By Cikui Liang and David A. Gallagher

Prediction of physical and chemical properties by quantitative structure-property relationships

THE PHYSICAL AND chemical properties of a compound are a function of its molecular structure. Structure-property relationships are developed by finding one or more molecular descriptors that explain variations in the physical or chemical properties of a group of congeners/analogs. While some descriptors can be determined experimentally, deriving them from either the two-dimensional (2-D) or three-dimensional (3-D) molecular structure is generally more convenient and practical. A relationship, once quantified, can be used to estimate the properties of other molecules simply from their structures and without the need for experimental determination or synthesis. This has resulted in the development of quantitative structure-property relationships (QSPRs)[1] as an important tool in chemical, biological, and environmental research. When a structure-property relationship is found, it may also provide insight into which aspect of the molecular structure influences the property. Such insight can facilitate a systematic approach to the design of new molecules with more desirable properties.

When applied to biologically active molecules such as drugs, QSPR is usually referred to as a quantitative structure-activity relationship (QSAR). Many commercial software packages are available for determining QSPRs and QSARs, including modules for structure input and for the calculation of physical, chemical, and electronic descriptors.

Background

Water solubility (S) is basic to the fundamental understanding of processes in almost all branches of basic and applied science.[2] Water solubility plays a key role in a range of applications such as drug dosage; anesthesiology; corrosion of metals; transport fate of pollutants in terrestrial, aquatic, and atmospheric ecosystems; deposition of minerals and radioactive fission products from ocean waters; composition of groundwaters; and availability of oxygen and other gases in life support systems.

The widespread relevance of water solubility data to many branches and disciplines of science, medicine, technology, and engineering has led to the development of several models to predict water solubility. Li et al.[3] used a UNIFAC model to predict aqueous solubility for polychlorinated biphenyls (PCBs) based on empirical data such as melting and boiling points. The need for empirical data limits the usefulness of this approach, which has only been demonstrated to work well for PCBs with less than seven chlorines. Zhang et al.[4] examined the relationship between molecular size and solubility focusing on thermodynamic properties related to water solubility. However, no general regression equation was derived for water solubility prediction. Bodor et al.[5] developed a model to estimate water solubilities for various classes of organic compounds using as many as 17 descriptors. The limitations of Bodor's model are discussed later.

This paper describes the development of a model to predict water solubilities for both analogs and broad ranges of compounds from only theoretically derived descriptors and with as few descriptors as possible to avoid overfitting.[6]

Procedure

The QSPR model was developed using the CAChe WorkSystem™ 3.8 (Oxford Molecular Group, Beaverton, OR) running on a Power Macintosh (Apple Computer Corp., Cupertino, CA). The system allows calculation of various physical, chemical, topological, and electronic descriptors directly from the structure of the compound.

All the molecules were built in a ProjectLeader™ table with the CAChe Editor (Oxford Molecular Group) and optimized using the standard procedures provided based on CAChe MOPAC PM3 (Oxford Molecular Group). The PM3 parameter set was used for all optimizations in the model development because it is pa-

Ms. Liang is with the Department of Environmental Science and Engineering, Oregon Graduate Institute, Portland, OR, U.S.A. Mr. Gallagher, C. Chem., is with CAChe Scientific/Oxford Molecular Group, P.O. Box 4003, Beaverton, OR 97076, U.S.A.; tel.: 503-526-5000; fax: 503-526-5099; e-mail: dgallagher@oxmol.com.

Table 1

Compounds by class used for development of water solubility model

Compound classes	Compound list
Acid	Benzoic acid
	Phenylacetic acid
Alcohol	Methanol
	Ethanol
	1-Propanol
	1-Butanol
	1-Pentanol
	2-Pentanol
	1-Hexanol
	Cyclohexanol
	1-Heptanol
	Cyclohexanol
	Heptanol
	1-Octanol
	1-Pentene-3-ol
	2,2-Dimethyl-3-pentanol
Alkane	Methane
	Ethane
	Propane
	2-Methyl propane
	Butane
	Pentane
	2,2-Dimethylpropane
	Cyclopentane
	Methyl cyclopentane
	Cyclohexane
	Cycloheptane
	Cyclooctane
	Diphenylmethane
	Methyl chloride
	Dichloromethane
	Chloroform
	1,1-Dibromethane
	1,2-Dibromethane
	1,1-Dichloroethane
	Chloroethane
	Pentochloroethane
	Isoamyl bromide
	Methyl fluoride
Amine	Aniline
	Diethylamine
	Dibutylamine
	N,N-dimethyl aniline
	Dipropylamine
	Ethylamine
	Heptylamine
	Hexylamine
	N-methyl aniline
	Octylamine
	Triethylamine
	Trimethylamine
	m-Toluidine
	o-Toluidine
Aldehyde and ketone	Benzaldehyde
	Butanal
	Hexanal
	Acetone
	2-Butanone
	2-Pentanone
	2-Heptanone
	4-Heptanone
Alkene	Ethylene
	1,3-Butadiene
	2-Methyl-2-butene
	3-Methyl-1-butene
	1,4-Pentadiene
	1-Hexene
	1-Octyne
	1-Cyclopentene
	1-Cyclohexene
	1-Cycloheptene
	1,3-Cycloheptadiene
	1,3,5-Cycloheptatriene

Table 1 (*cont)*

Compound classes	Compound list	Compound classes	Compound list
	Nitromethane		
	1 -Nitropropane		
	2-Nitropropane		
Aromatic	Acenaphthalene	Halogenated aromatic	Dibenzo-pdioxin
	Anthracene		I-CDD
	Benzene		2,3-DCDD
	Biphenyl		2,7-DCDD
	o-Cresol		2,8-DCDD
	m-Cresol		1,2,4-T3CDD
	Ethylbenzene		1,2,3,4-TCDD
	Naphthalene		1,2,3,7-TCDD
	Phenol		1,3,6,8-TCDD
	Phenanthrane		2,3,7,8-TCDD
	Pyrene		1,2,3,4,7-PCDD
	Pyridine		1,2,3,4,7,8-H6CDD
	Quinoline		1,2,3,4,6,7,8-H7CDD
	1,2,4-Trimethylbenzene		OCDD
	Styrene		
	Toluene		2-PCB
	o-Xylene		2,2′-PCB
	p-Xylene		2,4′-PCB
	Methoxybenzene		2,6-PCB
	Ethoxybenzene		3,3-PCB
			4,4′-PCB
			2,4,6-PCB
			2′3,4-PCB
	Nitrobenzene		
	o-Nitrotoluene		
	m-Nitrotoluene		1-Bromo-3-fluorobenzene
			1-Bromo-4-iodobenzene
			1-Chloro-2-fluorobenzene
			1-Chloro-2-iodobenzene
			1,2-Dibromobenzene
			1,2-Difluorobenzene
			1,2-Diiodobenzene
Ether	Butyl ethyl ether	Ester	Butyl acetate
	Diethyl ether		Ethyl acetate
	Dipropyl ether		Methyl acetate
	Ethyl isopropyl ether		Propyl acetate
	Ethyl propyl ether		Isobutyl formate
	Furan		Propyl formate
	Methyl isopropyl ether		
	Methyl propyl ether		
	Methyl sec-butyl ether		
	Methyl tert-butyl ether		
Nitrile	Acetonitrile		
	Benzonitrile		
	Butyronitrile		
	Propionitrile		

rameterized for a wide range of elements, and it also computes heat of formation with reasonable accuracy.[7]

Properties and descriptors quantifying different structural attributes were chosen from a list of over 100 parameters already provided in the CAChe WorkSystem. Selection was based on the results from similar work in the literature, and on any properties intuitively related to water solubility, which included:

Table 2

Summary regression results for dioxins

Chemical sample	Water solubilities (logS)	SAS	Regression predicted (logs)	Absolute errors
Dibenzo-pdioxin	-4.359	94.556	-4.346	0.013
1-CCD	-4.923	104.194	-5.063	0.160
2,3-DCDD	-5.855	113.775	-5.815	0.040
2,7-DCDD	-5.999	115.320	-5.933	0.066
2,8-DCDD	-5.934	155.067	-5.914	0.020
1,2,4-T3CDD	-6.493	122.807	-6.506	0.013
1,2,3,4-TCDD	-7.135	129.775	-7.039	0.096
1,2,3,7-TCDD	-7.431	132.072	-7.214	0.217
1,3,6,8-TCDD	-7.084	134.431	-7.395	0.311
2,3,7,8-TCDD	-7.453	133.020	-7.287	0.166
1,2,3,4,7-PCDD	-7.799	140.272	-7.841	0.042
1,2,3,4,7,8-H6CDD	-8.493	149.222	-8.526	0.033
1,2,3,4,6,7,8-H7CDD	-8.876	156.764	-9.102	0.226
CCDD	-9.854	164.409	-9.687	0.167

Table 3

Regression results with SAS descriptor

Compound class	LogS prediction equation	r^2	r_{cv}^2	*n*	Standard error of estimate
Alcohol	logS = –0.074SAS + 4.334	0.983	0.973	12	0.172
Dioxin	logS = –0.076SAS + 2.885	0.990	0.984	14	0.157
PCBs and halogenated benzenes	logS = –0.072SAS + 2.590	0.951	0.933	17	0.252
Dioxins, PCBs, and halogenated benzenes	logS = –0.079SAS + 3.306	0.989	0.988	30	0.198
Aromatics	logS = -0.081 SAS + 4.367	0.925	0.921	61	0.663
All classes	logS = –0.076SAS + 3.388	0.698	0.693	154	1.245

r^2 = Square of the regression coefficient.
r_{cv}^2 = Cross-validation, calculated by dividing the input data into three groups and running three separate regression calculations. each using ⅔ of the data to predict the other ⅓
n = Number of compounds used to develop the regression.

Table 4

Comparison of three-descriptor models

Descriptor(s) and #NH_2	r^2	r_{cv}^2	Standard error of estimate
SAS, log(dHF), and #NH_2	0.904	0.900	0.707
SAS, log(dHF), and sqrt (N+O)	0.877	0.871	0.798
SAS, log(dHF), #NO_2	0.876	0.872	0.804
SAS, log(dHF), and THA	0.865	0.854	0.839
SAS, log(dHF), and #OH	0.864	0.857	0.842
SAS, log(dHF), and TAA	0.860	0.847	0.853
SAS, log(dHF), and MW	0.859	0.854	0.853
SAS, log(dHF), and TAN	0.859	0.851	0.853

- Solvent accessible surface area (SAS)
- Difference of heat of formation in water and vacuum (dHF)
- Total atom number of C, Cl, N, 0, F, I, and Br (TAN)
- Dipole moment
- Polarizability
- Molecular weight (MW)
- Number of NH_2 groups (#NH_2)
- Number of NO_2 groups (#NO_2)
- Number of OH groups (#OH)
- Total number of halogen atoms-Cl, F, Br, and I (THA)
- Total number of electronegative atoms--Cl, F, Br, I, N, and 0 (TAA)
- Total number of nitrogen and oxygen atoms (TNO).

Different transformations of the above descriptors were also tested, including square, square root (sqrt), logarithm, and reciprocal.

The system was set to automatically calculate all the properties and fill in the table. Experimental data (S = mol/L) for water solubility were taken from Bodor et al." and Mackay et al.[8] Initially, investigations were limited to single sets of analogs, including alcohols, halogenated alkanes, and dioxins. The correlation of each calculated descriptor column against the experi-

Table 5

Summary of one-, two-, and three-descriptor models for all 154 compounds

Descriptor(s)	LogS prediction equation	r^2	r_{cv}^2	Standard error of estimate
SAS	logS = –0.076SAS + 3.388	0.698	0.693	1.245
SAS, log(dHF)	logS = –0.085SAS + 1.544log(dHF) + 3.339	0.859	0.854	0.853
SAS, log(dHF), and #NH_2	logS = –0.084SAS + 1.484log(dHF) + 1.722 #NH_2 + 3.098	0.904	0.900	0.707

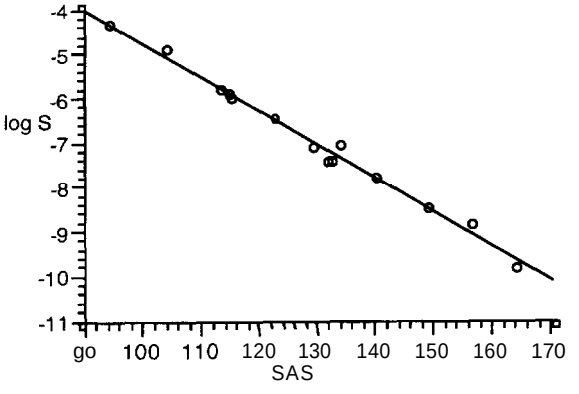

Figure 1 *Correlation between SAS and water solubility for dioxins*

mental water solubility (logS) was analyzed by viewing scatter plots. The scatter plot allowed visual inspection of the fit and the r^2 value. The descriptors that gave the best fits were used to generate regression analysis columns. Cross-validations and average errors from the regression columns were then compared.

Finally, all 154 molecules (listed in *Table 1*), including acids, alcohols, aldehydes and ketones, alkanes and halogenated alkanes, alkenes, aromatics and halogenated aromatics, ethers, esters, and nittiles, were entered into a single table in an attempt to establish a general QSPR across all classes of compounds. At this stage, it became necessary to use combinations of two descriptors or more, as discussed below.

Results and discussion

The SAS descriptor gave the best overall fit with r^2 values of 0.983, 0.990, 0.95 1, and 0.925 for alcohols, dioxins, PCBs and halogenated benzenes, and aromatic compounds, respectively. *Table 2* and *Figure 1* illustrate the results for dioxins. Detailed regression results for some of the compound classes are presented in *Table 3*. The slopes and intercepts for dioxins and for PCBs and halogenated benzenes were found to be close enough to allow combining these two groups into a single group without significantly lowering the r^2 value. The success of the SAS descriptor was consistent with the results published by Brezonik.[9]

MW gave the second best fit as a descriptor, with an r^2 value of 0.988 for the dioxins. A plot of MW against SAS showed them to be highly correlated, as expected, and thus MW was omitted as a second desciptor to avoid overfitting.[6] When the single SAS descriptor was applied to the total 154 compounds from all classes, the r^2 value decreased to 0.689. More descriptors were needed to improve the accuracy of this model when applied to broad classes of compounds.

Recent computational developments with molecular orbital package (MOPAC) PM3 and conductor-hyphen-like screening model (COSMO)[10] have facilitated the reasonably accurate prediction of heat of formation in water as well as heat of formation in vacuum. The difference between the two (dHF = heat of formation in vacuum-heat of formation in water) is related to the heat of water solvation, which should influence water solubility. This dHF descriptor could account for the difference in hydrophilicity between compound classes, and was thus tested in conjunction with the SAS descriptor. Although the dHF descriptor gave a very poor r^2 value by itself, the combination of SAS and log(dHF) gave the best r^2 value of 0.859 for all 154 compounds when compared to other possible combinations of their transformations.

Three-descriptor models were also tested by adding one of the remaining descriptors to different transformations of the best two-descriptor model (SAS and log[dHF]), as shown in *Table 4*. Amine solubilities were consistently underestimated in the two-descriptor model. Thus, not surprisingly, #NH_2 gave the best improvement on the r^2 value as the third descriptor. Since all log(dHF) calculations were performed on neutral amines, no allowance was made for the higher heat of solution of the protonated species. Thus, the #NH_2 descriptor may account for higher solubility arising from the presence of some protonated amine.

A comparison of the one-, two-, and three-descriptor models for water solubility prediction for all 154 compounds is shown in *Table 5*.

Some published models fail outside the range of

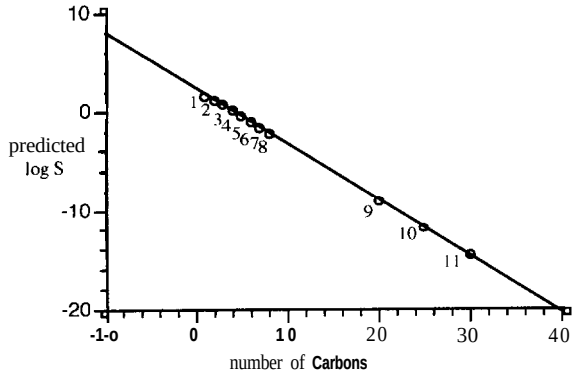

Figure 2 *Asymptopic test for single-descriptor SAS model: 1) methanol, 2) ethanol, 3) propanol, 4) butanol, 5) pentanol, 6) hexanol, 7) heptanol, 8) octanol, 9) eicosanol, 10) pentacosanol, 11) carnaubyl alcohol.*

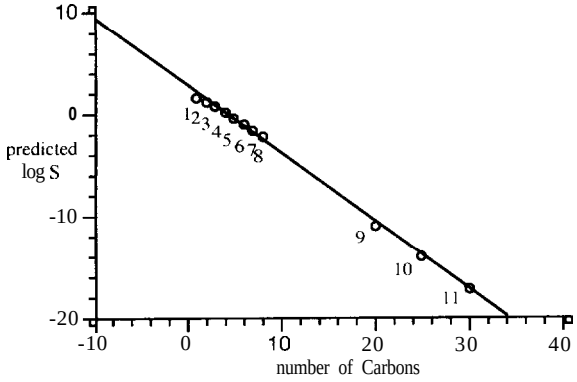

Figure 3 *Asymptopic test for two-descriptor SAS and log(dHF) model: 1) methanol, 2) ethanol, 3) propanol, 4) butanol, 5) pentanol, 6) hexanol, 7) heptanol, 8) octanol, 9) eicosanol, 10) pentacosanol, 11) carnaubyl alcohol.*

molecules with which they were calibrated. In one published water solubility mode,' the predicted water solubility goes through a minimum, then actually starts to increase with increasing chain length when tested with primary alcohols, contrary to intuition. This is presumably due to the square of a surface area term that overpowers the other descriptors in larger molecules. The asymptopic performance of the single-descriptor SAS model was tested by comparing the predicted water solubility for some high-molecular-weight alcohols with anticipated results in the absence of experimental data. *Figure* 2 shows that the predicted water solubility continues to decrease linearly (r^2 value of 1.000) with increasing chain length (eicosanol, C_{20}; pentacosanol, C_{25}; carnaubyl alcohol, C_{30}). The asymptopic performance of the two-descriptor model (SAS and log[dHF]) demonstrates similarly decreasing solubility in *Figure* 3. These trends support the validity of the SAS descriptor as a model for water solubility.

Conclusion

SAS is shown to be a remarkably accurate predictor of water solubility within each single class of compounds that was investigated.

Together, SAS and log(dHF) provide a reasonably accurate and useful predictor of water solubility across a wide range of different classes. The addition of the $\#NH_2$ descriptor significantly improves the prediction for amines.

There is often the need to predict nonhandbook properties such as environmental lifetimes of pollutants, carcinogenicity, or the leaching of monomers from packaging into intravenous saline solution. The QSPR approach described here can be easily adapted to the prediction of a wide range of such physical and chemical properties and biological activities. These structure-derived calculations are carried out in a fraction of the time that it would take to perform the necessary syntheses and experimental determinations.

References

1. Gupta SP. QSAR studies on drug acting at the central nervous system. Chem Rev 1989; 89:1765–1800.
2. Shaw DG. Solubility data series. New York: Pergamon Press, 1989:37.
3. Li A, Doucette WJ, Andren AW. Estimation of aqueous solubility, octanol/water partition coefficient, and Henry's law constant for polychlorinated biphenyls using UNIFAC. Chemosphere 1994: 29:657–69.
4. Zhang X. Gobas FAPC. A thermodynamic analysis of the relationships between molecular size, hydrophobicity, aqueous solubility and octanol-water partitioning of organic chemicals. Chemosphere 1995; 31:3501–21.
5. Bodor N, Huang MJ. A new method for the estimation of the aqueous solubility of organic compounds. J Pharm Sci 1992; 81:954–60.
6. **Shorter** J, Phil D. Correlation analysis of organic reactivity: with particular reference to multiple regression. Chichester. U.K.: Research Studies Press, 1982.
7. Stewart JP. Optimization of parameters for semiempirical methods II. J Comp Chem 1989; 10(2).
8. Mackay D, Shiu WY. Ma KC. Illustrated handbook of physical-chemical properties and environmental fate for organic chemicals. Boca Raton, FL: Lewis Publishers, 1992.
9. Brezonik PL. Chemical kinetics and process dynamic in aquatic systems. Boca Raton, FL: Lewis Publishers, 1994.
10. Klamt A, Schuurmann C. COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. J Chem Soc 1993. Perkin Trans; 2:799–805.