

Classification of the Carcinogenicity of N-Nitroso Compounds Based on Support Vector Machines and Linear Discriminant Analysis

Feng Luan,[†] Ruisheng Zhang,^{*,†,§} Chunyan Zhao,[†] Xiaojun Yao,[‡] Mancang Liu,[†] Zhide Hu,[†] and Botao Fan[‡]

Departments of Chemistry and Computer Science, Lanzhou University, Lanzhou, Gansu 730000, China, and Université Paris 7-Denis Diderot, ITODYS 1, Rue Guy de la Brosse, 75005 Paris, France

Received August 9, 2004

The support vector machine (SVM), as a novel type of learning machine, was used to develop a classification model of carcinogenic properties of 148 N-nitroso compounds. The seven descriptors calculated solely from the molecular structures of compounds selected by forward stepwise linear discriminant analysis (LDA) were used as inputs of the SVM model. The obtained results confirmed the discriminative capacity of the calculated descriptors. The result of SVM (total accuracy of 95.2%) is better than that of LDA (total accuracy of 89.8%).

Introduction

N-Nitroso compounds (NOCs)¹ have been known for more than 100 years since dimethylnitrosamine was prepared. The carcinogenic property of this kind of compounds was not described until 1956 by Magee and Barnes (1). Since then, extensive studies have been carried out in various countries on the occurrence and influence of NOCs. There seems to be two main reasons for this worldwide interest on this topic permanently. First, the NOCs, including nitrosamines and nitrosamides, are a class of potent and widespread environmental carcinogens, which are potentially important in the etiology of human cancer. They induce tumors in various vital organs causing pancreatic cancer (2), gastrointestinal cancer (3), and renal and childhood brain tumors, etc. (4, 5). They also influence a wide range of animals. For this reason, the presence of NOCs is a matter of concern. Second, they have been easily found in many substances such as the betel nut, in bacteria, and in smoke and foodstuffs. These substances are commonly in contact with us. Additionally, NOCs can be formed easily from the reaction of amines and nitrites. The amine precursors are normal constituents of food, drugs, pesticides, and food additives. Nitrite is abundantly present in the environment, in cured meat, and in human saliva and could be the reduction of nitrate. Therefore, it is highly likely that man is susceptible to its carcinogenic effect everyday and anywhere.

Relationships between the molecular structure of NOCs and their metabolism and/or their carcinogenic potential have been studied extensively. Many biochemical and physicochemical investigations have been directed toward establishing their structure–activity re-

lationships (SARs). Because the molecular geometry and conformational behavior critically influence the biological activity, quantitative SARs (QSARs) of N-nitrosamines have been extensively studied by different computational methods. Wishnok et al. reported, with some degree of confidence, an estimate of carcinogenic activity for 51 nitrosamines through SARs by correlating the number of carbon atoms with the carcinogenic activity (6). Later, they reported a quantitative Hansch–Taft SAR for nitrosamine carcinogenicity, which demonstrated that variation in carcinogenicity could be correlated with a number of molecular properties (7). Then, the same authors predicted organ specificity using physicochemical properties of N-nitrosodialkylamines. Partition coefficients, electronic factors, and a measure of steric hindrance gave a near perfect prediction of 19 compounds (8). Singer et al. linked liposolubility with nitrosamine carcinogenicity through QSAR (9). Chou et al. expanded the approach to SARs by applying computer-assisted mathematical and statistical methods to a large set of 144 NOCs (10). Dunn et al. used a pattern recognition technique called SIMCA to perform the classification of 61 NOCs (11, 12). They reported an 88% correct classification of the carcinogens. Peter et al. reported a pattern recognition method of 150 nitrosamines, and they reported a 97% correct classification using 22 descriptors (13). Dai et al. also reported a pattern recognition method of 153 nitrosamines, and they reported a 97% correct classification using 10 descriptors (14).

Because of computational bottlenecks in descriptor generation and statistical algorithms, most of the previous approaches are not satisfactory. Some models have been developed for a relatively small data set of compounds. Some models, however, include a large number of descriptors, leading to the difficulty of the explanation of the physical meaning of the descriptors. Usually, they only used the linear statistical method, which solved the highly nonlinear QSAR problems with difficulty.

Several nonlinear QSAR techniques have been proposed in recent years. One of them is support vector

* To whom correspondence should be addressed. Tel: +86-931-8912578. Fax: +86-931-8912582. E-mail: liumc@lzu.edu.cn.

[†] Department of Chemistry, Lanzhou University.

[§] Department of Computer Science, Lanzhou University.

[‡] Université Paris 7-Denis Diderot.

¹ Abbreviations: SVM, support vector machine; NOCs, N-nitroso compounds; LDA, linear discriminant analysis.

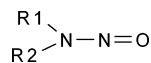


Figure 1. General structure of NOCs.

machine (SVM), which is a new method increasingly being used in pattern recognition studies. In the present study, for the first time, a novel modeling approach based on SVMs to classify NOCs is presented. A large number of descriptors were calculated by CODESSA software. The LDA method was also utilized to establish a linear classification model to compare the results with those obtained by SVM. The aim of this study is to establish an accurate classification model for the prediction of the carcinogenic property of NOCs and to seek the important structural features related to the carcinogenic property of NOCs.

Data Set and Molecular Descriptor Generation

Data Set. The data set of this investigation consisted of 148 NOCs, which were taken from the paper published by Dai (14). Of these compounds, 116 are carcinogenic compounds, and 32 are not carcinogenic ones. It is asymmetric with fewer inactive than active compounds. The general structure of the NOCs is shown in Figure 1. A complete list of the compounds' structures and their corresponding classification is in Table 1. In Table 1, "+" represents carcinogenic compounds and "-" represents noncarcinogenic compounds. The entire set of compounds was divided into two subsets: a training set, whose information was used to build the models, and a test set, consisting of molecules not found in the training set, which was used to validate the models once they were built. Members of each set were assigned randomly. The training set consisted of 118 compounds (79.7%), and the test set contained 30 compounds (20.3%). As an added precaution, it was verified that each set contained roughly the same percentage of noncarcinogenic compounds (training set = 22.0%, test set = 18.75%).

Molecular Descriptor Generation. The structures of the compounds were drawn with the ISIS DRAW 2.3 program (15). The final geometries were obtained with the semiempirical PM3 method in the HYPERCHEM 4.0 program (16). All calculations were carried out at a restricted Hartree-Fock level with no configuration interaction. The molecular structures were optimized using the Polak-Ribiere algorithm until the root-mean-square gradient was 0.001. Then, the resulting geometry was transferred into CODESSA software, developed by the Katritzky group (17, 18), which can calculate constitutional, topological, geometrical, electrostatic, and quantum chemical descriptors and has been successfully used in various QSPR and QSAR researches. Constitutional descriptors are related to the number of atoms and bonds in each molecule. Topological descriptors include valence and nonvalence molecular connectivity indices calculated from the hydrogen-suppressed formula of the molecule, encoding information about the size, composition, and the degree of branching of a molecule. The topological descriptors describe the atomic connectivity in the molecule. The geometrical descriptors describe the size of the molecule and require three-dimensional coordinates of the atoms in the given molecule. The electrostatic descriptors reflect characteristics of the charge distribution of the molecule. The quantum chemical descriptors offer

information about binding and formation energies, partial atom charge, dipole moment, and molecular orbital energy levels.

Methodology

LDA Model Development. The basic theory of linear discriminant analysis (LDA) is to classify the dependent by dividing an n -dimensional descriptor space into two regions that are separated by a hyperplane defined by a linear discriminant function (19, 20) as follows:

$$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_nX_n$$






where Y is a discriminant score, that is, the dependent variable; $X_1 - X_n$ represents the specific descriptor; and b corresponds to weights associated with the respective descriptor. The two regions formed by the hyperplane correspond to the two classes to which individual compounds are predicted to belong.

LDA was performed using the SPSS statistical software. For the purposes of modeling, a value of 1 was assigned to compounds with carcinogenic activity, and a value of 2 was assigned to those with no carcinogenic. The linear classifications were performed in a stepwise manner: At each step, the variable that adds the most to the separation of the groups is entered into (or the variable that adds the least is removed from) the discriminant function. The selection of the descriptors was based on the F parameter. In this study, the minimum partial F value to enter is set to 5.50 and the maximum partial F to remove is 2.71. The prior probabilities were computed from group size (0.784 and 0.216 for the carcinogenic activity and noncarcinogenic compounds, respectively).

Theory of SVM. Because there are a number of introductions into SVM (21-24), here, we only briefly summarized the main ideas of SVM for classification. The SVM method was proposed by the Vapnik group (25). The main advantage of SVM is that it adopts the structure risk minimization (SRM) principle, which has been shown to be superior to the traditional empirical risk minimization (ERM) principle (26), employed by conventional neural networks. SRM minimizes an upper bound of the generalization error on the Vapnik-Chernoverkis dimension, as opposed to ERM, which minimizes the training error. This method has proven to be very effective for addressing general purpose classification and regression problems (27-33). In most of these cases, the performance of SVM modeling either matches or is significantly better than that of traditional machine learning approaches, including artificial neural networks. The SVM method has a number of interesting properties, including an effective avoidance of overfitting, which improves its ability to build models using large numbers of molecular property descriptors with relatively few experimental results in the training set.

The objective of SVM for classification is to construct an "optimal hyperplane" as the decision surface such that the margin of separation between two different chemical substances is maximized. In the simplest form, the SVM is a linear classifier. However, if we cannot find a linear separator, data points are projected into a (usually) higher dimensional space where the data points effectively become linearly separable. In nonlinearly separable cases, SVM maps the input variable into a high

Table 1. Compounds' Structure and Its Corresponding Classification

| No. | NR ₁ R ₂ (R) ^a | | Experiment | Calculate | | No. | NR ₁ R ₂ (R) ^a | | Experiment | Calculate | |
|-----|---|---|------------|-----------|-----|------|--|--|------------|-----------|-----|
| | | | | LDA | SVM | | | | | LDA | SVM |
| 1 | C ₂ H ₅ | C ₂ H ₅ | + | + | + | 77* | CH ₃ | C ₆ H ₅ | + | -** | + |
| 2* | CH ₃ | C ₆ H ₅ CH ₂ | + | + | + | 78 | CH ₃ CH ₂ | (CH ₃) ₂ CH | + | + | + |
| 3 | CH ₃ | C ₆ H ₅ CH ₂ CH ₂ | + | + | + | 79 | CH ₃ (CH ₂) ₃ | CH ₃ (CH ₂) ₄ | + | + | + |
| 4 | CH ₃ | ClCH ₂ CH ₂ | + | + | + | 80 | | (CH ₂) ₄ | + | + | + |
| 5 | CH ₃ | CH ₂ =CH | + | + | + | 81 | CH ₃ | H | + | + | + |
| 6 | C ₂ H ₅ | NH ₂ CO | + | + | + | 82* | CH ₃ CH ₂ | HOCH ₂ CH ₂ | + | + | + |
| 7* | C ₂ H ₅ | CH ₂ =CH | + | + | + | 83 | CH ₃ (CH ₂) ₃ | HO(CH ₂) ₄ | + | + | + |
| 8 | CH ₃ | CH ₃ (CH ₂) ₄ | + | + | + | 84 | CH ₃ | CH ₃ OCOCH ₂ | + | + | + |
| 9 | CH ₃ | (CH ₂) ₅ CH | + | + | + | 85 | | (CH ₂) ₃ | + | + | + |
| 10 | | CH ₂ CH(CH ₃)OCH(CH ₃)CH ₂ | + | + | + | 86 | | CH ₂ CH=CHCH ₂ | + | + | + |
| 11 | | CH ₂ CHClCHClCH ₂ CH ₂ | + | + | + | 87* | | CH ₂ CH(CH ₃)CH(CH ₃)CH ₂ | + | + | + |
| 12* | | CH ₂ CHBrCHBrCH ₂ CH ₂ | + | + | + | 88 | | CH(CH ₃)-(CH ₂) ₄ | + | + | + |
| 13 | | CH ₂ CH=CHCH ₂ CH ₂ | + | + | + | 89 | | (CH ₂) ₂ S(CH ₂) ₂ | + | + | + |
| 14 | | (CH ₂) ₆ | + | + | + | 90 | | CH ₂ CH(CH ₃)OCH ₂ | + | + | + |
| 15 | | (CH ₂) ₇ | + | + | + | 91 | CH ₃ | CH ₃ (CH ₂) ₇ | + | + | + |
| 16 | | CH ₂ CHClCHClCH ₂ | + | + | + | 92* | CH ₃ | CH ₃ (CH ₂) ₈ | + | + | + |
| 17* | CH ₃ | (CH ₃) ₃ CCH ₂ | + | + | + | 93 | | CH ₂ CH(CH ₃)-(CH ₂) ₃ | + | + | + |
| 18 | | CH(CH ₃)CH ₂ N(NO)CH ₂ CH(CH ₃) | + | + | + | 94 | HO(CH ₂) ₄ | CH ₃ (CH ₂) ₂ | + | + | + |
| 19 | | CH ₂ CH ₂ N(NO)CH ₂ CH(CH ₃) | + | + | + | 95 | CH ₃ (CH ₂) ₃ | HO(CH ₂) ₂ | + | + | + |
| 20 | CH ₃ (CH ₂) ₃ | CH ₃ | + | + | + | 96 | CH ₃ CHCH ₂ | CH ₃ OCOCH | + | + | + |
| 21 | | CH(CH ₃)CH ₂ N(NO)CH(CH ₃)CH ₂ | + | + | + | 97* | CH ₃ (CH ₂) ₃ |  | + | + | + |
| 22* | | CH ₂ CH ₂ OCH(CH ₃)CH ₂ | + | + | + | 98 | CH ₃ (CH ₂) ₃ | CH ₃ (CH ₂) ₂ CH(OOH) | + | -** | + |
| 23 | CH ₃ | CH ₃ CH ₂ CH ₂ | + | + | + | 99 | (CH ₃) ₂ CHCH ₂ | (CH ₃) ₂ CHCH ₂ | + | + | + |
| 24 | CH ₃ | CH ₃ N(NO)(CH ₂) ₃ | + | + | + | 100 | CH ₃ (CH ₂) ₃ | HO(CH ₂) ₃ | + | + | + |
| 25 | | (CH ₂) ₂ N(NO)(CH ₂) ₃ | + | + | + | 101 | CH ₃ | C ₆ H ₅ CH(CH ₃) | + | + | + |
| 26 | | CH=CH(CH ₂) ₃ | + | + | + | 102* | | (CH ₂) ₁₂ | + | -** | + |
| 27* | | CH ₂ CHCl(CH ₂) ₃ | + | + | + | 103 | CH ₃ | <i>p</i> -CH ₃ C ₆ H ₄ CH ₂ | + | + | + |
| 28 | ClCH ₂ CH ₂ | ClCH ₂ CH ₂ | + | + | + | 104 | (CH ₃) ₂ CH | (CH ₃) ₂ CH | + | -** | + |
| 29 | (CH ₂) ₂ COCH(CH ₃)CH ₂ | | + | + | + | 105 | CH ₃ (CH ₂) ₄ | CH ₃ (CH ₂) ₄ | + | + | + |
| 30 | C ₂ H ₅ | C ₂ H ₅ N(NO)(CH ₂) ₂ | + | + | + | 106 | CH ₃ (CH ₂) ₂ | HO(CH ₂) ₂ | + | + | + |
| 31 | | CH ₂ N(NO)(CH ₂) ₃ | + | + | + | 107* | CH ₃ COOCH ₂ CH ₂ | CH ₃ COOCH ₂ CH ₂ | + | + | + |
| 32* | | (CH ₂) ₂ CHCl(CH ₂) ₃ | + | + | + | 108 | CH ₃ | HOCOCH ₂ | + | + | + |
| 33 | CH ₃ | CH ₃ COOCH ₂ | + | + | + | 109 | | (CH ₂) ₂ CH(C ₆ H ₅)-(CH ₂) ₂ | + | + | + |
| 34 | C ₂ H ₅ | CH ₃ COOCH ₂ | + | + | + | 110 | | (CH ₂) ₂ CH(C(CH ₃) ₃)-(CH ₂) ₂ | + | + | + |
| 35 | CH ₃ CH ₂ CH ₂ | CH ₃ COOCH ₂ | + | + | + | 111 | CH ₃ | CH ₃ (CH ₂) ₉ | + | + | + |
| 36 | CH ₃ (CH ₂) ₃ | CH ₃ (CH ₂) ₂ CO | + | + | + | 112* | CH ₃ | CH ₃ (CH ₂) ₁₀ | + | + | + |
| 37* | ClCH ₂ CH ₂ | NH ₂ CO | + | + | + | 113 | CH ₃ | CH ₃ (CH ₂) ₁₁ | + | + | + |
| 38 | | CH ₂ CH(OH)(CH ₂) ₃ | + | + | + | 114 | CH ₃ | CH ₃ (CH ₂) ₁₂ | + | + | + |
| 39 | | (CH ₂) ₂ CH(OH)(CH ₂) ₂ | + | + | + | 115 | CH ₃ | CH ₃ (CH ₂) ₁₃ | + | + | + |
| 40 | | (CH ₂) ₂ CO(CH ₂) ₂ | + | + | + | 116 | | CH ₂ CH(CH ₃)N(COC ₆ H ₅)CH(CH ₃)CH ₂ | + | + | + |
| 41 | | (CH ₂) ₃ | + | + | + | 117* | CH ₂ =CHCH ₂ | CH ₂ =CHCH ₂ | - | +** | +** |
| 42* | CH ₃ | CH ₃ | + | + | -** | 118 | CH ₃ CH ₂ | (CH ₃) ₃ C | - | - | +** |
| 43 | CH ₃ CH ₂ CH ₂ | CH ₃ CH ₂ CH ₂ | + | + | + | 119 | CYCLOHEXANE | CYCLOHEXANE | - | - | - |
| 44 | CH ₃ | C ₂ H ₅ | + | + | + | 120 | C ₆ H ₅ | C ₆ H ₅ | - | - | - |
| 45 | CH ₃ | CH ₂ =CHCH ₂ | + | + | + | 121 | C ₆ H ₅ CH ₂ | C ₆ H ₅ CH ₂ | - | - | - |
| 46 | CH ₃ | CH ₃ N(NO)(CH ₂) ₂ | + | + | + | 122* | CH ₃ | NCCH ₂ | - | - | - |
| 47* | C ₂ H ₅ | CH ₃ (CH ₂) ₂ | + | + | + | 123 | | (CH ₂) ₃ CH(COOC ₂ H ₅) | - | +** | - |
| 48 | | (CH ₂) ₅ | + | + | + | 124 | | CH ₂ N(NO)CH ₂ N(NO)CH ₂ | - | - | - |
| 49 | | (CH ₂) ₂ N(NO)(CH ₂) ₂ | + | + | + | 125 | CH ₃ | CH ₃ O | - | +** | - |
| 50 | | (CH ₂) ₂ O(CH ₂) ₂ | + | + | + | 126 | CH ₃ | <i>p</i> -C ₆ H ₄ CHO | - | - | - |
| 51 | CH ₃ |  | + | + | + | 127* | | CH(CH ₃)(CH ₂) ₃ CH(CH ₃) | - | +** | +** |
| 52* | CH ₃ | CH ₃ N(NO)CO-CO | + | + | + | 128 | | C(CH ₃) ₂ (CH ₂) ₃ C(CH ₃) ₂ | - | - | - |
| 53 | CH ₃ | CH ₃ CO | + | + | + | 129 | | (CH ₂) ₄ CH(COOH) | - | - | - |
| 54 | CH ₃ | CH ₃ OCO | + | + | + | 130 | | (CH ₂) ₂ CH(COOH)(CH ₂) ₂ | - | - | - |
| 55 | C ₂ H ₅ | CH ₃ OCO | + | + | + | 131 | | (CH ₂) ₂ CH=C(COOH)CH ₂ | - | - | - |
| 56 | CH ₃ | NH ₂ CO | + | + | + | 132* | | C(CH ₃) ₂ (CH ₂) ₂ C(CH ₃) ₂ | - | - | - |
| 57* | CH ₃ | CH ₃ NHCO | + | + | + | 133 | | (CH ₂) ₃ CH(COOH) | - | - | - |
| 58 | CH ₃ (CH ₂) ₃ | NH ₂ CO | + | + | + | 134 | | CH(COOH)CH ₂ CH(OH)CH ₂ | - | +** | +** |
| 59 | CH ₃ | O ₂ NNHC(=NH) | + | -** | + | 135 | | CH(CH ₃)CH(CH ₃)N(NO)CH(CH ₃)CH(CH ₃) | - | - | - |
| 60 | | (CH ₂) ₂ NHCO | + | + | + | 136 | | (CH ₂) ₂ NH(CH ₂) ₂ | - | +** | - |
| 61 | | (CH ₂) ₃ OCH ₂ | + | + | + | 137* | | (CH ₃ CH ₂ O)CHCH ₂ | - | - | - |
| 62* | | CH ₂ CH(OH)(CH ₂) ₂ | + | + | + | 138 | | CH ₃ CH ₂ CH(CH ₃) | - | - | - |
| 63 | | CH ₂ CH=CHCH(CH ₃)CH ₂ | + | + | + | 139 | CH ₃ (CH ₂) ₇ | CH ₃ (CH ₂) ₇ | - | +** | - |
| 64 | | (CH ₂) ₂ OCH ₂ | + | + | + | 140 | CH ₃ CO | CH ₃ CO | - | - | - |
| 65 | CH ₃ | (CH ₃) ₂ NCO | + | + | + | 141 | CH ₃ | CH ₃ COCH ₂ C(CH ₃) ₂ | - | +** | - |
| 66 | CH ₃ | CH ₃ (CH ₂) ₅ | + | + | + | 142* | CH ₃ | (CH ₃) ₃ C | - | - | +** |
| 67* | | CH ₂ CH(CH ₃)NHCH(CH ₃)CH ₂ | + | + | + | 143 | CH ₃ | β  | - | +** | - |
| 68 | | CH ₂ CH(CH ₃)N(CH ₃)CH(CH ₃)CH ₂ | + | + | + | 144 | CH ₃ | γ  | - | - | - |
| 69 | CH ₃ (CH ₂) ₃ | CH ₃ COOCH ₂ | + | + | + | 145 | CH ₃ | <i>p</i> -ON-C ₆ H ₄ | - | - | - |
| 70 | CH ₃ (CH ₂) ₃ | CH ₃ OCO(CH ₂) ₂ CO | + | + | + | 146 | |  | - | +** | - |
| 71 | CH ₃ OCO(CH ₂) ₃ | CH ₃ (CH ₂) ₂ CH(OOCCH ₃) | + | + | + | 147* | CH ₃ | <i>p</i> -CH ₃ -C ₆ H ₄ SO ₂ | - | - | - |
| 72* | CH ₃ (CH ₂) ₃ | CH ₃ (CH ₂) ₂ CH(OOCCH ₃) | + | + | + | 148 | CF ₃ CH ₂ | CF ₃ CH ₂ | - | - | - |
| 73 | | CH ₂ CH(CH ₃)NH(COCH ₃)CH(CH ₃)CH ₂ | + | + | + | | | | | | |
| 74 | | (CH ₃)CH(CH ₃)-(CH ₂) ₂ | + | + | -** | | | | | | |
| 75 | CH ₃ (CH ₂) ₃ | CH ₃ (CH ₂) ₃ | + | + | + | | | | | | |
| 76 | CH ₃ | CH ₃ (CH ₂) ₆ | + | + | + | | | | | | |

*Test set. **Misclassified one. ^a The compound is a cyclic compound.

dimensional feature space ($\Phi : R^N \mapsto F$) using a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$. As both the objective function and the

decision function are expressed in terms of dot products of data vectors x , the potentially computation intensive

Table 2. Seven Descriptors, the *F* Value, and Unstandardized Coefficient for the LDA

| | chemical meaning | <i>F</i> to remove | unstandardized coefficient |
|----------|---|--------------------|----------------------------|
| constant | constant | | -8.471 |
| RI(3) | Randic index (order 3) | 12.386 | 1.313 |
| SIC(2) | structural information content (order 2) | 62.521 | -0.347 |
| BI | Balaban index | 36.433 | 1.752 |
| RPCS | relative positively charged surface area | 25.188 | 0.423 |
| WNSA-1 | surface-weighted charged partial surface area | 12.076 | -0.040 |
| RNC | relative no. of C atoms | 8.401 | 8.767 |
| RI(2) | Randic index (order 2) | 7.397 | 0.815 |

Table 3. Correlation Matrix of the Seven Descriptors

| | RI(3) | SIC(2) | BI | RPCS | WNSA-1 | RNC | RI(2) |
|--------|--------|--------|--------|--------|--------|-------|-------|
| RI(3) | 1.000 | | | | | | |
| SIC(2) | 0.642 | 1.000 | | | | | |
| BI | -0.241 | 0.240 | 1.000 | | | | |
| RPCS | 0.605 | 0.446 | 0.132 | 1.000 | | | |
| WNSA-1 | -0.384 | -0.279 | -0.214 | -0.229 | 1.000 | | |
| RNC | 0.842 | 0.728 | 0.017 | 0.642 | -0.470 | 1.000 | |
| RI(2) | 0.527 | 0.397 | -0.267 | 0.390 | 0.212 | 0.455 | 1.000 |

Table 4. Results of Two Models

| | LDA | | SVM | |
|--------------------|------|------|------|------|
| | + | - | + | - |
| + | 111 | 5 | 114 | 2 |
| - | 10 | 22 | 5 | 27 |
| % + | 95.7 | 4.3 | 98.2 | 1.8 |
| - | 31.3 | 68.6 | 15.6 | 84.4 |
| total accuracy (%) | 89.8 | | 95.2 | |

mapping $\Phi(\cdot)$ does not need to be explicitly evaluated. SVM classifiers are generated by a two-step procedure: First, the sample data vectors are mapped to a very high-dimensional space. The dimension of this space is significantly larger than that of the original data space. Then, the SVM algorithm finds a hyperplane in this space with the largest margin separating classes of data.

All calculation programs implementing SVM were written in an R-file based on the R script for SVM. The scripts were compiled using an R 1.7.1 compiler running operating system on a Pentium IV PC with 256M RAM.

Results and Discussion

Results of LDA. After LDA, it can be seen that the best linear model contains seven molecular descriptors. The selected variables, their chemical meanings, their unstandardized coefficients, and their *F* values are shown in Table 2. The *F* value was the parameter for choosing descriptors. As we know, a larger *F* value means there are significant differences of each other. The larger an *F* value is, the descriptor has priority to enter in the model, indicating that the variable is better at discriminating between groups. The correlation matrix of the seven selected descriptors is shown in Table 3. From Table 3, it can be seen that the linear correlation coefficient value of each of the two descriptors is <0.85 , which means that the descriptors are independent in this LDA analysis. The LOO results of LDA model are listed in Table 1. It gave a total accuracy of 89.8%; see Table 4.

By interpreting the descriptors in the LDA model, it is possible to gain some insight into factors that are likely to relate to the carcinogenic property of the NOCs. NOCs are known to react with DNA, which leads them to form a DNA adduct (34–37). If the adducts persist, miscoding can occur during DNA replication, leading to permanent mutations and derangement of normal cellular growth,

ultimately, tumorigenesis. NOCs must be activated to exert their carcinogenic effects. In the carcinogenic process, hydroxylation of the carbons α - to the *N*-nitroso group is a key step. Following α -hydroxylation, the unstable α -hydroxy one decomposes to electrophilic intermediates that can react with nucleophilic DNA bases to yield adducts. Because of the diversity of the molecules studied in this work, the carcinogenic property of the compounds is related to the molecular structure in a complex way. Of the seven descriptors, one is constitutional, four are topological, and two are electrostatic descriptors. These descriptors encode different aspects of the molecular structure.

The relative number of C atoms (RNC) is a constitutional descriptor, which is calculated as the number of C atoms divided by the number of atoms. The RNC partially accounts for the steric hindrance effect. The size and shape of compounds influence their transport properties through a biological system as well as their steric hindrance at the reactive site. The larger the descriptor value is, the larger the steric hindrance is. Thus, an increase of the descriptor value leads to a decrease of the binding ability to DNA, indicating the noncarcinogenicity of the compounds.

The Balaban index (BI) (38), a topological descriptor, describes the atomic connectivity and branching information in the molecule and has some correlation with the hydrophobic interaction of the molecules. The other two topological descriptors are the Randic index (order 2) (RI2) and order 3 (RI3) (39), which encode the size, shape, and degree of branching in the compound and also relate to the dispersion interaction among molecules. Because of their positive coefficients in the linear model, increasing this descriptor also increases the discriminant score values, indicating the disfavor of the binding to the reactive site of DNA. Additionally, the large degree of branching and dispersion for molecules also gave a negative influence on the transport properties through a biological system. The fourth topological descriptor, structural information content (order 2), developed by Basak and co-workers based on the Shannon information theory (40, 41), takes into account all atoms in the constitutional formula (hydrogens also being included), and it considers the information content provided by various classes of atoms based on their topological neighborhood. It is not intercorrelated with other topological indexes. The negative coefficient in the model implies that increasing the value of this descriptor can lead to the carcinogenicity of the compounds.

Two electrostatic descriptors, relative positively charged surface area (RPCS) and surface-weighted charged partial surface area (WNSA-1), are both of the charged partial surface area (CPSA) type (42), which are based on the surface area of the whole molecule and on the charge distribution in the molecule, so they combine

shape and electronic information to characterize the molecule; therefore, they encode features responsible for polar interactions between molecules. RPCS (42, 43) is the product of the solvent accessible surface area of the most positive atom by the relative positive charge (RPCG). The chemical charges in the molecule are calculated using the approach proposed by Zefirov (44), based on the Sanderson's electronegativity scale. WNSA-1 (45) indicates the effect of negative charge distribution in the molecule, and it also encodes information about polar interactions. The charge distribution of the molecule is most likely an influence in the key step of α -hydroxylation. The larger the solvent accessible surface area produced by the relative positive charge is, the less chance for hydroxylation on the α carbon is. It implies that the molecule trends to no carcinogenic property. On the contrary, it gave a beneficial environment for the α -hydroxylation process. As can be seen from the coefficients in the linear model, the above two descriptors have opposite effects really.

From the above discussion, it can be seen that the steric and electric descriptors are likely two major factors in the process of carcinogenicity, and all of the descriptors involved in the model, which have explicit physical meanings, may account for the structural features responsible for carcinogenic properties of NOCs.

Results of SVM. After the establishment of the linear model, SVM was used to develop a nonlinear model based on the same subset of descriptors. Similar to other multivariate statistical models, the performances of SVM for classification depend on the combination of several parameters. The kernel functions should be decided first. There are a number of kernel functions, which have been found to provide good generalization capabilities. One has several possibilities for the choice of this function, including linear, polynomial, splines, and basis function. However, for classification tasks, a commonly used kernel function is the Gaussian radial basis function because of its good general performance and a few number of parameters (Bishop, 1997); the RBF is formulated as below:

$$\exp[-\gamma * (x - x_i)^2]$$

This function was used in the present work.

The other two parameters are capacity parameter C and γ . C is a regularization parameter that controls the tradeoff between maximizing the margin and minimizing the training error. If C is too small, then insufficient stress will be placed on fitting the training data. If C is too large, then the algorithm will overfit the training data. To make the learning process stable, a large value should be set up for C . According to our experience (46), in this study, C was set to 100 first. γ , the parameter of the kernel, controls the amplitude of the Gaussian function and, further, controls the generalization ability of SVM. Therefore, the models were obtained researching the effects of far going γ and C on the accuracy of LOO cross-validation of all training compounds and the maximum accuracy was chosen as the optimal condition. The accuracy of LOO cross-validation was plotted vs different γ (Figure 2) and C (Figure 3). The optimal γ was found as 0.037, and the final optimal value of C is 100.

Then, the test set data were tested with the built model, and the result of the test set was listed in Table 1. The misclassified samples (marked by double asterisk)

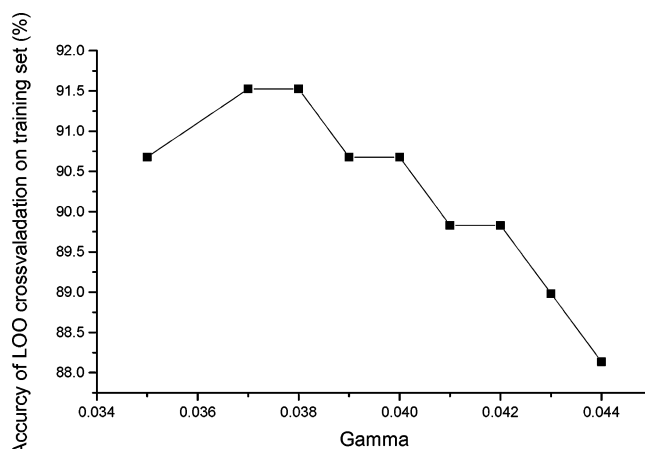


Figure 2. Accuracy of LOO cross-validation of training set vs γ .

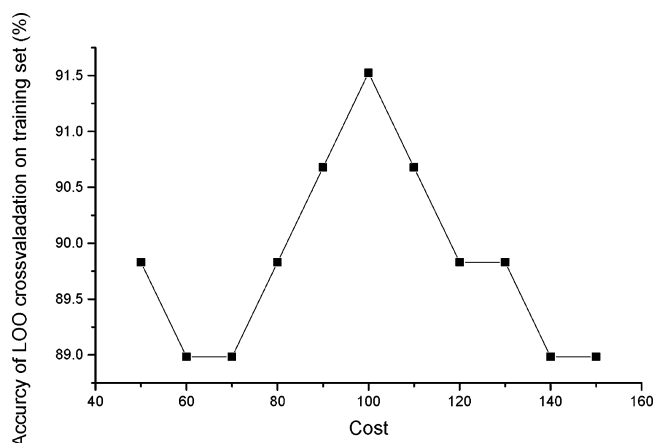


Figure 3. Accuracy of LOO cross-validation of training set vs C .

by LDA and SVM were listed also. The same misclassified ones of LDA and SVM were 117, 127, and 134. The accuracy of each class is shown in Table 4. The accuracy was 95.7% on the active group for LDA and 98.2% for SVM, and on the inactive group, the overall accuracy was 68.6% for LDA and 84.4% for SVM. The accuracy of the training set for SVM was 97.4%, and the test set was 86.6%. The total accuracy for SVM was 95.2%, which was higher than that of LDA (89.8%). From comparison of the two methods, it can be seen that performance of SVM was better than that of LDA, which implies that using the same descriptors, the SVM method is capable of recognizing highly nonlinear SARs; in contrast, LDA approaches can only capture linear relationships between molecular characteristics. It also can be seen from Table 4 that the accuracy of the inactive group is lower than the active group.

Conclusion

In this work, we applied LDA and support vectors machine for the prediction of the carcinogenic property of a set of 148 NOCs using descriptors calculated from the molecular structure alone. Satisfactory results were obtained with the proposed methods. The proposed LDA model could provide some insight into what structural features are related to the carcinogenic properties of NOCs. Additionally, using Gaussian kernel SVM produced even better classification models with a better predictive ability than LDA. The training procedure is also simple when using SVM because fewer parameters

are optimized, and only support vectors are used in the generalization process. Besides, the SVM exhibits the better whole performance due to embodying the SRM principle and some advantages over the other techniques. Furthermore, the proposed approach can also be extended in other QSPR/QSAR or classification investigations.

Acknowledgment. We thank the Association Franco-Chinoise pour la Recherche Scientifique & Technique (AFCRST) for supporting this study (Program PRA SI 02-03). We also thank the R Development Core Team for affording the free R1.7.1 software.

References

- Magee, P. N., and Barnes, J. M. (1956) The production of malignant primary hepatic tumors in the rat by feeding dimethylnitrosamine. *Br. J. Cancer* 10, 114–122.
- Capurso, G., Delle Fave, G., and Lemoine, N. (2004) Etiology of pancreatic cancer, with a hypothesis concerning the role of N-nitroso compounds and excess gastric acidity. *J. Natl. Cancer Inst.* 96 (1), 75.
- Chiara, C., and Parry, J. M. (2004) Comparative genomic hybridization analysis of N-methyl-N'-nitrosoguanidine-induced rat gastrointestinal tumors discloses a cytogenetic fingerprint. *Environ. Mol. Mutagen.* 43 (1), 20–27.
- Shiao, Y.-H., Ramakrishna, G., Anderson, L. M., Perantoni, A. O., Rice, J. M., and Diwan, B. A. (2002) Down-regulation of von Hippel-Lindau protein in N-nitroso compound-induced rat non-clear cell renal tumors. *Cancer Lett.* 179, 33–38.
- Huncharek, M., and Kupelnick, B. (2004) A meta-analysis of maternal cured meat consumption during pregnancy and the risk of childhood brain tumors. *Neuroepidemiology* 23, 78–84.
- Wishnok, J. S., and Archer, M. C. (1976) *Br. J. Cancer* 33, 307–311.
- Wishnok, J. S., Archer, M. C., Edelman, A. S., and Rand, W. M. (1978) Nitrosamine carcinogenicity: A quantitative Hansch-Taft structure–activity relationship. *Chem.-Biol. Interact.* 20, 43–54.
- Edelman, A. S., Kraft, P. L., Rand, W. M., and Wishnok, J. S. (1980) Nitrosamine carcinogenicity: A quantitative relationship between molecular structure and organ selectivity for a series of acyclic N-nitroso compounds. *Chem.-Biol. Interact.* 31, 81–92.
- Singer, G. M., Taylor, H. W., and Lijinsky, W. (1977) Liposolubility as an aspect of nitrosamine carcinogenicity: Quantitative correlations and qualitative observations. *Chem.-Biol. Interact.* 19, 133–142.
- Chou, J. T., and Jurs, P. C. (1979) Computer assisted structure–activity studies of chemical carcinogens. An N-nitroso compound data set. *J. Med. Chem.* 22, 792–797.
- Dunn, W. J., III, and Wold, S. J. (1981) An assessment of the carcinogenicity of N-nitroso compounds by the SIMCA method of pattern recognition. *J. Chem. Inf. Comput. Sci.* 21, 8–13.
- Dunn, W. J., III, and Wold, S. (1981) The carcinogenicity of N-nitroso compounds: A SIMCA pattern recognition study. *Bioorg. Chem.* 10, 29–45.
- Rose, S. L., and Jurs, P. C. (1982) Computer-assisted studies of structure–activity relationships of N-nitroso compounds using pattern recognition. *J. Med. Chem.* 25 (7), 769.
- Dai, Q. Y., Zhong, R. M., and Gao, X. M. (1987) structure–activity relationships of N-nitroso compounds using pattern recognition based on di-region theory. *Environ. Chem.* 6 (6), 1–11.
- ISIS Draw2.3 (1990–2000) MDL Information Systems, Inc.
- HyperChem, Release 4.0 for Windows (1995); Hypercube, Inc.
- Katritzky, A. R., Lobanov, V. S., and Karelson, M. (1995) CODESSA: Training Manual, University of Florida, Gainesville, FL.
- Katritzky, A. R., Lobanov, V. S., and Karelson, M. (1994) CODESSA: Reference Manual, University of Florida, Gainesville, FL.
- Kachigan, S. K. (1986) *Statistical Analysis*, Radius Press, New York.
- Fisher, R. A. (1936) The use of multiple measurements in axonomic problems. *Ann. Eugen.* 7, 179–188.
- Vapnik, V. N. (1995) *The Nature of Statistical Learning Theory*, Springer, New York.
- Christianini, N., and Shawe-Taylor, J. (2000) *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, New York.
- Herbrich, R. (2002) *Learning Kernel Classifiers: Theory and Algorithms*, MIT Press, Cambridge, MA.
- Schölkopf, B., and Smola, A. J. *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond*, The MIT Press, Cambridge, MA.
- Cortes, C., and Vapnik, V. (1995) Support-vector networks. *Machine Learn.* 20, 273–297.
- Burges, C. J. C. (1998) A tutorial on Support Vector Machine for pattern recognition. *Data Min. Knowl. Disc.* 2, 121–167.
- Zernov, V. V., Balakin, K. V., Ivaschenko, A. A., Savchuk, N. P., and Pletnev, I. V. (2003) Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *J. Chem. Inf. Comput. Sci.* 43, 2048–2056.
- Cai, C. Z., Han, L. Y., Ji, Z. L., and Chen, Y. Z. (2004) Enzyme family classification by support vector machines. *Struct., Funct., Bioinformatics* 55, 66–76.
- Byvatov, E., Fechner, U., Sadowski, J., and Schneider, G. (2003) Comparison of support vector machine and artificial neural network systems for drug/non-drug classification. *J. Chem. Inf. Comput. Sci.* 43, 1882–1889.
- Burbidge, R., Trotter, M., Buxton, B., and Holden, S. (2001) Drug design by machine learning: Support vector machines for pharmaceutical data analysis. *Comput. Chem.* 26, 5–14.
- Liu, H. X., Zhang, R. S., Yao, X. J., Liu, M. C., Hu, Z. D., and Fan, B. T. (2003) QSAR study of ethyl 2-[(3-methyl-2,5-dioxo(3-pyrrolinyl)amino]-4-(trifluoromethyl)pyrimidine-5-carboxylate: An inhibitor of AP-1 and NF- κ B mediated gene expression based on support vector machines. *J. Chem. Inf. Comput. Sci.* 43, 1288–1296.
- Liu, H. X., Zhang, R. S., Yao, X. J., Liu, M. C., Hu, Z. D., and Fan, B. T. (2004) Prediction of isoelectric point of amino acid based on GA-PLS and SVMs. *J. Chem. Inf. Comput. Sci.* 44 (1), 161–167.
- Xue, C. X., Zhang, R. S., Liu, M. C., Hu, Z. D., and Fan, B. T. (2004) Study of the quantitative structure–mobility relationship of carboxylic acids in capillary electrophoresis based on support vector machines. *J. Chem. Inf. Comput. Sci.* 44 (3), 950–957.
- Jalas, J. R., McIntee, E. J., Kenney, P. M. J., Upadhyaya, P., Peterson, L. A., and Hecht, S. S. (2003) Stereospecific deuterium substitution attenuates the tumorigenicity and metabolism of the tobacco-specific nitrosamine 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK). *Chem. Res. Toxicol.* 16, 794–806.
- Wang, M., Cheng, G., Sturla, S. J., Shi, Y., McIntee, E. J., Villalta, P. W., Upadhyaya, P., and Hecht, S. S. (2003) Identification of adducts formed by pyridyloxobutylation of deoxyguanosine and DNA by 4-(acetoxymethylnitrosamino)-1-(3-pyridyl)-1-butanone, a chemically activated form of tobacco specific carcinogens. *Chem. Res. Toxicol.* 16, 616–626.
- Wong, H. L., Murphy, S. E., and Hecht, S. S. (2003) Preferential metabolic activation of N-nitrosopiperidine as compared to its structural homologue N-nitrosopyrrolidine by rat nasal mucosal microsomes. *Chem. Res. Toxicol.* 16, 1298–1305.
- Schut, H. A. J., and Snyderwine, E. G. (1999) DNA adducts of heterocyclic amine food mutagens: Implications for mutagenesis and carcinogenesis. *Carcinogenesis* 20, 353–368.
- Balaban, A. T. (1982) Highly discriminating distance-based topological index. *Chem. Phys. Lett.* 89, 399–404.
- Randi, M. (1975) On characterization of molecular branching. *J. Am. Chem. Soc.* 97, 6609–6615.
- Katritzky, A. R., Perumal, S., and Petrukhin, R. (2001) A QSRR treatment of solvent effects on the decarboxylation 6-nitrobenzoxazole-3-carboxylates employing molecular descriptors. *J. Org. Chem.* 66, 4036–4040.
- Basak, S. C., Harriss, D. K., and Magnuson, V. R. (1984) *J. Pharm. Sci.* 73, 429.
- Stanton, D. T., and Jurs, P. C. (1990) Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure–property relationship studies. *Anal. Chem.* 62, 2323–2329.
- Stanton, D. T., Egolf, L. M., and Jurs, P. C. (1992) Computer-assisted prediction of normal boiling points of pyrans and pyrroles. *J. Chem. Inf. Comput. Sci.* 32, 306–316.
- Zefirov, N. S., Kirpichenok, M. A., Izmailov, F. F., and Trofimov, M. I. Calculation schemes for atomic electronegativities in molecular graphs within the framework of Sanderson principle. *Dokl. Akad. Nauk SSSR*.
- Stanton, D. T., and Jurs, P. C. (1990) Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure–property relationship studies. *Anal. Chem.* 62, 2323.
- Liu, H. X., Zhang, R. S., Luan, F., Yao, X. J., Liu, M. C., Hu, Z. D., and Fan, B. T. (2003) Diagnosing breast cancer based on support vector machines. *J. Chem. Inf. Comput. Sci.* 43 (3), 900–907.