QSAR & Combinatorial Science QCS

Applying Novel Three-Dimensional Holographic Vector of Atomic Interaction Field to QSAR Studies of Artemisinin Derivatives

Yanrong Ren[a,b], Guoping Chen[a]*, Zongli Hu[a], Xuqing Chen[a], Bo Yan[a]

[a] Department of Life Science and Chemistry College of Bioengineering, Chongqing University, Chongqing 400044, China
[b] Chongqing Education College, Chongqing 400067, China, E-mail: chenguoping@cqu.edu.cn

Keywords: Artemisinin derivative – Molecular structural characterization – Quantitative structure–activity relationship – Three-dimensional holographic vector of atomic interaction field

Received: December 31, 2006; Accepted: May 25, 2007

DOI: 10.1002/qsar.200630167

Abstract

Classifying common atoms in organic compounds in terms of families in periodic table of elements and hybridization states, a novel rotation–translation invariant Three-Dimensional Molecular Structural Characteristic (3D-MSC) method, Three-Dimensional Holographic Vector of Atom Interacting Field (3D-HoVAIF), is proposed by calculating three kinds of interatomic nonbonding interactions namely electrostatic, van der Waals, and hydrophobic interactions. In an attempt to apply 3D-HoVAIF into QSAR studies on antimalarial activities of 32 artemisinin derivatives, the resulting Genetic Algorithm-Partial Least Square (GA-PLS) model is confirmed to be stable and predictable by both modeling validation and methodological contrast. For external test set, the 3D-HoVAIF model has correlation coefficient (q_{ext}^2) and Root Mean Square Error of Prediction (RMSEP) of 0.751 and 0.372, respectively.

1 Introduction

Quantitative Structure–Activity Relationship (QSAR) plays an important role in Computer-Aided Drug Design (CADD) by providing internal relationship between drug molecular structures and bioactivities by mathematical and statistical methods, thus directing meaningful pharmacodynamic predictions on unknown compounds and structural modifications on lead compounds. Often QSAR studies involve in two keys: Molecular Structural Characterization (MSC) and construction of statistical model. MSC aims to transform molecular structural features into a group of numerical codes, and to endeavor to minimize information loss during this process. In early 1940s, Wiener [1] proposed the famous W index on the basis of two-dimensional graphical structures, and following that, a series of other topological indices were consequently developed, including Hosoya index (Z index) [2], Randic index (χ index) [3], Balaban index (J index) [4], Kier–Hall index ($^m\chi^v$ index) [5], *etc.* which achieved successful results on predictions of physicochemical properties for simple homologous compounds. Then the middle of 1960s witnessed prelude of QSAR studies, confirmed by the Hansch–Fujita analytical method (proposed by Hansch and Fujita [6] to relate molecular bioactivities to physicochemical properties) and Free–Wilson method (corporately created by Free and Wilson [7] to relate molecular bioactivities to indication variables) which were later successfully utilized in the drug design of pyridonecarboxylic [8], dihydrofolate reductase inhibitor [9], *etc.* However, based on molecular two-dimensional structures, the above-mentioned methods are all insufficient in providing information on three-dimensional drug–receptor interactions, thus being restricted with respect to further development. In contrast with 2D-MSC, Three-Dimensional Molecular Structural Characteristic (3D-MSC) methods are more amenable to physicochemical interpretations, enabling directly reflections of ligand–receptor binding manners and nonbonding interactions in drug molecules and thus obtaining wide generalization in QSAR fields in the last 20 years. Among many 3D-MSC methods, Comparative Molecular Field Analysis (CoMFA) [10], proposed on the basis of DYLOMMS [11] and PLS [12] by Cramer, has attracted considerable importance as a standard QSAR technique to be widely utilized in drug analysis and design. In addition to CoMFA, other 3D-MSC methods are also put forward one after one, with some pertaining to conformational alignment-dependent kinds (*e.g.*, CoMSIA [13], HASL [14], COMPASS [15], *etc.*) and some to conformational alignment-independent (*e.g.*, WHIM [16], COMMA [17], GRIND [18], *etc.*). The latter is often called Translationally and Rotationally Invariant Descriptor (TRI descriptor) [19] which has great

merits such as easy calculation, simple operation, fewer interferences, conformational insensitiveness *etc*., thus ensuring high-speed calculations and reproducibility to achieve wide applications in the quick screening of drug-related database and pharmacodynamic evaluation of lead compounds. Although many available references always report that qualities of TRI model are in no way inferior to that of CoMFA, TRI descriptor applications largely fall behind conformation alignment-dependent CoMFA methods since TRI models are less interpretable.

According to atomic valence number, atoms have always been classified into for four types by Liu *et al.* [20–23] who subsequently discussed electronic interactions at 2D electrotopological levels. Enlightened by this idea, common atoms in organic compounds are classified for ten types in terms of families in periodic table of elements and hybridization states, and following that, a further step is taken by calculating electrostatic, van der Waals and hydrophobic interactions among the ten atomic types based on molecular 3D structures, herein resulting in a novel TRI descriptor Three-Dimensional Holographic Vector of Atomic Interacting Field (3D-HoVAIF). Compared with traditional 3D-MSC, 3D-HoVAIF differs in the following points: (a) distributions of nonbonding interaction potential fields are indirectly embodied into intramolecular interatomic interactions; (b) that atoms are classified according to chemical properties elevates resolution abilities on molecular structures and interpretabilities on statistical model; (c) avoiding demerits such as conformation alignment, grid assignment, and probe setting in CoMFA, the calculating process is largely simplified. By applying the 3D-HoVAIF approach to systematic QSAR studies on 32 artemisinin derivatives, this method has been confirmed to be efficacious to indicate information on molecular steric nonbonding potential fields and to relate with bioactivities *via* strict external and internal validations, with model constructed of high qualities and good interpretabilities.

2 Principle and Methodology

2.1 Atomic Types

As is well known, a fundamental rule in QSARs is in that molecular properties depend upon structures. Thus, it is deemed that constituent atoms of a molecule provide insight into its external properties. Often common atoms in organic molecules are distributed over five families in periodic table of elements (*e.g.*, H in family IA, C and Si in family IVA, N and P in family VA, O and S in family VIA, F, Cl, Br, and I in family VIIA), so atoms are naturally classified into five types according to families. On further consideration, since the same atom may sometimes behave with distinct chemical properties in different hybridization states, the above-mentioned five atomic types are further divided into ten types according to hybridization states (Table 1).

Table 1. Ten atomic types and their 55 interactions in 3D-HoVAIFs.

No.	Atomic type	1	2	3	4	5	6	7	8	9	10
1	H	1–1	1–2	1–3	1–4	1–5	1–6	1–7	1–8	1–9	1–10
2	$C_{(sp3)}$		2–2	2–3	2–4	2–5	2–6	2–7	2–8	2–9	2–10
3	$C_{(sp2)}$			3–3	3–4	3–5	3–6	3–7	3–8	3–9	3–10
4	$C_{(sp)}$				4–4	4–5	4–6	4–7	4–8	4–9	4–10
5	$N_{(sp3)}$, $P_{(sp3)}$					5–5	5–6	5–7	5–8	5–9	5–10
6	$N_{(sp2)}$, $P_{(sp2)}$						6–6	6–7	6–8	6–9	6–10
7	$N_{(sp)}$, $P_{(sp)}$							7–7	7–8	7–9	7–10
8	$O_{(sp3)}$, $S_{(sp3)}$								8–8	8–9	8–10
9	$O_{(sp2)}$, $S_{(sp2)}$									9–9	9–10
10	F, Cl, Br, I										10–10

2.2 Cross-Interactions

As a chemical entity, the organic molecule has its internal atoms associated with each other by chemical bonds and other factors. In 3D-HoVAIF, ten atomic types are defined, thus resulting in 55 interactions (Table 1). Taking into account that drug–receptor binding is usually closely related to three nonbonding interactions, namely to electrostatic, van der Waals, and hydrophobic interactions, $3 \times 55 = 165$ interaction items ultimately correspond to an organic molecule as the total 3D-HoVAIF descriptors. Although not directly indicating ligand–receptor interactions, these 3D-HoVAIF descriptors are rich in information on molecular potential field distributions in many cases that receptor structures are unknown.

2.3 Atomic Interaction Potential Energies

Electrostatic interaction, as an important nonbonding interaction, follows Coulomb's law (Eq. 1). In this equation, d_{ij} denotes interatomic Euclid distance, with Å serving as its unit; e ($1.6021892 \times 10^{-19}$ C) represents elementary charge; ε_0 ($8.85418782 \times 10^{-12}$ $C^2/J \cdot m$) indicates dielectronic constant in vacuum; q is atom partial charges; m and n are atomic attributes.

$$V_{mn}^{\text{ele}} = \sum_{i \in m} \sum_{j \in n} \frac{e^2}{4\pi\varepsilon_0} \frac{q_i q_j}{d_{ij}} \quad (1 \le m \le 10,\ m \le n \le 10) \tag{1}$$

van der Waals interaction describes interatomic spatial nondipole–dipole or dipole-induced interactions, here expressed by the Lennard–Jones equation (Eq. 2), where $\varepsilon_{ij} = (\varepsilon_{ii} \cdot \varepsilon_{jj})^{1/2}$ is the potential well of atom pairs, with its value taken from Ref. [24]; $R_{ij}^* = (C_h R_{ii}^* + C_h R_{jj}^*)/2$ is the van der Waals radius for modified atom-pair, with cor-

rected factors $C(sp^3)=1.00$, $C(sp^2)=0.95$, and $C(s)=0.9$ [25].

$$V_{mn}^{ste}=\sum_{i\in m}\sum_{j\in n}\varepsilon_{ij}\left[\left(\frac{R_{ij}^*}{d_{ij}}\right)^{12}-2\left(\frac{R_{ij}^*}{d_{ij}}\right)^{6}\right] \quad (1\le m\le 10,\ m\le n\le 10) \tag{2}$$

Hydrophobic interaction plays very important roles for drug molecules binding to organisms. Indicating information on systematic entropic changes, this interaction is difficult to be uniformly described. In 3D-HoVAIF, hydrophobic interaction is expressed by Eq. 3 which is defined in method Hint proposed by Kellogg *et al.* [26]. In this equation, S is the atomic Solvent Accessible Surface Area (SASA) [27], indicating the surface area formed by a water-molecule probe rolling its center at an atom surface in a circle; a is the atomic hydrophobic constant, value taken from Ref. [28]; T is the discriminant function, denoting entropic changing orientation when different interatomic interactions take place.

$$V_{mn}^{hyd}=\sum_{i\in m}\sum_{j\in n}S_i a_i S_j a_j e^{-d_{ij}}T_{ij} \quad (1\le m\le 10,\ m\le n\le 10) \tag{3}$$

2.4 3D-HoVAIF Illustration

Scheme 1 illustrates the calculating process of 3D-HoVAIF descriptors for several hetero atoms included in a benzene substitution. As for atoms N(sp^3), O(sp^3), and F which pertain to the 5th, 8th, and 10th atomic types respectively, interactions among them are separately of 5–8, 5–10, and 8–10 3D-HoVAIF kinds. In the same way, interactions between atoms H–H, C(sp^2)–C(sp^2), and C(sp^2)–H, as well as these and hetero atoms, can also be calculated.

3 Results and Discussion

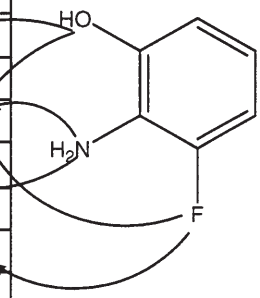
3.1 Dataset for Artemisinin Derivatives

Qinghao (*Artemisia annua*), a Chinese traditional medicine, lasted for more than 2000 years. In 1972, the valid ingredient artemisinin (Figure 1) of this plant was successfully isolated and identified, generating considerable interest worldwide due to its particular structure and high antimalarial activity but low toxicity [29, 30]. Catalyzed by heme–iron complex (Figure 2), artemisinin breaks its peroxy bond, dissociating into a series of free radicals which would disturb functions of membrane-bioblast of plasmodium [31, 32]. For the reason that this process includes many intricate steps, the antimalarial mechanism has not yet been completely elucidated in spite of extensive studies at molecular levels [33–36]. Thirty-two artemisinin derivatives are taken from refs. [37–41], with activities IC_{50} tested by *Plasmodium falciparum* D-6 clone. To reduce differences among different experiments, relative activities (RA) are employed, expressed as log(RA) = log[(artemisinin IC_{50}/analogue IC_{50}) × (analogue MW/artemisinin MW)] [42] (Table 2).

3.2 Calculations for 3D-HoVAIF

The original steric structure of heme is peeled off from the crystal structure of hemoglobin which was tested at 2.1 Å resolution by X-ray diffraction by Shaanan [43] (PDB ID: 1HHO) (Figure 3). Assignment of electronic charges to heme is implemented by quantum chemical software package Gaussian 98W at density function levels (B3LYP/6-31G**). While the artemisinin steric structure is constructed by molecular simulating software HyperChem 7.5 [44], it is then optimized at molecular mechanics levels (MM+ force field). Artemisinin–heme interaction is simulated by molecular docking software AutoDock 3.0 [45],

No.	AT	1	2	3	4	5	6	7	8	9	10
1	H	·····	·····	·····	·····	·····	·····	·····	·····	·····	·····
2	C(sp3)		·····	·····	·····	·····	·····	·····	·····	·····	·····
3	C(sp2)			·····	·····	·····	·····	·····	·····	·····	·····
4	C(sp)				·····	·····	·····	·····	·····	·····	·····
5	N(sp3)					·····	·····	·····	2.14	·····	6.75
6	N(sp2)						·····	·····	·····	·····	·····
7	N(sp)							·····	·····	·····	·····
8	O(sp3)								·····	·····	0.78
9	O(sp2)									·····	·····
10	F										·····



Scheme 1. Illustration of 3D-HoVAIF calculations.

Table 2. Molecular structures and bioactivities of 32 arteminisin derivatives.

No.[a]	Compound	R_1	R_2	R_3	log(RA)		
					Observed	Calculated (M1)	Calculated (M2)
1	a	$-H$	$-CH_3$	–	0.854	0.7716	1.1079
2*	a	$-CH_2COOCH_2CH_3$	$-CH_3$	–	0.689	0.0316	0.3687
3	a	$-(CH_2)_2COOCH_3$	$-CH_3$	–	0.202	0.1199	−0.1315
4*	a	$-CH_2C_6H_4COOCH_3$	$-CH_3$	–	0.580	−0.9663	0.9348
5	a	$-CH_2COO^-$	$-CH_3$	–	−1.264	−1.3123	−1.4090
6*	a	$-(CH_2)_2COO^-$	$-CH_3$	–	−1.463	−0.3084	−1.0063
7	a	$-(CH_2)_3COO^-$	$-CH_3$	–	−1.411	−1.4146	−1.2246
8	a	$-CH_2C_6H_4COO^-$	$-CH_3$	–	0.226	0.1975	−0.0044
9	a	$-(CH_2)_3COOH$	$-CH_3$	–	−0.786	−0.1032	0.4737
10	a	$-CH_3$	$-CH_3$	–	0.423	0.2050	−0.4210
11	a	$-C_2H_5$	$-CHO$	–	0.079	−0.3924	−0.1139
12	a	$-C_2H_5$	$-CH_3$	–	0.146	−0.1967	0.3688
13	b	Phenyl	–	–	−0.786	−0.5636	−0.5839
14	b	2-*N*-phenyl	–	–	−1.139	−0.8805	−1.2511
15	b	2,6-*N*-phenyl	–	–	−1.666	−0.9046	−1.3430
16*	c	$-CH_2CH_2CH_3$	$-CH_2C_6H_5$	–	−0.122	0.6355	0.1032
17	c	$-CH_2C_6H_5$	$-CH_2CH_2CH_3$	–	0.375	0.6347	0.7679
18*	c	$-C_6H_5$	$-COOCH_2CH_3$	–	0.904	0.3025	1.3650
19	c	$-COOCH_2CH_3$	$-C_6H_5$	–	0.655	0.4792	0.5825
20	c	$-CH_3$	$-C_6H_4-CF_3$(p)	–	0.199	−0.6203	−0.0234
21	c	$-CH_2COOCH_2CH_3$	$-C_6H_5$	–	0.667	0.6277	0.5976
22	c	$-C_6H_5$	$-CH_2COOCH_2CH_3$	–	0.700	0.6360	0.5973
23	c	$-CH_2COOCH_2CH_3$	$-C_6H_4-NO_2$(p)	–	0.612	0.5583	0.5986
24	c	$-C_6H_4$-NO_2(p)	$-CH_2COOCH_2CH_3$	–	0.971	0.6083	0.5971
25	c	$-C_6H_4$-$COOCH_3$(p)	$-CH_3$	–	0.522	0.3655	0.5438
26	c	$-CH_3$	$-C_6H_4-COOH$(p)	–	−0.399	−0.0054	−0.5169
27*	c	$-C_6H_4$-COOH(p)	$-CH_3$	–	−0.105	−0.4262	−0.3137
28	c	$-CH_2COOH$	$-C_6H_4-NO_2$(p)	–	−0.094	0.2172	0.3028
29	d	$-OH$	$-CH_3$	$-CH_2CF_3$	0.255	0.7037	−0.0441
30	d	$-CH_3$	$-OH$	$-CH_2CF_3$	−0.824	−0.8412	−0.3811
31*	d	$-OH$	$-CH_3$	$-CH_2CH_3$	−0.347	−0.9692	0.1301
32	d	$-CH_3$	$-OH$	$-CH_2CH_3$	−1.097	−0.5785	−0.8938

[a] "*" Superscript indicates that the compound was chosen to be a member of the test set.

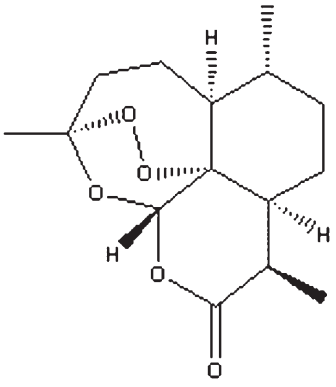


Figure 1. Molecular structure of arteminisin.

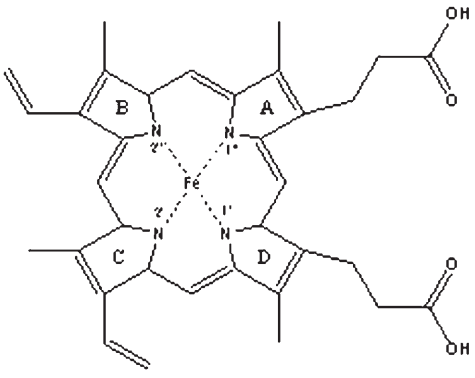


Figure 2. Molecular structure of heme.

with detailed parameter settings referring to reports by Tonmunphean *et al.* [46]. Figure 4 presents the most stable artemisinin – heme binding conformation among a series of candidates generated by molecular docking, indicating that peroxy bridge of artemisine gets close to iron Fe^{2+} of heme from its obverse side, with the two O-Fe^{2+} distance of 1.98 and 2.76 Å, respectively. Taking this docking complex as the conformation template, we construct and optimize molecular structures of 32 artemisine derivatives by molecular simulating software HyperChem 7.5 [44]. Then

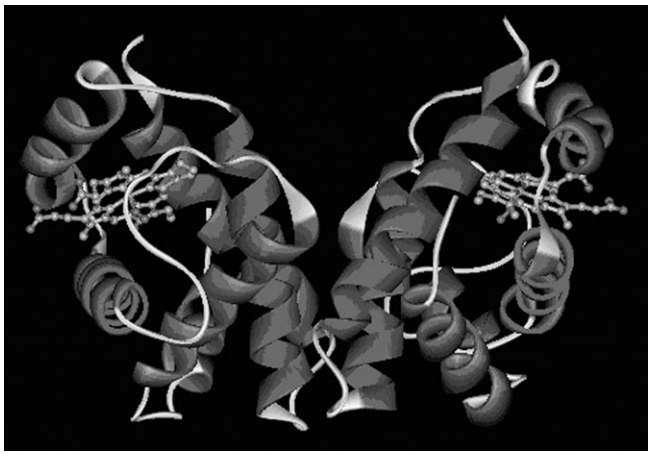


Figure 3. Crystal structure of hemoglobin.

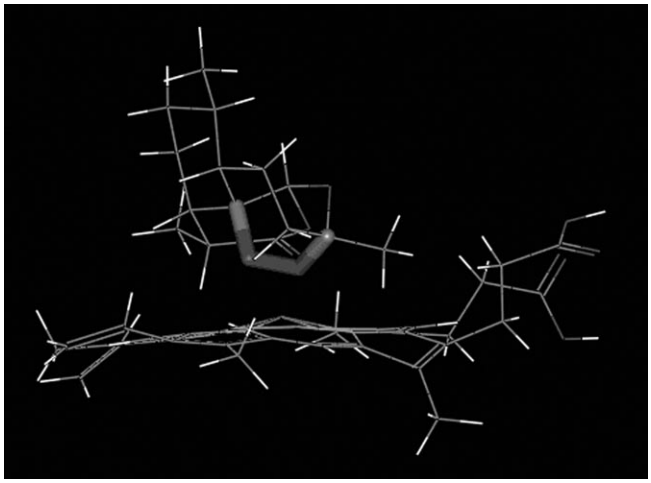


Figure 4. Artemisinin – heme binding conformation by molecular docking.

the optimum structures are selected and put into semi-experimental quantum chemical software MOPAC 6.0 [47], and following that, atomic Mülliken partial charge is calculated in the form of single point at PM3 levels for each molecule. Finally, Cartesian coordinates and partial charges for each atom are obtained and fed into the GET3D program (written in C language), generating the ultimate 3D-HoVAIF descriptors.

3.3 Partition of Sample Set and Validation of the Model

Recently, many researches indicate that only the cross-validation q^2 is insufficient in confirming validities of the QSAR model and it needs both rigorous statistical and external tests [48 – 50]. During this process, how to effectively divide the test set is very important. While the common method of random sampling is practically proved to be greatly arbitrary and incidental, it is infeasible to ensure training set space efficiently to cover the test set. In view of that, D-optimal (determinant-optimal) algorithm is utilized to divide the test set. D-optimal is an algorithm which allows for sampling space of the training set the mostly covers the test set *via* maximizing the determinant value of information matrix ($X'X$) of the training set, Here the details about D-optimal are introduced in Ref. [51, 52]. In this context, the D-optimal algorithm is implemented by Matlab 7.0 [53], generating 25 training and seven test samples (marked by symbol "*" in Table 2).

For the test set, modeling predictabilities are often evaluated by external correlation coefficient and Root Mean Square Error of Prediction (RMSEP), while recently several following parameters, reported by Tropsha *et al.* [54, 55], are deemed to be more convincible for such a purpose

$$q^2_{\text{ext}} = 1 - \frac{\sum_{i-1}^{n_{\text{ext}}} (Y^i_{\text{obsd}} - Y^i_{\text{pred}})^2}{\sum_{i-1}^{n_{\text{ext}}} (Y^i_{\text{obsd}} - \bar{Y}_{\text{tra}})^2} \tag{4}$$

or

$$\frac{r^2_{\text{ext}} - r'_{0,\text{ext}}2}{r^2_{\text{ext}}}$$

$$0.85 \le k \le 1.15 \text{ or } 0.85 \le k' \le 1.15 \tag{6}$$

where $q_{\text{ext}}{}^2$ (external q^2) is the external correlation coefficient indicating predictabilities on the test set by model. Y^i_{obs} denotes observed bioactivities on the test set while Y^i_{pred} is the predicted value by model for test samples. $\bar{Y}_{\text{tra}}$ represents the average observed bioactivities over training samples; $r_{\text{ext}}{}^2$ indicates the correlation coefficient of the observed-to-predicted regression for the test set, $r^2_{0,\text{ext}}$ and $r'_{0,\text{ext}}2$ are correlation coefficients of the origin-passed regression for the test set (predicted *vs.* observed activities $r^2_{0,\text{ext}}$, and observed *vs.* predicted activities$r'_{0,\text{ext}}2$), with k and k' corresponding to separate slopes.

3.4 QSAR Modeling and Analysis

For the reason that 32 artemisinin derivatives are lacking atoms P, S, C_{sp3}, N_{sp3}, *etc.*, 84 empty items occurr in 165 3D-HoVAIFs. Removing all these empty ones, we ultimately obtain 81 3D-HoVAIFs for each molecule, where electrostatic interactions are expressed by variables V1 – V27, van der Waals interactions by variables V28 – V54 and hydrophobic interactions by variables V55 – V81. Based on that, a linear relationship is obtained, by chemical quantum software SIMCA-P 10.0 [56], to relate 3D-HoVAIFs (X) with bioactivities (Y) for 25 training samples, and the resulting PLS model contains two prominent principal components which cumulatively account for 78.4% square error of variables Y and 55.2% by cross-validation. Statistics of this model are $r^2 = 0.784$, $q^2 = 0.552$, RMSEE = 0.359, and RMSCV = 0.548. Figure 5 shows the plot of the calculated *versus* observed activities for 25 training samples, with $r_{\text{fit}} = 0.886$ (indicating free linear fitness) and

slope $k_{fit}=0.750$, respectively. Thus, this model (called M1), in spite of some linearity drift, is suggested to favorably relate its calculated values with observed activities for the training set. However, M1 is practically a little unstable, confirmed by its low cross-validation ($q^2=0.552$) which falls below the recommended lower limit ($q^2>0.5$) as reported by Tropsha *et al.* [55]. Predicting statistics on seven test samples by model M1 are ${q_{ext}}^2=0.618$, ${r_{ext}}^2=0.674$, $r^2_{0,ext}=0.673$, $r'_{0,ext}2=0.673$, $k=0.628$, $k'=0.984$, and RMSEP$=0.462$ ($r^2_{ext}-r^2_{0,ext}/.r^2_{ext}=r^2_{ext}-r'_{0,ext}2/.r^2_{ext}=0.001$) (Figure 6), of which only the k does not satisfy Eq. 6. In summary, the model M1, although meeting basic demands of QSARs, is not very perfect, with both its stabilities for internal cross-validation and predictabilities for external samples being slightly low. The reason for this may be owing to too many 3D-HoVAIF descriptors. Although containing valuable structural information, some of them actually contribute little to artemisinin bioactivities, thus introducing noise and other interfering factors. So, some 3D-HoVAIF descriptors, unfavorably related to bioactivities, are filtered out prior to constructing the PLS model. The process of variable selection is implemented by Genetic Algorithm-Partial Least Square (GA-PLS) [57], with relative programs Gaot_Toolbox [58] and PLS_Toolbox [59] based on Matlab 7.0 environment. Parameter settings in GA are as follows: population size: 100, maximum iteration: 200, convergence standard: 80% of individuals achieve to an agreement, mutation probabilities: 0.5%, cross-over point: two points, cross-validation: leave-one-out, data pretreatment: autoscaling, and other settings just reserve the defaults. After such a screening, the optimal subset is composed of variables V6, V9, V10, V11, V12, V18, V24, V26, V27, V31, V32, V33, V36, V41, V42, V47, V48, V50, V52, V55, V56, V60, V64, V67, V71, V77, and of that there are nine electrostatic interactions, ten van der Waals items, and seven hydrophobic interactions. Generally speaking, this GA-PLS model has been largely advanced, with ultimately 26 variables from the overall 81 3D-HoVAIF descriptors to participate in modeling. By a further analysis of this dataset by SIMCA-P 10.0 [56], the resulting PLS model (called M2) gets two prominent principal components, with statistics as $r^2=0.852$, $q^2=0.778$, RMSEE$=0.297$, RMSCV$=0.368$, $r_{fit}=0.923$ (indicating correlativeness by free linear fitness) and slope $k_{fit}=0.849$ (Figure 7). Then, the normal probability of standardized residual [60] is tested for model M2 to validate its normal hypothesis. From Figure 9, most sample residues are found to obey a normal distribution, with only one standardized residual going beyond a ± 2 range, so hypothesis for M2 is believed to be true. Figure 10 shows scoring scatter of the 25 samples at the top two PLS principal component spaces [60], wherein most samples fall in the ellipse Hotelling T^2 with a 95% confidence. Besides, this figure also indicates that bioactivities of these samples are increasingly distributed from the left to the right, and compounds with similar bioactivities are favorably assembled together, suggesting the top two principal components are already sufficient to characterize activity distribution features for this group of samples. The values predicted by M2 for seven test samples are ${q_{ext}}^2=0.751$, ${r_{ext}}^2=0.831$, $r^2_{0,ext}=0.831$, $r'_{0,ext}2=0.831$, $k=0.882$, $k'=0.867$, and RMSEP$=0.372$ ($r^2_{ext}-r^2_{0,ext}/.r^2_{ext}=r^2_{ext}-r'_{0,ext}2/.r^2_{ext}=0$), meeting demands of Eqs. 5 and 6. Notice that r^2_{ext}, $r^2_{0,ext}$, and $r'_{0,ext}2$, indicating predictabilities on test set are all of 0.831, suggesting that this model is unbiased. Figure 8 presents the origin-passed regression line which almost equally goes through the sample area in an angle of 45° (slope $k=0.882$), with no obvious abnormal situations. To benefit a further comparison, Table 3 lists statistics of both model M1 and M2. In contrast with M1, M2 is obviously improved, especially with its q^2 (indicating internal stabilities) and q^2_{ext} (indicating external predictabilities) being prominently superior to that of M1.

3.5 Contrast Study

To further investigate the performance of 3D-HoVAIF descriptors, we make comparisons of this method with other molecular descriptors. In QSARs, large numbers of molecular structural characteristic methods already exist, approximately classified into several following types such as 2D descriptor (*e.g.*, E-state index [61], MEDV [21], CATS [62, 63], *etc.*), 3D TRI descriptors (*e.g.*, WHIM [16], GRIND [18], DiP [19], *etc.*), and conformation alignment-based 3D descriptors (*e.g.*, CoMFA [10], MSA [64], SOMFA [65], *etc.*). In this context, three typical molecular structural characteristic methods are separately selected from the above-mentioned three types, referred to as MEDV, WHIM, and SOMFA. The reason for such selections can be summarized as: (a) MEDV, a sort of electro-topological

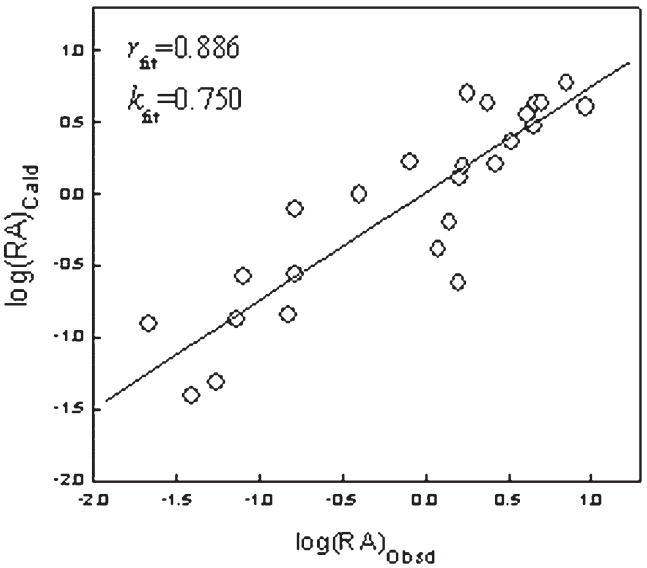


Figure 5. Plot of 3D-HoVAIF (M1) calculated *versus* observed activities for 25 artemisinin derivatives in training set.

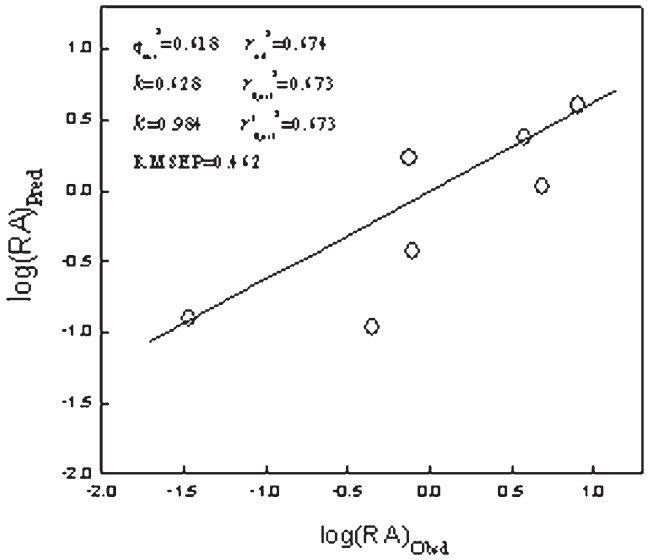


Figure 6. Plot of 3D-HoVAIF (M1) predicted *versus* observed activities for seven artemisinin derivatives in test set.

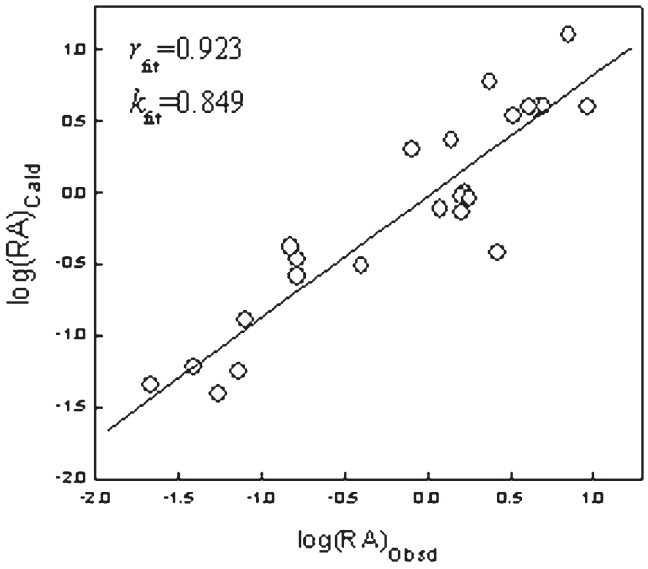


Figure 7. Plot of 3D-HoVAIF (M2) calculated *versus* observed activities for 25 artemisinin derivatives in training set.

index developed by Liu *et al.* [20–23], has its ideas of atomic classification manner and paired interaction integrated into the design of 3D-HoVAIF descriptors; (b) WHIM, being representative in TRI descriptors, has wide applications in many other research fields; (c) SOMFA, similar to CoMFA which is based upon molecular conformation alignment, is potent to investigate contributions of molecular shape and electrostatic features to bioactivities. Calculations of MEDV, WHIM, and SOMFA descriptors

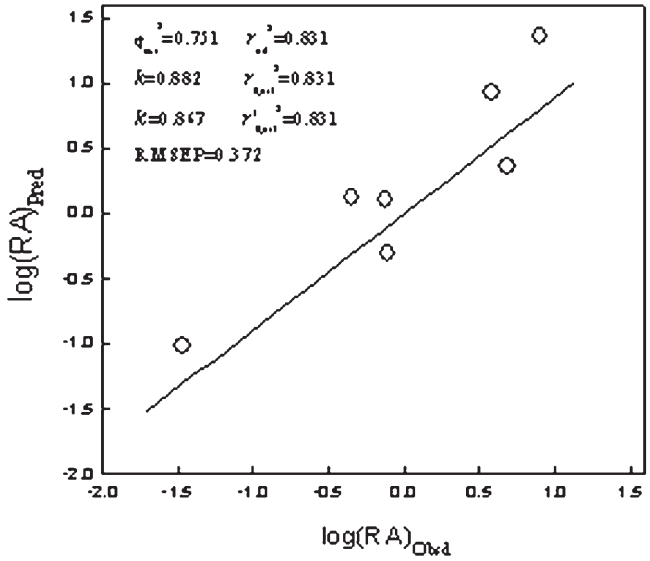


Figure 8. Plot of 3D-HoVAIF (M2) predicted *versus* observed activities for seven artemisinin derivatives in test set.

are implemented by software GITM 1.1, Dragon 5.0, and Somfa 2.0, respectively.

In MEDV, 13 types of atoms transmit electrostatic interactions by chemical bonds, ultimately yielding 91-dimensional vectors. Prior to creating model, all empty items are omitted. In WHIM and SOMFA, original molecular conformations of 32 artemisinin derivatives are just the same as those in 3D-HoVAIF. For SOMFA, molecular conformation alignment is fulfilled by module RMS Fit in Alchemy 2000 [66], and the optimal model gets its shape/electrostatic potential weight factor c1 of 0.47. To facilitate comparison with 3D-HoVAIF, the GA-PLS modeling method is utilized in MEDV and WHIM models; while for the SOMFA model, the default MLR modeling method is reserved here due to the particularity of SOMFA itself. Table 3 lists modeling results on this dataset separately by MEDV, WHIM, and SOMFA. From this table, the MEDV model is found to be the most inferior with respect to either its fitting or predicting abilities; the PLS model constructed by one principal component has correlation coefficient r^2 of 0.612 and external predicting q_{ext}^2 of 0.565 which just falls below the general standard ($q_{ext}^2 > 0.6$). Figure 11(a) delineates the plot of the predicted *versus* observed activities for seven test samples in the MEDV model, indicating that sample points are sporadic and one or two compounds even has large predicting errors. For the WHIM model which achieves good results on the training set ($r^2 = 0.831$), predictabilities on the test set are a little low ($q_{ext}^2 = 0.650$). From Figure 11(b), it is intuitively revealed that although all sample points are uniformly distributed along an origin-passed oblique line, they are far way from it, suggesting the model possesses no good stabilities and predictabilities. Both SOMFA and 3D-HoVAIF models provide good results, with calculations on both

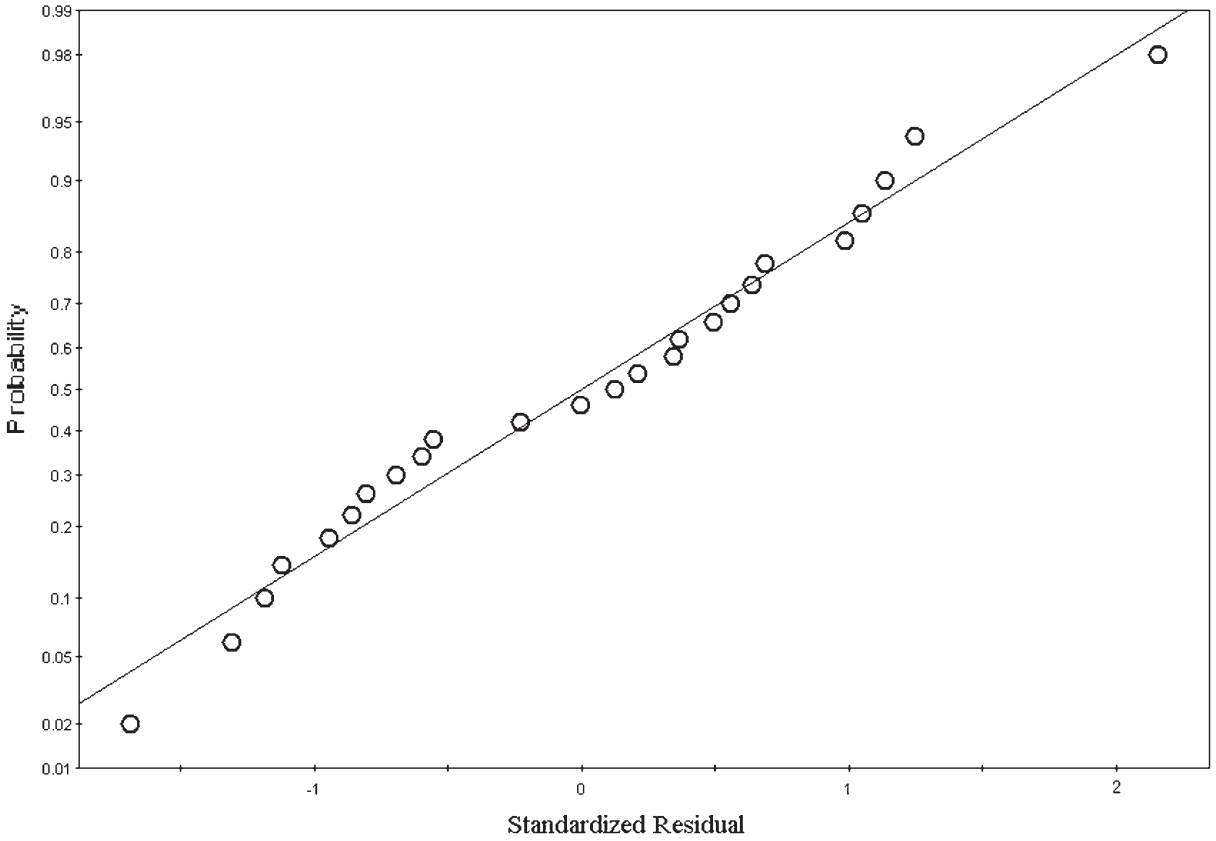


Figure 9. The normal probability plot of the *Y*-standardized residuals for 25 artemisinin derivatives.

training and test sets being above 0.8 and 0.75, where, predictabilities q_{ext}^2 of SOMFA are nearly equal to that of the 3D-HoVAIF model. By contrasting Figure 8 with Figure 11(c), an approximate result on seven test samples is found between 3D-HoVAIF and SOMFA models, with sample points of both positive and negative errors being5 and 2, respectively. Thus, both are deemed to be reasonable QSAR model.

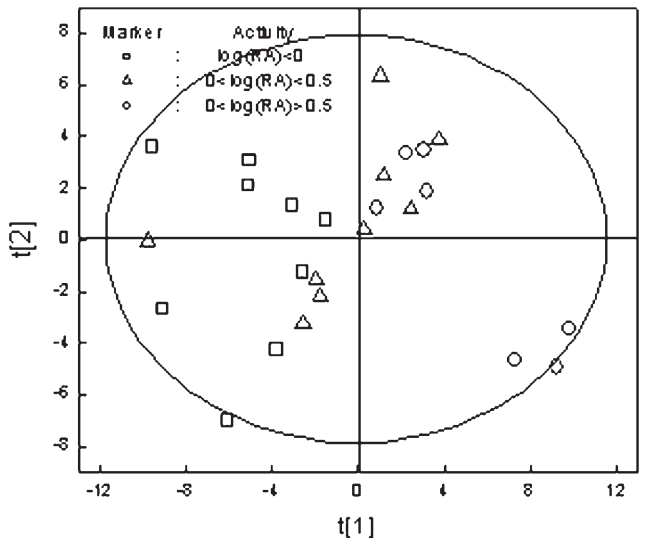


Figure 10. The PLS scores t1 and t2 for 25 artemisinin derivatives in training set.

In summary for this group of artemisinin derivatives, 2D MEDV descriptors are difficult to correctly reflect information on molecular steric conformation, thus yielding poor-quality model. WHIM index, pertaining to 3D TRI descriptors, is improved on the level of MEDV. However, lacking a straight reflection of information on nonbonding potential fields for drug molecules, the WHIM model still does not have its predictabilities largely advanced. 3D-HoVAIF and SOMFA, overcoming defects in MEDV and WHIM descriptors, enable an efficacious extraction of structural information directly related to bioactivities, thus performing favorably. But in contrast with SOMFA which requires molecular conformation alignment, 3D-HoVAIF has great merits such as easy calculation, simple operation, and highly reproducible performance as a sort of TRI descriptor.

4 Conclusions

By defining ten atomic types common in organic molecules and their 55 interaction items, a novel rotation–translation invariant 3D structure descriptor, 3D-HoVAIF,

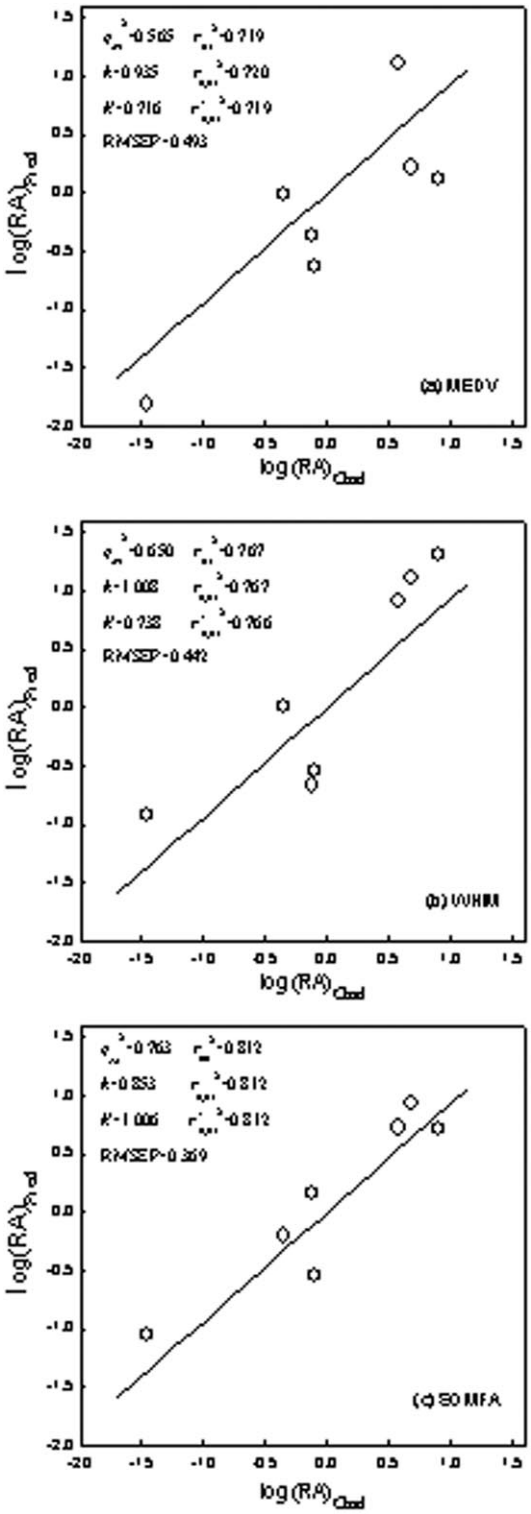


Figure 11. Plot of predicted *versus* observed activities for seven artemisinin derivatives in test set by (a) MEDV, (b) WHIM, and (c) SOMFA, respectively.

is derived from calculations of three nonbonding interactions directly related to drug bioactivities (*e.g.*, electrostatic, van der Waals, and hydrophobic interactions). Such a method has great merits such as experiment-free, easy calculation, and significant physicochemical meanings, avoiding many demerits such as molecular alignment and arbitrary grid partition in present 3D-QSAR fields. Besides, classifying atoms in terms of hybridization states and families in periodic table of elements also helps further generalize the 3D-HoVAIF approach into heteroatom-included systems. In this context, 3D-HoVAIF is employed to systematically perform QSAR studies on 32 artemisinin derivatives, with constructed model proving to be stable and predictable *via* double of internal and external tests. 3D-HoVAIF descriptor, favorably related to bioactivities of drug and biological molecules, is thus deemed to be promising in the screening of lead compounds and structural modification in future.

Acknowledgements

We thank the research group of Professor Zhiliang Li for providing aid on both software and technology. This study was supported by The National Natural Science Foundation of China (NSFC, grant number 30471180) and by Chongqing Science and Technology Committee (CSTC, grant number 2006BA5006).

References

[1] H. Wiener, *J. Am. Chem. Soc.* **1947**, *69*, 2636 – 2641.
[2] H. Hosoya, *Bull. Chem. Soc.* **1971**, *44*, 2332 – 2339.
[3] M. Randic, *J. Am. Chem. Soc.* **1975**, *97*, 6609 – 6615.
[4] A. T. Balaban, *Chem. Phys. Lett.* **1982**, *89*, 399 – 404.
[5] L. B. Kier, L. H. Hall, *Molecular Connectivity in Structure-Activity Analysis*, Wiley, New York, USA **1986**.
[6] C. Hansch, T. Fujita, *J. Am. Chem. Soc.* **1964**, *86*, 1616 – 1626.
[7] S. M. Free, J. B. Wilson, *J. Med. Chem.* **1964**, *7*, 395 – 399.
[8] H. Koga, *Kagaku No Ryoiki Zokan* **1982**, *126*, 177 – 202.
[9] C. D. Selassie, Z. X. Fang, R. L. Li, C. Hansch, G. Debnat, T. E. Klein, R. Langride, B. T. Kaufman, *J. Med. Chem.* **1989**, *32*, 1895 – 2824.
[10] R. D. Cramer, D. E. Patterson, J. D. Bunce, *J. Am. Chem. Soc.* **1988**, *110*, 5959 – 5967.
[11] M. Wise, R. D. Cramer, D. Smith, I. Exman, in: J C. Dearden (Ed.), *Quantitative Approaches to Drug Design*, (Proceedings of the 4th European Symposium on Chemical Structure-Biological Activity: Quantitative Approaches), Elsevier, Amsterdam, The Netherlands **1983**, pp. 145 – 146.
[12] P. Hoskuldsson, *J. Chemometr.* **1988**, *2*, 211 – 228.
[13] G. Klebe, U. Abraham, T. Mietzner, *J. Med. Chem.* **1994**, *37*, 4130 – 4146.
[14] A. M. Doweyko, *J. Med. Chem.* **1988**, *31*, 1396 – 1406.
[15] A. N. Jain, T. G. Dietterich, R. H. Lathrop, D. Chapman, R. E. Critchlow, T. A. Webster, T. Lozaoperez, *J. Comput. Aided Mol. Des.* **1994**, *8*, 635 – 652.
[16] R. Todeschini, P. Gramatice, R. Provenzani, *Chemom. Intell. Lab. Syst.* **1995**, *27*, 221 – 229.
[17] B. D. Silverman, D. E. Platt, *J. Med. Chem.* **1996**, *39*, 2129 – 2140.

[18] M. Pastor, G. Cruciani, I. McLay, S. Pickett, S. Clementi, *J. Med. Chem.* **2000**, *43*, 3233 – 3243.
[19] K. Baumann, *Quant. Struct. Act. Relat.* **2002**, *21*, 507 – 519.
[20] S. Liu, C. Cao, Z. Li, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 387 – 394.
[21] S. Liu, C. Yin, L. Wang. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 749 – 756.
[22] P. Zhou, H. Mei, Y. Zhou, F. Tian, Z. Li, *Chin. J. Anal. Chem.* **2006**, *34*, 200 – 204.
[23] P. Zhou, H. Zeng, F. Tian, B. Li, Z. Li, *QSAR Comb. Sci.* **2007**, *26*, 117 – 121.
[24] M. Levitt, *J. Mol. Biol.* **1983**, *170*, 723 – 764.
[25] M. Hahn, *J. Med. Chem.* **1995**, *38*, 2080 – 2090.
[26] G. E. Kellogg, S. F. Semus, D. J. Abraham, *J. Comput. Aided Mol. Des.* **1991**, *5*, 545 – 552.
[27] W. Hasel, T. F. Hendrikson, W. C. Still, *Tetrahed. Comp. Method.* **1988**, *1*, 103 – 116.
[28] J. Pei, Q. Wang, J. Zhou, L. Lai, *Proteins* **2004**, *57*, 651 – 664.
[29] R. K. Haynes, S. C. Vonwiller, *Acc. Chem. Res.* **1997**, *30*, 73 – 79.
[30] D. L. Klayman, *Science* **1985**, *228*, 1049 – 1055.
[31] Z. Ye, Z. Li, G. Li, X. Fu, H. Liu, M. Gao, *J. Trad. Chin. Med.* **1983**, *3*, 95 – 102.
[32] P. A. Berman, P. A. Adams, *Free Radic. Biol. Med.* **1997**, *22*, 1283 – 1288.
[33] A. R. Butler, B. C. Gilbert, P. Hulme, L. R. Irvine, *Free Radic. Res.* **1998**, *28*, 471 – 476.
[34] G. H. Posner, J. N. Cumming, P. Ploypradith, *J. Am. Chem. Soc.* **1995**, *117*, 5885 – 5886.
[35] W. M. Wu, Z. J. Yao, Y. L. Wu, K. Jiang, Y. F. Wang, H. B. Chen, F. Shan, Y. Li, *J. Chem. Soc. Chem. Commun.* **1996**, 2213 – 2214.
[36] W. M. Wu, Y. K. Wu, Y. L. Wu, Z. J. Yao, C. M. Zhou, Y. Li, F. Shan, *J. Am. Chem. Soc.* **1998**, *120*, 3316 – 3325.
[37] A. J. Lin, D. L. Klayman, W. K. Milhous, *J. Med. Chem.* **1987**, *30*, 2147 – 2150.
[38] A. J. Lin, L. Q. Li, D. L. Klayman, C. F. George, J. L. Flippen-Anderson, *J. Med. Chem.* **1990**, *33*, 2610 – 2614.
[39] A. J. Lin, R. E. Miller, *J. Med. Chem.* **1995**, *38*, 764 – 770.
[40] Y. M. Pu, D. S. Torok, H. Ziffer, X. Q. Pan, S. R. Meshnick, *J. Med. Chem.* **1995**, *38*, 4120 – 4124.
[41] M. A. Avery, S. Mehrotra, T. L. Johnson, J. D. Bonk, J. A. Vromn, R. Miller, *J. Med. Chem.* **1996**, *39*, 4149 – 4155.
[42] M. A. Avery, M. Alvim-Gaston, C. R. Rodrigues, E. J. Barreiro, F. E. Cohen, Y. A. Sabnis, J. R. Woolfrey, *J. Med. Chem.* **2002**, *45*, 292 – 303.
[43] B. Shaanan, *Nature* **1982**, *296*, 683 – 684.
[44] Hypercube Inc., HyperChem 7.5, **2004.** http://www.hyper.com.
[45] G. M. Morris, D. S. Goodsell, R. Huey, A. J. Olson, AutoDock Version 3.0, The Scripps Research Institute, Department of Molecular Biology, MB-5, LaJolla, California, USA.
[46] S. Tonmunphean, V. Parasuk, S. Kokpol, *Quant. Struct. Act. Relat.* **2000**, *19*, 475 – 483.
[47] J. J. P. Stewart, *J. Comput. Aided Mol. Des.* **1990**, *4*, 1 – 105.
[48] J. T. Leonard, K. Roy, *QSAR Comb. Sci.* **2006**, *25*, 235 – 251.
[49] C. Sźantai-Kis, I. Kövesdi, G. Kéri, L. Őrfi, *Mol. Divers.* **2003**, *7*, 37 – 43.
[50] P. Gramatica, P. Pilutti, E. Papa, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1794 – 1802.
[51] M. Baroni, S. Clement, G. Cruciani, S. Kettaneh-Wold, S. Wold, *Quant. Struct. Act. Relat.* **1993**, *12*, 225 – 231.
[52] P. F. de Aguiar, B. Bourguignon, M. S. Khots, D. L. Massart, R. Phan-Than-Luu, *Chemometr. Intell. Lab. Syst.* **1995**, *30*, 199 – 210.
[53] MathWorks Inc., Matlab 7.0, **2004.** http://www.mathworks.com.
[54] A. Golbraikh, A. Tropsha, *J. Mol. Graphics Mod.* **2002**, *20*, 269 – 276.
[55] A. Tropsha, P. Gramatica, V. K. Gombar, *QSAR Comb. Sci.* **2003**, *22*, 69 – 77.
[56] Umetrics Inc., SIMCA-P 10.0, **2002.** http://www.umetrics.com.
[57] R. Leardi, A. L. González, *Chemometr. Intell. Lab. Syst.* **1998**, *41*, 195 – 207.
[58] C. R. Houck, J. Joines, M. Kay, *ACM Transactions on Mathematical Software* **1996**.
[59] Eigenvector Research Inc., PLS Toolbox 3.0, **2003.** http://www.eigenvector.com.
[60] S. Wold, M. Sjöström, L. Eriksson, *Chemometr. Intell. Lab. Syst.* **2001**, *58*, 109 – 130.
[61] L. B. Kier, L. H. Hall, *Pharm. Res.* **1990**, *7*, 229 – 241.
[62] G. Schneider, W. Neidhart, T. Giller, G. Schmid, *Angew. Chem. Int. Ed.* **1999**, *38*, 2894 – 2896.
[63] G. Schneider, O. Clement-Chomienne, L. Hilfiger, P. Schneider, S. Kirsch, H. J. Bohm, W. Neidhart, *Angew. Chem. Int. Ed.* **2000**, *39*, 4130 – 4133.
[64] A. J. Hopfinger, *J. Am. Chem. Soc.* **1980**, *102*, 7196 – 7206.
[65] D. D. Robinson, P. J. Winn, P. D. Lyne, W. G. Richards, *J. Med. Chem.* **1999**, *42*, 573 – 583.
[66] Tripos, Inc., Alchemy 2000, **1996.** http://www.tripos.com.

