Semiempirical Quantum Chemical Method and Artificial Neural Networks Applied for λ_{max} Computation of Some Azo Dyes

Guo-Zheng Li,*,† Jie Yang,† Hai-Feng Song,‡ Shang-Sheng Yang,§ Wen-Cong Lu,§ and Nian-Yi Chen§

Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China 200030, Laboratory of Chemical Technology of Fine Chemicals, East China University of Science and Technology, Shanghai, China 200232, and Laboratory of Chemical Data Mining, Shanghai University, Shanghai, China 200436

Received February 10, 2004

The maximum absorption wavelengths of 31 azo dyes have been calculated by two comprehensive methods using the semiempirical quantum chemical method, PM3, and the weight decay based artificial neural network (WD-ANN) or the early stopping based artificial neural network (ES-ANN). The average absolute errors of WD-ANN and that of ES-ANN are 10.07 nm and 12.40 nm, respectively. These results are much better than the results using ZINDO/S with the default value (0.585) only.

1. INTRODUCTION

Computerized prediction of the maximum absorption wavelength, λ_{max}, is a crucial step of the molecular design for new dyestuff exploration. Since the quantum chemical PPP-MO method is only useful for the dyestuff molecules with planar structure, and many dyestuff molecules exhibit nonplanar geometry, a more sophisticated method, the ZINDO/S algorithm, has been used for the λ_{max} prediction of the dyestuff molecules with nonplanar geometry.[1] As a semiempirical method, the pi-pi overlap weighting factor (denoted by $OWF_{\pi-\pi}$, an empirical parameter to provide a mechanism to adjust the relative contributions of the sigma versus pi bonding in the semiempirical quantum chemical calculation) which is an empirical parameter to adjust relative contributions of sigma versus pi bonding has to be preassigned before computation. In some computation works, the default value (0.585) has been used.[2,3] However, Adachi and co-workers have found that the λ_{max} values of some azo dyes calculated by the ZINDO/S algorithm with $OWF_{\pi-\pi} = 0.585$ gave rise to poor results.[4] In this work, we have synthesized 31 samples of new azo dyes and found that these 31 samples by ZINDO/S with $OWF_{\pi-\pi} = 0.585$ also gave poor results of λ_{max} prediction.

To find more reliable λ_{max} prediction methods for azo dyes, we try to use some comprehensive strategies: to coordinate a quantum chemical method with weight decay based artificial neural network (WD-ANN) together. These strategies are performed by one of the following two ways: (1) Calculating λ_{max} directly by WD-ANN using quantum chemical parameters as its inputs: Since the values of λ_{max} of azo dyes should be dependent on quantum chemical parameters of azo dye molecules, we can use WD-ANN to find the relationship between the experimental λ_{max} and the quantum chemical parameters of training samples directly and use the trained WD-ANN to predict the λ_{max} of other azo dyes as test samples, without using the ZINDO/S algorithm. (2) Calculating $OWF_{\pi-\pi}$ by the trained WD-ANN and use the predicted values of $OWF_{\pi-\pi}$ in ZINDO/S method for λ_{max} computation: Based on the consideration that the $OWF_{\pi-\pi}$ of azo dye molecules should be some characteristic values different for different azo dye molecules. And the values of $OWF_{\pi-\pi}$ should be relevant to other quantum chemical parameters of azo dye molecules, so it is reasonable to find the suitable values of $OWF_{\pi-\pi}$ of azo dyes in the following way: At first we adjust the value of $OWF_{\pi-\pi}$ used in the ZINDO/S calculation to *fit* the experimental value of λ_{max} for every azo dye of the training set and then use these *adapted values* of $OWF_{\pi-\pi}$ as the output of WD-ANN having quantum chemical parameters as its inputs, to find the mathematical model describing $OWF_{\pi-\pi}$ with other quantum chemical parameters. Then the $OWF_{\pi-\pi}$ of the azo dyes of the test samples can be calculated by the trained WD-ANN. Finally, this predicted $OWF_{\pi-\pi}$ can be used for the computerized prediction of the λ_{max} by the ZINDO/S algorithm.

In both of these strategies, the quantum chemical parameters calculated by the PM3 algorithm are used as the inputs of WD-ANN for training. Beside WD-ANN, the early stopping based artificial neural network (ES-ANN) is also used in training, and the results are compared with those of WD-ANN.

The prediction abilities of the above-mentioned strategies are tested by the leave-one-out cross-validation method using the data of the 31 azo dyes synthesized and measured in this work.

2. EXPERIMENTAL DATA SET AND COMPUTATION METHODS

2.1. Method of Synthesis and λ_{max} Measurement. The preparation procedure of azo dyes has been reported in the literature.[5] Their λ_{max}s in ethanol solution (1×10^{-5} M) were

* Corresponding author fax: +86-21-62932035; e-mail: lgz@sjtu.edu.cn.
† Shanghai Jiao Tong University.
‡ East China University of Science and Technology.
§ Shanghai University.

10.1021/ci049941b CCC: $27.50
Published on Web 10/14/2004

recorded on a Shimadzu UV-260 spectrophotometer. The structures of the synthesized azo dyes with their experimental λ_{max} values are listed in Table 1 (Supporting Information).

2.2. Computation of the Quantum Chemical Parameters of Azo Dye Molecules by PM3 Method. Molecular mechanics algorithm has been used to optimize the configuration of the 31 azo dye molecules, and a semiempirical quantum chemical method, PM3, has been used to calculate the following quantum chemical parameters of the 31 azo dye molecules: HOMO, LUMO, second excitation energy level, the net charges of nitrogen atoms on N═N bond, N═N bond length, dipole moment, and the dihedral angle between two ring structures on both sides of N═N bond. Table 2 (Supporting Information) lists the quantum chemical parameters of 31 azo dye molecules calculated by the PM3 method.

In the second strategy of computation, the algorithm ZINDO/S has been used to calculate λ_{max} for the following two purposes: (1) to calculate the *adapted values* of $OWF_{\pi-\pi}$ of training samples by adjusting the $OWF_{\pi-\pi}$ values to fit the experimental values of λ_{max} and (2) to use the $OWF_{\pi-\pi}$ values predicted by the WD-ANN model to calculate the predicted λ_{max} of test samples in the leave-one-out method.

At the same time, the values of λ_{max} of the 31 azo dyes are also calculated by ZINDO/S with the default value of $OWF_{\pi-\pi} = 0.585$ to get the results for comparison.

The software Hyperchem[6] 7.0 is used for all the above-mentioned quantum chemical computation.

2.3. Weight Decay Based ANN and Early Stopping Based ANN Algorithms. The relationship between the experimental λ_{max} and the structural parameters obtained by the PM3 computation has been investigated by the multiple linear regression method, the correlation coefficient of predicted values using the leave-one-out cross-validation, and the experimental values is 0.7022. It has been found that their relationship is so nonlinear that artificial neural networks are used to process the data. Multilayer perceptron neural networks[7] with three layers are used in this work, and the weight decay or early stopping techniques are employed for the depression of overfitting in the computation. Here overfitting means the error of the test data set is much larger than that of the training data set and often occurs when the data processing method is powerful.[8]

The weight decay based artificial neural network (WD-ANN) is to add a regularization term to the object function during the training, and the object function *MSEreg* can be written as

$$MSEreg = (1 - \alpha)MSE + \alpha MSW$$

where α is used to control the contribution ratio between *MSE* and *MSW*. *MSE* is the object function in the traditional training, which is written as

$$MSE = \frac{1}{l}\sum_{i=1}^{l}(y_i^e - y_i)^2$$

where y_i^e is the prediction value, y_i is the real target value, and l is the total number of examples in the training data set.

Table 4. Statistical Results of the Prediction Error of $OWF_{\pi-\pi}$

algorithm	root mean square error (nm)	averaged absolute error (nm)	standard deviation
WD-ANN	0.0325	0.0204	0.0257
ES-ANN	0.0632	0.0539	0.0335

MSW is often defined[9] as

$$MSW = \frac{1}{M}\sum_{j=1}^{M}\omega_j^2$$

where ω is the weight in ANN, and M is the number of weights.

The new object function will make ANN have small weights, so the response will be more smooth, and the overfitting can be depressed which is often taking place in ANN computation and means that even the training error is small but the test error is much higher.

In chemometrics, Tekto et al.[10] used early stopping to overcome the overfitting of ANN and proposed the early stopping based artificial neural network (ES-ANN) which achieved better results than naive ANN did.

Both the above ANN algorithms have been implemented in MATLAB.[11] For WD-ANN, $\alpha = 0.895$, while for ES-ANN, the training data set is randomly split into two equally parts, one part is used as training data sets, and the other is as validation data sets. The number of nodes of input layer, hidden layer, and output layer is 3, 8, and 1, respectively. Other parameters are adopted as the default settings in MATLAB.

3. RESULTS OF COMPUTATION

3.1. Computation of λ_{max} Directly by WD-ANN Using Quantum Chemical Parameters Obtained from PM3 Algorithm. It has been found that the experimental λ_{max} values can be correlated by WD-ANN (or ES-ANN) using three inputs: ΔE (the difference of LUMO and HOMO), LUMO, and the net charge of N(1) as inputs. The prediction ability of the mathematical model built with these three parameters is rather good, as shown in Table 3 (Supporting Information). The averaged absolute errors of prediction is 10.07 nm in the leave-one-out cross-validation test for the λ_{max} of the 31 azo dyes

3.2. Computation of $OWF_{\pi-\pi}$ and Using It in ZINDO/S Algorithm for λ_{max} Prediction. Another method of λ_{max} prediction is to improve the ZINDO/S calculation by using $OWF_{\pi-\pi}$ predicted by the mathematical model obtained by WD-ANN using ΔE, LUMO, and the net charge of N(1) as inputs. The statistical results of the prediction of $OWF_{\pi-\pi}$ by this method are listed in Table 4, and the average absolute errors of prediction of λ_{max} in the leave-one-out cross-validation method test is 13.98 nm for the 31 azo dyes.

Figure 1 illustrates the comparison of the prediction results of λ_{max} by WD-ANN modeling of both the above-mentioned strategies. The experimental values and the predicted values by the ZINDO/S method using $OWF_{\pi-\pi} = 0.585$ are also plotted for comparison.

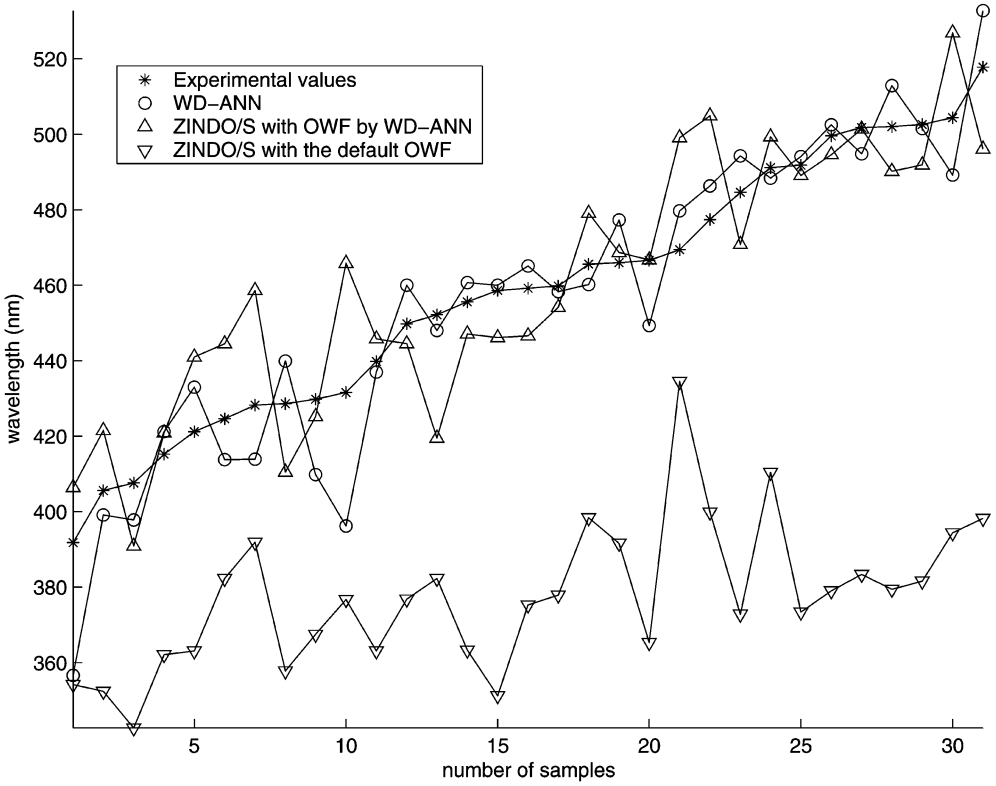

Figure 1. Comparison of the experimental λ_{max} and the predicted values of λ_{max} by different methods, in which samples are sorted (and labeled by consecutive numbers) according to the increasing number of the experimental wavelength absorption values.

Table 5. Comparison of the Prediction Errors of λ_{max} by Different Methods

algorithm	root mean square error (nm)	averaged absolute error (nm)	standard deviation
WD-ANN	12.99	10.07	8.33
ES-ANN	17.63	12.40	12.74
ZINDO/S[a]	16.9519	13.9806	9.7453
ZINDO/S	85.1540	80.5516	28.0725

[a] ZINDO/S with the predicted values of $OWF_{\pi-\pi}$ by WD-ANN.

Table 5 lists the statistical results on the prediction errors of all methods mentioned above.

4. DISCUSSION

According to the data processing in this work, the most important feature affecting the values of λ_{max} of azo dyes is the difference of HOMO and LUMO. It is easily understandable since the longest wavelength absorption band of azo dye molecule is chiefly dependent on the excitation from the highest occupied MO(HOMO) to the lowest unoccupied MO(LUMO). Besides, the correlation between the net charge of nitrogen atoms on the N═N bridge and λ_{max} may be explained as follows: it can be seen that the net charge values of nitrogen atoms on both sides of the N═N bridge, N(1) and N(2), are inversely correlated roughly, and their net charges can affect the easiness of the charge transfer through the N═N bridge. While the longest wavelength absorption for the azo dye is just the result of the intermolecular charge transfer of this type, therefore the net charges of nitrogen atoms of the N═N bridge are also relevant to the λ_{max} of azo dyes. So that the physical meaning of the mathematical model obtained by ANN is only reliable in the feature space not very far away from the domain of training samples. It may be necessary to be modified by data processing of more samples if it is necessary to predict some samples having quantum chemical parameters very far away from that used in the present work. At any rate, the coordination of WD-ANN and the semiempirical quantum chemical calculation can provide effective methods for azo dye λ_{max} estimation as the training samples with similar quantum chemical behavior are available.

Besides, it should be noticed that the neglect of solvent effect on the absorption spectra may be another source of error in the prediction of λ_{max} by this method.[12] Since the solvent effect can induce the change of energy levels of the $\pi-\pi^*$ transition, it makes the absorption spectra become non-Gaussian so that the position of maximum can be shifted to some degree.

The results of ES-ANN are much better than those by ZINDO/S with the default $OWF_{\pi-\pi}$, but they are slightly worse when compared with the results of WD-ANN. We think this may owe to the small number of samples, for ES-ANN, when the small data set is split into two data sets for training and validation simultaneously, the number of training samples is so small that bias is introduced into the neural networks. At the same time, WD-ANN uses the data set as long as possible and has not been biased.

ACKNOWLEDGMENT

This work is financially supported by Natural Science Foundation of China under grant number 50174038. Prof. Kong-Chang Chen, from the Laboratory of Chemical Technology of Fine Chemicals in East China University of Science and Technology, gave valuable advice for this work. Thanks are also due to two anonymous reviewers for their valuable advice.

Supporting Information Available: Structure and the observed λ_{max} of 31 synthesized azo dyes (Table 1), quantum chemical parameters of 31 azo dyes by the PM3 method (Table 2), and results and errors of prediction of λ_{max} of 31 azo dyes (Table 3). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Griffiths, J. *Color and constitution of organic molecules*; Academic Press: London, 1976.

(2) Zerner, M. C.; Loew, G. H.; Kirchner, R. F.; Mueller-Westerhoff, U. T. An Intermediate Neglect of Differential Overlap Technique for Spectroscopy of Transition-Metal Complexes. Ferrocene. *J. Am. Chem. Soc.* **1980**, *102*, 589−599.

(3) Starev, K. K.; Zerner, M. C. Outer-sphere charge-transfer effects on the spectroscopy of the $[Ru(NH_3)_5(py)]^{2+}$ complex. *J. Am. Chem. Soc.* **1995**, *117*, 8684−8685.

(4) Adachi, M.; Nakamura, S. Comparison of the INDO/S and CNDO/S method for the absorption wavelength calculation of organic dyes. *Dyes Pigments* **1991**, *17*, 287−296.

(5) Song, H.-F.; Chen, K.-C.; Tian H. Synthesis and absorption properties of some new azo-metal chelates and their ligands. *Dyes Pigments* **2004**, *60*, 111−119.

(6) Hyperchem 7.0, the Hypercube Inc., http://www.hyper.com/, 2004.7.

(7) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*, 2nd ed.; Wiley-VCH: Weinheim, 1999.

(8) Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification*, 2nd ed.; Wiley: New York, 2000.

(9) Moody, J. E. The Effective Number of Parameters: An Analysis of Generalization and Regularization in Nonlinear Learning Systems. In *Advances in Neural Information Processing Systems*; Moody, J. E., Hanson, S. J., Lippmann, R. P., Eds.; Morgan Kaufmann Publishers: 1992; Vol. 4, pp 847−854.

(10) Tetko, V. I.; Livingstone, J. D.; Luik, I. A. Neural Network Studies: 1. Comparison of Overfitting and Overtraining. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 826−833.

(11) Demuth, H.; Beale, M. *Neural Network Toolbox User's Guide for Use with MATLAB*, 4th ed.; The Mathworks, Inc.: Massachusetts, 2001; Chapter 5, pp 51−61.

(12) Luzhkov, V.; Warshel, A. Microscopic calculation of solvent effects on absorption spectra of conjugated molecules. *J. Am. Chem. Soc.* **1991**, *113*, 4491−4499.

CI049941B