Pergamon

Bioorganic & Medicinal Chemistry Letters 10 (2000) 1155–1158

BIOORGANIC & MEDICINAL CHEMISTRY LETTERS

Prediction of Drug Solubility from Monte Carlo Simulations

William L. Jorgensen [a,*] and Erin M. Duffy [b]

[a]*Department of Chemistry, Yale University, New Haven, CT 06520–8107, USA*
[b]*Central Research Division, Pfizer Inc., Groton, CT 06340, USA*

Received 28 January 2000; accepted 6 March 2000

Abstract—Monte Carlo statistical mechanics simulations have been carried out for 150 organic solutes in water. Physically significant descriptors such as the solvent-accessible surface area, numbers of hydrogen bonds, and indices for cohesive interactions in solids are correlated with pharmacologically important properties including octanol/water partition coefficient (log *P*) and aqueous solubility (log *S*). The regression equation for log *S* only requires five descriptors to provide a correlation coefficient, r^2, of 0.9 and rms error of 0.7 for the 150 solutes. The descriptors can form a basis for structural modifications to guide an analogue's properties into desired ranges.

The aqueous solubility (log *S*) and octanol/water partition coefficient (log *P*) of a drug are important factors in determining its bioavailability. Log *S* reflects the concentration *S* of the drug in mol/L for a saturated aqueous solution in equilibrium with the crystalline material, while log *P* gives the log of the concentration ratio of the drug at equilibrium partitioning between octanol and water phases. These quantities affect the ability of a drug to reach significant concentrations in the blood stream and to distribute into tissue. In view of their importance, numerous procedures have been developed for their estimation.[1–8] Most methods start with a structure drawing and have numerical increments associated with large numbers of molecular fragments. For example, the CLOGP procedure of Hansch and Leo uses more than 200 fragment and correction terms to predict log *P* values.[1]

We recently reported an alternative approach in which a Monte Carlo (MC) simulation is run for the solute in water.[9] Configurationally averaged results are obtained for physically significant quantities including the solute-water Coulomb and Lennard–Jones interaction energies, solvent-accessible surface area (SASA) and numbers of donor and acceptor hydrogen bonds. Correlations were obtained between these descriptors and gas to liquid free energies of solvation in hexadecane, octanol, and water, and log *P*. Linear regressions with only 3–4 descriptors yielded fits with correlation coefficients, r^2, of 0.9 in all cases. The regression equation for log *P* was developed using over 200 diverse compounds and only requires four descriptors to provide an rms error of 0.55, which is competitive with the best fragment-based methods. Extension of the method to log *S* is reported here using a database of 150 compounds including more than 80 drugs and related heterocycles.

Computational Methods

The computational details have been described in the earlier work.[9] Briefly, the MC calculations are performed for a single solute in a periodic cube with 500 TIP4P water[10] molecules at 25 °C and 1 atm. Each simulation consists of sampling 3.2 million configurations for equilibration and 10 million configurations during the averaging phase. The potential energy is represented by harmonic bond-stretching and angle-bending terms, a Fourier series for each dihedral angle, and Coulomb and Lennard–Jones non-bonded interactions. The parameters come from the OPLS-AA force field;[11] however, since OPLS-AA partial charges are not available for some functional groups, all partial charges are obtained from PM3 calculations using the CM1P procedure.[12] These charges, which are appropriate for the gas phase, are scaled by a factor of 1.3 for neutral molecules in the simulations to reflect the enhanced polarization in the liquid state. The TIP4P water molecules undergo only rigid-body translations and rotations, while the sampling for the solutes also covers all internal degrees of freedom. The MC calculations are run with the BOSS program[13] in an automated manner; only the atomic

*Corresponding author. Fax: +1-203-432-6299; e-mail: bill@adrik.chem.yale.edu

0960-894X/00/$ - see front matter
PII: S0960-894X(00)00172-4

numbers and a set of starting coordinates are required for the solute.

Eleven descriptors are averaged including the solute–water Coulomb (ESXC) and Lennard–Jones (ESXL) interaction energies, SASA and its hydrophobic, hydrophilic and aromatic components, and the numbers of solute as donor (HBDN) and acceptor (HBAC) hydrogen bonds.[9] Hydrogen bonds are defined using a geometric cutoff of 2.5 Å for solute H/water O and solute N, O, or S/water H distances.

Results were obtained for 150 compounds that have available experimental data for log S.[6–8] Emphasis was placed on representation by diverse structures, functionality, and drugs. The database was maintained and analyzed with the JMP program.[14] F ratios (regression model mean/error mean square) were used to establish the significance of the descriptors; the descriptors reported in the regression equations satisfy the condition that the probability of a greater F value occurring by chance (Prob $>F$) is less than 0.0001. Cross-validated r^2 values, q^2, were obtained by a leave-one-batch-out procedure using 15 batches of 10 randomly chosen compounds. The database was not split into training and test sets since this is only statistically meaningful for significantly larger data sets.

Results

Previously, we found that log P is well predicted by eq 1, where the dominant terms are the total surface area and the number of hydrogen bonds accepted by the solute. Corrections are included for the number of

$$\log P = 0.01448 \cdot \mathrm{SASA} - 0.7311 \cdot \mathrm{HBAC} - 1.064 \cdot \#\mathrm{amine} + 1.1718 \cdot \#(\mathrm{nitro} + \mathrm{acid}) - 1.772 \quad (1)$$

non-conjugated amine groups, #amine, and the total number of nitro and carboxylic acid groups, #(nitro + acid). The need for the corrections was traced to deficiencies in the CM1P charge distributions for these functional groups. Increasing size favors solvation in octanol or other organic solvents, while hydrogen-bond acceptor sites favor solvation in water.[3,9] The similar hydrogen-bond accepting ability of octanol and water eliminates the significance of a term for the number of donated hydrogen bonds (HBDN). This simple equation yielded an r^2 of 0.90, q^2 of 0.89, a rms error of 0.55, and a mean unsigned error of 0.44 log unit for the database of 200 compounds.[9]

For solubility, Yalkowsky has noted that log S correlates well with log P with an additional term involving the melting point (MP) for crystalline solutes, eq 2.[4] MP can be regarded as a gauge of

$$\log S = 0.8 - \log P - 0.01(\mathrm{MP} - 25) \quad (2)$$

cohesive interactions in the crystal such that a higher MP leads to lower solubility. Thus, we initially set out to supplement eq 1 with measures of the cohesive interactions, which could be extracted from the computed descriptors in water. Alternatively, it would be interesting to perform, for example, a gas-phase MC simulation for a solute dimer and use the average intermolecular interaction energy as a descriptor. None of the measures of the electrostatic interactions such as the Coulomb energy, ESXC, or the total number of hydrogen bonds, HBAC + HBDN, proved useful. However, ESXC/SASA is a statistically significant descriptor. It can be deemed the Coulomb tension and is large in magnitude for small, highly polar molecules, which have high melting points. This quickly led to eq 3 where SASA is replaced by the Lennard–Jones energy, which is an alternative

$$\log S = 0.3050 \cdot \mathrm{ESXL} + 0.5938 \cdot \mathrm{HBAC} + 2.055 \cdot \#\mathrm{amine} - 0.5811 \cdot \#(\mathrm{nitro} + \mathrm{acid}) + 17.18 \cdot \mathrm{ESXC/SASA} + 1.819 \quad (3)$$

measure of solute size and has a correlation coefficient of 0.93 with SASA.[9] For the data set of 150 compounds, eq 3 yields an r^2 of 0.82 and a rms error of 0.88. If SASA is used in place of ESXL, r^2 declines to 0.77. Since ESXL and ESXC are always negative numbers, both increasing size and Coulomb tension decrease solubility. Solubility is enhanced by increasing the number of hydrogen-bond acceptor sites or the saturated amine-content owing partly to protonation in water.

Analysis of the compounds with significant errors pointed especially to heteroaromatic molecules such as pyridines, pteridines, and cytosine, which have an excess of hydrogen-bond acceptor over donor sites. If the sites are not in balance and oriented properly, substantial hydrogen-bonding does not occur in the crystal. To reflect the needed balance, HBDN×HBAC was tried in place of ESXC/SASA in eq 3, but it did not improve the correlation. However, adjusting this for size with HBDN× HBAC/SASA yields an r^2 of 0.86 and rms error of 0.78. Significant outliers are then prostaglandin E2, chloramphenicol, and mannitol, which have unusually high numbers of hydrogen-bond donor and acceptor sites, and are predicted to have log S values that are too low by 2–3 units. With that many hydrogen-bonding sites, it is unlikely that they can all be satisfied simultaneously in the crystal. So, a saturating effect is expected. This can be introduced by applying a fractional power in the descriptor. We arrived at HBAC×HBDN$^{1/2}$/SASA as a reasonably simple and effective cohesive index, and the best five-descriptor equation that could be found is eq 4. The correction for carboxylic acids is

$$\log S = 0.3158 \cdot \mathrm{ESXL} + 0.6498 \cdot \mathrm{HBAC} + 2.192 \cdot \#\mathrm{amine} - 1.759 \cdot \#\mathrm{nitro} - 161.6 \cdot \mathrm{HBAC} \cdot \mathrm{HBDN}^{1/2}/\mathrm{SASA} + 1.181 \quad (4)$$

no longer significant and has been dropped. Eq 4 gives an r^2 of 0.88, q^2 of 0.87, a rms error of 0.72, and a mean unsigned error of 0.56 for the 150 compounds. Uncertainty in the experimental data makes it unlikely that predictive schemes for such diverse collections of compounds can yield rms errors below 0.5.[8] The results are recorded in Tables 1–3.

Though the data set contains predominantly solids at 25 °C and a few liquids, eq 4 works comparably well for

Table 1. Aqueous solubilities for reference organic molecules

Compds	log *S*—exptl[a]	log *S*—calcd[b]
pentane	−3.18	−2.16
hexane	−3.84	−2.46
cyclohexane	−3.10	−2.34
cyclohexene	−2.59	−2.14
pent-1-yne	−1.64	−1.10
1,1,1-trichloroethane	−2.00	−2.00
1-chloropropane	−1.47	−0.84
1,2-dichloroethane	−1.06	−0.53
trichloroethene	−1.96	−2.01
benzene	−1.64	−1.54
toluene	−2.21	−2.15
hexamethylbenzene	−5.23	−4.46
naphthalene	−3.60	−3.21
fluorene	−5.00	−4.33
pyrene	−6.18	−5.27
biphenyl	−4.35	−4.08
2,3,4,5,6-pentachlorobiphenyl	−7.78	−6.64
perchlorobiphenyl	−10.8	−9.20
fluorobenzene	−1.80	−1.73
chlorobenzene	−2.38	−1.82
bromobenzene	−2.55	−2.62
trifluoromethylbenzene	−2.51	−2.58
methanol	1.56	0.55
ethanol	1.10	0.29
1-propanol	0.62	−0.05
2-propanol	0.43	−0.18
2-methyl-2-propanol	0.63	−0.45
phenol	0.00	−1.02
p-cresol	−0.73	−1.39
2,3-dichlorophenol	−1.30	−2.27
p-*t*-butylphenol	−2.41	−2.72
2-naphthol	−2.28	−2.25
diethyl ether	−0.09	−0.09
tetrahydrofuran	0.49	0.14
dimethoxymethane	0.48	0.62
anisole	−1.85	−2.00
propanal	0.58	0.31
benzaldehyde	−1.19	−0.87
butanone	0.52	0.20
acetophenone	−1.28	−1.37
succinic acid	−0.19	−0.28
benzoic acid	−1.55	−0.90
m-nitrobenzoic acid	−1.68	−2.14
methyl acetate	0.46	0.63
methyl butyrate	−0.82	−0.58
methyl benzoate	−1.85	−1.58
ethyl acetate	−0.04	0.13
ethylamine	2.06	2.25
trimethylamine	1.32	1.13
aniline	−0.41	−1.30
p-chloroaniline	−1.66	−1.85
benzidine	−2.70	−3.07
propionitrile	0.28	0.85
benzonitrile	−1.00	−1.12
acetamide	1.58	−0.12
N,N-dimethylacetamide	1.11	0.09
benzamide	−0.96	−1.18
acetanilide	−1.33	−1.58
p-chloroacetanilide	−2.84	−2.23
nitromethane	0.26	0.55
nitroethane	−0.22	0.03
nitrobenzene	−1.80	−0.94
dimethylsulfide	−0.45	−0.62
p-toluenesulfonamide	−1.74	−0.44
morpholine	1.97	1.92

[a]Refs 6–8.
[b]Eq 4.

Table 2. Aqueous solubilities for drugs and drug-like molecules

Compds	log *S*— exptl[a]	log *S*—calc[b]
acetaminophen	−1.02	−1.60
alanine	0.25	1.38
allopurinol	−2.26	−1.05
aspirin	−1.72	−1.21
atropine	−2.12	−2.03
barbital	−1.42	−1.97
benzocaine	−2.32	−2.64
bifonazole	−5.95	−5.67
bromazepam	−3.48	−4.45
caffeine	−0.88	0.08
chloramphenicol	−1.94	−3.74
chlorpromazine	−5.10	−5.41
cocaine	−2.25	−1.51
codeine	−1.52	−1.98
corticosterone	−3.24	−4.62
desipramine	−3.66	−3.99
dexamethasone	−3.59	−4.50
diazepam	−3.75	−3.98
diethylstilbestrol	−4.07	−4.39
dimethylbarbiturate	−1.74	−1.08
ephedrine	−0.47	−0.72
estradiol	−5.03	−4.71
ethyl-*p*-hydroxybenzoate	−2.35	−1.94
fenbufen	−5.30	−4.08
fenclofenac	−3.85	−4.38
fluconazole	−1.80	−2.35
flurbiprofen	−3.74	−3.62
griseofulvin	−4.07	−2.31
hydrocortisone	−3.09	−4.08
ibuprofen	−3.76	−3.35
imipramine	−4.19	−4.83
indomethacin	−4.62	−4.81
indoprofen	−4.82	−4.22
ketoprofen	−3.16	−3.35
lidocaine	−1.71	−2.22
lorazepam	−3.60	−3.65
mannitol	0.06	−1.25
morphine	−3.28	−2.76
naproxen	−4.20	−2.98
nevirapine	−3.19[c]	−4.29
2-methyl-nevirapine	−4.27[c]	−4.38
nevirapine analogue 1	−5.15[c]	−4.47
nevirapine analogue 12	−2.62[c]	−3.06
nifedipine	−4.76	−4.01
nifuroxime (Z)	−2.19	−2.46
nitrofurantoin	−3.38	−3.01
oxazepam	−3.95	−3.72
perphenazine	−4.16	−3.90
phenacetin	−2.35	−2.49
phenobarbital	−2.29	−2.10
phenytoin	−3.99	−3.13
prednisone	−3.48	−2.93
procaine	−1.78	−2.24
progesterone	−4.42	−4.54
promazine	−4.30	−4.40
prostaglandin E2	−2.47	−4.01
salicylic acid	−1.82	−0.91
sulindac	−5.00	−3.98
testosterone	−4.02	−5.02
theophylline	−1.39	−0.86
thioridazine	−5.82	−6.20
trifluoperazine	−4.52	−4.23
triflupromazine	−5.30	−5.39
warfarin	−4.26	−5.28

[a]Refs 6–8.
[b]Eq 4.
[c]Ref 15.

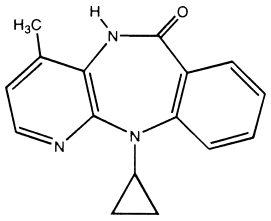
Table 3. Aqueous solubilities for heterocycles, pesticides, and herbicides

Compds	log *S*—exptl[a]	log *S*—calc[b]
adenine	−2.18	−1.22
guanine	−3.58	−1.99
uracil	−1.49	−0.71
cytosine	−1.16	−0.91
pyridine	0.76	−0.48
quinoline	−1.30	−2.67
7-methylpteridine	−0.85	−1.80
indole	−1.21	−2.37
furan	−0.82	−0.56
dibenzofuran	−4.60	−3.89
thiophene	−1.33	−1.54
atrazine	−3.55	−3.02
camphor	−1.96	−2.24
carvone	−2.06	−2.22
DDT	−7.15	−7.38
desmedipham	−4.63	−4.71
diuron	−3.05	−3.81
fenoxycarb	−4.70	−4.88
limonene	−4.00	−4.02
lindane	−4.60	−5.16
menthone	−2.35	−2.60

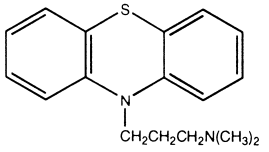
[a]Refs 6–8.
[b]Eq 4.

gases because they generally have no hydrogen-bonding groups and the cohesive term is zero in eq 4. Thus, all lower alkanes are predicted to be too soluble by ca. 1 log unit. Alkanes seem to be effectively larger in aqueous solution than expected, possibly associated with formation of clathrate-like water structures.[9] Similarly, PCBs and polycyclic aromatic hydrocarbons are predicted to be too soluble. ESXL is the only non-zero descriptor for these molecules. A separate fit for the 23 non-polar molecules, log $S = 0.3489 \cdot ESXL + 1.142$, gives an r^2 of 0.96 and rms error of 0.49. Note the larger coefficient for ESXL. With eq 4, the largest negative error is for the antifungal griseofulvin. It is an unusual drug because it has no hydrogen-bond donor groups, but HBAC = 7.1; the cohesive index is then zero and it is predicted to be too soluble by 1.76 log units. The results for PGE2, mannitol, and chloramphenicol still err in the opposite direction owing to their large values for the cohesive term. For more accurate predictions for polyols, the number of saturated alcohol groups can be added to eq 4. This descriptor is fully significant and brings r^2 to 0.90 and the rms error to 0.66.

Among specific series, four nevirapine (**1**) analogues were considered. Their log *S* values are predicted in the correct order, though the range is compressed from the observed[15] 2.5 to 1.2 log units. Furthermore, the results for promazine (**2**) analogues are particularly good; the errors are all less than 0.4 for promazine, chlorpromazine, perphenazine, thioridazine, trifluoperazine, and triflupromazine.



1



2

In summary, log *P* and log *S* can both be predicted well using regression equations with only four or five descriptors. The descriptors are extracted from routine MC simulations for the solute in water and correspond to easily interpreted quantities. They suggest changes that can be made in a structure to guide an analogue's properties into a desired range. The current method is applicable to any neutral molecule with atoms having PM3 parameters, (i.e., H, C, N, O, F, Al, Si, P, S, Cl, Br, and I). Improvements are possible through the addition of new descriptors, performance of simulations in different media, and use of alternative partial charges. The descriptors can also be applied to develop correlations for other properties or for refined analyses of narrower classes of compounds.

Acknowledgements

Gratititude is expressed to the National Science Foundation for support of this research.

References and Notes

1. Hansch, C.; Leo, A. *Exploring QSAR—Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.
2. Sangster, J. *Octanol–Water Partition Coefficients: Fundamentals and Physical Chemistry*; Wiley: Chichester, 1997.
3. Buchwald, P.; Bodor, N. *Curr. Med. Chem.* **1998**, *5*, 353.
4. Yalkowsky, S. H. *Solubility and Solubilization in Aqueous Media*; Oxford University: Oxford, 1999.
5. Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1.
6. Huuskonen, J.; Salo, M.; Taskinen, J. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 450.
7. Mitchell, B. E.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 489.
8. Abraham, M. H.; Le, J. *J. Pharm. Sci.* **1999**, *89*, 868.
9. Duffy, E. M.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2000**, 122, 2878.
10. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
11. Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225.
12. Storer, J. W.; Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. *J. Comput. Aided Mol. Des.* **1995**, *9*, 87.
13. Jorgensen, W. L. *BOSS Version 4.2*; Yale University: New Haven, CT, 2000.
14. SAS. *JMP Version 3*; SAS Institute Inc.: Cary, NC, 1995.
15. Morelock, M. M.; Choi, L. L.; Bell, G. L.; Wright, J. L. *J. Pharm. Sci.* **1994**, *83*, 948.