

Determination of the composition of sialoliths composed of carbonate apatite and albumin using artificial neural networks

I. Kuzmanovski^{a,*}, M. Ristova^a, B. Šoptrajanov^{a,b}, V. Stefov^a, V. Popovski^c

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

^b *Makedonska akademija na naukite i umetnostite, 1000 Skopje, Macedonia*

^c *Klinika za maksilofacijalna hirurgija, Univerzitet "Sv. Kiril i Metodij", 1001 Skopje, Macedonia*

Received 4 June 2003; received in revised form 28 September 2003; accepted 9 October 2003

Abstract

The determination of the components of the sialoliths is important both from the point of view of chances for a successful medical treatment of the patients and because the prevention of further re-occurrence of sialolithiasis depends upon the knowledge of the nature of the constituents of the concretions. Despite the fact that infrared spectroscopy is widely used for the determination of the composition of sialoliths, urinary calculi and bladder stones, we found no data for any chemometric method developed for such purposes. Here, a method is presented for quantitative determination of the content of salivary calculi composed of albumin and carbonate apatite (one of the most often found constituents in the analyzed calculi from the patients from Macedonia) using artificial neural networks (ANN). The results were checked on real samples using the standard addition method. The precision of the method was estimated using the relative standard deviation, which shows that it is suitable for routine analysis.

© 2003 Elsevier B.V. All rights reserved.

Keywords: Salivary calculi; Sialoliths; Artificial neural networks; Analysis; Infrared spectroscopy; Carbonate apatite; Albumin

1. Introduction

Sialoliths are common in submandibular glands and, to a lesser degree, in the parotid gland where they are formed as a result of crystallization of the salivary solutes. A number of studies of the structure and composition of sialoliths were found in the literature [1–6]. An accurate determination of their composition could help finding the reason for their appearance and preventing further formation of stones.

In order to obtain more reliable results, it is desirable to use, whenever possible, automated analytical procedures, which could minimize the role of the analyst. Although the present-day laboratory equipment gives an opportunity to collect digital signals suitable for further treatment, we did not find any data about chemometric techniques applied to the determination of the quantitative composition of salivary calculi. This is why, as an extension of our studies in which infrared spectroscopy coupled to chemometric treat-

ment was applied to another type of concretions (urinary calculi) [7,8], we tried to use the same instrumental and mathematical techniques for the analysis of salivary calculi. As a first step in that direction, we decided to investigate a simple (yet, realistic) system consisting of carbonate apatite and albumin.¹

The proposed method is based on the fact that infrared spectroscopy can provide information about the exact chemical individuality of the constituents since the IR spectrum is characteristic for a given compound. In addition to that, the duration of the analysis is about 15 min and only about 1 mg of sample is required (if necessary it can be reduced down to 20–30 µg), this being particularly important for studying the nucleation processes.

In the present work, the infrared spectra of artificial mixtures of the two mentioned components and of real samples were treated using artificial neural networks (ANNs). Application of different types of ANNs in the field of chemistry and related sciences varies from function ap-

* Corresponding author. Tel.: +389-2-3117-055; fax: +389-2-3226-865.

E-mail address: shigor@iunona.pmf.ukim.edu.mk (I. Kuzmanovski).

¹ Carbonate apatite and albumin are the two most frequent constituents of the sialoliths analyzed in our laboratory.

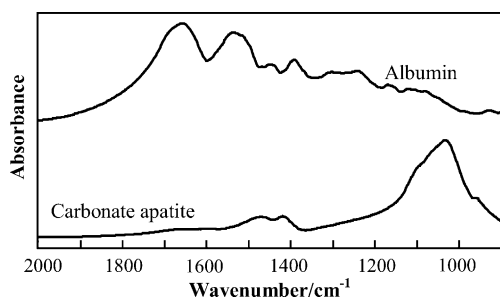


Fig. 1. Infrared spectra of albumin and carbonate apatite in the region of 2000–900 cm^{-1} .

proximation, classification, quantitative structure retention relations, quantitative structure properties/activation relations, determination of protein structure to protein folding, etc. [9–14]. Detailed descriptions of ANNs and their application in the above noted fields could be found elsewhere [9,10] so here we only mention that ANNs are capable of better modeling the non-linear relationships between the spectral response and the sample composition than the alternative methods—the principal component regression and partial least squares (PLS) [8]. Some of the possible factors that could cause deviations from linear relationship between spectral responses and concentrations are discussed by Gemperline et al. [15].

In this paper, we present the results of our attempt to develop a method for the analysis of salivary calculi using ANNs and the comparison of the outcome of such an analysis with the results obtained by another chemometric technique—the partial least-squares regression [16].

2. Experimental

Carbonate apatite was prepared in our laboratory according to a procedure found in the literature [17] while albumin was a Merck product. The spectra of both substances were compared with spectra from the database [18] and showed no significant differences. Mixtures of solid carbonate apatite and albumin were prepared in order to optimize the networks for the prediction of the composition of real samples.

The spectra of the pure substances (Fig. 1) as well as those of the mixtures were recorded in the 2000–900 cm^{-1} region on an FTIR Perkin-Elmer System 2000 instrument from KBr pellets containing 1 mg sample and 250 mg KBr. The resolution was 4 cm^{-1} and the sampling interval was 1 cm^{-1} (this gives a total number of 1100 absorbance values) using 32 scans for both the samples and the background (a pellet consisting solely of 250 mg KBr).

2.1. Data analysis

Thirty mixtures were prepared according to the mixture design presented in Fig. 2. The collected spectra were exported in ASCII format and stored in a single data matrix.

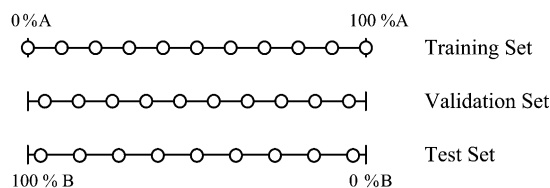


Fig. 2. Mixture design for the samples used for optimization of the neural networks (A: albumin; B: carbonate apatite).

The spectra were then offset corrected so that the minimum value for the absorbance in the region becomes zero. The normalization of the spectra to unit area was also carried out in order to minimize the influence on the absorbance of the measured mass of the sample. To make the training procedure faster and to reduce the complexity of the model, principal component analysis (PCA) was applied to the data matrix consisting of normalized spectra. The principal components (PCs) capturing 100% of the variance in the normalized data matrix were used as an input data to the neural networks. The mass fractions of the constituents were stored in another data matrix.

As shown in Fig. 2, the data were divided into three subsets: a training, a validation and a test set. This is necessary when early stopping procedure, as in this case, is applied for control of the generalization abilities of the network [19]. The first two sets were used during the optimization of the network. The training set was used for computing and adjusting the weights and biases. The validation set was used to supervise the performances of the networks during the training process in order to avoid overfit of the data—an effect that could give rise to trained networks with poor generalization abilities. At the beginning of the training, the sum-squared error (*SSE*) in both sets decreases. However, when the network starts to overfit the data, the *SSE* in the validation set starts to increase. When the latter happens in five consecutive steps, the training is stopped and the weights and biases, which correspond to the minimum of the *SSE* in the validation set, are returned.

The optimizations of the three-layered cascade-forward neural networks were performed with the Levenberg–Marquardt algorithm [20] for back-propagation of errors. The training of the networks was stopped when the performance goal of *SSE* equal to 10^{-3} was reached or when the error in the validation set increased for five consecutive iterations.

In order to select the optimal number of principal components in PLS regression model, training and validation sets were used for calibration (using cross-validation with leave-one-out technique). In order to find optimal number of PCs for prediction of the constituents of the calculi predictive residual sum of squares (*PRESS*) was calculated:

$$\text{PRESS} = \sum_{i=1}^n (\bar{w}_i - w_i)^2$$

where n is the number of samples in the set; \bar{w}_i and w_i are the predicted and actual mass fraction of that constituent in

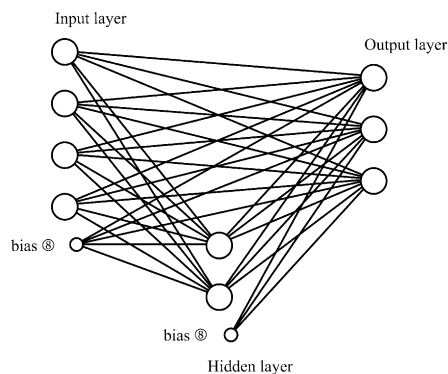


Fig. 3. Cascade-forward neural network.

the i -th sample. The minimum of the *PRESS* as a function of number of PCs indicates the optimal number of principal components, which were used for detection of composition of the samples.

The predictive abilities of the PLS regression model and of different optimized ANN architectures were checked calculating, for the test set, the standard error of prediction (*SEP*) which is defined as:

$$SEP = \left[\sum_{i=1}^n \frac{(\bar{w}_i - w_i)^2}{n} \right]^{1/2}$$

where \bar{w}_i is the predicted mass fraction of the analyzed constituent in the i -th sample, w_i the actual mass fraction of that constituent in the i -th sample, and n is the number of the samples in the test set.

The MATLAB software package was used for data processing. The ANN were optimized using its Neural Network Toolbox [21]. The PLS regressions model was developed using the Chemometric Toolbox for MATLAB.

3. Results and discussion

3.1. Optimization of the neural networks

The three-layered cascade-forward neural network architecture (Fig. 3) with best performances was sought out by changing the number of input neurons (from 1 to 11) and the number of hidden neurons (from 1 to 10).

The cascade-forward neural networks could be, as a rule, easily optimized with a smaller number of neurons in the hidden layer in comparison with the feed-forward neural networks [22]. With direct connections from input to output neurons with a linear transfer function, the cascade-forward neural networks behave as a multivariate linear regression. On the other hand, the neurons with a sigmoid transfer function in the hidden layer are capable of modeling the present non-linearities not captured by the input data and the linear part of the ANN.

The optimization of each of the network architectures was repeated 10 times in order to decrease the influ-

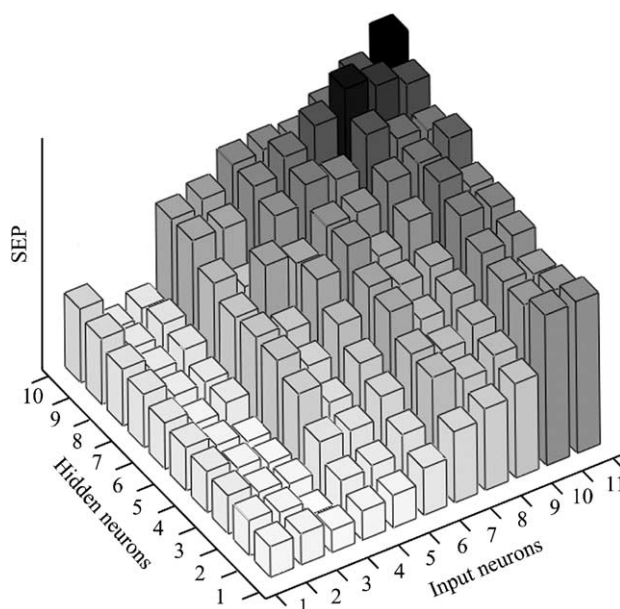


Fig. 4. Standard error of prediction for the test set obtained by different network architectures as a function of the number of neurons in the input layer and hidden layer (the lighter the color, the better the performance and the smaller *SEP*).

ence of the random initialization of weight and biases on *SEP*.

The training process was most often stopped as a consequence of the increase of the *SSE* for the validation set. During the training of the different network architectures, as a rule, with the increase of the number of neurons, the number of iteration cycles necessary to reach the optimal values of the weight and biases, also increases but this parameter is not monitored when early stopping is used.

The network architecture giving the best performances was chosen from Fig. 4 where the average *SEP* (calculated from repetitions of the trainings of all different architectures) as a function of the number of input neurons and hidden neurons is presented. The ANN architecture giving the best performances is that trained with three input neurons and with two neuron in the hidden layer ($SEP_{\text{albumin}} = 0.037$ and $SEP_{\text{carbonateapatite}} = 0.038$). Although the early stopping procedure used here for control of generalization abilities of the ANNs may be criticized for reduction of the number of data available for direct training [23] we have experienced that in cases where the deviations from linear behavior are not so enhanced, good results are obtained [8,22,24].

3.2. PLS regression optimization

Using the same data set and PLS1 algorithm the optimal number of principal components for PLS regression model was determined using cross-validation. The optimal number of PCs for albumin is three. The corresponding *SEP* for the samples in the test set for albumin, using three principal components is 0.055. The optimal number of PCs for de-

Table 1
Predicted and calculated mass fractions for the component of some of the analyzed calculi using ANN with best performances

Sample no.	1	2	3	4	5
Predicted mass fractions after standard addition					
Albumin	0.262	0.352	0.664	0.355	0.385
Carbonate apatite	0.738	0.648	0.336	0.645	0.615
Calculated mass fractions after standard addition					
Albumin	0.257	0.304	0.712	0.429	0.415
Carbonate apatite	0.737	0.688	0.243	0.601	0.589
Difference between calculated and predicted mass fractions of the samples after standard addition					
Albumin	0.005	0.047	−0.048	−0.074	−0.029
Carbonate apatite	0.001	−0.039	0.093	0.044	0.026

termination of carbonate apatite is two, and corresponding SEP value for carbonate apatite is 0.078.

The standard error of prediction for the constituents of the samples in the test set obtained using the cascade-forward ANN are better than that obtained by PLS regression which, as previously stated, proves that the ANN are a superior tool for modeling the relationship between absorbance and mass fraction of constituents [8].

3.3. Analysis of real samples

The network architecture with the best performance was used to analyze the real samples. The obtained results were checked using the standard addition method. All analyzed salivary calculi were homogenized before recording their infrared spectra. After the spectra have been recorded, to each of the five selected samples (with the exactly known mass—about 100–150 mg) a portion of the standard (albumin or carbonate apatite) with a precisely known mass (10–20 mg) was added. The samples were again homogenized and their infrared spectra recorded.

The mass fractions of the constituents in the samples before and after the standard addition were determined by same neural network. The calculated mass fractions of the samples after standard addition and predicted mass fractions of these samples are presented in Table 1. The absolute values of the differences between the predicted and calculated mass fraction in most cases are lower than 0.05. The largest discrepancies for the mass fraction of carbonate apatite were found in the sample #3, whereas in the sample #4 such is the case with albumin.

The precision of the method was estimated by the values of the relative standard deviation over three analysis of each sample (the mass fractions were obtained using three recorded infrared spectra from the same calculi samples us-

ing same neural network—the one with best performances) presented in Table 2. In most cases, the values of the relative standard deviation are lower than 5%. The values of the relative standard deviation satisfy the criteria for analytical methods used for routine analyses according to Püschel [25].

4. Conclusion

Cascade-forward artificial neural networks were used for predicting the composition of salivary calculi consisting of albumin and carbonate apatite. In this case, the networks with two neurons in the hidden layer and three input neurons exhibit the best performance in predicting the content of albumin and carbonate apatite in the calculi composed of these two substances. It was found that ANNs have better prediction abilities than the method based on PLS regression. The procedure described here is consisted of many steps of data treatment, which could make this method complicated for routine use in clinical laboratories, but it could be relatively easily automated if specialized software with suitable graphical interface is developed.

It is planned to extend the work to other combinations of constituents known to be present in the salivary calculi.

Acknowledgements

The financial support by the Ministry of Education and Science of Republic of Macedonia is gratefully acknowledged. One of the authors, Igor Kuzmanovski, is very grateful to Prof. Dr. Marcia M.C. Ferreira for the support provided.

References

- [1] G. Anneroth, C.M. Eneroth, G. Isacson, P.G. Lundquist, Scand. J. Dent. Res. 86 (1978) 182.
- [2] J. Lustmann, E. Regev, Y. Melamed, Int. J. Oral. Maxillofac. Surg. 19 (1990) 135.
- [3] G. Anneroth, C.M. Eneroth, G. Isacson, J. Oral Pathol. 4 (1975) 266.

Table 2
Relative standard deviation (%) for the samples

Sample number	1	2	3	4	5
Albumin	4.61	2.88	3.33	4.87	6.29
Carbonate apatite	2.89	1.58	6.58	2.64	2.24

- [4] G. Strubel, V. Rzepka-Glinder, *J. Clin. Chem. Clin. Biochem.* 27 (1989) 244.
- [5] T. Sakae, H. Yamamoto, G. Hirai, *J. Dent. Res.* 60 (1981) 842.
- [6] M. Blatt, M. Denning, H. Zumberge, H. Maxwell, *Ann. Otol.* 67 (1958) 595.
- [7] I. Kuzmanovski, M. Trpkovska, B. Šoptrajanov, V. Stefov, *Vib. Spectrosc.* 19 (1999) 249.
- [8] I. Kuzmanovski, Z. Zografski, M. Trpkovska, B. Šoptrajanov, V. Stefov, *Fresenius J. Anal. Chem.* 370 (2001) 919.
- [9] J. Zupan, J. Gasteiger, *Neural Networks in Chemistry and Drug Design*, Wiley, Weinheim, 1999.
- [10] A. Bos, M. Bos, W.E. van der Linden, *Anal. Chim. Acta* 256 (1992) 133.
- [11] K. Lin, A.C.W. May, W.R. Taylor, *J. Theor. Biol.* 216 (2002) 361.
- [12] L. Edler, J. Grassmann, S. Suhai, *Math. Comput. Model.* 33 (2001) 1401.
- [13] D.T. Jones, *J. Mol. Biol.* 292 (1999) 195.
- [14] J.H. Kim, J.R. Roche, *J. Comput. Syst. Sci.* 56 (1998) 223.
- [15] P.J. Gemperline, J.R. Long, V.G. Gregoriou, *Anal. Chem.* 63 (1991) 2313.
- [16] I.S. Helland, *Chemom. Intell. Lab. Syst.* 58 (2001) 97.
- [17] M. Santos, P.F. González-Díaz, *Inorg. Chem.* 16 (1977) 2131.
- [18] N.Q. Dao, M. Daudon, *Infrared and Raman Spectra of Calcium*, Elsevier, Paris, 1997.
- [19] H. Demuth, M. Beale, *Neural Network Toolbox User's Guide*, Mathworks, Natick, 1997.
- [20] M.T. Hagan, M. Menhaj, *IEEE Trans. Neural Netw.* 5 (1994) 989.
- [21] MATLAB 5.2, Mathworks, Natick, 1984–1998.
- [22] I. Kuzmanovski, S. Aleksovská, *Chemom. Intell. Lab. Syst.* 67 (2003) 167.
- [23] C.M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, 1995.
- [24] I. Kuzmanovski, M. Trpkovska, B. Šoptrajanov, V. Stefov, *Anal. Chim. Acta* 491 (2003) 211.
- [25] R. Püschel, *Mikrochim. Acta* 4 (1968) 783.