[Article] www.whxb.pku.edu.cn

一种确定反应中间态几何特征和能量的综合性方法

郑 铮[§] 刘振明[§,*] 张亮仁
(北京大学药学院药物化学系, 天然药物及仿生药物国家重点实验室, 北京 100191)

摘要: 通过综合使用传统的过渡态优化算法、数学统计工具以及人工神经网络算法(ANN)找到一种不依赖于反应物起始构象而得到化学反应中过渡态结构和能量的方法. 在两个反应物互相接近的过程中, 每一步的几何构象都对应着一个系统能量值. 本研究的目的是尽可能地收集处在反应能量面上的这种能量点值. 通过采用几何参数作为自变量对势能面进行模拟研究, 得到了势能面上对应过渡态结构的一阶鞍点. 采用乙醛负离子和甲醛作为反应物, 对经典的醛醇缩合反应中的亲核进攻步骤进行了研究. 对内禀反应坐标(IRC)路径的计算是从反应物的三组不同起始构象出发, 最终获得了反应势能面上的 96 个点. 本研究中的势能面采用人工神经网络算法进行模拟研究, 并利用交叉验证方法评估得到的结果, 避免了采用人工神经网络算法时过度拟合情况的发生.

关键词: 反应过渡态; 反应物几何构象; 人工神经网络; 反应势能面; 一阶鞍点; 交叉验证
中图分类号: O641

A Combined Method for Determining Reaction Transition State Geometry and Energy

ZHENG Zheng[§] LIU Zhen-Ming[§,*] ZHANG Liang-Ren
(*State Key Laboratory of Natural and Biomimetic Drugs, Department of Medicinal Chemistry, School of Pharmaceutical Sciences, Peking University, Beijing 100191, P. R. China*)

Abstract: We found an alternative method for the derivation of transition state structure energy in chemical reactions which would be less dependent on the starting geometry of reactants by combining a mathematical tool and artificial neural networks (ANN) with conventional transition state optimization algorithms. When two reactants approach each other, every geometric structure corresponds to a system energy value. The purpose of this investigation was to collect as many energy values on the reaction energy surface as possible. By simulating the energy surface using the geometric parameters as independent variables, the first order saddle point in the energy surface corresponding to the transition state structure was derived. The nucleophilic attack step of a classical Aldol reaction was studied using acetaldehyde anion and formaldehyde as reactants. The intrinsic reaction coordinate (IRC) path calculation started with 3 different sets of starting reactant geometries and 96 points on the reaction energy surface were derived. The energy surface was simulated using ANN. Cross-validation was applied to evaluate the result and avoided a possible overfitting of the ANN.

Key Words: Reaction transition state; Geometry of reactant; Artificial neural network; Reaction energy surface; First order saddle point; Cross-validation

In the investigation of reaction dynamics, finding transition state structure is a difficult task for many reasons[1−3]. The most serious problem is that the starting geometry is easy to make the algorithms fail to derive the correct transition state structure and

Received: February 16, 2009; Revised: April 14, 2009; Published on Web: May 20, 2009.

*Corresponding author. Email: zmliu@bjmu.edu.cn; Tel: +86-10-82805514; Fax: +86-10-82802724.
§These authors contributed equally to this work.

The project was supported by the National High-Tech Research and Development Program of China (863) (2006AA02Z337).
国家高技术研究发展计划(863) (2006AA02Z337)资助

get the local minimum in the reaction energy surface. With Quasi-Newton technique, the optimization will only be able to find the correct geometry if the starting geometry is sufficiently close to the transition state geometry. The starting geometry should also be closer to the reaction transition state structure than to any other structures satisfying the same mathematical criteria. As the results of conventional transition state optimization algorithms were to a great extent bound to the starting geometry of reactants in the reaction dynamic research of many reactions, sorts of combined methods that were more reliable and accurate were raised for finding the transition state structures and the intrinsic reaction coordinate paths these years[4–9].

Artificial neural networks (ANN) was introduced in calculation of finding transition state. ANN is a learning system based on a computational technique, which attempts to simulate the neurological processing ability of the brain[10–12]. ANN could be applied to quantify a non-linear relationship between the causal factors and outputs by means of iterative training of the input data. Generalization capacity of trained networks is far exceeding that of poly multiple regression. Back propagation (BP) algorithm was employed in the current investigation[13].

The nucleophilic attack step of classical Aldol reaction was selected (Fig.1). Acetaldehyde anion and formaldehyde acted as reactants and 3-hydroxy-propionaldehyde anion was the product.

1 Methods

In the current investigation, three starting geometries were set firstly by fixing the distance between reactive spots on two optimized reactants to van der Waals distance (D), D+0.005 nm, D+0.01 nm, D+0.015 nm, and D+0.02 nm. Quadratic synchronous transit method (QST3) was performed and hence five local minimum energy values were derived. Then HF calculation was performed based on the five structures corresponding to these local energy minima in order to find the reaction paths from the starting geometries leading to these corresponding structures. Ninety-six points on the reaction energy surface and their corresponding geometric parameters were derived. To ensure the simulation ability of ANN, input variables should be orthogonal and significant. In the current investigation, distance of C—C bond of acetaldehyde anion (D_1) and that of C═O double bond of formaldehyde (D_2) were selected to act as inputs to the ANN for simulating the reaction energy surface. Saddle point of the surface can then be calculated.

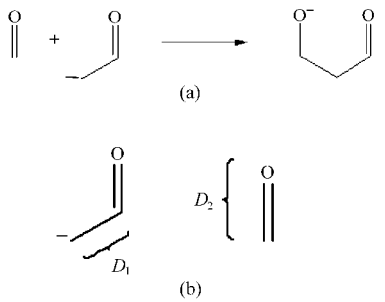

Fig.1 (a) Nucleophilic attack step of classical Aldol reaction; (b) distance of C—C bond of acetaldehyde anion (D_1) and C═O double bond of formaldehyde (D_2)

D_1 and D_2 are selected to act as inputs to the ANN for simulating the reaction energy surface.

2 Results and discussion

2.1 Data preparation and standardization

Firstly, the two reactants, acetaldehyde anion and formaldehyde, were optimized using RHF/6-31G with Gaussian 98W[14]. As second step, distance between α-carbon of acetaldehyde anion and carbon of formaldehyde was set to van der Waals distance (D), D+0.01 nm, and D+0.02 nm, respectively. Using QST3 method and IRC algorithm performed by Gaussian 98W, ninety-six eigenvalues of frontier molecular orbitals derived from D_1, D_2 and system energies were collected (listed in Supporting Information, Table S1, which is available free of charge *via* the internet at http://www.whxb.pku.edu.cn).

Data standardization is crucial for both input and output data in BP algorithm. The purpose of data standardization of causal factors is to avoid the plateau phenomenon in training and accelerating the learning speed of the network[15,16]. Activation functions between biases are saturated nonlinear functions; hence saturation of training readily occurs in the training process, which makes the accuracy of training increase slowly or even hardly. Meanwhile, any input of nodes in the secondary or final bias is the weighted sum of all the outputs of the nodes connected to it in the former bias. Range restriction of the causal factor data can reduce the difficulties in training the weights of all connections and avoid the saturation of training. For data of the biological activity, widely ranged data are readily causing dead links in training[17]. In polynomial simulation, standardization is crucial as well. Quite often the data sets are a mixture of various measurements made on different scales and/or in different units. Standardization is to eliminate the measurement units and minimize the influence of one component with very large magnitudes as opposed to other data components that are small in magnitude.

2.2 BP algorithm models, artificial neural networks training and cross-validation

In the current investigation, standardization is performed as follows:

$$x' = \frac{x - x_{\min} + 0.01}{x_{\max} - x_{\min} + 0.01} \quad (1)$$

$$y' = \frac{y - y_{\min} + 0.01}{y_{\max} - y_{\min} + 0.01} \quad (2)$$

where, $x_{\max}$ and $x_{\min}$ are frontier orbital eigenvalues in the x direction on the maximum distance and minimum distance, respectively; x is the characteristics of the orbital in the front line in the x direction of the value of the actual distance; x' is frontier orbital eigenvalues in the x direction after normalized values. $y_{\max}$ and $y_{\min}$ are frontier orbital eigenvalues in the y direction on the maximum distance and minimum distance, respectively; y is the characteristics of the orbit in the front line in the y direction of the value of the actual distance; y' is frontier orbital eigenvalues

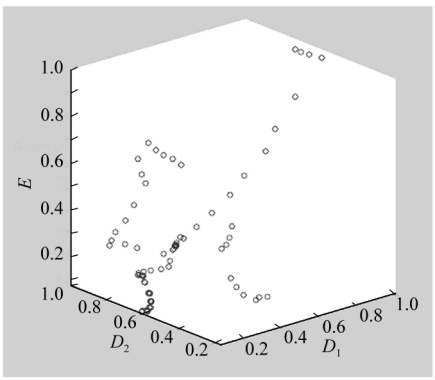

Fig.2 Data plot of the system energy related with LUMO and HOMO

HOMO: the highest occupied molocular orbital, LUMO: the lowest unoccupied molecular orbital; E, D_1 and D_2 are normalized

in the y direction after normalized values. This method made the distribution of data between 0 and 1. We added 0.01 in both numerator and denominator in order to avoid the minimum value in data of inputs or outputs turning to zero (Fig.2).

Theoretically, BP algorithm is a model of nonlinear fitting process. As a useful module of Matlab 6.5 software package, ANN was performed in the current investigation. Briefly, the model of BP algorithm is composed of connections of the processing elements (nodes). The processing elements transfer data from one bias to the next through activation functions until the output bias. Three-bias BP model was used under the current investigation, for its higher tolerance to mistakes and a comparatively simple structure for training.

The number of hidden nodes has a particularly large effect on the generalization capability of the network. The principle for determining the number is to use as few hidden neurons as possible on the prerequisite to maintain the precision of training. According to Kolmogorov′s theorem[18], networks with 1 hidden neuron to 7 were all tested. The network with 6 hidden neurons was found to be the simplest in structure while with comparatively fast decent velocity of error.

The network passes through activation functions defined as the sigmoid function. Under the current investigation, tansig(x) is selected as the activation function that defines the transference from input bias to hidden bias. lgsig(x) is selected as the activation function that defines the transference from hidden bias to output bias.

TRAINGDX, a network training function that updates weight and bias values according to gradient descent momentum and an adaptive learning rate, was selected to the model that deals with the relationship between the data of inputs and the data of outputs. Comparing with other training strategies, TRAINGDX had the highest convergent velocity in training in the current investigation. Performance goal was set to 0.001 in the current investigation and minimum performance gradient to 10^{-10}. Cross-validation (leave-1-out method) of the training results was employed in order to monitor and avoid overfitting in training.

Cross-validation (leave-n-out method) of the training results was employed in order to monitor and avoid the phenomenon of overfitting. Leave-1-out method was introduced under the current investigation to judge the ending of training, because test set with one datum makes the test more objective and more acceptable. From all the compounds, test set is chosen one by one.

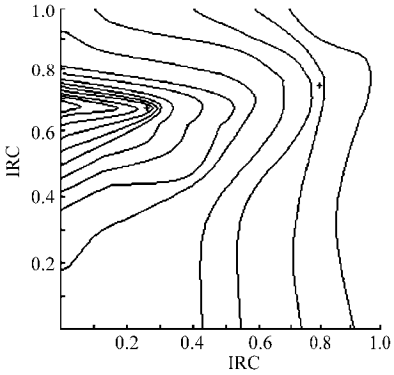

Fig.3 Contour line of reaction energy surface

The result of cross-validation was listed in Supporting Information (Table S2), R^2 reached 0.979. This proved that surface of reaction energy simulated by BP algorithm of ANN was convincing.

Hence a quantitative relationship was built between the geometric parameters and the system energy in reaction using BP algorithm. The first order saddle point that was an energy maximum in one direction and a minimum in all others in the surface should correspond to the transition state. Based on the trained networks, two stagnation points existed in the surface and one of them was saddle point. The energy of transition state was 0.28357 (standardized datum) or −266.15721 hartree, its corresponding D_1 was 0.2634 (standardized datum) or 0.1545382 nm; eigenvalue of the length according to LUMO was 0.78466 (standardized datum) or 0.1338472 nm.

Contour line was ploted in Fig.3. Two stagnation points were marked with cross in the figure with the left one indicating the energy of transition state.

3 Conclusions

In conclusion, the current investigation provided a method to search the energy of transition state structure and the corresponding geometric character. The more points on the reaction energy surface were found, the more convincing and accurate the simulation of ANN would be. Geometric parameters are the best chose to act as input variables to the ANN for the simulation of system energy, because they are orthogonal and independent to each other, and because they are the most direct and exact descriptors to define a structure, which increases the simulation accuracy of ANN. In the current investigation, two significant geometric descriptors were selected as the independent variables, for the simplification. However, the disadvantage of this method is that reaction with complex reagents would require a large number of geometric descriptors. Large amount of input variables would add to the difficulty of ANN training. It would lower the accuracy of the simulation of ANN, and take much

longer CPU time to complete the work.

Acknowledgments: The authors would like to thank Prof. Lai, Luhua in College of Chemistry and Molecular Engineering, Peking University and Prof. Qian, Minping in School of Mathematical Science, Peking University for related conversation and suggestions.

Supporting Information Available: Eigenvalues of frontier molecular orbitals and values of system energy, and cross-validation monitored ANN training result have been included. This information is available free of charge *via* the internet at http://www.whxb.pku.edu.cn.

References

1 Ishihara, K.; Kondo, S.; Kurihara, H.; Yamamoto, H.; Ohashi, S.; Inagaki, S. *J. Org. Chem.*, **1997, 62**(10): 3026

2 Caramella, P.; Quadrelli, P.; Toma, L. *J. Am. Chem. Soc.*, **2002, 124**(7): 1130

3 Bongiorno, A.; Pasquarello, A.; Hybertsen, M. S.; Feldman, L. C. *Phys. Rev. Lett.*, **2003, 90**(18): 186101

4 Vanquickenborne, L. G.; Vinckier, A. E.; Pierloot, K. *Inorg. Chem.*, **1996, 35**(5): 1305

5 Li, W. S.; Chung, W. S.; Chao, I. *Chemistry*, **2003, 9**(4): 951

6 Orlova, G.; Goddard, J. D.; Brovko, L. Y. *J. Am. Chem. Soc.*, **2003, 125**(23): 6962

7 Su, K. H.; Wei, J.; Hu, X. L.; Yu, H.; Lü, L.; Wang, Y. B.; Wen, Z. Y. *Acta Phys. -Chim. Sin.*, **2000, 16**(7): 643 [苏克和, 魏 俊, 胡小玲, 岳 红, 吕 玲, 王育彬, 文振翼. 物理化学学报, **2000, 16**(7): 643]

8 Zhang, X.; Du, H.; Wang, Z.; Wu, Y. D.; Ding, K. *J. Org. Chem.*, **2006, 71**(7): 2862

9 Berski, S.; Andrés, J.; Silvi, B.; Domingo, L. R. *J. Phys. Chem. A*, **2006, 110**(51): 13939

10 Zou, J.; Han, Y.; So, S. S. *Methods Mol. Biol.*, **2008, 458**: 15

11 Cartwright, H. M. *Methods Mol. Biol.*, **2008, 458**: 1

12 Hampson, S. *Prog. Neurobiol.*, **1991, 37**(5): 383

13 Koene, R. A.; Takane, Y. *Neural Comput.*, **1999, 11**(3): 783

14 Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; *et al*. Gaussian 98, Revision A.01. Pittsburgh, PA: Gaussian Inc., 2001

15 Fukumi, M.; Omatu, S. *IEEE Trans Neural Netw.*, **1991, 2**(5): 535

16 Phansalkar, V. V.; Sastry, P. S. *IEEE Trans. Neural. Netw.*, **1994, 5**(3): 505

17 Tafeit, E.; Reibnegger, G. *Clin. Chem. Lab. Med.*, **1999, 37**(9): 845

18 Qian, M. P.; Gong, G. L. Stochastic processes. Beijing: Peking University Press, 1992, Chapter 1: 3–7 [钱敏平, 龚光鲁. 随机过程论. 北京: 北京大学出版社, 1992, 第一章: 3–7]