Estimation of Aqueous Solubility in Drug Design

Jarmo Huuskonen*

Division of Pharmaceutical Chemistry, Department of Pharmacy, POB 56, IN-00014 University of Helsinki, Finland

Abstract: The solubility of drugs in water is of central importance in the process of drug discovery and development from molecular design to pharmaceutical formulation and biopharmacy. The ability to estimate the aqueous solubility and other properties of a promising lead compound affecting its pharmacokinetics is a prerequisite to rational drug design, although it has received much less attention than the prediction of drug-receptor interactions. In this review, methods for the estimation of aqueous solubility of organic compounds are described and limited to approaches, which might be used in the early stage of drug design and development.

INTRODUCTION

The aqueous solubility of drug compounds is one of the most important factors in determining their biological activity. In many cases, drugs that show a good activity when administered by the parenteral route may be totally inactive when given orally. In such cases poor oral activity is often due to the fact that a sufficient amount of drug to achieve the desired response has not reached at the site of action. Hence an insufficient aqueous solubility is likely to hamper bioavailability of the drugs. In recent years, high throughput screening (HTS), where collections of thousands of compounds are screened with the intention of finding relevant biological activity, has proven valuable in finding new lead compounds [1]. It has been noted that the synthesis of combinatorial libraries tends to result in compounds with higher molecular weights and higher lipophilicity, and presumably lower aqueous solubility, than with conventional synthetic strategies [2]. For this reason, computational screens and experimental design have been suggested and used to select sub-libraries with relevant physicochemical properties of the orally active drugs, such as lipophilicity and solubility, [2-6]. Although experimental HTS "ranking" screens have been developed and used to evaluate the solubility in a 96 well format, these methods require a sample of compound [7]. Hence there is much interest in fast, reliable, and generally applicable structure-based methods for the prediction of aqueous solubility of new drugs before a promising drug candidate has even been synthesized.

Numerous different methods for the prediction of aqueous solubility have been developed and summarized by Yalkowsky and Banerjee [8]. These methods can be classified in three categories: (i) correlations with physicochemical properties (usually experimentally determined) such as partition coefficient, melting point, boiling point, etc; (ii) group contribution approaches; (iii) and parameters calculated solely from molecular structure such as molar volume, molecular surface area, shape of molecules, topological indices, etc. Although numerous methods have been developed for the estimation of aqueous solubility, only few of them have been tested with drug compounds with relatively complex chemical structures. This point was clearly stated by Lipinski *et al.* [2], who concluded that none of the available methods could be employed for the relatively accurate prediction of solubility of complex drug compounds. One reason for this might be that the training set used

*Address correspondence to this author at the Division of Pharmaceutical Chemistry, Department of Pharmacy, POB 56, IN-00014 University of Helsinki, Finland; Tel: 358 9 19159170; FAX: 358 9 19159556; e-mail jarmo.huuskonen@helsinki.fi

1386-2073/01 $28.00+.00

to derive predictive models was constructed from simple, monofunctional compounds and usually with compounds with an environmental interest avoiding drug and related compounds with multifunctional molecular structures. Usually approaches that are constructed from structural analogues yield more accurate predictive models. According to Yalkowsky and Banerjee [8], there are two approaches, i.e., correlation with partition coefficient (and melting point) and the group contribution approach, which meet the requirement of a general predictive model of aqueous solubility.

METHODS

1. Correlation with Partition Coefficient

The aqueous solubility of solid compounds is governed by interactions between molecules in the crystal lattice, interactions in the solution, intermolecular interactions in the solution, and the entropy changes accompanying fusion and dissolution [8]. The semi-empirical approach of Yalkowsky and Valvani [9] was based on estimation of the thermodynamic activity in water and the effect of the crystal structure. The model parameters were the 1-octanol/water partition coefficient, entropy of melting, and the melting point. The coefficients of the model were originally estimated on the basis of theoretical considerations, and the values obtained by fitting experimental data were in close agreement with the theoretical values. Therefore, the method was expected to be generally applicable. For rigid and short-chain non-electrolytes, the following equation resulted. The effect of crystal structure was accounted for by the melting point only:

$$\log S = -\log P - 0.01\, mp + 1.05 \qquad (1)$$

where S is the molar solubility, log P is the octanol/water partition coefficient, and mp is the melting point in degrees Celsius. All the compounds studied have been considered sufficiently rigid and to possess chains which are short enough to accommodate this model. The problem of this approach is that although the partition coefficients can be estimated with a reasonable accuracy, the melting points have to be measured. After analyzing a diverse set of 300 organic compounds, Isnard and Lambert [10] showed that the melting point correction for solid compounds was justified to only a limited extent (5-12%). The log P values alone (if they are known) gave accurate solubility estimations for most compounds and squared correlation coefficient, $r^2 = 0.95$, and standard deviation, $s = 0.67$, of estimations were obtained. Recently Meylan *et al.*, examined a very large and diverse set of organic compounds and obtained excellent results for the estimation of water solubility using calculated partition coefficients, molecular weight, melting points and 15 simple correction factors for some compound groups [11]. The statistics for the training set of 1450 compounds were $r^2 = 0.95$ and $s = 0.51$, respectively. However, this model needs one experimentally determined data point, the melting point.

2. Solvatochromic Parameters

High quality correlation equations to predict aqueous solubility have been obtained by the linear solvation energy relationship (LSER) approach originally proposed by Kamlet *et al.* [12,13]. This approach is based on the solvatochromic descriptors, and a following linear regression equation has been used in the estimation of several solubility dependent properties (*SP*), including aqueous solubility:

$$\log SP = c + rR_2 + s\pi_2^H + a\Sigma\alpha_2^H + b\Sigma\beta_2^H + vV_x \qquad (2)$$

where *SP* is a set of solute properties in a given system, for example, in water, and the independent variables are solute descriptors as follows: R_2 is an excess molar refraction, π_2^H is the dipolarity/polarizability, $\Sigma\alpha_2^H$ and $\Sigma\beta_2^H$ are the overall hydrogen bonding acidity and basicity, and V_x is the McGowan characteristic volume. Recently, Abraham and Le [14] employed this equation for the estimation of a large and diverse set of 659 organic compounds with quite impressive results ($r^2 = 0.92$ and $s = 0.56$). The drawback of this model is that almost all independent variables must be experimentally determined. However, a scheme for calculation of

these variables from molecular structure is in progress [15].

3. Group Contribution Methods

The earliest attempts to estimate solubility from chemical structure were based on the group contribution approach. In this scheme a compound is divided into basic fragments and log *S* values are estimated by the summation of the aqueous solubility contributions of these fragments. To improve the estimation accuracy of his general model (eq. 1), Yalkowsky and co-workers have developed the AQUAFAC method, in which the aqueous activity coefficient is calculated by a group contribution scheme [16,17]. However, the applicability of the AQUAFAC method is limited because it requires an experimentally determined melting point. The validation of this approach for complex chemical structures is also inadequate at this time because mainly monofunctional chemical compounds have been used. Group contribution approaches (with or without melting point correction for solid compounds) have also been proposed by Wakita [18], Klopman [19] and Kühne [20]. One way to evaluate the predictive ability of the model is to use the test set designed by Yalkowsky [8]. This test set is compiled from 21 pharmaceuticals and environmentally interesting compounds, like pesticides, with relatively complex structures (Table 1.). Klopman *et al.* used only basic group contributions in their

Table 1. Comparison of the Aqueous Solubility Estimation Methods for the Test Set

No	Compound	log S_{exp}	ANN1[a]	MLR2[b]	ANN2[b]	Klopman[c]	Kuhne[d]
1	Antipyrine	0.39	-0.34	-0.86	-1.13	-2.76	-1.90
2	Theophylline	-1.39	-1.76	-0.02	-1.23	-1.07	0.54
3	Acetylsalicylic acid	-1.72	-1.74	-1.68	-1.79	-1.52	-1.93
4	Benzocaine	-2.32	-2.57	-1.76	-1.78	-1.71	-1.75
5	Phenobarbital	-2.32	-2.71	-2.53	-3.14	-2.08	-2.41
6	Prostaglandin E2	-2.47	-1.59	-4.70	-3.51	-4.21	na
7	Phenolphthalein	-2.90	-3.25	-4.33	-4.18	-4.48	-4.61
8	Malathion	-3.37	-2.05	-3.36	-3.72	-2.94	-3.48
9	Nitrofurantoin	-3.38	-2.49	-1.72	-2.56	-2.19	-2.62
10	Diazinon	-3.64	-2.45	-4.30	-4.12	-5.29	-4.98
11	Diazepam	-3.76	-3.73	-4.37	-4.16	-5.54	-4.51
12	Diuron	-3.80	-4.19	-2.89	-2.92	-2.85	-3.38
13	Atrazine	-3.85	-2.81	-2.13	-3.66	-3.05	-3.95
14	Phenytoin	-3.90	-3.84	-3.48	-3.78	-3.47	-5.25
15	Testosterone	-4.09	-4.05	-4.33	-4.25	-5.17	-4.62
16	Lindane	-4.64	-2.65	-5.67	-5.24	-4.88	-5.08
17	Parathion	-4.66	-3.40	-3.92	-4.34	-3.94	-4.59
18	Chlorpyriphos	-5.49	-4.97	-5.34	-5.84	-5.77	-3.75
19	a-Chlordane	-6.86	-3.01	-8.18	-7.73	-7.55	-6.51
20	2.2´.4.5.5´-PCB	-7.89	-4.83	-7.33	-7.50	-7.90	-7.47
21	p.p´-DDT	-8.08	-4.76	-8.11	-7.51	-8.00	-7.75
		r^2	0.67	0.78	0.89	0.72	0.76
		s	1.25	1.05	0.68	1.13	1.05
		n	21	21	21	21	20

[a] Accoring to Huuskonen *et al.* [25]. [b] According to Huuskonen [28], ANN = artificial neural networks, MLR = multiple linear regression. [c] According to Klopman *et al.* [19]. [d] According to Kuhne *et al.* [20].

model, while Kühne *et al.* included melting points. The latter method estimated more accurately the log *S* values in the training set, but when both models were employed to estimate log *S* values in a test set of 21 compounds, the results were comparable. Hence we could also ask if the correction term for the melting point of solid compounds is really necessary for group contribution approaches or other approaches as well.

4. Topological Indices

Molecular connectivity indices are a series of descriptors based on chemical graph theory and extensively developed by Kier and Hall [21]. These indices have been shown to encode information pertaining to molecular size, branching and polarizability. By using connectivity indices ($^0\chi$ and $^0\chi^v$) along with a polarizability term (Φ) Nirmalakhandan and Speece [22] were able to estimate log S values for a diverse set of 470 organic compounds with a reasonable accuracy (r^2 = 0.98 and s = 0.33). The polarizability term, Φ, was calculated by a certain group contribution method. However, this data set was compiled mainly from environmentally interesting compounds. Patil [23] used the same approach to estimate log *S* values for a diverse set of 52 pesticides solely from molecular structure with a reasonable accuracy (r^2 = 0.81 and s = 0.72). Huuskonen *et al.* [24] used these descriptors in a neural network modeling of solubility for three sets of drug compounds (steroids, barbiturates and reverse transcriptase inhibitors). This approach was later extended for a diverse set of 211 drugs and related compounds, for a test set of 51 compounds the statistics were r^2 = 0.86 and s = 0.53, respectively [25]. This method was also employed to the test set of 21 compounds described above, and the results are given in Table 1. In order to further develop the general applicability of these approaches, a diverse set of 1297 organic compounds with accurately determined log *S* values were extracted from the AQUASOL dATAbASE [26] and SCR´s PHYSPROP Database [27]. The data set was divided into a training set of 884 compounds and a randomly chosen test set of 413 compounds. The structural parameters in a 30-12-1 neural network included 24 atom type electrotopological state (E-state) indices and 6 other topological indices, and for the test set, a predictive r^2 = 0.92 and s = 0.60 were achieved [28]. The results of this method for the test set of 21 compounds are given in Table 1. This approach can be kept as an extension of the method proposed by Nirmalakhandan and Speece [22], except the polarizability term is calculated by a group contribution scheme using the atom type E-state indices [29], and no experimentally determined data points were used in the estimation of log *S* values.

5. Quantum Chemical Parameters

The aqueous solubility values, log *S*, of a set of 331 halogenated and oxygenated hydrocarbons were correlated with 18 descriptors including semi-empirically derived charge descriptors with a standard deviation of s = 0.30 by Bodor and Huang [30]. Katritzky and co-workers [31] were able to derive a general six-parameter correlation equation for a diverse set of 411 organic compounds with correlation coefficient r^2 = 0.88 and standard error s = 0.57. The descriptors utilized were related to the polarizability of the molecule, size and shape, and specific solute-solvent interactions. In study of Mitchell and Jurs [32], regression analysis and neural networks were utilized to derive mathematical models to relate the structures of a diverse set of 332 organic compounds to their log *S* values. Topological, geometric and electronic descriptors were used to numerically represent the structural features of the data set compounds. Genetic algorithm and simulated annealing routines were used to select subsets of descriptors, which accurately relate to aqueous solubility. A nine-descriptor model was developed that has a root mean square error of 0.39 for the training set, which span a log *S* range from $-12 \geq \log S \geq 2$. All three methods work well inside the training set but the predictive ability outside the model (for example against the test set of 21 compounds designed by Yalkowsky) is questionable and should be evaluated more carefully. In addition, these data sets were compiled mainly from monofunctional compounds lacking drug-like representatives. The adequate

design of the training and test sets, and sufficient validation of the accuracy should be resolved before these methods could be used as tools in drug design.

CONCLUSIONS

According to Yalkowsky and Banerjee [8], two approaches meet the criteria for general applicability to the estimation of aqueous solubility, i.e., correlation with partition coefficient (and melting point) and group contribution approaches. This situation has not changed in the recent ten years although significant advances have been made in the development of methods for estimation of aqueous solubility. Empirical modeling with the group contribution method utilizing the atom type E-state indices for coding the solute structures is promising. On the other hand, the use of topological, geometric and electronic descriptors has been found to correlate very well with aqueous solubility. In most cases multiple linear regression analysis was employed to derive the predictive model. However, it is possible that there are some nonlinear dependencies between the structural descriptors and the aqueous solubility. Thus, an application of a nonlinear method of data analysis, like back-propagation neural networks, might provide a better modeling of the data. In addition, neural network modeling with atomic group contributions, like the atom type E-state indices, might take into account specific interactions of the solute-solvent interactions, like crystallinaty, and might make the use of a correction term for melting points of solid compounds unnecessary. Overall, the group contribution methods seem to be preferred for the accurate estimation of aqueous solubility, log *S*, of a wide variety of compounds as they are for the estimation of the partition coefficient, log *P*, in drug design and development settings.

REFERENCES

[1] Gillet, V.J.; Willet, P.; Bradshaw, J. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165.

[2] Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeney, P.J. Adv.Drug:del.Rev. **1997**, *23*, 3.

[3] Milne, G.W.A.; Wang, S.; Nicklaus, M.C. *J. Chem. Inf. Comput*. Sci. **1996**, *36*, 726.

[4] Ferguson, A.M.; Patterson, D.E.; Garr, C.D.; Underinger, T.L. *J. Biomol. Screen*. **1996**, *1*, 65.

[5] Ghose, A.K.; Viswanadhan, V.N.; Wendoloski, J.J. *J.Comb.Chem*. **1999**, *1*, 55.

[6] Blaney, J.M., Martin, E.J. *Curr.Opin.Chem.Biol*. **1997**, *1*, 54.

[7] Quarterman, C.P.; Bonham, N.M.; Irwin, A.K. *Eur.Pharm.Rev.* **1998**, *3*, 27.

[8] Yalkowsky, S.H.; Banerjee, S. *Aqueous Solubility. Methods of Estimation for Organic Compounds*, Marcel Dekker Inc.: New York, **1992**.

[9] Yalkowsky, S.H.; Valvani, S.C. *J. Pharm. Sci.* **1980**, *69*, 912.

[10] Isnard, P.; Lambert, S. *Chemosphere* **1989**, *18*, 1837.

[11] Meylan, W.M.; Howard, P.H.; Boethling, R.S. *Environ. Toxicol. Chem*. **1996**, *15*, 100.

[12] Taft, R.W.; Abraham, M.H.; Doherty, R.M.; Kamlet, M.J. *Nature* **1985**, *313*, 384.

[13] Kamlet, M.J.; Doherty, R.M.; Abraham, M.H.; Carr, P.W.; Doherty, R.F.; Taft, R.W. *J. Phys. Chem.* **1987**, *91*, 1996.

[14] Abraham, M.H.; Le, J. *J. Pharm. Sci.* **1999**, *88*, 868.

[15] Platts, J.A.; Butina, D.; Abraham, M.H.; Hersey, A. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 835.

[16] Myrdal, P.B.; Manka, A.M.; Yalkowsky, S.H. *Chemosphere* **1995**, *30*, 1619.

[17] Yung-Chi, L.; Myrdal, P.B.; Yalkowsky, S.H. *Chemosphere* **1996**, *33*, 2129.

[18] Wakita, K.; Yosimoto, M.; Myamoto, S.; Watanabe, H. *Chem. Pharm. Bull*. **1986**, *34*, 4663.

[19] Klopman, G.; Wang, S.; Balthasar, D.M. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 474.

[20] Kühne, R.; Ebert, R-U.; Kleint, F.; Schmidt, G.; Schuurmann, G. *Chemosphere* **1995**, 30, 2061.

[21] Kier, L.B.; Hall, L.H. Molecular Connectivity in Structure-Activity Analysis, Wiley; New York, **1986**.

[22] Nirmalakhandan, N.N.; Speece, R.E. *Environ. Sci. Technol.* **1989**, *23*, 708.

[23] Patil, G.S. *J. Hazard. Mater*. **1994**, *36*, 35.

[24] Huuskonen, J.; Salo, M.; Taskinen, J. *J. Pharm. Sci.* **1997**, *86*, 450.

[25] Huuskonen, J.; Salo, M.; Taskinen, J. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 450.

[26] Yalkowsky, S.H.; Dannelfelser, R.M. *The ARIZONA dATAbASE of Aqueous Solubility*, College of Pharmacy, University of Arizona; Tucson, AZ, **1990**.

[27] Syracuse Research Corporation. *Physical/Chemical Property Database* (*PHYSPROP*), SRC Environmental Science Center; Syracuse, NY, **1994**.

[28] Huuskonen, J. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 773.

[29] Kier, L.B.; Hall, L.H. *Molecular Structure Description. The Electrotopological State*, Academic Press; San Diego, CA, **1999**.

[30] Bodor, N.; Huang, M-J. *J. Pharm. Sci*. **1992**, *81*, 954.

[31] Katritzky, A.R.; Wang, Y.; Sild, S.; Tamm, T.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 720.

[32] Mitchell, B.E.; Jurs, P.C. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 489.