Comparison of Two Methods for Predicting Aqueous Solubility

Debra L. Peterson* and Samuel H. Yalkowsky

Department of Pharmaceutical Sciences, College of Pharmacy, The University of Arizona, 1703 E. Mabel Street, Tucson, Arizona 85721

Received May 18, 2001

This study compares the solubility predictions of the two parameter general solubility equation (GSE) of Jain and Yalkowsky with the 171 parameter Klopman group contribution approach. Melting points and partition coefficients were obtained for each of the compounds from Klopman's test set. Using these two variables, the solubility of each compound was calculated by the GSE and compared to the values predicted by Klopman. Both methods give reasonable solubility predictions. The data of Klopman produced an average absolute error (AAE) of 0.71 and a root-mean-square error (RMSE) of 0.86, while the GSE had an AAE of 0.64 and a RMSE of 0.92.

INTRODUCTION

Knowledge of the aqueous solubility of a compound is extremely useful for predicting its behavior in different environments. The transport, release, and absorption of a drug can be effected by its aqueous solubility.[1] Since the aqueous solubility strongly influences the behaviors of the drug, it is essential to have an accurate means for predicting it.

Although there are many methods available for this estimation,[1−8] the general solubility equation (GSE) of Jain and Yalkowsky[2] is the most widely used method in the pharmaceutical field. The appeal of this method lies in its simplicity, reliability, and relative accuracy. The GSE is based on two well-defined parameters: the melting point (MP) and the octanol−water partition coefficient (K_{ow})[2]. These two parameters can provide an indication as to what effect a solute's crystal structure and polarity will have on its solubility. The octanol−water partition coefficient can be determined experimentally or estimated using software programs such as CLOGP.[9] For some drugs the melting point can also be predicted,[10,11] but more often it must be measured. The molar solubility can then be determined by performing a simple calculation with these two variables.

Several methods for predicting solubility based on group contributions have been proposed. The structural components and/or features of a compound are used to calculate the solubility. Recently Klopman[3] used a group contribution approach comprised of 171 parameters. In this manuscript we will compare the solubility values obtained by Klopman with those calculated by the GSE.

METHODS

The 120 compounds studied by Klopman[3] and evaluated by the GSE are listed in Table 1. For each compound a melting point and octanol−water partition coefficient were obtained. The melting point data came from several sources: the Merck Index, Chemfinder website, Aquasol database,[12] and the CRC Handbook of Chemistry and Physics. CLOGP software determined the calculated partition coefficients. Using these two variables, the solubility of each compound was calculated by the GSE:

$$\log S_w = 0.5 - 0.01(\mathrm{MP} - 25) - \log K_{ow} \qquad (1)$$

The melting points, calculated partition coefficients, and solubilities predicted by Klopman and the GSE can also be found in Table 1. The average absolute error (AAE) and the root-mean-square error (RMSE) were determined for each method. These values were used to evaluate the fit of each model.

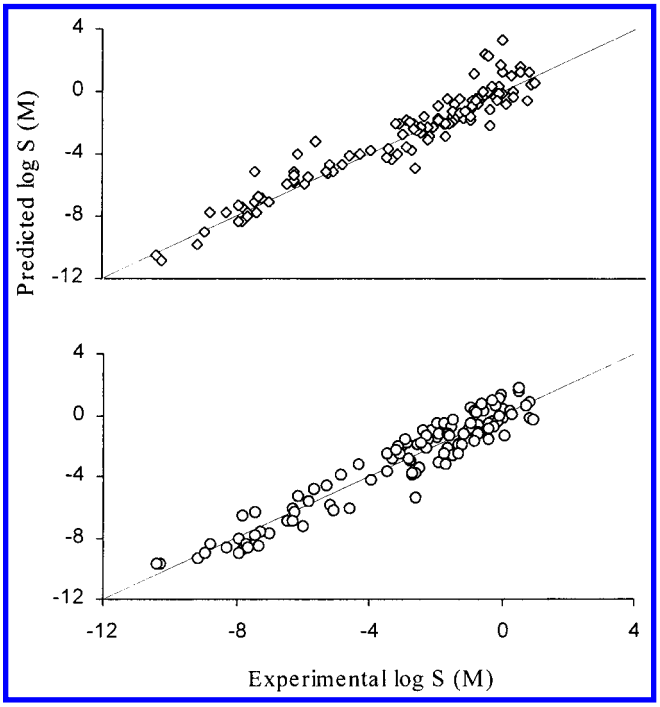

Figure 1. Experimental versus predicted log *S* values for both the GSE (◊) and Klopman (*o*) methods.

* Corresponding author phone: (520)626-4309; fax: (520)626-4063; e-mail: peterson@pharmacy.arizona.edu.

10.1021/ci010298s CCC: $20.00
Published on Web 10/23/2001

Table 1. Melting Points, Calculated Partition Coefficients, and Resulting Molar Solubilities of the Compounds Evaluated by Both the Klopman Approach and the GSE

			log *S* (M)		
	MP (C)	ClogP	GSE	exptl	Klopman
carbamic acid, ethyl ester	48	−0.18	0.45	0.85	−0.12
benzamide	129	0.65	−1.19	−0.96	−0.93
glycine	233	−3.21	1.63	0.52	1.52
L-serine	222	−2.74	1.27	−0.02	1.29
L-glutamine	185	−3.46	2.36	−0.55	0.26
7,12-dimethylbenz[a]anthracene	122	6.66	−7.13	−7.02	−7.76
lindane	113	3.75	−4.13	−4.59	−6.04
L-leucine	286	−1.67	−0.44	−0.80	−0.91
L-methionine	25	−1.73	2.23	−0.42	−0.62
L-phenylalanine	283	−1.56	−0.52	−0.92	−1.00
L-valine	315	−2.29	−0.11	−0.30	−0.46
endrin	228	3.63	−5.16	−6.29	−6.04
L-tryptophan	282	−1.57	−0.50	−1.28	−1.84
L-isoleucine	288	−1.76	−0.37	−0.59	−0.91
4-chlorobenzoic acid	243	2.70	−4.38	−3.31	−2.84
L-arginine	223	−4.79	3.31	0.00	0.42
codeine	155	0.82	−1.78	−1.52	−2.53
2-hydroxy-1,2,3-propanetricarboxylic acid	153	−2.00	1.22	0.51	1.77
2-propenamide	85	−0.61	0.52	0.95	−0.24
2-methyl-2-propenoic acid	25	0.66	−0.16	0.00	−0.17
2-methyl-2-propenoic acid, methyl ester	25	1.11	−0.61	−0.80	−1.17
1,1-dioxide-1,2-benzisothiazol-3(2H)-one	229	0.52	−2.06	−1.64	−2.16
2-amino-1-naphthalenesulfonic acid[b]		−0.70		−1.70	−2.02
9,10-anthracenedione	286	2.62	−4.73	−5.19	−5.87
1,2-benzenedicarboxylic acid, butyl phenylmethyl ester	25	3.70	−2.12	−5.64	−4.77
9H-carbazole	247	3.52	−5.24	−5.27	−4.62
2-nitrobenzenamine	72	1.92	−1.89	−1.96	−1.43
1,2-benzenedicarboxylic acid	230	0.73	−2.28	−2.11	−1.02
2-methoxyphenol	30	1.32	−0.87	−1.96	−0.51
1-naphthalenol	96	2.65	−2.86	−2.22	−1.59
1,2-dicyanobenzene	139	1.01	−1.65	−2.38	−0.97
N,*N*-diethylbenzenamine	25	3.23	−2.73	−3.03	−1.98
biphenyl	70	4.03	−3.98	−4.30	−3.24
1,1′-biphenyl-4-ol	167	3.36	−4.28	−3.48	−2.53
10H-phenothiazine	186	4.06	−5.17	−5.10	−6.17
1,1′-biphenyl-4,4′-diamine	128	1.58	−2.11	−2.70	−3.84
1,2-benzenediamine	101	−0.31	0.05	−0.42	−0.50
1,3-dichloro-2-propanol	25	0.20	0.30	−0.11	−0.41
2-propenoic acid, methyl ester	25	0.80	−0.30	−0.22	−0.76
2-imidazolidinethione	203	−0.66	−0.62	−0.71	−0.64
2-furancarboxaldehyde	25	0.67	−0.17	−0.10	0.00
1,3,5-trinitro-benzene	123	1.37	−1.85	−2.89	−1.77
1,2,3-propanetriol, triacetate	25	0.45	0.05	−0.60	0.73
diphenyldiazine	68	3.85	−3.78	−2.75	−2.90
N-phenylacetamide	114	1.16	−1.55	−1.33	−1.89
N,*N*′-diethylthiourea	76	0.79	−0.80	−1.46	−2.56
2-propenoic acid, 2-methylpropyl ester[a]	25	2.16	−1.66	−1.21	−1.87
2-aminoethanesulfonic acid	300	−3.91	1.66	−0.09	1.07
4-methyl-2-pentanone	25	1.25	−0.75	−0.74	−1.05
2-pentene	25	2.86	−2.36	−2.54	−1.86
butanedioic acid	183	−0.53	−0.55	−0.20	0.65
(E,E)-2,4-hexadienoic acid	135	1.51	−2.11	−1.77	−0.50
1,1′-iminobis-2-propanol	33	−0.85	1.27	0.81	0.91
endosulfan	106	3.65	−3.96	−6.15	−5.25
o-anthranilic acid	145	1.27	−1.97	−1.52	−0.75
5-amino-2-naphthalenesulfonic acid[b]		−0.70		−2.35	−2.53
2,4-dinitrotoluene	69	2.05	−1.99	−2.82	−2.85
1,2-diphenylhydrazine	131	2.97	−3.53	−2.92	−1.57
4-hydroxybenzaldehyde	117	1.44	−1.86	−0.96	0.48
4-methoxybenzaldehyde	25	1.78	−1.28	−1.49	−0.27
4-heptanone	25	1.91	−1.54	−1.30	−2.52
2-butenal	25	0.52	−0.02	0.32	0.09
3-methyl-1-butanol acetate	25	2.17	−1.67	−1.92	−1.27
1-naphthaleneamine	50	2.09	−1.84	−1.92	−3.04
2-naphthalenol	122	2.65	−3.12	−2.28	−2.09
2-(hydroxymethyl) phenyl-D-glucopyranoside	197	−2.38	1.16	−0.85	0.26
3-phenyl-2-propenoic acid	133	2.09	−2.67	−2.48	−3.40
2-propenoic acid, ethyl ester	25	1.33	−0.83	−0.74	−1.24
2,3-dihydro-2-thioxo-4-pyrimidinone	340	−0.37	−2.28	−2.26	−1.33
acetic acid, hexyl ester	25	2.83	−2.33	−2.46	−1.79
2-mercaptobenzothiazole	179	2.95	−3.99	−3.15	−2.03

Table 1 (Continued)

	MP (C)	ClogP	log *S* (M) GSE	log *S* (M) exptl	log *S* (M) Klopman
4-aminobenzoic acid	189	1.04	−2.18	−0.40	−1.52
acenaphthylene	94	3.62	−3.81	−3.96	−4.27
dibenzo-p-dioxin	122	4.62	−5.09	−5.31	−4.63
2,2,2-trichloro-1,1-ethanediol	57	0.71	−0.53	0.72	0.63
DL-alanine	289	−3.12	0.98	0.26	0.31
decanoic acid	31	4.04	−3.60	−3.44	−3.70
1,1,1-triflouro-2-propanol[a]	25	0.83	−0.33	0.30	0.03
cyanoguanidine	210	−1.68	−0.20	−0.31	1.02
5-nonanone	25	2.97	−2.47	−2.59	−3.80
1,2-dinitrobenzene	118	1.63	−2.06	−3.10	−2.55
2,3-dichloro-2-methylbutane[a]	25	2.91	−2.41	−2.69	−3.78
1,2-diiodoethylene	25	2.51	−2.01	−3.22	−2.28
3-methyl-3-hexanol[a]	25	2.06	−1.56	−1.00	−0.88
1,2-diethoxyethane	25	0.93	−0.43	−0.77	0.16
4-methylpentanol[a]	25	1.75	−1.25	−1.14	−1.16
1-phenylethanol	25	1.41	−0.91	−0.92	−0.56
1-hexen-3-one[a]	25	1.04	−0.54	−0.83	−1.73
1,2,3,6,7,8-hexahydropyrene	133	5.34	−5.92	−5.96	−7.25
dicamba	98	2.63	−2.86	−1.70	−3.17
dodine acetate	136	4.32	−4.93	−2.63	−5.34
3,4-dichlorobiphenyl	50	5.34	−5.09	−7.44	−6.30
asulam	144	−0.26	−0.43	−1.66	−1.21
O-*tert*-butyl carbamate	105	0.53	−0.83	0.10	−1.34
3-methyl-3-heptanol	25	2.59	−2.09	−1.60	−1.38
2,4′,5-PCB	67	5.92	−5.84	−6.25	−6.33
2,3-dimethyl-1-butanol	25	1.62	−1.12	−0.39	−0.83
ditolyl ether[a]	25	5.24	−4.74	−4.85	−3.84
3-methyl-2-heptanol	25	2.59	−2.09	−1.72	−2.45
2′,3,4,4′,5′-pentachlorobiphenyl[b]		7.10		−7.39	−8.25
2,3′,4′,5-tetrachlorobiphenyl	106	6.51	−6.82	−7.25	−7.63
2,7-dichlorodibenzo-p-dioxin	201	6.12	−7.38	−7.82	−6.59
2,2′,3,4,4′,5′-hexachlorobiphenyl	80	7.69	−7.74	−8.32	−8.62
2,2′,3,3′,4,4′,5,5′-octachlorobiphenyl	156	8.99	−9.80	−9.16	−9.33
2,2′,3,4,5′-pentachlorobiphenyl	112	6.97	−7.34	−7.91	−8.07
2,3,3′,4,4′,5-hexachlorobiphenyl	127	7.82	−8.34	−7.82	−8.76
2,3,4′-trichlorobiphenyl	69	5.80	−5.74	−6.26	−6.35
2-chlorodibenzo-p-dioxin	89	5.39	−5.53	−5.82	−5.66
2,2′,3,3′,4,4′,5,5′,6-nonachlorobiphenyl	205	9.58	−10.88	−10.26	−9.65
2,2′,3,5′-tetrachlorobiphenyl	47	6.26	−5.98	−6.47	−6.92
2,2′,3,5,5′,6-hexachlorobiphenyl	100	7.56	−7.81	−7.42	−8.45
2,2′,3,4,4′,5′,6-heptachlorobiphenyl	83	8.27	−8.35	−7.92	−8.93
2,2′,3,3′,4,5,5′,6,6′-nonachlorobiphenyl	182	9.45	−10.52	−10.41	−9.65
2,2′,3,4,5,5′-hexachlorobiphenyl	85	7.69	−7.79	−7.68	−8.45
2,2′,3,4,5,5′,6-heptachlorobiphenyl	147	8.27	−8.99	−8.94	−8.93
2,2′,3,4,6-pentachlorobiphenyl	100	6.84	−7.09	−7.43	−7.87
2,2′,3,3′,4,5-hexachlorobiphenyl	85	7.69	−7.79	−8.78	−8.45
2,3,6-trichlorobiphenyl	49	5.67	−5.41	−6.29	−6.86
2,2′,4,6,6′-pentachlorobiphenyl	85	6.59	−6.69	−7.32	−8.55
2,3,3′,4,4′,6-hexachlorobiphenyl	107	7.69	−8.01	−7.66	−8.61

[a] These compounds were assumed to be liquids, therefore having a melting point of 25 °C. [b] Melting points for these compounds were not found.

The AAE was determined by

$$\mathrm{AAE} = \frac{\sum|\log S_{\mathrm{calc}} - \log S_{\mathrm{obs}}|}{N} \qquad N = \text{number of compounds} \quad (2)$$

and the RMSE by

$$\mathrm{RMSE} = \sqrt{\frac{\sum(\log S_{\mathrm{calc}} - \log S_{\mathrm{obs}})^2}{N}} \qquad N = \text{number of compounds} \quad (3)$$

(Those compounds which have unknown melting points or partition coefficients were not included in the comparison).

RESULTS AND DISCUSSION

The comparison of the experimental versus predicted solubilities for both the Klopman model and the GSE is shown in Figure 1. It can be seen from the data provided that both methods can give reasonable solubility predictions. The AAE of Klopman is 0.71 and that of the GSE is 0.64. Likewise, the RMSE are very similar, Klopman being 0.86 and the GSE being 0.92.

Although the predictions are comparable, the approach of each method is quite different. The Klopman model is a group contribution approach based on 171 parameters that were obtained by fitting a training set of 1168 compounds. In contrast, the GSE does not require the use of a training set and utilizes only two parameters. On the other hand, the

GSE requires a measured melting point where the Klopman group contribution approach does not.

Even though the Klopman model is based on a large number of parameters there are still compounds for which solubility prediction can be difficult. If a structural feature of a molecule is not characterized by the members of the training set the error in the solubility estimation is increased. This can occur when trying to estimate the solubility of newly synthesized compounds, very complex molecules, or even isomers. For example, anthracene and phenanthrene are isomers which contain the same groups but have a 10-fold difference in their solubilities. The Klopman model created a parameter to distinguish between the two molecules and then was able to predict the solubility. However, to create a parameter for every existing isomer would be quite cumbersome. The GSE, on the other hand, can easily distinguish between isomers and accurately predict their solubility. This is because the use of the melting point by the GSE accounts for crystallinity.

CONCLUSION

This study shows that the solubility predictions by the Klopman approach and the GSE are comparable. However, when estimating the solubility of complex molecules or isomers the use of the GSE is advantageous. Although both methods can be used for the estimation of solubility, the GSE is simpler and more elegant.

REFERENCES AND NOTES

(1) Mitchell, B. E.; Jurs, P. C. Prediction of aqueous solubility of organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 489−496.

(2) Jain, N.; Yalkowsky, S. H. Estimation of the aqueous solubility I: Application to organic nonelectrolytes. *J. Pharm. Sci.* **2001**, *90*, 234−252.

(3) Klopman, G.; Zhu, H. Estimation of the aqueous solubility of organic molecules by the group contribution approach. *J. Chem. Inf. Comput. Sci.* **2001**, *41*(2), 439−445.

(4) Yalkowsky, S. H.; *Solubility and solubilization in aqueous media*; Oxford University Press: New York, 1999.

(5) Klopman, G.; Wang, S.; Balthasar, D. M. Estimation of aqueous solubility of organic molecules by the group contribution approach: Application to the study of biodegradation. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 474−482.

(6) Yalkowsky, S, H,; Pinal, R. Estimation of the aqueous solubility of complex organic molecules. *Chemosphere* **1993**, *26*(7), 1239−1261.

(7) Jorgensen, W. L.; Duffy, E. M. Prediction of Drug Solubility from Monte Carlo Simulations. *Bioorg. Medicinal Chem. Lett.* **2000**, *1*, 1155−1158.

(8) Huuskonen, J. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773−777.

(9) Heekman D. ClogP for windows version 2.0.0b; Biobyte Corp.: 1995−1997.

(10) Simamora, P.; Yalkowsky, S. H.; Group contribution methods for predicting the melting points and boiling points of aromatic compounds. *Ind. Eng. Chem. Res.* **1994**, *33*, 1405−1409.

(11) Krzyzaniak, J. F.; Myrdal, P. B.; Simamora, P.; Yalkowsky, S. H. Boiling point and melting point prediction for non-hydrogen bonding compounds. *Ind. Eng. Chem. Res.* **1995**, *34*, 2531−2535.

(12) Aquasol Database; University of Arizona, Tucson, AZ, 2001.

CI010298S