Perspectives in Drug Discovery and Design, **19:** 47–66, 2000.
KLUWER/ESCOM

Calculating partition coefficient by atom-additive method

RENXIAO WANG, YING GAO and LUHUA LAI*
Institute of Physical Chemistry, Peking University, Beijing 100871, People's Republic of China

Received 15 August 1999; Accepted 12 September 1999

Summary. A new atom-additive method is presented for calculating octanol/water partition coefficient (log *P*) of organic compounds. The method, XLOGP v2.0, gives log *P* values by summing the contributions of component atoms and correction factors. Altogether 90 atom types are used to classify carbon, nitrogen, oxygen, sulfur, phosphorus and halogen atoms, and 10 correction factors are used for some special substructures. The contributions of each atom type and correction factor are derived by multivariate regression analysis of 1853 organic compounds with known experimental log *P* values. The correlation coefficient (*r*) for fitting the whole set is 0.973 and the standard deviation (*s*) is 0.349 log units. Comparison of various log *P* calculation procedures demonstrates that our method gives much better results than other atom-additive approaches and is at least comparable to fragmental approaches. Because of the simple methodology, the 'missing fragment' problem does not occur in our method.

Key words: atom-additive, atom type, correction factor, partition coefficient

Introduction

The logarithm of the partition coefficient between *n*-octanol and water (log *P*) is often used to represent molecular hydrophobicity. Since the pioneering work of Hansch and Fujita [1], log *P* has become a valuable parameter in many quantitative structure–activity relationship (QSAR) approaches that have been developed for pharmaceutical, environmental, biochemical, and toxicological sciences. Studies have shown that log *P* is useful for correlating various properties of drug molecules, such as the transport process, ligand–receptor interaction, biological and toxic effects [2,3]. Therefore, accessibility to accurate log *P* values for compounds of interest may be essential

* To whom correspondence should be addressed. E-mail: lai@mdl.ipc.pku.edu.cn. *Supplementary information available from the authors:* the training set (including an index and 1853 organic compounds represented in Mol2 format), the test set (including an index and 138 organic compounds represented in Mol2 format), and the source codes of program XLOGP v2.0.

for the correct prediction of their biological properties. Although log *P* can be measured reliably for a given compound, the experimental process might be time-consuming and expensive. This problem becomes critical when a large number of candidate molecules, which sometimes are just virtual, require screening during a drug design and discovery procedure. Thus, there is a clear need for calculation procedures that can give reliable estimations of log *P* based merely on the chemical structure of a given compound.

During the past three decades, many methods of calculating log *P* have been reported in the literature [4]. At present, the most widely accepted method is classified as the 'additive method', where a molecule is dissected into basic fragments (functional groups or atoms) and its log *P* value is obtained by summing the contributions of each fragment. 'Correction factors' are also introduced to rectify the calculated log *P* value when some special substructures occur in the molecule. This method originated with Rekker and co-workers [5,6]. Current popular fragment-additive methods include CLOGP [7,8]), KLOGP [9], KOWWIN [10], CHEMICALC-2 [11], etc. Atom-additive methods include MOLCAD [12], ALOGP [13], and SMILOGP [14]. There are also methods that try to incorporate molecular properties into the calculation, such as HINT [15] and ASCLOGP [16]. A more detailed description of these methods is beyond the scope of this paper.

We have recently developed a new atom-additive method, XLOGP, for log *P* calculation [17]. In that approach, we used 80 basic atom types to classify carbon, nitrogen, oxygen, sulfur, phosphorus, and halogen atoms. In addition, we also introduced five correction factors to account for some intramolecular interactions, such as internal hydrogen bonding. The final model was obtained by multivariate regression analysis of 1831 organic compounds. The correlation coefficient for fitting the whole set (*r*) was 0.968 and the standard deviation (*s*) was 0.37 log units. XLOGP gives much better results than other atom-additive methods. There are a number of users of XLOGP all over the world.

In this paper, we will describe our improvements to the XLOGP method. First, we have enlarged the training set to include more phosphorus-containing compounds. Second, we have adopted a new scheme for atom classification. Third, we have redefined the correction factors, which efficiently improves the accuracy of log *P* estimation. The new model, XLOGP v2.0, yields better regression results than the previous one while still retaining the simple methodology.

Methods

Training set

Constructing a reliable training set is crucial for developing an empirical model for log *P* calculation. We have inherited the previous data set [17]. It contains 1831 diverse organic compounds. Since there were only four phosphorus-containing compounds in that set, we have added 22 additional phosphorus-containing compounds. Note that we do not include metalloorganic, organosilicon or organic salts in the training set. The experimental log *P* values of all 1853 compounds are from Hansch and Leo's compilation [18].

Molecular models of all 1853 compounds are constructed with SYBYL [19] on an SGI O2 workstation and minimized using the Tripos force field. Atomic partial charges are calculated using the MNDO method. The models are then saved in Mol2 format for further analysis. All these molecular models as well as their log *P* values are available in the supplementary information.

Atom classification

We use 90 atom types to classify carbon, nitrogen, oxygen, sulfur, phosphorus, and halogen atoms in neutral organic compounds (Table 1). The classification scheme differentiates atoms according to (i) element, (ii) hybridization state, (iii) accessibility to the solvent (represented by the number of attached hydrogen atoms), (iv) nature of the neighboring atoms, and (v) adjacency to π-systems. Thus, atoms belonging to the same atom type generally have similar solvent accessible surfaces and charge densities. This establishes the rough theoretical support for the assumption that a certain type of atom has a specific contribution to the partition coefficient.

Note that we do not use an additional atom type for hydrogen atoms because they are included in our atom classification scheme implicitly. In other words, the atom types in our method are 'united' atoms that already include the attached hydrogen atoms.

Correction factors

The whole is often more than the sum of its parts. It has become apparent in log *P* calculations that, for various compounds, the log *P* values obtained by summing the atom/fragment contributions alone deviate significantly from the experimental values. This is usually explained by the intramolecular group–group interactions. The term 'correction factor' is appropriate because they are derived from analyzing the differences between the calculated and the experimentally measured log *P* values.

We use 10 correction factors (Table 2). Since the atom types defined in our method take the neighboring atoms into account, these correction factors aim

Table 1. Atom types used in XLOGP v2.0

No.	Description[a]	HB[b]	Compound	Occurrence	Contribution
sp^3 carbon in					
1	CH_3R (π=0)	N	458	747	0.528
2	CH_3R (π=1)	N	324	413	0.267
3	CH_3X	N	327	461	–0.032
4	CH_2R_2 (π=0)	N	354	793	0.358
5	CH_2R_2 (π=1)	N	237	299	–0.008
6	CH_2R_2 (π=2)	N	55	57	–0.185
7	$CH_2R_nX_{2-n}$(π=0)	N	375	564	0.137
8	$CH_2R_nX_{2-n}$(π=1)	N	151	154	–0.303
9	$CH_2R_nX_{2-n}$(π=2)	N	8	8	–0.815
10	CHR_3 (π=0)	N	64	120	0.127
11	CHR_3 (π=1)	N	61	68	–0.243
12	CHR_3 ($\pi \geq 2$)	N	13	13	–0.499
13	CHR_nX_{3-n} (π=0)	N	149	297	–0.205
14	CHR_nX_{3-n} (π=1)	N	43	55	–0.305
15	CHR_nX_{3-n} ($\pi \geq 2$)	N	12	12	–0.709
16	CR_4 (π=0)	N	42	46	–0.006
17	CR_4 (π=1)	N	94	100	–0.570
18	CR_4 ($\pi \geq 2$)	N	19	19	–0.317
19	CR_nX_{4-n} (π=0)	N	18	20	–0.316
20	CR_nX_{4-n} ($\pi > 0$)	N	13	13	–0.723
sp^2 carbon in					
21	$A{=}CH_2$	N	41	52	0.420
22	$A{=}CHR$ (π=0)	N	77	106	0.466
23	$A{=}CHR$ (π=1)	N	137	191	0.136
24	$A{=}CHX$ (π=0)	N	114	124	0.001
25	$A{=}CHX$ (π=1)	N	34	34	–0.310
26	$A{=}CR_2$ (π=0)	N	37	41	0.050
27	$A{=}CR_2$ ($\pi > 0$)	N	178	228	0.013
28	$A{=}CRX$ (π=0)	N	397	448	–0.030
29	$A{=}CRX$ ($\pi > 0$)	N	307	333	–0.027
30	$A{=}CX_2$ (π=0)	N	219	222	0.005
31	$A{=}CX_2$ ($\pi > 0$)	N	30	30	–0.315

Table 1. (continued)

No.	Description[a]	HB[b]	Compound	Occurrence	Contribution
Aromatic carbon in					
32	C...C(H)...C	N	1355	6045	0.337
33	A...C(H)...N	N	203	296	0.126
34	C...C(R)...C	N	1106	1800	0.296
35	C...C(X)...C	N	925	1291	–0.151
36	A...C(R)...N	N	128	166	0.174
37	A...C(X)...N	N	60	93	0.366
sp carbon in					
38	R≡CH	N	4	4	0.209
39	A≡C–A	N	85	92	0.330
40	A=C=A	N	23	24	2.073
sp^3 nitrogen in					
41	R-NH$_2$ (π=0)	D	92	94	–0.534
42	R-NH$_2$ (π=1)	D	178	192	–0.329
43	X-NH$_2$	D	16	16	–1.082
44	R-NH-R (π=0)	D	41	42	–0.112
45	R-NH-R ($\pi>0$)	D	56	56	0.166
46	R-NH-R (ring)[c]	D	55	56	0.545
47	A-NH-X	D	40	41	0.324
48	A-NH-X (ring)	D	9	9	0.153
49	NR$_3$ (π=0)	N	46	57	0.159
50	NR$_3$ ($\pi>0$)	N	29	31	0.761
51	NR$_3$ (ring)	N	50	50	0.881
52	NR$_n$X$_{3-n}$	N	26	29	–0.239
53	NR$_n$X$_{3-n}$(ring)	N	11	11	–0.010
Amide nitrogen in					
54	–NH$_2$	D	93	99	–0.646
55	–NHR	D	283	324	–0.096
56	–NHX	D	14	14	–0.044
57	–NR$_2$	N	75	79	0.078
58	–NRX	N	24	26	–0.118

Table 1. (continued)

No.	Description[a]	HB[b]	Compound	Occurrence	Contribution
sp^2 nitrogen in					
59	C=N–R (π=0)	N	19	20	0.007
60	C=N–R (π=1)	N	128	136	–0.275
61	C=N–X (π=0)	N	56	56	0.366
62	C=N–X (π=1)	N	13	24	0.251
63	N=N–R	N	15	19	0.536
64	N=N–X	N	14	17	–0.597
65	A-NO	N	37	40	0.427
66	A-NO_2	N	138	157	1.178
Aromatic nitrogen in					
67	A...N...A[d]	N	79	84	–0.493
sp nitrogen in					
68	–C≡N	N	239	301	–0.566
sp^3 oxygen in					
69	R–OH (π=0)	D/A	199	285	–0.467
70	R–OH (π=1)	D/A	404	451	0.082
71	X–OH	D/A	11	11	–0.522
72	R–O–R (π=0)	N	113	137	0.084
73	R–O–R ($\pi>0$)	N	469	549	0.435
74	R–O–X	N	4	4	0.105
sp^2 oxygen in					
75	A=O	A	1074	1597	–0.399
sp^3 sulfur in					
76	A–SH	N	5	5	0.419
77	A–S–A	N	77	85	0.255
sp^2 sulfur in					
78	A=S	N	46	47	–0.148
Sulfoxide sulfur in					
79	A–SO–A	N	5	5	–1.375

Table 1. (continued)

No.	Description[a]	HB[b]	Compound	Occurrence	Contribution
Sulfone sulfur in					
80	A–SO_2–A	N	81	88	–0.168
Phosphorus in					
81	O=PA_3	N	20	20	–0.477
82	S=PA_3	N	9	9	1.253
Fluorine in					
83	–F (π=0)	N	90	272	0.375
84	–F (π=1)	N	50	61	0.202
Chlorine in					
85	–Cl (π=0)	N	50	124	0.512
86	–Cl (π=1)	N	190	297	0.663
Bromine in					
87	–Br (π=0)	N	21	23	0.850
88	–Br (π=1)	N	67	87	0.839
Iodine in					
89	–I (π=0)	N	4	4	1.050
90	–I (π=1)	N	36	37	1.109

[a] –: single bond; =: double bond; ≡: triple bond; ...: aromatic bond; X: any nitrogen or oxygen atom; R: any other atom; A: any atom. π represents any π-system, such as double bond, triple bond, and aromatic ring. A note of 'π=*n*' indicates that there are *n* π-systems in the neighborhood.
[b] This represents the role that this atom may play in a hydrogen bond. D: donor atom; A: acceptor atom; D/A: either donor or acceptor; N: 'none', 'nothing', or 'no way'.
[c] This atom locates in a conjugated ring, such as the nitrogen atom in a pyrrole ring.
[d] This represents the nitrogen atom in a six-membered aromatic ring, such as the nitrogen atom in a pyridine ring.

at 1–3, 1–4, or even further intramolecular interactions, as explained in detail below.

(1) *Hydrophobic carbon:* We have observed that the log *P* values of compounds with hydrocarbon chains are often underestimated by atom addition. Such compounds tend to be more flexible and easier to aggregate in the aqueous phase. We define an sp^3 or sp^2 hybridized carbon atom as a 'hydro-

Table 2. Correction factors used in XLOGP v2.0

Factor[a]	Contribution	Compound	Occurrence
Hydrophobic carbon	0.211	369	883
Internal H-bond	0.429	157	160
Halogen 1–3 pair	0.137	92	274
Aromatic nitrogen 1–4 pair	0.485	15	15
Ortho sp^3 oxygen pair	–0.268	26	26
Para donor pair	–0.423	33	34
sp^2 Oxygen 1–5 pair	0.580	64	66
α-Amino acid	–2.166	16	16
Salicylic acid	0.554	26	26
p-Amino sulfonic acid	–0.501	20	20

[a] Please refer to the text for a detailed description.

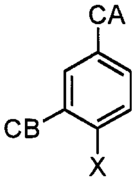

Figure 1. Illustration of hydrophobic carbons. Here X represents a heteroatom. According to our algorithm, CA is a hydrophobic carbon while CB is not because a heteroatom is within four atoms.

phobic carbon atom' if there is no heteroatom (any atom other than carbon) within the 1–4 relationship (Figure 1). Such a carbon atom locates in a hydrophobic micro-environment and thus its hydrophobicity is reinforced. The total number of hydrophobic carbon atoms in the given compound is counted. Note that our definition of hydrophobic carbons does not include aromatic carbon atoms.

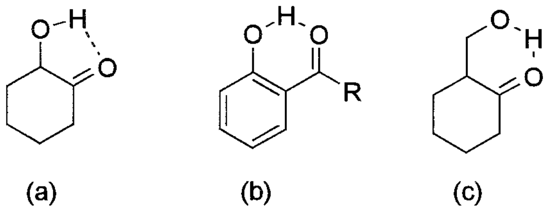

Figure 2. Examples of internal hydrogen bonds. (a) Both the donor and the acceptor are linked to a ring. (b) The donor is linked to a ring while the acceptor is not. (c) The acceptor is linked to a ring while the donor is not.

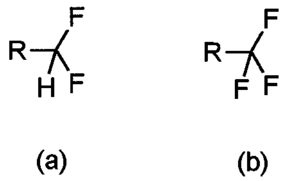

Figure 3. Examples of halogen 1–3 pairs. (a) One pair. (b) Three pairs.

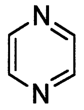

Figure 4. Illustration of aromatic nitrogen 1–4 pair.

(2) *Internal hydrogen bond:* Although it is indisputable that internal hydrogen bonding can increase the molecular hydrophobicity, identifying the existence of an internal hydrogen bond merely from the chemical structure is not straightforward. Careful conformation analysis of the given compound will help to solve this problem but, unfortunately, it is not likely to be possible for a quick log P estimation procedure. We have adopted the following definition to detect an internal hydrogen bond. (i) The donor atom could be any nitrogen or oxygen atom with at least one hydrogen atom attached, while the acceptor atom could be any sp^2 oxygen atom or sp^3 oxygen atom in a hydroxy group (see Table 1). (ii) Either the donor or the acceptor atom should be linked directly to a ring. The ring serves to immobilize the orientations of the donor and the acceptor. (iii) If both the donor and the acceptor are linked to a ring, they should be of 1–4 relationship (see Figure 2a). (iv) If only the donor or the acceptor is linked to a ring, they should be of 1–5 relationship (see Figures 2b and 2c). By using such definitions, we take only ‘reliable’ internal hydrogen bonds into account.

(3) *Halogen 1–3 pair:* When two or more halogen atoms are attached to the same atom, the properties of those atoms will change because of dipole shielding. We count the number of halogen–halogen 1–3 pairs as a correction factor (see Figure 3).

(4) *Aromatic nitrogen 1–4 pair:* This correction factor is triggered when two aromatic nitrogen atoms are in the same aromatic ring and separated by two other atoms (see Figure 4).

(5) *Ortho sp^3 oxygen pair:* This correction factor is used for compounds as illustrated in Figure 5.

(6) *Para donor pair:* This correction factor is used for compounds with two para hydrogen bond donors on an aromatic ring (see Figure 6).

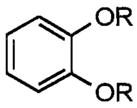

Figure 5. Illustration of ortho sp^3 oxygen pair.

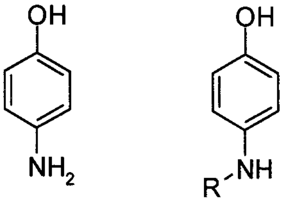

Figure 6. Examples of paralleled donor pair.

(7) *sp^2 oxygen 1–5 pair:* This correction factor is used for the chemical structure as illustrated in Figure 7. It is not counted if such a sub-structure is part of a ring.

(8) *Indicator for α-amino acid:* It is well known that α-amino acids do not have free amino and carboxylic acid groups but rather exist as zwitterions (see Figure 8). The log *P* value of an α-amino acid will be largely overestimated by atom addition alone. Therefore we use a special indicator variable for α-amino acids. This indicator could be 0 or 1.

(9) *Indicator for salicylic acid:* Salicylic acid and its derivatives are generally more hydrophobic than we have predicted by atom addition alone. Even after we add the correction of the internal hydrogen bond, the log *P* values of such compounds are still underestimated. It seems that the internal hydrogen bond existing in such a compound is much stronger than an 'average' internal hydrogen bond. Therefore, we use an additional indicator variable for salicylic acid and its derivatives (see Figure 9). The indicator could be 0 or 1.

(10) *Indicator for p-amino sulfonic acid:* The log *P* values of p-amino sulfonic acid and its derivatives are generally overestimated if predicted by atom addition alone. Therefore, we introduce another indicator variable for this kind of compound (see Figure 10). The indicator could be 0 or 1.

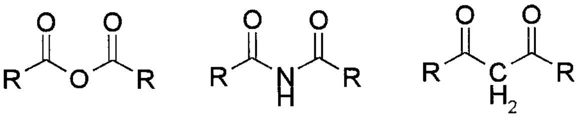

Figure 7. Examples of sp^2 oxygen 1–5 pair.

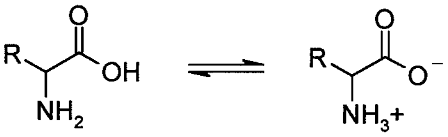

Figure 8. Illustration of alpha-amino acid.

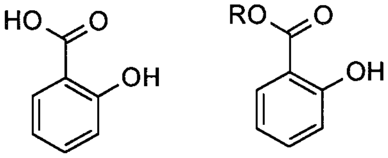

Figure 9. Illustration of salicylic acid.

Results

Regression analysis

The model for log *P* calculation includes additive (atom types) and constitutive (correction factors) terms:

$$\log P = \sum_i a_i A_i + \sum_j b_j B_j \tag{1}$$

where A_i is the occurrence of the *i*th atom type and B_j is the occurrence of the *j*th correction factor; a_i is the contribution of the *i*th atom type and b_j is the contribution of the *j*th correction factor. There are 100 terms in this equation (90 atom types plus 10 correction factors).

The contributions of each atom type and correction factor are obtained by using Equation 1 to perform multivariate regression analysis on the training set (see Tables 1 and 2). The regression analysis yields $n = 1853$, $r = 0.973$, $s = 0.349$, $F(100,1752) = 312.4$. The standard deviation of 0.349 log units is within the experimental error range that is generally considered to be 0.4 log units. Figure 11 shows the correlation between the experimental and calculated log *P* values. The slope and intercept of the regression line are 0.948 and 0.091, respectively. Figure 12 shows the distribution of the calculation errors in which a zero-centering near-Gaussian distribution is observed.

We have also performed leave-one-out cross-validation on the whole training set, which yields a correlation coefficient between the experimental and the predicted log *P* values (*r*) of 0.966 and a standard deviation in prediction (*s*) of 0.373 log units. These results demonstrate the excellent predictive ability of our method for log *P* calculation.

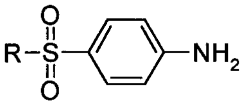

Figure 10. Illustration of p-amino sulfonic acid.

Programming

Based on the final regression model, we have developed a computer program, XLOGP v2.0, in C++. This program reads the given compound (represented in SYBYL/Mol2 format), performs atom classification, detects correction factors, and then calculates the log *P* value using the parameters listed in Tables 1 and 2. Because of the simple methodology we have developed, this program is very fast. If run on an SGI O2/R10000 workstation, it can process approximately 100 medium-sized compounds per second. The program is available in the supplementary information.

Comparison to other log P calculation procedures

As mentioned in the Introduction section, many approaches have already been developed for log *P* calculation. Some of them offer results comparable to experimental measurement. As far as the cost is concerned, they are even superior. However, routine application of log *P* calculation procedures demands a continuous check of their validity by comparing with experimental data. In an interesting paper [20], Mannhold and co-workers compared 14 calculation procedures using a test set of 138 organic compounds. Although the value of this comparison should not be overestimated because the number of compounds being tested is rather limited, it is remarkable that they have collected so many log *P* calculation procedures. To check the validity of our method, we also add XLOGP v2.0 to this comparison.

The test set is cited from Mannhold's paper [20]. It is made up of 138 organic compounds, including 90 simple molecules and 48 chemically complex drug molecules. The molecular models of the 138 compounds were built using SYBYL, stored in Mol2 format, and then subjected to calculation. The test set is available in the supplementary information. The correlation between the experimental and the calculated log *P* values of the test set is shown in Figure 13.

Fourteen log *P* calculation procedures were studied by Mannhold. All are well-established, commercially available procedures that can be roughly grouped into three categories: fragmental, atom-based, and conformation-dependent approaches (see Table 3). Evaluation of all procedures, including XLOGP v2.0, is performed as follows: (i) The individual estimation errors are grouped using Mannhold's criteria: errors less than ±0.50 are considered as acceptable; errors greater than ±0.50 and less than ±1.00 are considered

Table 3. Comparison of 15 log *P* calculation procedures

Program	Acceptable[a]	Disputable[b]	Unacceptable[c]	Uncalculated[d]	MSD[e]	r[f]	s[g]	*F*[h]	Reference
Fragmental methods									
PROLOGP_cdr 5.1	76.8	16.7	5.1	1.4	0.199	0.957	0.448	1472	5
Σf-SYBYL	81.9	13.8	4.3	0.0	0.200	0.959	0.444	1583	6
SANALOGP	79.7	15.2	3.6	1.4	0.167	0.967	0.402	1919	6
PROLOGP_comb 5.1	81.2	15.2	2.2	1.4	0.184	0.960	0.387	1582	6,13
CLOGP 4.34	84.8	10.1	3.6	1.4	0.156	0.965	0.398	1849	7,8
KLOGP	84.1	13.8	0.7	1.4	0.134	0.966	0.362	1859	9
KOWWIN	90.6	5.8	3.6	0.0	0.113	0.974	0.334	2517	10
CHEMICALC-2	68.8	17.4	13.8	0.0	0.418	0.926	0.535	827	11
Atom-based methods									
MOLCAD	68.1	20.3	11.6	0.0	0.334	0.932	0.439	911	12
Tsar2.2	68.1	20.3	11.6	0.0	0.345	0.937	0.438	987	13
PROLOGP_atom 5.1	76.8	14.5	7.2	1.4	0.262	0.947	0.431	1164	13
SMILOGP	49.3	24.6	18.8	7.2	0.551	0.917	0.588	660	14
XLOGP 2.0	91.3	8.0	0.7	0.0	0.120	0.971	0.333	2341	
Conformation-dependent methods									
HINT	68.1	15.9	13.8	2.2	0.454	0.912	0.682	665	15
ASCLOGP	55.1	28.3	15.2	1.4	0.583	0.873	0.771	431	16

[a] Percentage of acceptable results (estimation error <±0.50). [b] Percentage of disputable results (estimation error >±0.50 and <±1.00). [c] Percentage of unacceptable results (estimation error >±1.00). [d] Percentage of uncalculated results. [e] Mean squared deviations. [f] Correlation coefficient between the experimental and calculated log *P* values. [g] Standard deviations. [h] Fisher values.

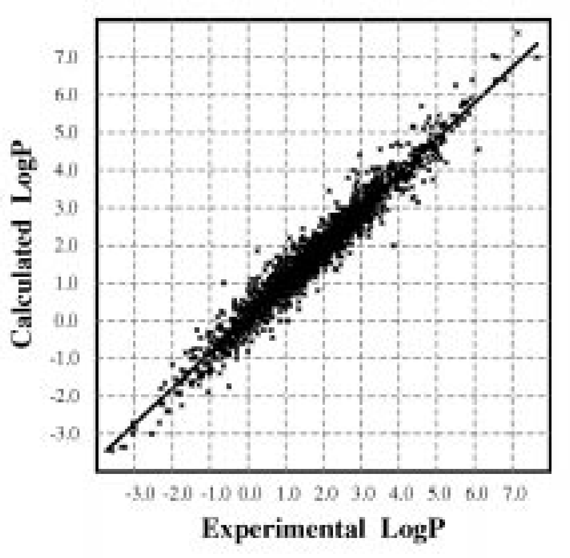

Figure 11. Correlation between the experimental and the calculated log *P* values of the training set (n = 1853, r = 0.973, s = 0.349).

as disputable; and errors exceeding ±1.00 are considered as unacceptable. The missing calculations are also counted. All these results are given as a percentage of the entire test set. (ii) The experimental and the calculated log *P* values are correlated using regression analysis. The statistical results (i.e. *r*, *s*, and *F*-value) are recorded. The mean squared deviations (MSD) are also calculated. All the results are summarized in Table 3. Except for XLOGP, data for the other 14 calculation procedures are cited from Mannhold's paper [20].

Figure 14 presents a scatter diagram in which the horizontal coordinate is the correlation coefficient (*r*) while the vertical coordinate is the acceptable percentage. According to Figure 14, the 15 procedures fall roughly into two classes. The first occupies the upper-right corner with acceptable percentages of 77% up to 91% while the correlation coefficients are generally higher than 0.950. Most members in the first class are fragmental methods. All the other procedures fall into the second class. They give acceptable percentages lower than 69% and are less successful for this test set. Although XLOGP is an atom-based approach, it is not in the second class. On the contrary, the performance of XLOGP is, if not the best, among the several best procedures in the first class.

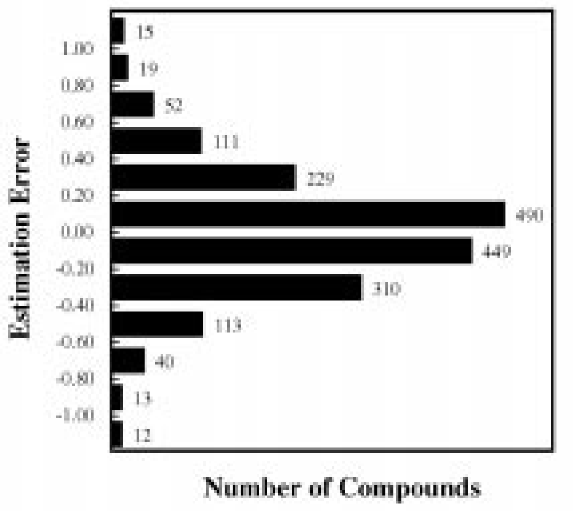

Figure 12. Distribution histogram of the calculation errors.

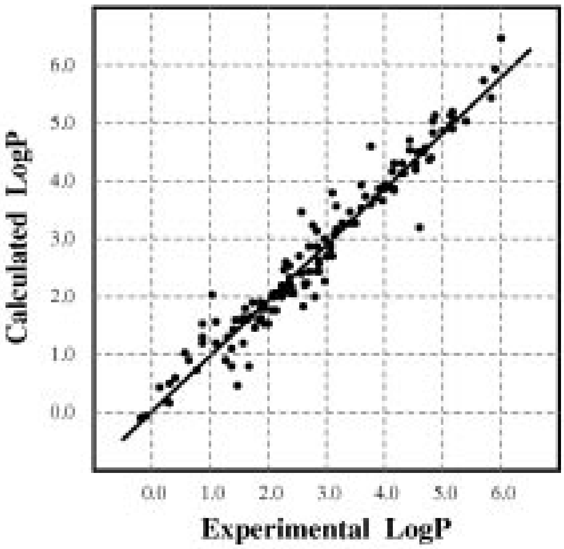

Figure 13. Correlation between the experimental and the calculated log *P* values of the test set (n = 138, r = 0.971, s = 0.333).

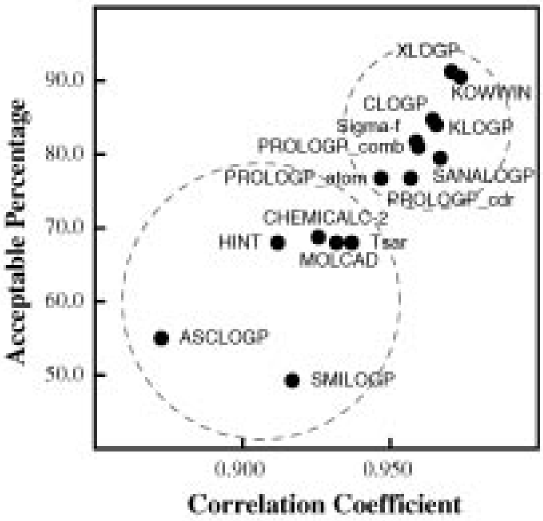

Figure 14. Scattering graph of 15 log P calculation procedures being compared.

Discussion

Atom typing

Any additive method, either by fragment or atom, needs a relevant scheme for fragment/atom classification. The quality of such a classification scheme can be evaluated by how well the calculated log P values agree with their experimental counterparts. To some extent, an additive method is the art of fragment/atom classification.

We have developed a set of 90 atom types in our new method. Compared to the atom classification scheme that we used previously [17], the new scheme is more systematic and easier to understand. It pays more attention to whether an atom is adjacent to any π-system, which proves to be important for affecting the charge densities. Furthermore, by using the new scheme we no longer need any ‘pseudo atom type’ for terminal groups such as cyano, isothiocyano, nitroso, and nitro groups. On the negative side, the new atom classification scheme uses 90 rather than 80 atom types. However, this number is still smaller than Ghose’s set of 110 [13] and much smaller than Broto’s set of 222 [12]. Using fewer parameters does not weaken the power of our method. In fact, our method yields better statistical results than Ghose’s and Broto’s approach even by atom addition alone.

It is very informative to study the hydrophobic contributions of each atom type. Although it is not proper to compare these coefficients with the ones

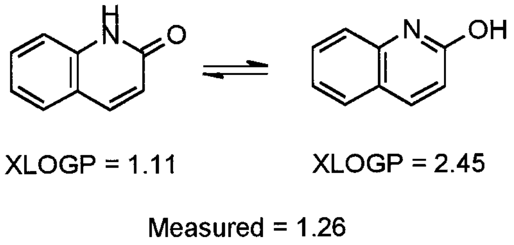

Figure 15. Illustration of tautomerization effect.

given by other additive methods (since other methods use different atom/fragment classification schemes), some common conclusions can still be established. For example, the hydrophobicity of carbon atoms clearly decreases in the order: -CH_3, -CH_2-, -CH<, >C<. The presence of a nitrogen or oxygen atom generally lowers the hydrophobicity, while the presence of a halogen atom increases the hydrophobicity by I > Br > Cl > F. If linked to a π-system, a hydrophobic atom (e.g. carbon) will become less hydrophobic, while a hydrophilic atom (e.g. nitrogen or oxygen) will become less hydrophilic. As the investigation of the charge densities proves, this happens probably because the partial charge on such an atom is dispersed by the π-system.

The major problem that our atom classification scheme will encounter is tautomerism. Since such compounds can be represented in different forms, the estimated log *P* values could be dramatically different (see Figure 15). Although one of the tautomers usually will give an estimation close to the experimental value, it is impossible to judge in advance which tautomer is the right one. This problem also exists in the other additive methods. Perhaps the ratio of all tautomers at equilibrium is necessary for the reliable prediction of the log *P* value. However, this is definitely beyond the reach of a fast log *P* calculation procedure.

Correction factors

Since the atom classification scheme adopted in our method takes only the nearest neighboring atoms into account, significant error in estimation may occur when long-range intramolecular interactions exist within the given molecule. Many other approaches have demonstrated that, in such cases, using correction factors is an efficient way to improve the accuracy of log *P* estimation.

In our method, we use 10 correction factors. They fall into two categories: representing either general intramolecular interactions or specific chemical structures. ‘Hydrophobic carbon’ and ‘internal hydrogen bond’ belong to the first category while all the other correction factors belong to the second category.

In our previous approach [17], 'hydrophobic carbon' is defined as the carbon atoms in a hydrocarbon compound. Being restricted to hydrocarbon, the application of this correction factor is rather limited. In the current approach, we have assigned this correction factor using a completely new concept: if there is no heteroatom at a certain range, a carbon atom is a 'hydrophobic carbon'. By adopting the new definition, the application of this correction factor has been extended to all kinds of organic compounds.

'Internal hydrogen bond' has also been introduced in our previous approach. In the current approach, we have reformed the algorithm of detecting internal hydrogen bonds. The new algorithm requires that the donor or acceptor atom be immobilized by a ring. This will no doubt overlook some plausible internal hydrogen bonds. However, we believe it is necessary considering that a detailed conformation analysis is impossible for a fast log *P* calculation procedure.

Certain classes of compounds will show systematic errors in log *P* estimation if predicted by atom addition alone. The other eight correction factors deal with such compounds. These correction factors are concerned with the interactions between two electronegative atoms or are simply indicator variables for certain chemical structures.

However, we must admit that finding correction factors is a tedious process. In fact, it is just as difficult as for either a 'constructionist approach' or a 'reductionist approach'. Some fragment-additive methods, such as KOWWIN [10], use hundreds of correction factors. Almost all their correction factors are indicators for specific chemical structures, which are found by 'trial-and-error' efforts. We believe this is a misleading strategy because one simply cannot figure out all the possible combinations of chemical structures. Just imagine using a double-sized training set – this will probably also double the number of correction factors! Further improvement on the accuracy of log *P* estimation should not come from using more parameters. Therefore, we strongly suggest that the correction factor should be general rather than specific. This is possible because, in theory, categories of intramolecular interactions are limited. In our method, we have attempted to introduce some 'general' correction factors, such as 'hydrophobic carbon' and 'internal hydrogen bond', and we found they work well. We are currently working on establishing a new system for defining correction factors and our ultimate goal is to eliminate all the 'specific' correction factors.

Comparison to other log P calculation methods

Our method, XLOGP, is basically an atom-based approach. Compared to other atom-based approaches, XLOGP gives much better statistical results. This is demonstrated by the comparison we have made above and also by the feedback from many XLOGP users. The superiority comes from the application of a new atom classification scheme and, more importantly, the application of correction factors. There is no reason why an atom-additive method must not use any correction factor. In fact, our method has brought this idea to atom-based approaches for the first time.

Many believe that fragmental approaches are generally more accurate than atom-based approaches, just as Mannhold concluded in his paper [20]. However, the results with XLOGP contradict this. Table 3 and Figure 14 demonstrate that XLOGP is at least comparable to other popular fragmental methods. This may indicate that the only difference left between atom-based methods and fragmental methods is whether the basic units used in addition are fragments or atoms. We do not see any reason why a fragmental method is theoretically more ‘correct’.

We have four reasons for choosing to work on an atom-additive method:

(1) An additive method will not be able to do the calculation for any compound containing a ‘missing fragment’. This often happens to fragment-additive methods. Table 3 shows that there are quite a few procedures which cannot calculate the entire test set. This problem will become severe when a large database of thousands of compounds needs to be screened. The greatest advantage of atom addition is that, in principle, atom typing can always describe the infinite variety of chemical structures.

(2) Some fragmental approaches produce ambiguous results because there are different ways to dissect the given compound into fragments. This will not happen to an atom-additive procedure since atoms are the elementary building blocks of a molecule.

(3) In some applications of hydrophobic parameters, such as the molecular lipophilicity potential approaches [21], atom-centered parameters are preferred. Therefore, the parameters derived from an atom-additive approach will be incorporated into such applications straightforwardly. Although in our method we have used correction factors in addition to basic atom types, the contribution of a correction factor can also be distributed onto the atoms involved in that correction factor and thus the ‘purity’ of atom addition is still maintained.

(4) Atom-additive methods are conceptually concise. For example, our method could be explained explicitly and thoroughly in a few pages while other complicated fragmental methods, such as CLOGP, may need a whole chapter. Atom-additive methods are also easier to program.

Conclusions

In this paper we have described our new method for log *P* calculation, XLOGP v2.0. It calculates the log *P* value by summing the contributions of component atoms and correction factors. The contributions of each atom type and correction factor are derived from regression analysis of a large number of organic compounds. The final model yields satisfactory statistical results. A comparison of various log *P* calculation procedures demonstrates that XLOGP v2.0 gives much better results than other atom-additive approaches and is at least comparable to other fragmental approaches. It is also very easy to program and applicable to QSAR studies.

References

1. Hansch, C. and Fujita, T., J. Am. Chem. Soc., 86 (1964) 1616.
2. Hansch, C., Bjorkroth, J.P. and Leo, A., J. Pharm. Sci., 76 (1987) 663.
3. Hansch, C. and Muir, R.M., Nature, 194 (1964) 178.
4. Leo, A., Chem. Rev., 93 (1993) 1281.
5. Rekker, R.F., The Hydrophobic Fragment Constant, Elsevier, New York, NY, 1977.
6. Rekker, R.F. and Mannhold, R., Calculation of Drug Lipophilicity, VCH, Weinheim, 1992.
7. Hansch, C. and Leo, A., Substituent Constants for Correlation Analysis in Chemistry and Biology, Wiley, New York, NY, 1979.
8. Leo, A., Comprehensive Medicinal Chemistry, Vol. 4, Pergamon, Oxford, 1990.
9. Klopman, G., Li, J.-Y., Wang, S. and Dimayuga, M., J. Chem. Inf. Comput. Sci., 34 (1994) 752.
10. Meylan, W.M. and Howard, P.H., J. Pharm. Sci., 84 (1995) 83.
11. Suzuki, T. and Kudo, Y., J. Comput.-Aided Mol. Design, 4 (1990) 155.
12. Broto, P., Moreau, G. and Vandycke, C., Eur. J. Med. Chem., 19 (1984) 71.
13. Ghose, A.K., Pritchett, A. and Crippen, G.M., J. Comput. Chem., 9 (1988) 80.
14. Convard, T., Dubost, J.-P. and Le Solleu, H., Quant. Struct.-Act. Relat., 13 (1994) 34.
15. Kellogg, G.E., Semus, S.F. and Abraham, D.J., J. Comput.-Aided Mol. Design, 5 (1991) 545.
16. Abraham, D.J. and Kellogg, G.E., J. Comput.-Aided Mol. Design, 8 (1994) 41.
17. Wang, R., Fu, Y. and Lai, L., J. Chem. Inf. Comput. Sci., 37 (1997) 615.
18. Hansch, C., Leo, A. and Hoekman, D. Exploring QSAR: Hydrophobic, Electronic, and Steric Constants, Vol. 2, American Chemical Society, Washington, DC, 1995.
19. SYBYL v6.4, Tripos Associates, St. Louis, MO, 1998. http://www.tripos.com/
20. Mannhold, R. and Dross, K., Quant. Struct.–Act. Relat., 15 (1996) 403.
21. Furet, P., Sele, A. and Cohen, N.C., J. Mol. Graph., 6 (1998) 182.