Quantitative prediction of substituted products based on quantum-chemical parameters and neural network method

WANG, Xue-Ye*(王学业) SONG, Huang(宋锽)

College of Chemistry and Chemical Engineering, Xiangtan University, Xiangtan, Hunan 411105, China

The criterion of orientating group of electrophilic aromatic nitration was discussed by means of pattern recognition method with quantum-chemical parameters as features, and the product ratios of the reactions were quantitatively calculated using artificial neural network (ANN) method with the same parameters as inputs, based on the *ab initio* calculation of quantum chemistry. The quantum-chemical parameters involved orbital energy, orbital electron population, atomic total electron density and atomic net charge. The predicted values are in agreement with experimental results and the predicted error of the ANN with quantum-chemical parameters for the reaction is the smallest among the all methods.

Keywords Product ratio, quantitative structure-activity relationship, artificial neural network, quantum-chemical parameters

Introduction

Chemists can usually predict the major products in electrophilic aromatic substitution simply by examining the substituent and classifying it as either *ortho-para* or *meta* directing group without any attempt to predict percentages of each product. It is usually implied that the products formed are only *ortho-para* or only *meta*, but it is apparent from examining the chemical literature that many intermediate cases exist which give mixtures of all three products. In these cases, it would be useful to have a prediction of the *ortho-para* to *meta* product ratio.

Electrophilic aromatic substitution involves the substitution of a hydrogen atom on an aromatic ring by an electrophile such as nitric acid when the aromatic ring has a substituent, the electrophilic substitution will lead to *ortho*-, *meta*-, or *para*-substituted products. The ratio of isomers formed depends mainly on the nature of the substituent X and little on the electrophile Y. Substituents may be divided into two classes, *ortho-para* directors, which tend to be electron donors, and *meta* directors, electron acceptors. Resonance effects of the substituents can reinforce or oppose these inductive effects and thereby affect the product ratio. Steric hindrance by large substituents can also affect the reaction by blocking the adjacent positions.

Frontier molecular orbital theory[1] involves perturbation expression of the energy arising from stabilizing interaction of the highest occupied molecular orbital (HOMO) on one fragment and the lowest unoccupied molecular orbital (LUMO) on the other, and vice versa. The orbital coefficients, orbital energies and others decide the interaction energy.

Gasteiger *et al.*[2] generated quantitative parameters to predict complex reactions by the application of pattern recognition methods. Artificial neural network (ANN) is a promising new method for solving chemical problems by virtue of its ability to construct an internal representation of the problem which allows predictions to be made for similar problems. It may be particularly useful in cases where it is difficult to specify exact rules governing reactivity or where several overlapping or seemingly opposing rules apply.[3] ANN is an effective method for the investigation and prediction of complicated phenomena affected by many factors, so it would seem reasonable to use ANN and quantum-chemical parameters for the calculation and prediction of the ratio of the *ortho-para* to *meta* product. We will discuss the criterion of orientating group by means of pattern recognition method with quan-

* Received August 11, 1999; accepted March 30, 2000.
Project (No. 99C113) supported by the Eduction Commission of Hunan Province.

tum chemical-parameters as features, and the quantitative prediction of product ratio, using ANN method with the same parameters as inputs based on the calculation of quantum chemistry.

Method of computation

Quantum-chemical parameters were obtained by using HF method with STO-3G basis set and standard geometrical configuration.[4] The quantum-chemical parametrs are orbital energy, orbital coefficient, net atomic charge, and atomic electron population, based on the mechanism of electrophilic aromatic substitution and the formula of the interaction energy between two molecules or fragments. Then we can investigate the activity and product ratio by means of pattern recognition and ANN methods with these parameters.

We used quantum-chemical parameters to span a multi-dimensional space and find semi-empirical rules from experimental data by pattern recognition in this space. The rules found can be used for computerized prediction. The pattern recognition methods used here are computational methods mapping the patterns in multi-dimensional space to two-dimensional figures, along with some techniques for mapping the two-dimensional figures back to original multi-dimensional space. The principal component analysis (PCA)[5] of pattern recognition was used. Artificial neural network is a new type of information processing system based on modeling the neural system structures of human brain.[6] It has some remarkable properties such as self-learning and adaptation, a resistance to noise, a high degree of fault tolerance, which make it suitable for nonlinear problems with complex factors. It is powerful for exploiting information from a vast amount of experimental data through learning, and is especially useful for quantitative prediction. The network consists in general of an input layer, an output layer, and any number of intermediate layers, called hidden layers. Each unit in the network is influenced by those units to which it is connected, the degree of influence being dictated by the values of the links or connections. The overall behavior of the system can be modified by adjusting the values of the connections, or weights, through the repeated application of a learning algorithm. One of the most popular algorithms is the back propagation (BP) algorithm.[6] In this work, a three-layered neural network and back propagation algorithm were used with chemical bond parameters. The transfer function $f(x)$ is usually a nonlinear function, and we selected hyperbolic tangent function as transfer function, $f(x) = (e^x - e^{-x})/(e^x + e^{-x})$. To avoid local minimum, a simulated annealing technique was used.

Results of computation

Criterion of orientating group

There were twenty-seven reaction systems in the sample set of electrophilicaromatic nitrating substitution for C_6H_5X, and the substituent X is listed as follows: (1) F, (2) OH, (3) $NHCOCH_3$, (4) OCH_3, (5) $CH_2CH_2OCH_3$, (6) CH_3, (7) CH_2NH_2, (8) CH_2OCH_3, (9) CH_2F, (10) CH_2COOH, (11) CH_2NO_2, (12) $CH_2N^+H_2CH_3$, (13) $COOCH_2CH_3$, (14) $COOCH_3$, (15) $N^+H_2CH_3$, (16) $CONH_2$, (17) $COCH_3$, (18) CHO, (19) COOH, (20) CN, (21) NO_2, (22) $CH_2SO_2O^-$, (23) CH_2Cl, (24) $CH_2SO_2NH_2$, (25) $CHCl_2$, (26) CH_2SO_2Cl, (27) CCl_3.

The calculated results of *ab initio* (STO-3G) method are listed in Table 1. The orbital energy for HOMO (E_{HOMO}, unit: a.u.), atomic orbital electron population (p), atomic total electron density (Q), atomic net charge (q) and atomic charge with hydrogen summed into carbon atom (q') were selected in order to describe the activity of the reactions. Finally, we used the $-E_{HOMO}$, $p_2 - p_3(\times 10^2)$, p_4, $Q_2 - Q_3(\times 10^2)$, $q_3 - q_2(\times 10^2)$, $-q_4(\times 10^2)$, $q'_3 - q'_2(\times 10^2)$ and $-q'_4(\times 10^2)$ parameters as the features affecting the reactions. The subscript represents the number code of carbon atom. We used the average value between carbon-2 and carbon-6 as the corresponding value for carbon-2 (*ortho*), and the average value between carbon-3 and carbon-5 as the value of carbon-3 (*meta*). In Table 1, the ratios of *meta* to *ortho-para* products (ρ) were quoted from Ref.7.

The substituents can be classified into two classes: the first class, $\rho < 0.50$, called as the first orientating group (*ortho - meta* group), and the second class, $\rho \geq 0.50$, called as the second orientating group (*meta group*). We deleted the parameters, which have a linear relation to each other and have a little influence on activity by PCA method. Fig. 1 is the linear mapping on the plane with the first PC *vs*. the second PC from four-

dimensional space spanned by $-E_{HOMO}$, p_4, Q_2-Q_3 and q_3-q_2 quantum-chemical parameters. It illustrates that the representative points of the first orientating groups and ones of the second orientating groups are distributed in different regions. We obtained a criterion for the orientating groups.

$$D = 17.21 - 1.36(-E_{HOMO}) - 16.93p_4 - 0.12(Q_2 - Q_3) + 0.14(q_3 - q_2)$$

$D < 0$ denotes the *ortho-para* orientating group, $D > 0$ the *meta* orientating group.

The relative activity of a group is mainly decided by the electron population on carbob-4, based on the regression equation. So, we obtained the order of reactive activity for groups from *ortho-para* to *meta* substitution, $CH_2SO_2O^-$, OCH_3, OH, CH_2OCH_3, $NHCOCH_3$, CH_2NH_2, CH_2F, F, $CH_2CH_2OCH_3$, $CHCl_2$, CH_3, CH_2Cl, CH_2COOH, $CH_2SO_2NH_2$, $COOCH_2CH_3$, CH_2SO_2Cl, $CONH_2$, $COOCH_3$, CH_2NO_2, $COCH_3$, CHO, COOH,

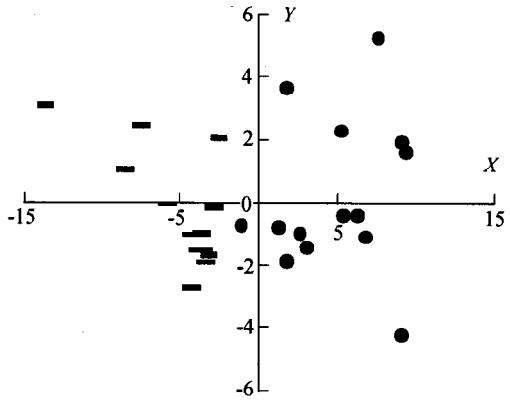

Fig. 1 Classification of orientating groups (PCA method). (●: $\rho < 0.50$ -: $\rho \geq 0.50$) $X = -0.15(-E_{HOMO}) + 0.95p_4 + 0.24(Q_2 - Q_3) + 0.15(q_3 - q_2)$, $Y = 0.33(-E_{HOMO}) - 0.22p_4 + 0.77(Q_2 - Q_3) + 0.50(q_3 - q_2)$.

Table 1 Calculated results of *ab intio* method and product ratios

No.	$-E_{HOMO}$	$p_2 - p_3$	p_4	$Q_2 - Q_3$	$q_3 - q_2$	$-q_4$	$q'_3 - q'_2$	$-q'_4$	ρ_{exp}
1	0.26523	5.834	1.02046	3.5411	3.4991	7.0714	3.2437	-0.7052	0.00
2	0.24443	9.152	1.03895	3.5405	4.2995	7.8291	4.7180	-1.8434	0.00
3	0.21477	2.063	1.02642	6.1769	4.2346	7.7479	2.4665	-2.5560	0.02
4	0.22965	6.893	1.03988	3.8442	3.5942	7.9179	2.6029	-2.3369	0.02
5	0.25602	2.191	1.01508	1.4706	1.3227	7.0106	1.0850	-1.1593	0.03
6	0.26675	1.776	1.01002	1.1222	0.8605	6.7486	1.1603	-0.6775	0.04
7	0.22310	4.993	1.02793	2.4166	2.3309	7.5062	2.8841	-1.8146	0.10
8	0.21702	4.338	1.03133	2.1149	2.1389	7.6988	1.9609	-2.2678	0.12
9	0.24254	4.492	1.02431	2.1639	2.3410	7.3459	2.5904	-1.5081	0.18
10	0.27693	4.603	1.00216	5.1996	1.5436	6.3342	1.9567	0.3414	0.22
11	0.27146	-3.210	0.98394	2.7905	1.9773	5.9231	-3.0533	0.3797	0.56
12	0.41132	4.426	0.96300	1.4735	2.1062	4.4782	4.7111	4.6705	0.60
13	0.27029	-2.403	0.98639	-0.3321	-0.8440	5.7401	-2.1679	0.8158	0.68
14	0.27472	-2.424	0.98439	-0.3501	-0.8666	5.6422	-2.1335	1.0150	0.68
15	0.44088	1.748	0.93358	1.1764	1.3352	3.1514	2.5637	7.1103	0.70
16	0.26758	-1.592	0.98555	-0.1070	-0.8381	5.6520	-1.2084	1.0630	0.70
17	0.26869	-3.604	0.98304	-0.5090	-1.3311	5.6552	-3.0691	0.8327	0.72
18	0.27784	-2.201	0.98241	-0.2332	-0.9199	5.5472	-1.5622	1.2152	0.72
19	0.28190	-2.836	0.97996	-0.4844	-1.0260	5.4464	-2.3682	1.3668	0.80
20	0.29461	-2.320	0.97177	-0.5698	-1.5878	4.9857	-2.2802	2.3540	0.82
21	0.26820	-4.406	0.95706	-0.9385	-1.7899	4.3593	-3.9407	3.3926	0.93
22	0.00324	0.786	1.04222	1.7242	0.8703	8.2445	-0.7443	-4.2976	0.14
23	0.26872	2.259	1.00566	1.5145	1.4506	6.5454	0.7148	-0.2030	0.16
24	0.25834	1.507	0.99334	1.3502	1.0881	5.8922	1.6263	1.0155	0.31
25	0.14528	2.802	1.01011	2.0021	1.7289	6.8214	1.3782	-0.6916	0.34
26	0.29970	2.286	0.98613	1.3193	1.1226	5.5282	2.0155	1.8142	0.51
27	0.11819	-1.246	0.97941	0.3407	-0.2234	5.3216	-0.5989	1.7425	0.64

CCl_3, CN, $CH_2N^+H_2CH_3$, NO_2, $N^+H_2CH_3$. These results are basically in agreement with the experiment.[8] The criterion shows that the decrease of orbital energy E_{HOMO}, the increase of the electron population on carbon-4, the difference of the total electron densities between carbon-2 and carbon-3, and the difference of the net charges between carbon-2 and carbon-3 favor to form *ortho-para* products. The influenced level on substitution depend on four factors being p_4, E_{HOMO}, $Q_2 - Q_3$ and $q_3 - q_2$ from big to small, based on the product of the average value and regression coefficient for every factor. The large atomic orbital electron population and the large atomic net charge favor the substitution by electrophile. These results are in agreement with the mechanism of electrophilic aromatic substitution.[9]

Quantitative calculations of product ratios

In order to compare the calculated results with other methods, the training set and testing set of the samples are the same as Ref .7. The seven parameters, $-E_{HOMO}$, $p_2 - p_3$, p_4, $Q_2 - Q_3$, $q_3 - q_2$, $q'_3 - q'_2$ and $-q'_4$ in Table 1, were used as the inputs of ANN (p_4 has a linear relation to $-q_4$, and the relative coefficient is 0.99, so we deleted $-q_4$ parameter). The ratio of *meta* to *ortho-para* product was used as the output of ANN. A three-layered ANN (two hidden nodes) was used for learning the training set. The calculated results for training set are listed in Table 2. After training, the trained ANN was used for predicting the product ratios of the "six unknown" systems which were not included in the training set, and their predicted results are listed in Table 3.

Table 2 Calculated results of product ratios for training set

No.	ρ_{exp}	ρ_{reg}	ρ_{pre}	No.	ρ_{exp}	ρ_{reg}	ρ_{pre}	No.	ρ_{exp}	ρ_{reg}	ρ_{pre}
1	0.00	0.03	0.00	10	0.22	0.22	0.46	21	0.93	0.93	0.88
3	0.02	0.03	0.02	13	0.68	0.66	0.59	22	0.14	0.22	0.18
4	0.02	0.05	0.04	15	0.70	0.68	0.50	23	0.16	0.16	0.36
5	0.03	0.11	0.09	17	0.72	0.75	0.61	24	0.31	0.34	0.45
6	0.04	0.13	0.26	18	0.72	0.72	0.66	25	0.34	0.19	0.28
8	0.12	0.10	0.20	19	0.80	0.78	0.72	26	0.51	0.49	0.46
9	0.18	0.07	0.11	20	0.82	0.83	0.72	27	0.64	0.67	0.47

Note: ρ_{reg} is the learning result, ρ_{pre} is the predicted result by leave-two-out method.

Table 3 Predicted results of product ratios for testing set

No.	ρ_{exp}	ρ_{pre}	No.	ρ_{exp}	ρ_{pre}
2	0.00	0.11	12	0.60	0.50
7	0.10	0.10	14	0.68	0.66
11	0.56	0.47	17	0.70	0.61

The residual errors of the learning and predicted results of product ratios for electrophilic aromatic substitution with quantum-chemical parameter-ANN method in this paper, the connection table-ANN method,[7] charge vector-ANN method,[7] and CAMEO (chemical expert system) method[10] are listed in Table 4. It shows that the predicted error using quantum chemical method is the smallest among four methods, the residual error of this method for learning is larger than the one of connection table-ANN method, that may be caused by the overfitting. In Ref. 7, the authors used four hidden nodes, twenty-five inputs and two outputs, but the predicted error is larger than the one of our method.

Table 4 Calculated residual error for nitration activity

Method	Learning result		Predicted result	
	Max error	Average error	Max error	Average error
Quantum Chem.-ANN	0.15	0.035	0.11	0.071
Charge Vector-ANN	0.18	0.052	0.55	0.198
CAMEO	0.90	0.180	0.60	0.236
Connection-ANN	0.04	0.003	0.44	0.121

Therefore, if some suitable quantum-chemical parameters are selected, based on quantum chemical calculation, the pattern recognition and ANN method can be used correctly to determine the types of orientating group and quantitatively predict the ratios of *meta* to *ortho-para* products for electrophilic aromatic substitutions. In addition, the predicted error of the ANN with quantum chemical parameters as inputs for the reactions is the smallest in comparison with other methods.

References

1. Kahn, S.D.; Pau, C.F.; Overman, L.E.; Hedre, W. J., *J. Am. Chem. Soc.*, **108**, 7381(1986).
2. Gasteiger, J.; Saller, H.;Löw, P., *Anal. Chim. Acta*, **191**, 111(1986).
3. Elrod, D.W.; Maggiora, G.M.; Trenary, R.G., *J. Chem. Inf. Comput. Sci.*, **30**, 477(1990).
4. Liao, M.Z.; Wu G.S.; Liu, H.L., *Ab Initio Method in Quantum Chemistry*, Tsinghua Unieversity Press, Beijing, 1984, p.176.
5. Jolliffe, I.T., *Principal Component Analysis*, Springer Verlag, New York, 1986, p.18.
6. Lippmann, L.P., *IEEE ASSP Magazine*, **4**, 4(1987).
7. Elrod, D.W.; Maggiora, G.M.; Trenary, R.G., *Tetrahedron Comput. Method.*, **3**, 163(1990).
8. Organic Chemistry Group of Tianjin University, *Organic Chemistry*, People Eduction Press, Beijing, 1978, p.105.
9. Wang, J.T,. *Advanced Organic Chemistry*, People Education Press, Beijing, 1980, p.181.
10. Gushurst, A.J.; Jorgensen, W.L., *J. Org. Chem.*, **53**, 3397(1988).

(E9908094 SONG, J.P.; DONG, L.J.)