The Role of Solvent-Accessible Surface Area in Determining Partition Coefficients

W. J. Dunn III,* M. G. Koehler, and Stelian Grigoras†

Department of Medicinal Chemistry and Pharmacognosy, The University of Illinois at Chicago, Chicago, Illinois 60612.

Received September 10, 1986

The logarithm of the partition coefficient (log P) of low-molecular-weight organic compounds is a physicochemical parameter used extensively in structure–biological activity studies to model interactions of the compounds with nonpolar phases in vitro and in vivo. The partition coefficient can be determined between water and a number of nonpolar solvents. The most common nonpolar solvent is 1-octanol, but solvents such as benzene, carbon tetrachloride, and chloroform are frequently used as models for the nonpolar phases. The functional relationship between chemical structure and partitioning is not well-understood. In this paper, partition coefficient data for 50 solutes in six nonpolar solvent systems are analyzed by using principal components analysis. The objective of the work is to explore the relationship between solute structure and partitioning behavior for simple organic compounds. Two structural factors are found to be important, with the isotropic surface area being the most important. The isotropic surface area can be used to estimate log P in some solvents and as an independent variable in quantitative structure–activity relationships (QSAR). This is illustrated by estimating the rate of epidermal diffusion of steroids.

There have been a number of theoretical studies reported dealing with the relationship between solute structure and the solution properties of low-molecular-weight organic compounds.[1,2,3,4,5,6] Hermann[1,2] has proposed that the partition coefficient and aqueous solubility are related to the size of the solute. According to Hermann, the solvent-accessible surface area or cavity surface area can be used as the size-related parameter. Harris et al.[6] used relative solute surface area, approximated by attaching spheres to space-filling models of the solute, to relate solute size to free energies of transfer of the solute from water to a nonpolar phase. A similar approach was used by Leo et al.[7] to estimate the 1-octanol/water partition coefficients of low-molecular-weight compounds. Calculated solvent-accessible surface areas were used by Yalkowsky and Valvani[3,4] to estimate partition coefficients of hydrocarbons.

The results of these studies were of limited utility since the treatment could not be extended to functionalized solutes. Attempts to treat saturated and unsaturated hydrocarbons in a single relationship were not successful, suggesting that other factors, in addition to size, might be significant in determining partitioning behavior.[6]

Here we report the results of work to determine the structural factors that determine the partitioning behavior of simple organic compounds. A preliminary report of this work has been published.[8]

Methods

It is assumed, in this work, that the logarithm of the partition coefficient, log P, of solute k in nonpolar solvent i can be expressed as a linear combination of parameters which can be obtained from the three-dimensional arrangement of the atoms in the solute. This is expressed in eq 1 where A is the number of parameters.

$$\log P_{ki} = \sum_{a=0}^{A} t_{ka} b_{ai} \qquad (1)$$

This expression is in the form of a principal components model in which the t's are solute-specific terms or principal components and the b's are solvent-specific terms or loadings. Such a model can be derived from data as in Table I in which log P data are obtained in various nonpolar solvents. In these data, information about the change in log P with solute structure is contained in the columns while information about the change in log P with nonpolar solvent is in the rows. Therefore, solution for the t's in eq 1 will yield an abstract expression of the relationship between partitioning and chemical structure. This treats the structure–log P problem as a "multivariable" one.

Table I. Data Table for a Structure–log P Problem

	log P						
solute	1	2	3	...	i	...	p
1	log P_{11}						
2	log P_{21}	log P_{22}					
3							
⋮							
k					log P_{ki}		
⋮							
n							

This is an expression or extension of the work of Collander,[9] who showed that for a small number of nonpolar solvents and a limited number of solutes, log P was a linear free energy related parameter.[10]

In order to obtain a solution to the parameters in eq 1, a matrix of log P data for 50 solutes in six nonpolar solvents was obtained from the literature. These data are given in Table II and were taken from the Pomona College Medicinal Chemistry Data Bank.

For several of the compounds in Table II, more than one log P value was reported in the data bank. Therefore, some data selection was necessary. Data were selected according to the following criteria: (1) values reported by the Hansch group at Pomona College, (2) values published by researchers trained in the Hansch laboratory, and (3) values reported by groups that were shown to obtain the same results as the Hansch group at Pomona College.

To determine the number of structural features that determine the partitioning properties of a solute, the data

† Present address: Dow Corning, Midland, MI.

[1] Hermann, R. B. *J. Phys. Chem.* **1971**, *75*, 363.
[2] Hermann, R. B. *J. Phys. Chem.* **1972**, *76*, 2754.
[3] Yalkowsky, S. H.; Valvani, S. C. *J. Med. Chem.* **1976**, *19*, 727.
[4] Yalkowsky, S. H.; Valvani, S. C. *J. Chem. Eng. Data* **1979**, *24*, 127.
[5] Bultsma, T. *Eur. J. Med. Chem.—Chim. Ther.* **1980**, *15*, 371.
[6] Harris, M. J.; Higuchi, T.; Rytting, J. H. *J. Phys. Chem.* **1973**, *77*, 2694.
[7] Leo, A. J.; Hansch, C.; Yow, P. Y. C. *J. Med. Chem.* **1976**, *19*, 611.
[8] Dunn, W. J., III; Grigoras, S.; Johannson, E. In *Partition Coefficient: Theory and Estimation*; Dunn, W. J., III, Block, J. S., Pearlman, R. S., Eds.; Pergamon: New York, 1986; Chapter 2.
[9] Collander, R. *Acta Chem. Scand.* **1949**, *3*, 717; **1950**, *4*, 1085; **1951**, *5*, 774. Collander, R. *Physiol. Plant.* **1954**, *7*, 420.
[10] Leffler, J. E.; Grunwald, G. *Rates and Equilibria of Organic Reactions*; Wiley: New York, 1963.

0022-2623/87/1830-1121$01.50/0 © 1987 American Chemical Society

Table II. log *P* Data for Solutes in Nonpolar Solvents

solute	1-octanol	ether	chloroform	benzene	carbon tetrachloride	hexane
1. methanol	−0.77	−1.15	−1.26	−1.89	−2.10	−2.80
2. ethanol	−0.31	−0.57	−0.85	−1.62	−1.40	−2.10
3. propanol	0.25	−0.02	−0.40	−0.70	−0.82	−1.52
4. butanol	0.88	0.89	0.45	−0.12	−0.40	−0.70
5. pentanol	1.56	1.20	1.05	0.62	0.40	−0.40
6. hexanol	2.03	1.80	1.69	1.30	0.99	0.46
7. heptanol	2.41	2.40	2.41	1.91	1.67	1.01
8. acetic acid	−0.17	−0.34	−1.60	−2.26	−2.45	−3.06
9. propionic acid	0.33	0.27	−0.96	−1.35	−1.60	−2.14
10. butyric acid	0.79	0.61	−0.27	−0.96	−0.97	−1.76
11. hexanoic acid	1.92	1.95	1.15	0.30	0.57	−0.46
12. pentanoic acid	1.39	1.00	0.28	−0.10	−0.42	−1.00
13. trichloroacetic acid	1.33	1.21	−0.69	−1.30	−1.66	−2.63
14. dichloroacetic acid	0.92	1.31	−0.89	−1.40	−2.31	−2.72
15. chloroacetic acid	0.22	0.37	−1.92	−1.60	−2.56	−3.14
16. methyl acetate	0.18	0.43	1.16	0.53	0.32	−0.26
17. ethyl acetate	0.73	0.93	1.80	1.01	0.95	0.29
18. acetone	−0.24	−0.21	0.24	−0.05	−0.30	−0.91
19. ethylamine	−0.30	−1.18	−0.35	−1.30	−1.27	−1.77
20. propylamine	0.28	−0.54	0.26	−0.52	−0.59	−1.00
21. trimethylamine	0.27	−0.26	0.54	−0.29	−0.09	−0.48
22. *n*-butylamine	0.74	0.11	0.56	−0.08	−0.04	−0.62
23. diethylamine	0.57	−0.07	0.81	−0.05	0.03	−0.48
24. pyridine	0.65	0.08	1.43	0.41	0.23	−0.21
25. aniline	0.90	0.85	1.42	1.00	0.60	−0.30
26. phenol	1.46	1.64	0.37	0.36	−0.36	−0.70
27. benzoic acid	1.87	1.89	0.50	0.21	−0.22	−0.72
28. benzamide	0.64	−0.22	0.11	−0.71	−1.54	−2.30
29. 2-naphthol	2.70	1.77	1.74	1.74	0.99	0.30
30. hydroquinone	0.59	0.38	0.23	0.15	0.04	0.05
31. *p*-hydroxybenzaldehyde	1.35	1.10	−0.12	−0.55	−1.70	−0.95
32. *o*-hydroxybenzoic acid	2.26	2.37	0.58	0.50	0.00	−0.57
33. *p*-hydroxybenzoic acid	1.58	1.42	−0.50	−1.07	−1.38	−1.82
34. *o*-hydroxyanisole	1.32	1.44	1.70	1.32	0.98	0.36
35. *p*-hydroxyanisole	1.34	1.47	0.23	0.27	−0.34	−0.76
36. *o*-nitrophenol	1.79	2.18	2.35	2.32	1.91	1.40
37. *m*-nitrophenol	2.00	2.18	0.60	0.48	−0.64	−1.40
38. *p*-nitrophenol	1.91	2.01	0.20	0.17	−0.92	−2.00
39. *m*-nitrobenzoic acid	1.83	1.97	0.48	0.21	0.15	−1.22
40. *o*-aminobenzoic acid	1.21	1.43	−1.15	−0.40	−1.10	−2.12
41. *p*-aminobenzoic acid	0.83	0.88	−1.52	−1.46	−2.48	−3.74
42. *m*-nitroaniline	1.37	1.71	1.61	1.31	0.45	−0.62
43. *o*-nitroaniline	1.85	1.95	2.13	1.78	1.08	0.25
44. *p*-nitroaniline	1.39	1.48	1.23	0.93	−0.14	−0.14
45. vanillin	1.21	0.96	1.42	0.82	0.20	−0.72
46. *o*-vanillin	1.37	1.35	2.30	1.87	1.40	0.53
47. isovanillin	0.97	0.82	1.18	0.74	0.04	−0.85
48. isobutyl alcohol	0.65	0.53	0.34	−0.11	−0.32	−0.60
49. phenobarbital	1.71	1.51	0.62	−0.01	−0.63	−2.22
50. pentobarbital	2.10	1.28	1.38	0.74	−0.03	−1.30

in Table II were submitted to principal components analysis. This technique has been discussed adequately in the literature,[11] so only a brief discussion is presented here.

The objective of principal components analysis is to factor a data block, **X**, into a systematic part, **TB**, where T is a solute-specific vector and **B** is a solvent-specific vector. T corresponds to the *t*'s in eq 1, and **B** corresponds to the *b*'s. The difference in the block, **X**, and the systematic part is a residual, **E**, as shown in eq 2. The

$$\mathbf{X} - \mathbf{TB} = \mathbf{E} \qquad (2)$$

residuals **E** can be made as small as desired by increasing the number of components.

The number of columns, *A*, in T is the rank of **X**. In this application *A* is expected to be small as it is the number of significant structural features that determine log *P*. Therefore a critical aspect of the analysis is the choice of *A*. Here, we have used a cross-validation[12] procedure to determine the rank of **X**.

To illustrate how this procedure works, assume that a one-component model, *A* = 1, has been extracted from **X** and a decision must be made as to whether a second principal component is significant. According to this method, beginning for example with *A* = 1, elements of **X** are deleted and T* and **B*** are calculated for the reduced matrix. Here * is used to indicate that these are reduced matrices. T* and **B*** are used to calculate residuals for the deleted objects. This is repeated until each element of **X** is deleted once and only once. The sums of squared prediction errors of the deleted elements of **X** are calculated and compared with those from the less complex model (**A** = 1). If the ratio of prediction errors from the two models is less than 1, when corrected for differences in degrees of freedom, the second component is considered significant and added to the model. When the addition of a component to the model results in this ratio being greater than 1, the analysis is stopped and the model is adopted. Principal components analysis was carried out

[11] Joreskog, K. G.; Klovan, J. E.; Reyment, R. A. *Geological Factor Analysis*; Elsevier Scientific: Amsterdam, 1976.

[12] Wold, S. *Technometrics* **1978**, *20*, 397.

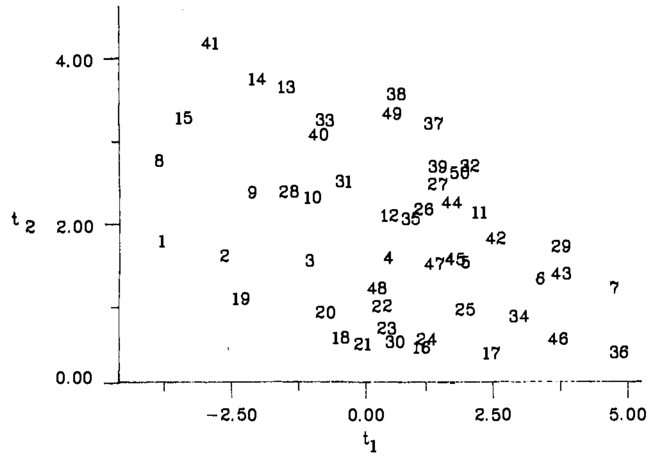

Figure 1. Principal components or eigenvector plot of the log *P* data showing that the principal components are uncorrelated.

with the SIMCA software package[13] on a desktop computer.

Solvent-accessible surface areas were calculated by using software provided by Pearlman[15] and adapted in this laboratory to run on a VAX 11/750.

Results and Discussion

The application of principal components analysis to the unscaled data in Table II revealed three principal components in the data by cross-validation. The first component accounted for 80% of the variance in the log *P* data, the second accounted for an additional 15% of the variance, and the third accounted for 2% of the variance. The principal component loadings show that the first component is approximately equally weighted in the log *P* data in all of the solvents, the second component is present in all solvents except benzene, and the third component is mainly observed in the solvent hexane. The principal components and loadings, which are normalized to unit length, are given in Tables III and IV. Due to the relatively small contribution of the third component in the partitioning process, it will be disregarded in the discussion that follows. A plot of the first two component vectors in Figure 1 shows that the components are orthogonal.

An overall standard deviation of fit, SD = 0.26, of the log *P* data to the two-term model is obtained. None of the solutes had a standard deviation twice that of the data set, and the solute *p*-hydroxybenzaldehyde (no. 31) had a standard deviation greater than 0.50. This was due to high residuals in carbon tetrachloride and hexane. None of the data were deleted from the analysis.

The results from the principal components analysis show that partitioning is determined by two features of the solute structure. One interpretation of the principal components, at this point, is that they are abstract representations of these structural features. In the next phase of the study it is necessary to attempt to identify the two features and develop methods for their estimation from structural considerations. The remainder of this paper deals with an interpretation of the first factor in terms of solute structure.

From the work of Hermann,[1,2] Yalkowsky and Valvani,[3,4] Harris et al.,[6] and Leo et al.,[7] solute size must be considered to be one of the most significant factors determining partitioning properties. This is also suggested from Figure 1. For the homologous alcohols there is a systematic increase in the value of the first principal component with carbon number. There is also a corresponding increase in log *P* associated with an increase in the first component, as would be expected if this component is a reflection of solute size.

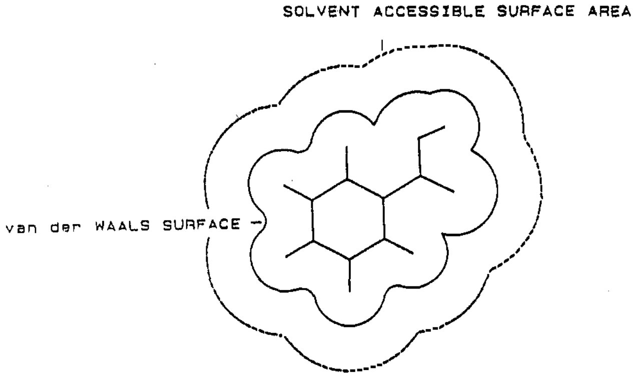

Figure 2. Van der Waals and solvent-accessible surface areas for the solute benzoic acid.

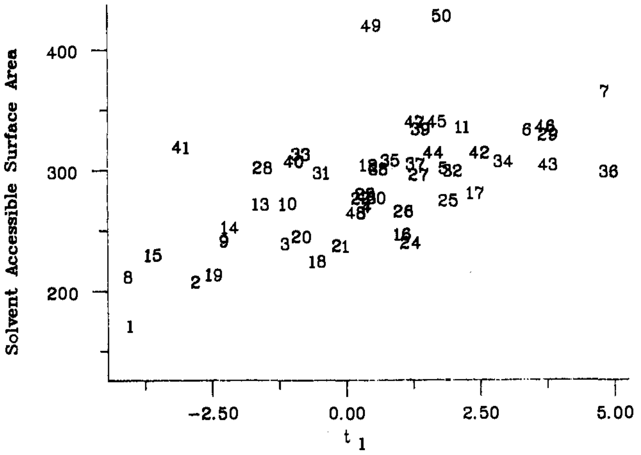

Figure 3. Plot of solvent-accessible surface area vs. the first principal component.

A number of parameters have been used as a measure of molecular size. The parameter we have used is solvent-accessible surface area first discussed by Lee and Richards.[14] These workers developed this parameter for peptides and proteins to interpret conformational properties of biopolymers. Hermann[1,2] used it as a measure of the size of the cavity formed in a solvent to accommodate a solute molecule. The solvent-accessible surface area (with the solvent water) of benzoic acid is shown in Figure 2.

Solvent-accessible surface areas, with water as solvent, were calculated for the solutes in Table II by using a program provided by Pearlman[15] and adapted in this laboratory to run on a DEC VAX 11/750. A plot of the surface areas vs. the first principal component (t_1) is given in Figure 3.

The correlation between t_1 and solvent-accessible surface area is only suggestive of a relationship. The lipophilicity of a compound is thought to be related to its nonpolar, or hydrocarbon, surface. However, the correlation between this surface and t_1 is no better than that with solvent-accessible surface area.

One pattern that appears in the plot is that unfunctionalized solutes, such as the alcohols, simple esters, and ketones, appear on the diagonal of the plot. The higher molecular weight, more highly functionalized solutes appear to have smaller t_1 values, or, alternatively, larger surface areas than would be expected if there were a

[13] Wold, S. *Pattern Recognit.* 1976, 8, 127.
[14] Lee, B.; Richards, F. M. *J. Mol. Biol.* 1971, *55*, 379.
[15] Pearlman, R. S. *QCPE Bull.* 1981, *1*, 16.

Table III. Principal Components for the Solutes and Isotropic Surface Areas (Å^2)

solute	t_1	t_2	isotropic surface area
1	-3.97	1.67	107.66
2	-2.75	1.49	142.68
3	-1.09	1.43	174.58
4	0.43	1.47	206.42
5	1.85	1.47	238.32
6	3.40	1.21	270.16
7	4.85	1.09	302.07
8	-4.02	2.65	128.25
9	-2.22	2.26	153.00
10	-1.03	2.20	190.30
11	2.21	2.02	253.95
12	0.47	1.97	222.12
13	-1.54	3.55	191.84
14	-2.09	3.65	172.00
15	-3.53	3.18	150.61
16	1.09	0.36	225.29
17	2.45	0.30	261.92
18	-0.48	0.48	194.91
19	-2.41	0.96	150.03
20	-0.78	0.79	182.74
21	-0.06	0.41	208.51
22	0.33	0.87	213.69
23	0.41	0.60	211.03
24	1.19	0.46	226.80
25	1.95	0.83	244.24
26	1.12	2.05	202.23
27	1.40	2.36	219.62
28	-1.50	2.27	161.45
29	3.80	1.60	263.81
30	0.57	0.43	147.62
31	-0.42	2.40	168.55
32	2.03	2.59	218.86
33	-0.79	3.15	157.55
34	2.98	0.75	279.52
35	0.87	1.94	210.47
36	4.93	0.30	302.25
37	1.34	3.11	202.16
38	0.60	3.47	201.80
39	1.40	2.58	221.55
40	-0.92	2.97	178.88
41	-3.02	4.09	158.27
42	2.52	1.70	244.20
43	3.80	1.27	276.32
44	1.67	2.14	244.00
45	1.73	1.45	251.36
46	3.74	0.47	272.94
47	1.32	1.39	250.83
48	0.22	1.09	203.32
49	0.51	3.23	237.90
50	1.83	2.50	244.23

functional relationship between the two parameters. This latter group of solutes contain polar substituents capable of specific solute–solvent interactions, namely, hydrogen bonding and/or hydration with the solvent water.

Work by Beveridge and co-workers[16] has shown that solvation of functionalized solutes can be specific. Monte Carlo simulation studies of the structure of dilute aqueous solutions of methane predict a coordination number of 19–23 with water at a radial distance of 5.3 Å. Formaldehyde, due to the carbonyl group, is predicted to interact differently with solvent water. Two solvent water molecules are predicted to be associated with this group, one in a position consistent with hyrogen bonding and the other in a position consistent with a dipole–dipole interaction. More recently reported Monte Carlo simulations of dilute aqueous solutions of formamide, *N*-methylacetamide, and dimethylformamide by Jorgensen and Swenson[17] predict that each amide group forms two hydrogen bonds with water at the carbonyl oxygen and that the NH_2 group of formamide has 1–2 waters attached as hydrogen-bond acceptors. Also, caging by water occurs around the hydrophobic methyl groups in *N*-methylacetamide and dimethylformamide. These results are consistent with the observed hydration state of the amide groups in proteins in the crystalline state.[18]

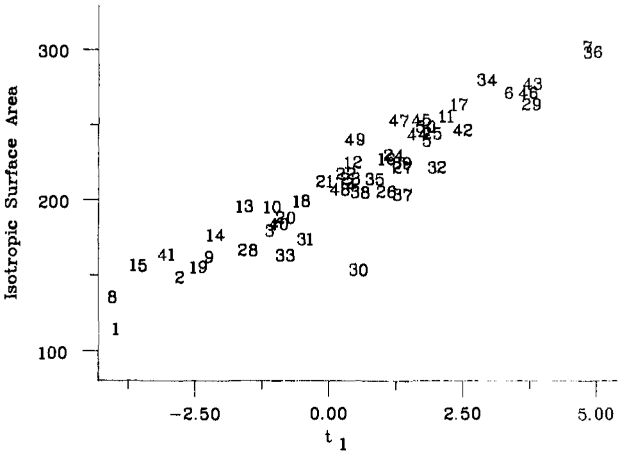

Figure 4. Plot of the isotropic surface area vs. the first principal component for the solutes.

To account for solute–solvent interactions that may result in specific hydration of functional groups, waters of hydration were added to functional groups capable of specific hydration. Hydrated solutes, or supermolecules, were generated and their solvent-accessible surface areas calculated by using geometries observed[18] and predicted[19] for hydration of groups such as the amide, amino, carboxyl, and hydroxyl. A summary of the number of waters of hydration and their geometries relative to the functional groups is given in Table V.

The hydrophobicity of a compound or a substituent is considered to be a function of the solvent-accessible surface area of nonpolar regions of the solute or substituent.[14,21] In aqueous solutions of methane, the microscopic solvent environment, while structured, is much less structured than that of solutions of formaldehyde.[14,15] This is due to stronger solvent–solvent interactions compared to solvent–solute interactions in regions of a solute's surface incapable of specific hydration. Therefore, we have assumed that the solvent-accessible surface area of each supermolecule can be partitioned into that associated with the hydrated functional groups and that due to the nonpolar, or isotropic, surface. The intersection of the two regions is taken as the boundary of the areas. The isotropic solvent-accessible surface areas of the supermolecules are given in Table III.

A plot of the isotropic surface areas vs. t_1 is given in Figure 4. It can be seen that the relationship is significant and predictive in the sense that it accounts for the major structural contribution to partitioning in the six nonpolar solvents considered.

The predictive utility of the isotropic surface area can be seen in Figure 5, in which the observed and predicted

[16] Swaminathan, B.; Whitehead, R. J.; Guth, E.; Beveridge, D. L. *J. Am. Chem. Soc.* **1977**, *99*, 7817. Swaminathan, B.; Harrison, S. W.; Beveridge, D. L. *J. Am. Chem. Soc.* **1978**, *100*, 5705.

[17] Jorgensen, W. L.; Swenson, C. J. *J. Am. Chem. Soc.* **1985**, *107*, 1489.

[18] Yang, C.; Brown, J. N.; Kopple, K. D. *Int. J. Pept. Protein Res.* **1979**, *14*, 12.

[19] Kabisch, G.; Pollmer, K. *J. Mol. Struct.* **1982**, *81*, 35.

[20] Port, G. N. J.; Pullman, A. *Int. J. Quantum Chem., Quantum Biol. Symp.* **1974**, *No. 1*, 21.

[21] Tanford, C. *The Hydrophobic Effect*, 2nd ed.; Wiley: New York, 1980.

Table IV. Principal Component Loadings[a]

	octanol	ether	chloroform	benzene	carbon tetrachloride	hexane
b_1	0.36	0.37	0.49	0.46	0.41	0.32
b_2	0.51	0.46	0.05	−0.10	−0.32	−0.64

[a] Normalized to length of 1.

Table V. Geometries of Hydrogen-Bonding Water Molecules in Hydrated Complexes

	H acceptor		H donor		
polar group	R-X-O[a]	X-O[b]	R-X-O[a]	X-O[b]	no. of waters
ROH	109.0	2.85			1
			109.0	2.85	1
RCOOH	130.0	2.85			2
			104.0	2.85	1
RCOOR′	150.0	4.09			1
RCOR′	120.0	2.85			1
RNH_2, RNHR′	109.0	2.85			1
			109.0	3.15	1
RNR′R″	109.0	2.85			1′
sp^3 N	120.0	4.50			1
ArOH	109.0	2.90			1
			104.0	2.90	1
ArCOOH	130.0	2.85			2
			104.0	2.85	1
$ArNH_2$	109.0	2.85			1
$ArCONH_2$ (O)	120.0	2.95			2
$ArCONH_2$ (N)	110.0	2.95			1
			110.0	3.15	1
$ArCOCH_3$	109.0	2.90			1
ArCHO	130.0	2.85			2
$ArNO_2$			120.0	4.51	1
$ArNH_3^+$			109.0	2.85	3
$ArCOO^-$	130.0	2.65			2
	109.0	2.65			1
barbiturate (O)	130.0	2.85			6
ring N			120.0	2.75	2

[a] Degrees. [b] Angstroms.

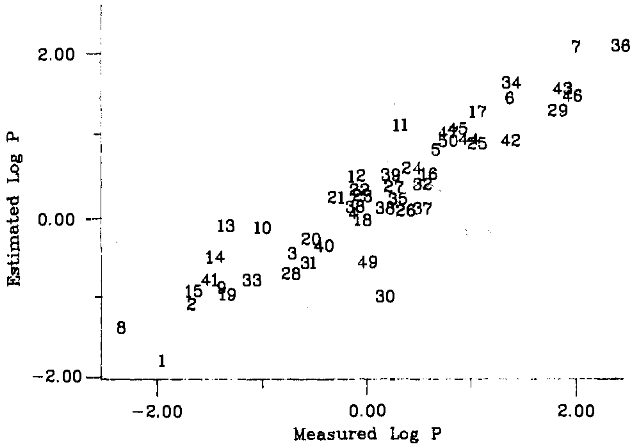

Figure 5. Estimated vs. experimental log P values for the 50 solutes in the nonpolar solvent benzene.

log P values in the nonpolar solvent, benzene, are plotted for the 50 compounds in our basis set of solutes. Similar results are obtained for the estimation of log P data for the solutes in chloroform and carbon tetrachloride since the second component is quite small in these solvent systems.

Utility of Isotropic Surface Area in QSAR Studies

The results above show that the partition coefficient is a parameter that is a linear combination of two structural features, the main one being the isotropic surface area. It has been suggested that a number of nonspecific types of biological activity of compounds can be modeled by their log P values.[24] To illustrate the utility of the isotropic surface area in modeling this type of biological activity, the isotropic surface areas of 11 steroids were calculated and correlated with their diffusion coefficients for permeability of stratum corneum.[25] The data are given in Table VI, and the results are shown in Figure 6.

Table VI. Diffusion Constants and Isotropic Surface Areas (Å^2) of Steroids

steroid	log k	isotropic surface area
1. progesterone	3.18	483.3
2. hydroxyprogesterone	2.78	422.5
3. cortexone	2.65	430.0
4. cortexolone	1.88	365.8
5. cortisone	1.00	335.1
6. cortisol	0.48	290.7
7. testosterone	2.60	412.8
8. pregnenolone	3.17	458.2
9. corticosterone	1.78	346.6
10. aldosterone	0.48	333.9
11. hydroxypregnenolone	2.78	397.5

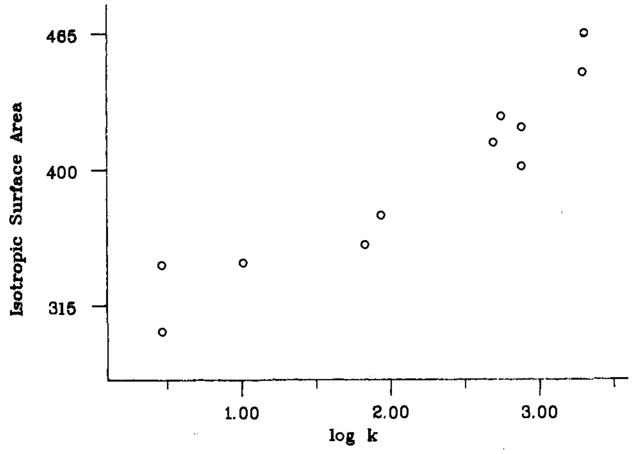

Figure 6. Plot of the isotropic surface areas for steroids vs. the logarithm of rates of diffusion, k, across stratum corneum.

The fact that the isotropic surface area can be used to model the change in diffusion data indicates that this is a nonspecific activity that is a function of one structural feature of the molecule. Further, the effect of change in structure on diffusion can be studied from one theoretically derived descriptor. Since the log P is a complex parameter, the interaction(s) that can determine biological activity may not be obvious when this variable is used in QSAR studies.

Conclusions

Principal components analysis of partitioning data for low-molecular-weight solutes in six nonpolar solvents shows that there are two major structural features of the solute that determine partitioning behavior. The major factor

(22) Huot, J.; Jolicoeur, C. In *The Chemical Physics of Solvation*; Dogonadze, R. R., Kalman, E., Kornyshev, A. A., Ulstrup, J., Ed.; Elsevier: Amsterdam, 1985; Chapter 11.
(23) Pearlman, R. S. In *Partition Coefficient: Determination and Estimation*; Dunn, W. J., III, Block, J. H., Pearlman, R. S., Eds.; Pergamon: New York, 1986; Chapter 1.
(24) Hansch, C.; Dunn, W. J., III *J. Pharm. Sci.* 1971, *61*, 1.
(25) Scheuplein, R. J.; Blank, I. H.; Brauner, G. J.; MacFarlane, D. J. *J. Invest. Dermatol.* 1969, *52*, 63.

accounts for 80% of the variation in partitioning data while the other factor accounts for an additional 15%. Solute size is identified as the major factor. This can be parameterized by what we term the isotropic surface area of a solute. The isotropic surface area is that area of the solute that interacts with the solvent water in a "nonspecific" manner.

The log P of a compound, which is considered a measure of its "lipophilicity", is proportional to the free energy of distribution of the solute between water and the nonpolar phase. The analysis carried out here deconvolutes this free energy into two components, a result consistent with cavity-based theoretical treatments of aqueous solubility and partitioning from an aqueous phaase into a nonpolar phase.[22,23] The driving force for this latter process is considered to be the increase in entropy associated with desolvation of the solute on transfer from the aqueous to the nonpolar phase.[6,21] The use of the isotropic surface area of a solute supermolecule to represent solute structure is consistent with this view.

The treatment presented considers solute hydration, and therefore intramolecular hydrogen bonding, explicitly. The number and positions of waters of hydration in the supermolecule are treated, at this point, as adjustable parameters. In order to generalize the approach, a suitable function must be developed that will theoretically determine the number and positions of hydration in a given solute. This is presently under study.

An advantage of this approach is that the isotropic surface area is a function of solute conformation. For the limited set of solutes on which this first report is based, the members are of limited flexibility. In order to fully understand the role of solute size on partitioning, it will be necessary to consider size as a function of conformation.

The second factor, while relatively small in its general contribution to partitioning, must also be identified in order to completely understand structural effects of the solute on partitioning behavior. Work on these aspects of the problem are under investigation.

Registry No. Methanol, 67-56-1; ethanol, 64-17-5; propanol, 71-23-8; butanol, 71-36-3; pentanol, 71-41-0; hexanol, 111-27-3; heptanol, 111-70-6; acetic acid, 64-19-7; propionic acid, 79-09-4; butyric acid, 107-92-6; hexanoic acid, 142-62-1; pentanoic acid, 109-52-4; trichloroacetic acid, 76-03-9; dichloroacetic acid, 79-43-6; chloroacetic acid, 79-11-8; methyl acetate, 79-20-9; ethyl acetate, 141-78-6; acetone, 67-64-1; ethylamine, 75-04-7; propylamine, 107-10-8; trimethylamine, 75-50-3; *n*-butylamine, 109-73-9; diethylamine, 109-89-7; pyridine, 110-86-1; aniline, 62-53-3; phenol, 108-95-2; benzoic acid, 65-85-0; benzamide, 55-21-0; 2-naphthol, 135-19-3; hydroquinone, 123-31-9; *p*-hydroxybenzaldehyde, 123-08-0; *o*-hydroxybenzoic acid, 69-72-7; *p*-hydroxybenzoic acid, 99-96-7; *o*-hydroxyanisole, 90-05-1; *p*-hydroxyanisole, 150-76-5; *o*-nitrophenol, 150-76-5; *m*-nitrophenol, 554-84-7; *p*-nitrophenol, 100-02-7; *m*-nitrobenzoic acid, 121-92-6; *o*-aminobenzoic acid, 118-92-3; *p*-aminobenzoic acid, 150-13-0; *m*-nitroaniline, 99-09-2; *o*-nitroaniline, 88-74-4; *p*-nitroaniline, 100-01-6; vanillin, 121-33-5; *o*-vanillin, 148-53-8; isovanillin, 621-59-0; isobutyl alcohol, 78-83-1; phenobarbital, 50-06-6; pentobarbital, 76-74-4; octanol, 111-87-5; ether, 60-29-7; chloroform, 67-66-3; benzene, 71-43-2; carbon tetrachloride, 56-23-5; hexane, 110-54-3; progesterone, 57-83-0; hydroxyprogesterone, 68-96-2; cortexone, 64-85-7; cortexolone, 152-58-9; cortisone, 53-06-5; cortisol, 50-23-7; testosterone, 58-22-0; pregnenolone, 145-13-1; corticosterone, 50-22-6; aldosterone, 52-39-1; hydroxypregnenolone, 12041-98-4; water, 7732-18-5.

Peptide Quantitative Structure–Activity Relationships, a Multivariate Approach

Sven Hellberg, Michael Sjöström,* Bert Skagerberg, and Svante Wold

Research Group for Chemometrics, Umeå University, S-901 87 Umeå, Sweden. *Received March 3, 1986*

The variation in amino acid sequence within sets of peptides is described by three principal properties, z_1, z_2, and z_3, per varied amino acid position. These principal properties are derived from a principal components analysis of a matrix of 29 physicochemical variables for the 20 coded (in mRNA) amino acids. The scales z_1, z_2, and z_3 are used to construct informative sets of analogues for exploring and developing quantitative structure–activity relationships (QSAR) of peptides. For the QSARs, the multivariate partial least squares (PLS) method is used. Multivariate QSARs are developed for four families of peptides, and it is shown how these QSARs can predict the activity of new peptide analogues.

Peptides are of central importance in all living systems. Hence, they may be considered to be the drugs of the future. In drug development, quantitative structure–activity relationships (QSARs) are essential to optimize the structure to give desired biological activities. Here we present a strategy for developing peptide QSAR.

The quantitative description of amino acids is crucial for QSARs of peptides. In a pioneering work Sneath[1] derived amino acid descriptors from qualitative (interval) data for the 20 coded amino acids. In a recent paper[2] we extended the multivariate approach of Sneath to continuous amino acid properties. The scales derived from this matrix are relevant in peptide QSAR.[3] Here we have further expanded the property matrix by including nine HPLC measurements of dansylated amino acids at different pH and eluent mixtures[4] (see Table I). The new multiproperty matrix (available as supplementary mate-

[1] Sneath, P. H. A. *J. Theoret. Biol.* **1966**, *12*, 157.

[2] Sjöström, M.; Wold, S. *J. Mol. Evol.* **1985**, *22*, 272.

[3] Hellberg, S.; Sjöström, M.; Wold, S. *Acta Chem. Scand., Ser. B* **1986**, *40*, 135.

[4] Skagerberg, B.; Sjöström, M.; Wold, S., manuscript in preparation.

[5] *The Merck Index*, 9th ed., 1977.

[6] *Handbook of Biochemistry*; CRC: Baca Raton, FL, 1968.

[7] Seydel, J. K.; Schaper, K.-J. *Chemische Struktur und biologische Aktivität von Wirkstoffen*; Verlag Chemie: Weinheim, 1979.

[8] Roberts, G. C. K.; Jardetzky, O. *Adv. Protein Chem.* **1970**, *24*, 447.

[9] Horsley, W.; Sternlicht, H.; Cohen, J. S. *J. Am. Chem. Soc.* **1970**, *92*, 680.

[10] Rosenthal, S. N.; Fendler, J. H. *Adv. Phys. Org. Chem.* **1976**, *13*, 279.

[11] Aboderin, A. A. *Int. J. Biochem.* **1971**, *2*, 537.

[12] Woese, C. R.; Drugre, D. H.; Saxinger, S. A. *Proc. Natl. Acad. Sci. U.S.A.* **1966**, *55*, 966.

[13] Jones, D. D. *J. Theor. Biol.* **1975**, *50*, 167.

[14] Wolfenden, R.; Andersson, L.; Cullis, P. M.; Southgate, C. C. B. *Biochemistry* **1981**, *20*, 849.

0022-2623/87/1830-1126$01.50/0 © 1987 American Chemical Society