Prediction of Molar Absorptivities of Color Reagents during Their Color Reactions with Cerium Using Multiple Regression Analysis and Neural Network[†]

Ming Guo, Lu Xu^{††}, Hua Li and Chang-Yu Hu^{††}

Applied Spectroscopy Laboratory, Changchun Institute of Applied Chemistry, Chinese Academy of Sciences, Changchun 130022, People's Republic of China

A quantitative structure–property study has been made on the relationship between molar absorptivities (ϵ) of asymmetrical phosphone bisazo derivatives of chromotropic acid and their color reactions with cerium by multiple regression analysis and neural network. The new topological indices $A_{x1} - A_{x3}$ suggested in our laboratory and molecular connectivity indices of 43 compounds have been calculated. The results obtained from the two methods are compared. The neural network model is superior to the regression analysis technique and gave a prediction which was sufficiently accurate to estimate the molar absorptivities of color reagents during their color reactions with cerium.

Keywords Quantitative structure–property relationship, regression analysis, neural network, topological index, cerium, asymmetrical phosphone bisazo derivative, chromotropic acid

Determination of rare earth elements has attracted increasing interest in recent years, because the rare earth elements have been exploited extensively. Although many approaches have been developed, spectrophotometric approach as a routine method still hold a important position for its simplicity and general effectiveness.^{1–4} A key step in the spectrophotometric method is the selection of a sensitive, high selectivity color reagent and of suitable analytical conditions. Of the many color reagents, asymmetrical phosphone bisazo derivatives of chromotropic acid, owing to their high sensitivity, have been synthesized and used for determination of the rare earth elements in China. Despite the fact that new reagents based on asymmetrical phosphone bisazo derivatives of chromotropic acid are still being created and encountered in analytical practice, they are treated mainly by empirical variation of the functional group without rigorous consideration of the chromotropic acid system in terms of quantitative structure–property relationship studies.

Quantitative structure–activity/property studies have been applied in many fields, especially in the field of designing drugs and pesticides, but very few studies have been done on the design of color reagents.

Different ways have been suggested for quantitative structure–activity/property relationship studies; the topological index is one of the commonly used methods. The topological index is a numeric quantity that is

mathematically derived from the structure graph of a molecule. Many topological indices, such as the Wiener index⁵, Randic index⁶, Hosoya index⁷, Balaban index⁸, and the general a_N -index⁹, etc., have been proposed to convert chemical structure to numerical values, and further used to express the properties of a compound. The topological indices $A_{x1} - A_{x3}$ based on the extended distance-matrices devised recently by our laboratory have been successfully employed to studies on structure–activity relationship for compounds of alkanes, alcohols and barbiturates.¹⁰ In this work, we attempted to establishing a relationship between topological indices $A_{x1} - A_{x3}$, molecular connectivity indices¹¹ and the molar absorptivities of asymmetrical phosphone bisazo derivatives of chromotropic acid during their color reactions with cerium using multiple regression analysis and neural networks; satisfactory results were obtained.

Experimental

The neural network software included two modules: back-propagation^{12,13} and quasi-Newton^{14–19} neural networks, which were written in FORTRAN 77 and installed on a micro VAX II. All computations were performed using the multivariate statistic analysis programs package (MSAPP) developed by us, which consists of multiple-regression analysis, multiple stepwise regression analysis, pattern recognition and calculations of topological indices, etc.

[†] Presented at the ASIANALYSIS III, August 20 – 24, 1995, Seoul, Korea.

^{††} To whom correspondence should be addressed.

Structure of color reagents

The structures of asymmetrical phosphone bisazo derivatives of chromotropic acid are shown in Scheme 1.

The experimental conditions (pH values) employed and color reactions with cerium (*i.e.*, molar absorptivities) are listed in Table 1.

Scheme 1

Calculation of descriptors

The topological indices A_{x1} – A_{x3} and the molecular connectivity index ${}^mX_1^v$ of 43 asymmetrical phosphone bisazo derivatives of chromotropic acid have been calculated. A total of 21 topological indices (0X_p – 6X_p , 3X_c – 6X_c , ${}^3X_{ch}$ – ${}^6X_{ch}$, ${}^4X_{pc}$ – ${}^6X_{pc}$, A_{x1}^* – A_{x3}^*) were generated for each color reagent. Indices with asterisks A_{x1}^* – A_{x3}^* are the modified A_{x1} – A_{x3} ¹⁰ indices.

Objective feature selection

In order to determine which set of descriptors was important in computed topological indices, the multiple

Table 1 The structure of color reagents and molar absorptivities of their color reaction with cerium

No.	Structure				System	pH	$\epsilon \times 10^{-4}$
	X	R ₁	R ₂	R ₃			
1	–Cl			–NO ₂	KCl/HCl	1	7.10
2	–Cl		–NO ₂		KCl/HCl	1	8.20
3	–Cl		–COCH ₃		KCl/HCl	1	8.80
4	–Cl			–COCH ₃	KCl/HCl	1	7.30
5	–Cl		–Br		KCl/HCl	1	7.50
6	–Cl			–Br	KCl/HCl	1	6.60
7	–Cl		–COOH		H ₂ SO ₄ /HNO ₃	0.5–1.2	8.10
8	–Cl			–COOH	H ₂ SO ₄ /HNO ₃	0.4–1.5	7.90
9	–Cl		–COOH	–COOH	H ₂ SO ₄ /HNO ₃	0.4–1.5	7.80
10	–Cl			–PO ₃ H ₂	HCl	0.2–1.5	8.40
11	–Cl		–Cl	–NO ₂	KCl/HCl	2	3.20
12	–Cl	–CH ₃		–NO ₂	KCl/HCl	2	5.90
13	–Cl	–NO ₂		–NO ₂	KCl/HCl	1	3.50
14	–Cl		–NO ₂		KCl/HCl	2	1.50
15	–Cl	–NO ₂			KCl/HCl	2	1.00
16	–Cl	–Cl			KCl/HCl	1	1.00
17	–Cl		–Cl		KCl/HCl	1	7.10
18	–Cl			–Cl	KCl/HCl	1	6.50
19	–Cl	–Br			KCl/HCl	3	1.20
20	–Cl	–COOH			KCl/HCl	2	1.20
21	–Cl		–COOH		KCl/HCl	1	6.40
22	–Cl		–OCH ₃		KCl/HCl	1	6.40
23	–Cl			–OCH ₃	KCl/HCl	2	5.90
24	–Cl		–SO ₃ H		HCl	0.8–1.8	7.40
25	–Cl			–SO ₃ H	HCl	0.8–1.5	5.50
26	–Cl		–OH		KCl/HCl	2	4.90
27	–Cl			–OH	KCl/HCl	1	3.80
28	–Cl		–CH ₃		KCl/HCl	2	6.50
29	–Cl			–CH ₃	KCl/HCl	2	5.60
30	–Cl				KCl/HCl	2	6.40
31	–Br		–Br		HCl	1	11.20
32	–Br			–Br	HCl	1.5	10.50
33	–Br			–CHO	HCl	1.2	9.60
34	–Br			–NO ₂	HCl	1.2	8.20
35	–Br		–NO ₂		HCl	1.0	9.10
36	–Br		–COCH ₃		HCl	0.8	10.40
37	–Br			–COCH ₃	HCl	1	9.30
38	–Br		–COOH		HCl	1.1	9.50
39	–Br			–COOH	HCl	1.2	9.90
40	–Br		–CH ₃		HCl	0.8	8.80
41	–Br			–CH ₃	HCl	1.2	10.10
42	–Br			–SO ₃ H	HCl	0.8	10.50
43	–Br			–SO ₂ NH ₂	HCl	1	9.90

Note: R₄-th position substituent in structures 16 and 17 are –Cl.

stepwise regression analysis was necessary to refine the topological indices so as to reduce the possibility for chance correlation, and to weed out the less useful topological indices. A model possessing a high multiple correlation value, a small standard deviation of regression and as few topological indices as possible is considered. Thus A_{x1}^* , A_{x2}^* , 3X_c , $^6X_{pc}$ were selected and will be taken as the variables in the subsequent sections.

Results and Discussion

Regression analysis

The total 43 compounds were randomly divided into two groups: a training set containing 36 compounds and a test set containing seven (compounds labeled in footnote by "a" in Table 2). A regression model was developed with the 36 color reagents set:

$$\log \varepsilon = -68.5736 + 3743.5754A_{x1}^* - 3590.5771A_{x2}^* + 3.9464^3X_c - 1.573^6X_{pc} \quad (1)$$

$$R = 0.9621, F = 96.5310, \text{RMS} = 0.2173, n = 36.$$

Where R is the correlation coefficient, F denotes the F -test value, RMS is the root-mean-square error and n represents the number of samples.

The molar absorptivities calculated by Eq. (1) for the training set and the test set are listed in Table 2, column $\log \varepsilon$ (reg). The RMS for the test set is 0.6108.

Results with neural network

The architecture of the neural network in this study is as follows.

The experiments revealed that the results obtained by using three layer network were not satisfactory. The

Table 2 The topological indices A_{x1}^* , A_{x2}^* , 3X_c indices and $\log \varepsilon$ of color reagents

No.	A_{x1}^*	A_{x2}^*	3X_c	$^6X_{pc}$	$\log \varepsilon$	$\log \varepsilon$ (reg)	$\log \varepsilon$ (neural)
1	0.17028	0.15513	2.88428	11.57624	4.8513	5.0529	4.8362
2	0.17028	0.15518	2.88428	11.62758	4.9138	4.7783	4.8448
3	0.17060	0.15539	2.98933	11.98189	4.9445	5.1126	4.9295
4 a	0.17060	0.15533	2.98933	11.86632	4.8633	5.4955	4.8932
5	0.17505	0.16028	3.29993	12.23887	4.8751	5.0126	4.9032
6	0.17505	0.16023	3.29993	12.22771	4.8195	5.2059	4.9036
7	0.17044	0.15528	2.92419	11.74470	4.9085	4.9957	4.8719
8	0.17044	0.15523	2.92419	11.67760	4.8976	5.3023	4.8608
9 a	0.16437	0.14897	3.06826	12.50642	4.8921	4.2986	5.1000
10	0.16832	0.15325	3.03868	11.84950	4.9243	4.6237	4.9085
11	0.16828	0.15314	3.14641	12.46842	4.5051	4.3495	4.5009
12	0.16828	0.15312	3.09836	12.29574	4.7709	4.5033	4.7728
13	0.16407	0.14876	2.99816	12.10715	4.5441	4.2815	4.5384
14 a	0.16407	0.14871	3.04141	12.08663	4.1761	4.6740	4.1777
15	0.16828	0.15322	3.18855	12.38040	4.0000	4.3493	4.0544
16	0.17499	0.16027	2.98927	11.95244	4.0000	4.0596	4.0143
17	0.17498	0.16024	3.07467	11.82278	4.8513	4.6860	4.7998
18	0.17498	0.16019	3.07467	11.81742	4.8129	4.8740	4.8000
19	0.17505	0.16031	3.17319	12.46315	4.0792	4.0706	3.9417
20	0.17045	0.15540	2.87607	11.85084	4.0792	4.2232	4.0756
21	0.17044	0.15528	2.92419	11.74470	4.8062	4.9957	4.8719
22	0.17266	0.15762	2.84499	11.60925	4.8062	4.7899	4.8393
23	0.17266	0.15757	2.84499	11.55425	4.7709	5.0631	4.8044
24 a	0.16826	0.15327	3.00077	11.86267	4.8692	4.1719	4.9051
25	0.16829	0.15323	3.41699	12.67011	4.7404	4.7818	4.7463
26	0.17481	0.16012	2.85624	11.41932	4.6902	4.2131	4.6807
27	0.17481	0.16008	2.85624	11.41959	4.5798	4.3814	4.6817
28	0.17498	0.16024	3.01582	11.71408	4.8129	4.6247	4.7605
29	0.17498	0.16019	3.01582	11.71024	4.7482	4.8104	4.7603
30 a	0.17724	0.16264	2.72714	11.22387	4.8062	4.0375	4.7540
31	0.17511	0.16032	3.52519	12.89255	5.0492	4.9617	5.0446
32	0.17511	0.16027	3.52519	12.88138	5.0212	5.1552	5.0298
33	0.17275	0.15764	3.11907	12.23127	4.9823	5.1946	4.8164
34	0.17034	0.15516	3.10954	12.22992	4.9138	4.9906	4.9913
35	0.17034	0.15522	3.10954	12.28126	4.9590	4.7197	4.9948
36	0.17066	0.15543	3.21459	12.63557	5.0170	5.0540	5.0327
37 a	0.17066	0.15537	3.21459	12.51999	4.9685	5.4369	5.1071
38	0.17050	0.15532	3.14945	12.39838	4.9777	4.9370	4.9694
39	0.17049	0.15527	3.14945	12.33127	4.9956	5.2399	5.0414
40	0.17505	0.16028	3.24108	12.36775	4.9445	4.5740	4.9311
41	0.17505	0.16023	3.24108	12.36391	5.0043	4.7522	4.9353
42 a	0.16832	0.15325	3.22604	12.42742	5.0086	4.4468	4.9608
43	0.16839	0.15330	3.72772	13.46868	4.9956	4.8778	5.1075

a. Member of the test set.

results were improved by using a four-layer network, *i.e.*, one input layer, two hidden layers and one output layer. The input neurons are also the four variables mentioned in the regression analysis section: A_{x1}^* , A_{x2}^* , 3X_c and $^6X_{pc}$.

Usually, the value of each neuron is defined between 0 and 1; thus the input data should be scaled within the defined region. Note that if the value of a neuron in the input layer is zero, the connections from this neuron are always zero, *i.e.*, the information from that neuron cannot be propagated to the following layers. To avoid this situation, the values were set between 0.05 and 0.95 in our research. The total number of weights including biases should be as few as possible. Finding a compromise between the performance and the training time of the network, the optimum numbers of neurons in the hidden layers for this application were 4 and 3, respectively.

The number of output neurons is one. There are two ways to express the output neurons. Typically, single output neuron networks are used usually to predict continuous values. When multiple neurons are used in the output layer, the neuron networks are employed for pattern recognition. Therefore, in this paper, single output neural networks were used. To sum up, the architecture of the neural network in this research is 4:4:3:1.

Although both back-propagation and BFGS neural networks were found to be sensitive to the random starting weights, BFGS nets were especially so. When one is training a neural network, the objective is to find the global minimum on an surface, where the surface is multidimensional. The dimensions on the error surface correspond to the number of adjustable parameters in a given neural network. Because this error surface is so complex, it is very easy to get caught in local minima. The random weights assigned prior to training determine where on the error surface training begins. Because some sets of random weight provide better starting point than others, the results can vary depending on the random values chosen. Therefore, when one is training a neural network, it is essential to try various sets of random starting weights, several or many times, in searching for the global minimum.

Due to the tremendous number of adjustable parameters, extreme caution must be taken to avoid over-training. This problem has been overcome by two methods in the past: by using the method of weight elimination and by using an external cross-validation set.²⁰ In this experiment, we chose the test set as a monitor to control the training processing. In this way, the performance of the neural network can be enhanced markedly.²¹

The 43 compounds were divided into the same two sets as in the regression analysis above. The calculated results by the best model obtained using the network are also listed in Table 2, column $\log \epsilon$ (neural). The RMS for training set and test set are 0.05669 and 0.09621,

respectively; all these values are less than the deviations obtained by using regression analysis. The performance of the neural network is more satisfactory than and favorably compared with the regression analysis method.

The topological indices $A_{x1} - A_{x3}$ recently introduced by us and the molecular connectivity indices have been successfully used to model the structure property relationships between color reagents and their color reactions with cerium. The experiments have shown in this study that quasi-Newton neural networks give a superior performance to that given by a regression analysis. The study demonstrates convincingly that $A_{x1} - A_{x3}$ are useful topological indices. These results also demonstrate that the neural network can be used successfully in the prediction of molar absorptivities of color reactions.

The financial support of the National Science Foundation of China is gratefully acknowledged.

References

1. S. B. Savvin, T. V. Petrova and P. N. Romanov, *Talanta*, **19**, 1439 (1971).
2. T. Taketatsu, M. Kaneko and N. Kono, *Talanta*, **21**, 87 (1974).
3. N. U. Perisic-Janjic, A. A. Muk and V. D. Canic, *Anal. Chem.*, **45**, 798 (1973).
4. C. G. Su, C. S. Hu, X. P. Jia and J. M. Pan, *Anal. Chim. Acta*, **124**, 177 (1981).
5. H. Wiener, *J. Am. Chem. Soc.*, **69**, 17 (1947).
6. M. Randic, *J. Am. Chem. Soc.*, **97**, 6609 (1975).
7. H. Hosoya, *Bull. Chem. Soc. Jpn.*, **44**, 2332 (1971).
8. A. T. Balaban, *Pure Appl. Chem.*, **55**, 199 (1983).
9. L. Xu, H. Y. Wang and Q. Su, *Comput. Chem.*, **16**, 187 (1992).
10. Y. Y. Yao, L. Xu, Y. Q. Yang and X. S. Yuan, *J. Chem. Inf. Comput. Sci.*, **33**, 590 (1993); M. Guo, L. Xu, *Acta Chim. Sinica*, in press.
11. L. B. Kier and L. H. Hall, "Molecular Connectivity in Chemistry and Drug Research", Academic Press, New York, 1976.
12. J. Zupan and J. Gasteiger, *Anal. Chim. Acta*, **248**, 1 (1991).
13. P. A. Janson, *Anal. Chem.*, **63**, 357A (1991).
14. C. G. Broyden, *Inst. Math. Appl.*, **76**, 222 (1970).
15. R. Fletcher, *Comp. J.*, **13**, 317 (1970).
16. D. Goldfarb, *Math. Comp.*, **24**, 23 (1970).
17. D. F. Shanno, *Math. Comp.*, **24**, 647 (1970).
18. R. Fletcher, "Practical Methods of Optimization", Vol. I, John Wiley and Sons, New York, 1980.
19. L. Xu, J. W. Ball, S. L. Dixon and P. C. Jurs, *Environ. Toxicol. Chem.*, **5**, 841 (1994).
20. A. S. Weigend, B. S. Huberman and D. E. Rumelhart, *Int. J. Neural. Syst.*, **1**, 193 (1990).
21. L. Xu, Y. Q. Yang and C. Y. Hu, *J. Rare Earths*, **3**, 168 (1994).

(Received September 16, 1995)

(Accepted January 29, 1996)