Simple Method of Calculating Octanol/Water Partition Coefficient

Ikuo MORIGUCHI,* Shuichi HIRONO, Qian LIU, Izumi NAKAGOME, and Yasuo MATSUSHITA

School of Pharmaceutical Sciences, Kitasato University, 5-9-1 Shirokane, Minato-ku, Tokyo 108, Japan. *Received July 23, 1991*

A simple method of calculating log *P* (partition coefficient in octanol/water) has been developed based on the quantitative structure-log *P* relationship for 1230 organic molecules having a wide variety of structures. The 1230 organic compounds investigated included general aliphatic, aromatic, and heterocyclic molecules together with various drugs and agrochemicals. The predictive structure–log *P* model obtained by multiple regression analysis involved only 13 parameters for hydrophobic atoms, hydrophilic atoms, their proximity effects, unsaturated bonds, amphoteric property, and several specific functionalities. A saturation effect was recognized in the parameters for hydrophobic and hydrophilic atoms, and unsaturated bonds. The structure–log *P* relationship was highly significant as the *F*-statistics = 900.4. This simple method appears accurate enough for semiquantitative uses in structure–activity rating studies, especially for quantitative structure–activity relationship in toxicity.

Keywords partition coefficient; octanol/water partition; hydrophobicity; multiple regression analysis; quantitative structure–activity relationship; predictive model

Introduction

Many diverse biochemical, pharmacological, and toxicological processes involved in drug action are known to be dependent on the hydrophobic property of drug molecules. Parametrization of the hydrophobicity is one of the important aspects in quantitative structure–activity relationship (QSAR) studies. As a parameter for the hydrophobicity, Hansch and Fujita[1] successfully introduced the logarithm of partition coefficient between octanol and water, log *P*, to regression analysis of biological activities to establish QSAR. Since then, there has been ever-increasing need for prediction of log *P* for various structures, especially those for which experimental values are not available.

Hansch *et al.*[2,3] and Rekker and his colleague[4,5] empirically calculated log *P* using some fragment constants and correction terms. Fully computerized systems such as CLOGP[6] based on the empirical method of Hansch *et al.* are in widespread use. The computerized empirical methods work well for a number of compounds; however, difficulties have sometimes arisen in decomposing the structure into appropriate fragments whose constants are available.

For compounds having simple structures, more sophisticated methods of estimating log *P* were proposed by Rogers and Cammarata,[7] Hopfinger and Battershell,[8] Klopman and Iroff,[9] Iwase *et al.*,[10] Kasai *et al.*,[11] and Sasaki *et al.*[12] Although these methods may be theoretically interesting, they do not seem practically applicable to complex structures of drugs and agrochemicals.

Recently, QSAR's in toxicity for large sets of data have been studied and their use attempted by regulatory agencies and industry to screen compounds for possible health and environmental hazards. For this purpose, a simple method of calculating log *P* for any type of molecule is strongly desired. Since such toxicity data are generally collected from many different sources, observed potencies are usually classified into several ratings and treated semiquantitatively.

In this study, we have attempted to develop a simple method of approximating log *P* for organic molecules of diverse and complex structures. The method is based on the structure–log *P* relationship obtained from the multiple regression analysis (MRA) of 1230 organic molecules including general aliphatic, aromatic, and heterocyclic compounds together with complex drugs and agrochemicals. The predictive structure–log *P* model involves only 13 structural parameters, and appears accurate enough for semiquantitative use in structure–activity studies.

Methods

Compounds and log *P* Data for MRA The 1230 compounds used for the structure–log *P* relationship analysis have diverse structures including C, H, N, O, S, P, F, Cl, Br, and/or I atoms listed in Table I. Their observed log *P* values were cited from the literature.[3]

Structural Parameters for Multiple Regression Models Parameters for hydrophobic atoms, hydrophilic atoms, their proximity effects, unsaturated bonds, intramolecular hydrogen bonds, ring structures, amphoteric property, and several specific functionalities were used and are listed in Table II. For three of these parameters, *CX*, *NO*, and *UB*, their saturation effects were investigated using nonlinear forms, $(CX)^a$, $(NO)^a$, and $(UB)^a$ $(0.5 \leq a < 1.0)$ as well as the original parameters. *CX* is the summation of weighted numbers of carbon and halogen atoms, and the weight values were taken to be simple but approximately proportional to the van der Waals volume of atomic spheres, since the van der Waals volume is well correlated with hydrophobicity for apolar structures.[13] For *POL*, the upper value was limited to 4.0, since, with this limitation, the contribution to the regression with log *P* was found to be best. Other values for estimation of parameters such as those of *PRX*, *HB*, *AMP*, *QN*, and *NCS* were also empirically evaluated.

Calculation All calculations were carried out on a Sony NWS-830 computer and a Kobe Steel KTR-BO8 transputer attached to an Epson PC-286VF microcomputer using a self-written MRA program.

Results and Discussion

Multiple Regression Studies Using 1230 log *P* Data It is generally thought that log *P* of a molecule can be estimated from the contribution of its hydrophobic and hydrophil-

TABLE I. Composition of Compounds for Multiple Regression Analysis

Atom	Number	Compound with the max. number of the atom
C	1—24	Triamcinolone acetonide
H	0—34	Prostaglandin E-1
N	0— 5	2,4-Diamino-6-dimethylaminopyrimidine-3-oxide
O	0— 7	Sulbenicillin
S	0— 2	Sulbenicillin, *etc.*
P	0— 1	Parathion, *etc.*
F	0— 7	2,2,3,3,4,4,4-Heptafluorobutanol
Cl	0— 6	γBHC, *etc.*
Br	0— 3	2-(2,4,6-Tribromophenoxy)ethanol
I	0— 1	5-Iodo-uracil, *etc.*

TABLE II. Parameters Used

Parameter	Type[a]	Description
CX	N	Summation of numbers of carbon and halogen atoms weighted by C: 1.0, F: 0.5, Cl: 1.0, Br: 1.5, and I: 2.0
NO	N	Total number of N and O atoms
PRX	N	Proximity effect of N/O; X–Y : 2.0, X–A-Y : 1.0 (X, Y : N/O, A: C, S, or P) with a correction (−1) for carboxamide/sulfonamide
UB	N	Total number of unsaturated bonds except those in NO_2
HB	D	Dummy variable for the presence of intramolecular hydrogen bond as *ortho*–OH and –CO–R, –OH and $-NH_2$, $-NH_2$ and –COOH, or 8-OH/NH_2 in quinolines, 5 or 8-OH/NH_2 in quinoxalines, *etc.*
POL	N	Number of aromatic polar substituents (aromatic substituents excluding Ar–CX_2– and Ar–CX= C<, X: C or H)
AMP	N	Amphoteric property; α-aminoacid: 1.0, aminobenzoic acid: 0.5, pyridinecarboxylic acid: 0.5
ALK	D	Dummy variable for alkane, alkene, cycloalkane, or cycloalkene (hydrocarbons with 0 or 1 double bond)
RNG	D	Dummy variable for the presence of ring structures except benzene and its condenced rings (aromatic, heteroaromatic, and hydrocarbon rings)
QN	N	Quaternary nitrogen: $>\overset{+}{N}<$, 1.0; N oxide, 0.5
NO2	N	Number of nitro groups
NCS	N	Isothiocyanato (–N=C=S), 1.0; thiocyanato (–S–C≡N), 0.5
BLM	D	Dummy variable for the presence of β-lactam

a) N, numerical variable; D, dummy variable.

ic substructures. As basic parameters, we used *CX*, the summation of empirically weighted numbers of carbon and halogen atoms for primary contribution of hydrophobic atoms, and *NO*, the total number of nitrogen and oxygen atoms for primary contribution of hydrophilic atoms, for MRA of 1230 log *P* data. The resultant two-parameter equation is shown as Eq. 1,

$$\log P = 0.246CX - 0.386NO + 0.466 \quad (1)$$
$$(t=32.0)\quad (t=25.6)$$

$$n=1230,\quad r=0.730,\quad s=0.912,\quad F_0(2,1227)=700.3$$

where t = *t*-statistics for the coefficient, n = number of compounds, s = standard deviation of the estimation error, and F_0 = *F*-statistics for the correlation. The equation, showing a positive contribution of *CX* and a negative contribution of *NO* to log *P*, is entirely consistent with the general image of the log *P* model.

The contributions of *CX* and *NO* to log *P* were considered not simply linear but, instead, there seemed to be a saturation with greater values of *CX* and *NO*. So, the saturation effect was investigated using nonlinear forms of the parameters, $(CX)^a$ and $(NO)^a$ $(0.5 \leq a < 1.0)$, and the following equation with an improved correlation (F_0 = 810.7) was derived.

$$\log P = 1.001(CX)^{0.6} - 0.479(NO)^{0.9} + 0.754 \quad (2)$$
$$(t=34.5)\quad (t=27.2)$$

$$n=1230,\quad r=0.754,\quad s=0.876,\quad F_0(2,1227)=810.7$$

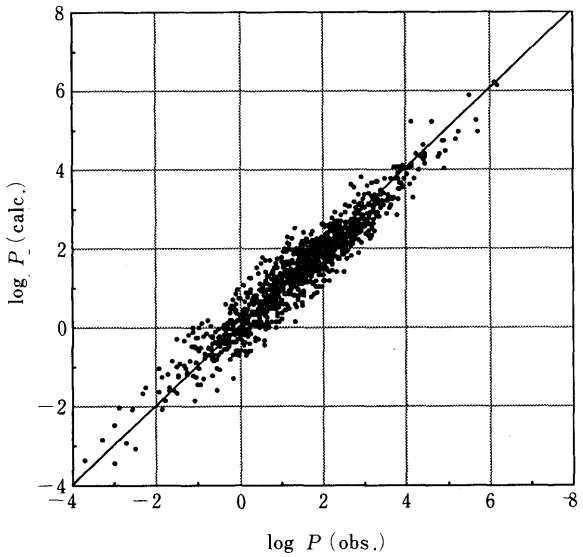


Fig. 1. Correlation between Observed log *P* and Calculated log *P* from Eq. 4 for 1230 Compounds

The proximity effect of nitrogen and/or oxygen atoms was also considered important as a correction for the electronic structure. Incorporating an empirical parameter for proximity, *PRX*, into Eq. 2 resulted in a remarkable improvement in the correlations as shown in Eq. 3.

$$\log P = 1.241(CX)^{0.6} - 1.071(NO)^{0.9} + 0.463PRX - 1.155 \quad (3)$$
$$(t=49.6)\quad (t=40.1)\quad (t=26.1)$$

$$n=1230,\quad r=0.850,\quad s=0.703,\quad F_0(3,1226)=1067.6$$

In this equation, *F*-statistic is very high, but the *s* value is not low enough for practical use. The effects of various substructures of molecules were therefore further investigated for addition to Eq. 3 using parameters such as those listed in Table II. Finally, we obtained the following equation with 13 simple parameters using MRA for the entire set of 1230 compounds.

$$\log P = 1.244(CX)^{0.6} - 1.017(NO)^{0.9} + 0.406PRX \quad (4)$$
$$(t=60.5)\quad (t=58.5)\quad (t=33.8)$$
$$-0.145(UB)^{0.8} + 0.511HB + 0.268POL - 2.215AMP$$
$$(t=9.5)\quad (t=5.9)\quad (t=19.6)\quad (t=19.5)$$
$$+0.912ALK - 0.392RNG - 3.684QN + 0.474NO2$$
$$(t=9.5)\quad (t=13.1)\quad (t=22.1)\quad (t=10.8)$$
$$+1.582NCS + 0.773BLM - 1.041$$
$$(t=16.4)\quad (t=5.0)$$

$$n=1230,\quad r=0.952,\quad s=0.411,\quad F_0(13,1216)=900.4$$

The squared correlation matrix for the parameters included in Eq. 4 is listed in Table III. There seemed tobe no possibility of chance correlation from this matrix and the *t*-values for regression coefficients given in Eq. 4.

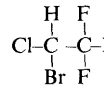
The relation of log *P* values calculated using Eq. 4 and the corresponding experimental values is drawn in Fig. 1. This shows a good fit, in spite of a large number of diverse molecules and a small number of straightforward parameters. In Eq. 4, the values of the *t*-statistics indicate that the parameters for hydrophobic and hydrophilic atoms, $(CX)^{0.6}$ and $(NO)^{0.9}$, provide dominant contributions as

TABLE III. Squared Cross-Correlation Matrix of Parameters Used in Eq. 4

	$(CX)^{0.6}$	$(NO)^{0.9}$	PRX	$(UB)^{0.8}$	HB	POL	AMP	ALK	RNG	QN	$NO2$	NCS	BLM
$(CX)^{0.6}$	1.00												
$(NO)^{0.9}$	0.20	1.00											
PRX	−0.04	0.82	1.00										
$(UB)^{0.8}$	0.71	0.37	0.19	1.00									
HB	0.07	0.11	−0.05	0.11	1.00								
POL	0.36	0.36	0.25	0.51	0.16	1.00							
AMP	−0.03	0.08	0.04	0.00	0.09	0.05	1.00						
ALK	−0.15	−0.20	−0.10	−0.21	−0.02	−0.12	−0.02	1.00					
RNG	0.07	0.26	0.17	0.04	−0.05	−0.09	−0.03	−0.02	1.00				
QN	−0.03	0.02	0.05	0.00	−0.01	0.00	−0.01	−0.01	0.09	1.00			
$NO2$	−0.05	0.42	0.58	0.04	−0.04	0.21	−0.03	−0.04	−0.12	−0.01	1.00		
NCS	0.06	−0.06	−0.07	0.16	−0.02	0.03	−0.02	−0.02	−0.07	−0.01	−0.01	1.00	
BLM	0.41	0.24	0.05	−0.06	0.05	−0.14	0.00	−0.01	0.26	0.00	−0.09	−0.05	1.00

$t=60.5$ and 58.5, respectively. In this and other respects, Eq. 4 seems a reasonable model for structure–log P relationship. Further, the value of s indicates that the estimation of log P using this simple equation is accurate enough for semiquantitative uses in structure–activity rating studies.

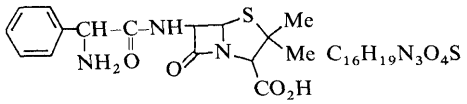
Four examples are shown below to illustrate the calculation of log P using this method. For comparison, log P values calculated by CLOGP are also listed. The first two are examples giving results of reasonable accuracy. Examples 3 and 4 give rather poor accuracy, possibly owing to electronic and topological properties peculiar to some lactone and fused ring structures. Our simple method does not cope effectively with such special cases, however, this shortcoming appears common to CLOGP.

Ex. 1. Halothane  $C_2HBrClF_3$

$CX=1.0\times2$ (for C_2) $+0.5\times3$ (for F_3) $+1.0\times1$ (for Cl) $+1.5\times1$ (for Br) $=6.0$

$\log P=1.244\times(6.0)^{0.6}-1.041=2.60$

measured $=2.30$; Calcd (CLOGP) $=2.45$[14]

Ex. 2. Ampicillin  $C_{16}H_{19}N_3O_4S$

$CX=1.0\times16$ (for C_{16}) $=16.0$

$NO=7.0$ (for N_3O_4)

$PRX=1.0$ (for –CO–NH–) $+1.0$ (for –CO–N<) $+2.0$ (for –CO–OH) $=4.0$

$UB=6.0$ (for 6 double bonds)

$RNG=1.0$ (for ring)

$BLM=1.0$ (for β-lactam)

$\log P=1.244\times(16.0)^{0.6}-1.017\times(7.0)^{0.9}+0.406\times4.0$
$-0.145\times(6.0)^{0.8}-0.392\times1.0+0.773\times1.0-1.041$
$=1.06$

measured $=1.35$; Calcd (CLOGP) $=1.00$[14]

Ex. 3. δ-Valerolactone  $C_5H_8O_2$

$CX=5.0$ (for C_5)

$NO=2.0$ (for O_2)

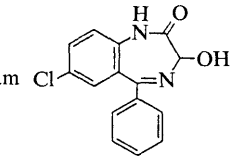
$PRX=2.0$ (for –CO–O–)

$UB=1.0$ (for a double bond)

$RNG=1.0$ (for ring)

$\log P=1.244\times(5.0)^{0.6}-1.017\times(2.0)^{0.9}+0.406\times2.0$
$-0.145\times(1.0)^{0.8}-0.392\times1.0-1.041$
$=0.60$

measured $=-0.35$; Calcd (CLOGP) $=0.66$[14]

EX. 4. Oxazepam  $C_{15}H_{11}ClN_2O_2$

$CX=1.0\times15$ (for C_{15}) $+1.0\times1$ (for Cl) $=16.0$

$NO=4.0$ (for N_2O_2)

$PRX=1.0$ (for –NH–CO–) $+2.0$ (for =N–C(–)–OH) $=3.0$

$UB=8.0$ (for 8 double bonds)

$POL=4.0$ (for Ph–Cl, Ph–NH–, $2\times$Ph–C(–)=N–)

$RNG=1.0$ (for ring)

$\log P=1.244\times(16.0)^{0.6}-1.017\times(4.0)^{0.9}+0.406\times3.0$
$-0.145\times(8.0)^{0.8}+0.268\times4.0-0.392\times1.0-1.041$
$=3.12$

measured $=2.25$; Calcd (CLOGP) $=3.33$[14]

Comparison of the Present Method with Other Studies In a similar study with less diverse structures, Klopman *et al.*[15] reported simple correlation models for calculating log P. They studied a set of 195 general organic molecules including C, H, N, O, and/or Cl atoms. The fit for the 195 compounds using seven or nine parameters was almost as good, with a correlation coefficient of $r=0.962$ and 0.974, respectively. The seven descriptors were the number of carbon, hydrogen, nitrogen, oxygen, and chlorine atoms and the number of acid/ester and nitro functional groups. The additional two descriptors were the number of methylene or methyl substituents attached to a phenyl ring and a descriptor to indicate the aliphatic hydrocarbons from the rest.

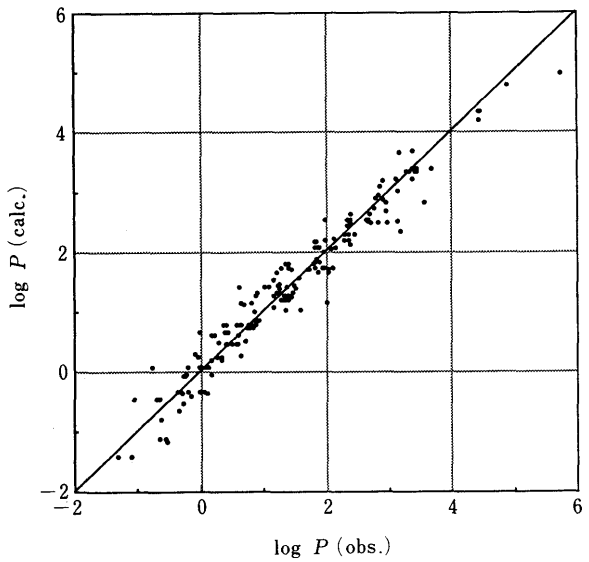


Fig. 2. Correlation between Observed log P and Calculated log P from Eq. 5 for 195 Compounds

For comparison, we investigated the structure–log P relationship for the same set of compounds using the 13 parameters appearing in Eq. 4 as candidate descriptors. The resultant equation was as follows:

$$\log P = 1.464(CX)^{0.6} - 1.221(NO)^{0.9} \times 0.653PRX \quad (5)$$
$$(t=31.6) \qquad (t=27.1) \qquad (t=21.9)$$
$$-0.300(UB)^{0.8} + 0.335POL + 0.726ALK$$
$$(t=10.0) \qquad (t=9.4) \qquad (t=5.8)$$
$$-0.269RNG - 1.358$$
$$(t=3.8)$$
$$n=195, \quad r=0.975, \quad s=0.290, \quad F_0(7,187)=512.2$$

Seven parameters sufficed to describe the relationship, since molecular structures of the 195 compounds were rather simple compared with those of the 1230 molecules. The high correlation is shown in Fig. 2, indicating a good fit.

In conclusion, our new procedure gives comparatively better results in the estimation of log P for diverse structures. The method is very simple and applicable to almost any type of organic molecules, although it is not precise enough to differentiate log P among geometrical isomers. It is hoped that the present method will be widely used for structure–activity rating studies, especially for QSAR in toxicity.

References

1) C. Hansch and T. Fujita, *J. Am. Chem. Soc.*, **86**, 1616 (1964).
2) A. Leo, C. Hansch, and D. Elkins, *Chem. Rev.*, **71**, 525 (1971).
3) C. Hansch and A. Leo, "Substituent Constants for Correlation Analysis in Chemistry and Biology," John Wiley and Sons, New York, 1979.
4) G. G. Nys and R. F. Rekker, *Chim. Therap.*, **8**, 521 (1973).
5) R. F. Rekker, "The Hydrophobic Fragmental Constant," Elsevier, Amsterdam, 1977.
6) MEDCHEM Software, Daylight Chemical Information Systems, 3591, Claremont St., Irvine, CA 92714, U.S.A.
7) K. S. Rogers and A. Cammarata, *Biochim. Biophys. Acta*, **193**, 22 (1969).
8) A. J. Hopfinger and R. D. Battershell, *J. Med. Chem.*, **19**, 569 (1976).
9) G. Klopman and L. Iroff, *J. Comput. Chem.*, **2**, 157 (1981).
10) K. Iwase, K. Komatsu, S. Hirono, S. Nakagawa, and I. Moriguchi, *Chem. Pharm. Bull.*, **33**, 2114 (1985).
11) K. Kasai, H. Umeyama, and A. Tomonaga, *Bull. Chem. Soc. Jpn.*, **61**, 2701 (1988).
12) Y. Sasaki, H. Kubodera, T. Matsuzaki, and H. Umeyama, *J. Pharmacobio-Dyn.*, **14**, 207 (1991).
13) I. Moriguchi, Y. Kanada, and K. Komatsu, *Chem. Pharm. Bull.*, **24**, 1799 (1976).
14) A. J. Leo, "Comprehensive Medicinal Chemistry," Vol. 4, ed. by C. Hansch, Pergamon Press, 1990, pp. 295—319.
15) G. Klopman, K. Namboodiri, and M. Schochet, *J. Comput. Chem.*, **6**, 28 (1985).