Journal of Environmental Sciences Vol.16, No.3, pp.423—427, 2004

ISSN 1001—0742
CN11—2629/X

Article ID: 1001-0742(2004)03-0423-05 **CLC number**: X174 **Document code**: A

Holographic quantitative structure-activity relationship for prediction acute toxicity of benzene derivatives to the guppy(*poecilia reticulata*)

HUANG Hong, WANG Xiao-dong, DAI Xuan-li, YU Ya-juan, WANG Lian-sheng*

(State Key Laboratory of Pollution Control and Resources Reuse, School of the Environment, Nanjing University, Nanjing 210093, China. E-mail: environment_hh75@yahoo.com.cn)

Abstract: Holographic quantitative structure-activity relationship (HQSAR) is an emerging QSAR technique with the combined application of molecular hologram, which encoded the frequency of occurrence of various molecular fragment types, and the subsequent partial least squares (PLS) regression analysis. In this paper, the acute toxicity data to the guppy (*poecilia reticulata*) for a series of 56 substituted benzenes, phenols, aromatic amines and nitro-aromatics were subjected and this resulted in a model with a high predictive ability. The influence of fragment size and fragment distinction parameters on the quality of HQSAR model was investigated. The robustness and predictive ability of the model were also validated by leave-one-out (LOO) cross-validation procedure and external testing data set.

Keywords: benzene derivatives; HQSAR; molecular hologram; acute toxicity; guppy (*poecilia reticulata*)

Introduction

It is widely recognized that knowledge on the emissions, environmental fate and acute or chronic toxicity of pollutants are basic needs in environmental risk assessment. However, because of time and monetary constraints, hazards have been assessed for only a small percentage of these chemical compounds (Ren, 2002). Increasing concern over the use of animals in toxicity testing, allied to the cost of these tests, has made the search for and validation of alternative methods to predict the hazard a priority. In the past several decades, quantitative structure-activity relationships (QSARs) have been used widely to practice the hazard of untested chemicals with already tested chemicals by developing statistical relationships between molecular physicochemical descriptors and biological activity (Kurup, 2001). Most promising QSARs were developed and subsequently used as prediction tool, for compounds with the same or similar mode of toxic action (Mckim, 1987; Lipnick, 1989). However, due to the complexity of molecular structures of chemicals and the diverse factors involved in the complicated interaction between xenobiotics and bio-systems, it is not an easy task to correctly assign a mode of toxic action.

QSARs are now acknowledged to be in the heart of the long-term task as systematically evaluation of existing chemicals (Blum, 1990). At present, the challenge is to improve the accuracy and predictability of QSAR by taking into account the structural and physicochemical features of the tested compounds. A comparative molecular field analysis (CoMFA) program is in keeping with the general pattern of searching for these new descriptors, where steric and electrostatic fields of tested molecules are mapped by probe atom. Since its introduction in the year of 1988, the utility of CoMFA has rapidly been demonstrated in a wide range of applications (Briens, 1995; Tong, 1998; Debnath, 1999). However, CoMFA requires some knowledge or hypothesis regarding the functionally active conformations of the molecules and molecular superposition as a prerequisite for structural alignment. Moreover, care must be exercised when constructing molecular alignments because slight differences in alignment can lead to wide variation in the resultant CoMFA model (Agarwal, 1993; Hasegawa, 1999).

Holographic QSAR (HQSAR) is a newly developed QSAR technique, which relates biological activity to structural molecular composition, where molecular composition is described in terms of patterns of sub-structural fragments eliminates the need for generation of 3D structure, putative binding conformations, and molecular alignment. In HQSAR, each molecule in the database is divided into a set of unique overlapping structural fragments and sorted to form a molecular hologram, unlike other fragment-based fingerprinting methods, which encodes more information, such as branched and cyclic fragments and overlapping fragments as well as stereochemistry, and maintains a count of the number of times about each fragment occurs (Park, 2001; Rodrigues, 2002; Cha, 2003). With the combined application of molecular hologram and subsequent partial least squares (PLS) regression analysis, highly predictive QSARs are developed and validated with cross-validation procedure. No 3D molecular structure and molecular alignment are needed for the generation of hologram. With partial least square (PLS) regression analysis, the problem of co-linearity among parameters is avoided. In addition the molecular descriptors can be created automatically and quickly and

Foundation item: The National Natural Science Foundation of China (No. 20177008); the National High-Tech R&D Program of China (No. 2003CB415001; 2003CB415002; 2001AA646010-4) and European Commission International Scientific Cooperative Project (ICA4-CT-2001-10039); * Corresponding author

avoid the selection and calculation or measurement of physicochemical descriptors required by traditional QSAR (Michael, 1999). Thus, it provides promising screening tools for large scale of dataset. In the present work, we use the HQSAR technique to generate molecular representation and derive QSAR model, aiming to develop robust, highly predictive QSAR models for predictive use.

1 Materials and methods

1.1 Biological data

Toxicity data of a series of 56 benzene derivatives are taken from literature (Verhaar, 1992). The chemicals investigated include anilines, phenols, nitro-aromatics, alkyl- and / or chloro- substituted benzenes. The toxicological endpoint was defined as the negative logarithmic form of 50% lethal concentration(log $1/LC_{50}$, mmol/L).

1.2 Generation of molecular hologram

The novel molecular hologram representation designed by Tripos Associates as generated by the HQSAR package (HQSAR Software ver 1.0, Tripos Associates) is used for HQSAR analysis. Generally HQSAR analysis includes three main steps: (1) the generation of the sub-structural fragments for each molecule in the dataset; (2) the encoding of these fragments in holograms; and (3) the correlation of the structure with activity/property.

A molecular hologram is generated in much the same way as fingerprints generated by UNITY (UNITY Reference Manual, Tripos Inc., St. Louis, MO, 1995) except for a major difference. Within UNITY, each corresponding fragment is mapped to a pseudo-random integer in the 0 to 2^{31} using the CRC (cyclic redundancy check) algorithm. The integer generated by the CRC algorithm is unique and reproducible for each unique SLN (SYBYL Line Notation) string(Ash, 1997). Then the hashing occurs by folding the pseudo-random integer for a particular SLN string into the bin range defined. A molecular hologram retains a count of the number of times each bin is set rather than using a binary bit string containing either 0 or 1 in each bin. As a result, a molecular hologram is presented as a string of integers, just as follows:

Hashed fingerprints 0 0 1 1 0 0 0 1 1 1 1 0 0 0 0
Molecular hologram 0 0 6 18 0 0 0 12 5 14 42 0 0 0 0

In the above example, the chemical structure contains 97(6 + 18 + 12 + 5 + 14 + 42) fragments, which are hashed into the occupied bins as shown.

1.3 HQSAR building and regression method

All molecular modeling and statistical analyses were performed on SGI INDIG O_2 workstations using SYBYL 6.7 molecular modeling software (Tripos Inc. 2001). The 2D molecular structures of all investigated benzene derivatives were built by using the sketch option and then energy minimized with Tripos standard force field and Gasteiger-Huckel charge, with a 0.01 kcal/mol energy gradient convergence criterion.

HQSAR models were done using the following options: Fragment size: the molecular hologram generation was carried out for several fragment length size ranges, including 1 – 1, 1 – 3, 3 – 5, 3 – 10 and the default 4 – 7. Hologram length: 6 predetermined prime numbers from 97 to 353. Fragment distinction: atomic numbers (A), bond types (B), and atomic connections (C). According to the quality of the models, firstly we determined the better range of fragment size range and hologram length, then the molecular hologram generation for the better size range of fragment was processed, detailed additional description of parameters was considered, including other prime numbers for hologram length, donor and acceptor atoms (D) and inclusion of hydrogen atoms(H) for fragment distinction.

In order to get a predictive statistical model, the method of partial least squares (PLS) is used to construct the correlation between biological activities and molecular hologram. The PLS algorithm is initially used with the leave-one-out(LOO) cross-validation option to establish the optimal number of components needed for the analysis. In the leave-one-out cross-validation, each compound is systematically excluded from the dataset, and its biological activity is predicted by the model based the rest of tested compounds. This process determines the number of optimal components corresponding to the smallest standards error of prediction. Using the number of optimal components, the final PLS analysis is carried out with non-cross-validation to generate a predictive QSAR model with a conventional coefficient r^2. In this study, a cross-validated q^2 and a standard error of prediction($S.E._p$), non-cross-validated r^2, and a standard error of estimate ($S.E.$) were used for the model performance characterization.

2 Results and discussion

2.1 HQSAR model building

Toxicity data of 56 benzene derivatives are listed in Table 1. We selected the hologram length from the predetermined prime numbers: 97, 151, 199, 257, 307, and 353. Fragment distinction factors include atomic numbers (A), bond types (B), and atomic connections (C). An initial HQSAR runs using 10 components and leave-one-out cross-validation to determine the number of optimal components, then using the determined number of optimal component runs no-cross-validation procedure to yield the r^2 and standard error of estimate ($S.E.$). The results of HQSAR analyses are summarized in Table 2. According to the quality of the models based on the lowest standard error associated with the cross-validation analysis, the optimal model with a fragment size 3 – 5 from a best hologram length of 199 with 4 components was obtained. The cross-validated q^2 was 0.851 and the standard error was 0.263. The final(non-cross-validated) r^2 was 0.930 and the standard error was 0.180. Then for the better fragment

size range 3 – 5 of fragment, detailed fragment size of the molecular hologram generation was processed, another 6 predetermined prime numbers and other prime numbers near 199 for Hologram length were input for HQSAR analyses. Through these calculations the optimal model with a fragment size 4 – 5 from a best hologram length of 199 with 3 components was achieved. The cross-validated q^2 was 0.868 and the standard error was 0.246. The final (non-cross-validated) r^2 was 0.932 and the standard error was 0.176. At last additional description of fragment distinction parameters: donor and acceptor atoms(D) and inclusion of hydrogen atoms (H) was considered. The results denoted inclusion of donor and acceptor atoms (D) improved the quality, the most promising model was obtained with a fragment size 4 – 5 from a best hologram length of 199 with 4 components. The cross-validated q^2 was 0.878 and the standard error was 0.238. The final(non-cross-validated) r^2 was 0.951 and the standard error was 0.151. However inclusion of hydrogen atom(H) degraded the quality of HQSAR model, resulting in a model with a fragment size 4 – 5 from a best hologram length of 199 with 5 components. The cross-validated q^2 was 0.860 and the standard error was 0.257. The final(non-cross-validated) r^2 was 0.953 and the standard error was 0.149.

Table 1 Acute toxicity(log $1/LC_{50}$, mmol) of 56 benzene derivatives to the guppy and calculated /cross-validation predicted toxicity by HQSAR model

Chemicals	log $1/LC_{50}$, mmol			Chemicals	log $1/LC_{50}$, mmol		
	HQSAR				HQSAR		
	Obser.	Cal.	Pred.		Obser.	Cal.	Pred.
1,2,3,4-tetrachlorobenzene	0.57	0.47	0.54	3,4-dichloroaniline	1.59	1.54	1.52
1,2,3,5-tetrachlorobenzene	0.57	0.63	0.69	3,4-dichlorotoluene	1.50	1.41	1.39
1,2,3-trichlorobenzene	1.11	1.05	1.08	3,4-dimethylnitrobenzene	1.79	1.89	1.97
1,2,4-trichlorobenzene	1.12	1.12	1.14	3,5-dichloroaniline	1.38	1.31	1.64
1,2-dichlorobenzene	1.60	1.72	1.78	3,5-dichloronitrobenzene	1.47	1.60	1.67
1,3,5-trichlorobenzene	1.26	1.32	1.31	3,5-dichlorophenol	1.22	1.19	1.15
1,3-dichlorobenzene	1.70	1.74	1.75	3-chloroaniline	2.02	2.17	2.20
1,4-dichlorobenzene	1.43	1.38	1.47	3-chloronitrobenzene	1.99	2.14	2.17
2,3,4,5-tetrachloroaniline	0.19	0.17	0.30	3-chlorophenol	1.70	1.82	1.87
2,3,4-trichloroaniline	0.85	0.81	0.88	3-chlorotoluene	2.16	1.98	1.93
2,3-dichloronitrobenzene	1.34	1.26	1.29	3-ethylaniline	2.35	2.49	2.51
2,3-dimethylnitrobenzene	1.61	1.67	1.80	3-methylaniline	2.53	2.37	2.32
2,4,5-trichloroaniline	1.00	0.97	0.85	3-nitroaniline	2.57	2.5	2.44
2,4,5-trichlorotoluene	0.94	1.09	1.10	3-nitrotoluene	2.34	2.37	2.36
2,4-dichloroaniline	1.59	1.53	1.51	4-chloroaniline	2.31	2.06	1.95
2,4-dichloronitrobenzene	1.54	1.49	1.45	4-chloro-2-nitrotoluene	1.56	1.61	1.66
2,4-dichlorophenol	1.41	1.24	1.20	4-chloronitrobenzene	1.58	1.80	1.93
2,4-dichlorotoluene	1.46	1.47	1.48	4-chlorotoluene	1.67	1.72	1.78
2,5-dichloroaniline	1.01	1.20	1.36	4-ethylaniline	2.38	2.43	2.37
2,5-dichloronitrobenzene	1.41	1.36	1.35	4-methylaniline	2.00	2.41	2.54
2-chloroaniline	1.69	2.09	2.25	4-nitroaniline	2.59	2.48	2.35
2-chloro-6-nitrotoluene	1.48	1.43	1.49	4-nitrotoluene	2.43	2.14	1.99
2-chloronitrobenzene	2.28	2.07	1.96	Aniline	3.13	3.03	2.89
2-chlorophenol	1.94	1.86	1.82	Benzene	2.91	2.98	2.62
2-ethylaniline	2.79	2.61	2.34	Monochlorobenzene	2.23	2.27	2.28
2-methylaniline	2.88	2.55	2.31	Nitrobenzene	2.70	2.79	2.76
2-nitroaniline	1.85	2.08	2.27	Phenol	2.50	2.55	2.47
2-nitrotoluene	2.38	2.33	2.28	Toluene	2.87	2.74	2.64

Notes: Obs. observed log $1/LC_{50}$ to guppy of benzene derivatives; HQSAR _ Cal. calculated log $1/LC_{50}$ to guppy of benzene derivatives by no cross-validated analysis; HQSAR _ Pred. predicted log $1/LC_{50}$ to guppy of benzene derivatives by cross-validation analysis

As a result, the model with a fragment size of 4 – 5 resulted from a hologram length of 199 with 4 components was the best model. HQSAR method yielded high r^2 value(0.951) and reasonably high cross validation value q^2(0.878) for the fitting of all 56 compounds. A plot of calculated vs. observed toxicity (a) and cross-validation predicted vs. observed toxicity; (b) by the final model is shown in Fig.1. Fragment size parameters control the minimum and maximum length of fragments to be included in the hologram fingerprint. As mentioned previously, molecular holograms are formed by the generation of all linear, branched, overlapping fragments between M and N atoms in size. The parameters M and N can be changed to include smaller or larger fragments in the holograms. Perhaps default fragment lengths of $M = 4$ and $N = 7$ can not produce the optimal HQSAR model. The HQSAR results for several different sizes indicated either q^2 or r^2 is sensitive to fragment size, and changing the default parameters such as fragment size, molecular hologram length, and

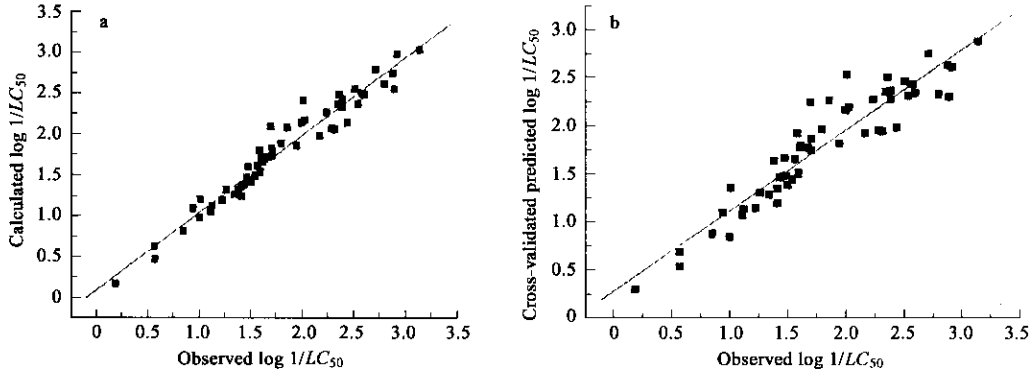

Fig.1 Scatter plot of calculated values (a) and cross-validation predicted values (b) from HQSAR analysis versus observed log $1/LC_{50}$ value of 56 benzene derivatives to guppy

fragment distinction factors will improve the quality of generated QSAR models based on molecular hologram. The inclusion of donor and acceptor atoms (D) for fragment distinction had slight effects on the quality of the HQSAR models, while inclusion of hydrogen atom (H) even lowered the quality of HQSAR model.

Table 2 Result of HQSAR for toxicity to guppy of 56 benzene derivatives

Fragment size	Hologram length	r^2	$S.E.$	q^2	$S.E._p$	Components	Fragment distinction
1—1	257	0.861	0.247	0.848	0.258	1	A,B,C
1—3	97	0.866	0.242	0.847	0.259	1	A,B,C
3—5	199	0.930	0.180	0.851	0.263	4	A,B,C
4—7	199	0.955	0.144	0.836	0.276	4	A,B,C
3—10	151	0.961	0.136	0.802	0.306	5	A,B,C
2—5	199	0.931	0.179	0.859	0.255	4	A,B,C
3—5	71	0.953	0.151	0.866	0.254	4	A,B,C
4—5	199	0.932	0.176	0.868	0.246	3	A,B,C
4—7	199	0.955	0.144	0.836	0.276	4	A,B,C
5—5	71	0.952	0.151	0.851	0.265	5	A,B,C
3—5	199	0.945	0.160	0.865	0.250	4	A,B,C,D
4—5	199	0.951	0.151	0.878	0.238	4	A,B,C,D
4—5	199	0.953	0.149	0.860	0.257	5	A,B,C,D,H

Notes: r^2. square of correlation coefficient; $S.E.$ standard error of estimation; q^2. leave-one-out cross-validated r^2; $S.E._p$. leave-one-out cross-validated standard error of prediction; components. the number of optimal components to derive the HQSAR model; A. atomic numbers; B. bond types; C. atomic connections; D. donor and acceptor atoms; H. inclusion of hydrogen atoms

2.2 Predictions for test set

Perhaps a more convincing test is to use the HQSAR models to predict the values of a biological activity for an entirely new set of compounds, and to examine how well these predictions compare with experimental values. As an effort to examine the prediction of the HQSAR models for acute aquatic toxicity to guppy based on the molecular holograms and PLS technique, 16 compounds were excluded randomly from data set and acted as a testing set. These 16 compounds are shown in Table 3. The data set containing the rest 40 compounds was employed as the training set. HQSAR_PLS analyses were re-performed to derive a HQSAR model based on the training set. The derived HQSAR models were then used to predict the values of the chemicals of the testing sets. It can be inspected that the model quality of the HQSAR model based on the training set was almost identical to that of the HQSAR model based on the whole data set for acute toxicity to the guppy investigated. The model with a fragment size of 4 – 5 resulted from a hologram length of 199 with 4 components was developed. The cross-validated q^2 was 0.865 and the standard error was 0.236. The final (non-cross-validated) r^2 was 0.940 and the standard error was 0.201. The results predicted from HQSAR model for the test compounds are shown in Table 3. A comparison of the predicted values from the training set HQSAR model with the experimental acute toxicity to the guppy of the testing set shows that they are very close. Statistical characteristics of the model are the following: $n = 16$, $r^2 = 0.849$, $S.E. = 0.225$, $F = 78.181$, $p < 0.0001$. The differences between

Table 3 Observed toxicity to guppy and prediction from HQSAR model for the testing set

Chemicals	Log $1/LC_{50}$, mmol			Chemicals	log $1/LC_{50}$, mmol		
	HQSAR				HQSAR		
	Obser.	Pred.	Residual		Obser.	Pred.	Residual
1,2,3-trichlorobenzene	1.11	1.22	-0.11	2-methylaniline	2.88	2.41	0.47
1,2-dichlorobenzene	1.60	1.76	-0.16	3,4-dichloroaniline	1.59	1.46	0.13
1,4-dichlorobenzene	1.43	1.62	-0.19	3,5-dichloroaniline	1.38	1.82	-0.44
2,3-dichloronitrobenzene	1.34	1.37	-0.03	3-chloroaniline	2.02	2.24	-0.22
2,4,5-trichlorotoluene	0.94	0.82	0.12	3-chlorotoluene	2.16	2.07	0.09
2,4-dichlorophenol	1.41	1.19	0.22	3-nitroaniline	2.57	2.47	0.10
2,5-dichloronitrobenzene	1.41	1.33	0.08	4-chloro-2-nitrotoluene	1.56	1.62	-0.06
2-chloronitrobenzene	2.28	2.01	0.27	4-ethylaniline	2.38	2.30	0.08

Notes: HQSAR_Pred = predicted acute toxicity (log $1/LC_{50}$) to guppy of benzene derivatives by HQSAR model; Residual = Observ. – Pred.

predicted values and the experimental acute toxicity to the guppy are listed in Table 3. The residuals of 16 test compounds are very low. The results confirmed the excellent prediction and the robustness of the QSAR models derived from molecular holograms and PLS analysis.

3 Conclusions

In this paper, the newly developed QSAR method based on the molecular hologram was employed to predict acute toxicity of benzene derivatives to guppy. The results showed this new HQSAR approach present highly predictive models for aquatic toxicity of pollutants. The predicted acute aquatic toxicity to the guppy (*poecilia reticulata*) for benzene derivatives is very close to the experimental values. Furthermore, based on molecular hologram, alignment-free QSAR models could be rapidly and easily developed with highly statistical significance and predictive ability, so HQSAR technique provides promising tool for the screening and prediction of large datasets of contaminants or pollutants.

References:

Agarwal A, Taylor E W, 1993. 3-D QSAR for intrinsic activity of 5-HT_{1A} receptor ligands by the method of comparative molecular field analysis[J]. J Computat Chem, 14: 237—245.

Ash S, Cline M A, Homer R W *et al*., 1997. SYBYL line notation(SLN): a versatile language for chemical structure representation [J]. J Chem Inf Comput Sci, 37: 71—79.

Blum D J W, Speece R E, 1990. Determining chemical toxicity to aquatic species [J]. Environ Sci Technol, 24: 284—293.

Briens F, Bureau R, Rault S *et al*., 1995. Applicability of CoMFA in ecotoxicology: a critical study on chlorophenols [J]. Ecotoxicol Environ Safety, 31: 37—48.

Cha M Y, Lee I Y, Cha J H *et al*., 2003. QSAR studies on piperazinylalkylisoxazole analogues selectively acting on dopamine D_3 receptor by HQSAR and CoMFA[J]. Bioorg Med Chem, 11: 1293—1298.

Debnath A K, 1999. Three-dimension quantitative structure-activity relationship study on cyclic urea derivatives as HIV-1 protease inhibitors: application of comparative molecular field analysis[J]. J Med Chem, 42: 249—259.

Hasegawa K, Arakawa M, Funatsu K, 1999. Rational choice of bioactive conformations through use of conformation analysis and 3-way partial least squares modeling[J]. Chemometrics and Intelligent Laboratory System, 50: 253—261.

Kurup A, Garg R, Carini D J *et al*., 2001. Comparative QSAR: angiotensin II antagonists[J]. Chem Rev, 101: 2727—2750.

Lipnick R L, 1989. Base-line toxicity predicted by quantitative structure-activity relationships as a probe for molecular mechanism of toxicity[J]. Environ Sci Technol, 24: 284—293.

Lowis D, 1997. HQSAR: A new, highly predictive QSAR technique[M]. In: Tripos technical notes. Vol 1, Number 5. HQSAR. Tripos Inc.

Mckim J M, Bradbury S P, Niemi G J, 1987. Fish acute toxicity syndromes and their use in the QSAR approach to hazard assessment[J]. Environ Health Persp, 71:171—186.

Michael S, David B T, Peter W, 1999. Effect of parameter variations on the effectiveness of HQSAR analyses[J]. Quant Struct-Act Relat, 18: 245—252.

Park Choo H Y, Lim J S, Kam Y *et al*., 2001. A comparative study of quantitative structure activity relationship methods based on antitumor diarylsulfonylureas[J]. Eur J Med Chem, 36: 829—836.

Ren S, Schultz S R, 2002. Identifying the mechanism of aquatic toxicity of selected compounds by hydrophobicity and electrophilicity descriptors[J]. Toxicology Letters, 129: 151—160.

Rodrigues C R, Flaherty T M, Springer C *et al*., 2002. CoMFA and HQSAR of acylhydrazide cruzain inhibitors[J]. Bioorg Med Chem Lett, 12: 1537—1541.

SYBYL 6.7 user guide, 1999. Tripos Inc. St. Louis[Z]. USA.

Tong W, Lowis D R, Perkins R *et al*., 1998. Evaluation of quantitative structure-activity relationship methods for large-scale prediction of chemicals binding to the estrogen descriptor[J]. J Chem Inf Comput Sci, 38: 669—677.

Verhaar H J M, Van Leeuwen C J, Hermens J L M, 1992. Classifying environmental pollutants: Part 1. Structure-activity relationships for prediction of aquatic toxicity[J]. Chemosphere, 25: 471—491.

(Received for review April 25, 2003. Accepted July 14, 2003)