.L. McConnell, G. Rhodes, U. Watson, M. Novotny, J. Chromatogr. 162 (1979) 495.

[15] A. Zlatkis, K.Y. Lee, C.F. Poole, G. Holzer, J. Chromatogr. 163 (1979) 125.

[16] H.A. Scoble, J.L. Fasching, P.R. Brown, Anal. Chim. Acta 150 (1983) 171.

[17] C.G. Looney, Neurocomputing 10 (1996) 7.

[18] M. Jalali-Heravi, M.H. Fatemi, J. Chromatogr. 915 (2001) 177.

[19] M. Jalali-Heravi, Z. Garkani-Nejad, J. Chromatogr. 927 (2001) 211.

[20] H. Zhang, R. Zhang, M. Liu, Z. Hu, Chin. J. Anal. Chem. 28 (2000) 1336.

[21] H. Liu, X. Cao, R. Xu, N. Chen, Anal. Chim. Acta 342 (1997) 223.

[22] S. Dong, J. Yao, K. Yu, H. Tang, H. Gao, Chin. J. Anal. Chem. 28 (2000) 1025.

[23] P.K. Hopke, X. Song, Anal. Chim. Acta 348 (1997) 375.

[24] S.R. Johnson, J.M. Sutter, H.L. Engelhardt, Anal. Chem. 69 (1997) 4641.

[25] J. Zupan, Anal. Chim. Acta 248 (1991) 1.

[26] N.J. Nilsson, Artificial Intelligence—A New Synthesis, Prentice Press, p. 54.