Prediction of Pharmaceutically Important Properties 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
E-mail: william.jorgensen@yale.edu; eduffy@achillion.com

Received: 12th June 2000 / Published 11th May 2001

Abstract

Monte Carlo statistical mechanics simulations have been carried out for more than 250 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 the octanol/water partition coefficient (log P), aqueous solubility (log S), and brain/blood concentration ratio (log BB). The regression equations for log P and log S only require 4 - 5 descriptors to provide correlation coefficients, r^2, of 0.9 and rms errors of 0.7. The descriptors can form a basis for structural modifications to guide an analog's properties into desired ranges. For more rapid application, a program that estimates the significant descriptors, QikProp, has been created. It can be used to predict the properties for ca. 1 compound/sec. with no loss of accuracy.

Introduction

The aqueous solubility (log S), octanol/water partition coefficient (log P), and brain/blood concentration ratio (log BB) 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 and log BB give the log of the concentration ratio of the drug at equilibrium partitioning between octanol and water phases or brain and blood. 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. [9a] 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

Published in "Chemical Data Analysis in the Large: The Challenge of the Automation Age", Martin G. Hicks (Ed.), Proceedings of the Beilstein-Institut Workshop, May 22nd – 26th, 2000, Bozen, Italy

http://www.beilstein-institut.de/bozen2000/proceedings/jorgensen/jorgensen.pdf

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 was reported using a database of 150 compounds including more than 80 drugs and related heterocycles. [9b] A more rapid procedure, QikProp, has been developed, which uses algorithms to estimate the significant descriptors including the hydrogen-bond counts. No degradation in the quality of the results is found in comparison to the full simulation results. Its application is illustrated here including for predictions of log BB.

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 [10a] 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; [10b] 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. [11] 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 [12] in an automated manner; only the atomic numbers and a set of starting coordinates are required for the solute.

Twelve descriptors are averaged including the solute-water Coulomb (ESXC) and Lennard-Jones (ESXL) interaction energies, the number of freely rotatable bonds other than for CXYZ (X,Y,Z = H, halogen) groups, 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 more than 250 compounds for log P, [9a] 150 compounds for log S, [9b] and 61 compounds for log BB, [13] that have available experimental data. [1-8,13] 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.

Monte Carlo Results

From the Monte Carlo simulations, it was 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 non-conjugated amine groups, #amine, and the total number of nitro and carboxylic acid groups, #(nitro+acid).

log P = 0.01448•SASA – 0.7311•HBAC – 1.064•#amine + 1.1718•#(nitro+acid) –1.772 **(1)**

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 250 compounds.

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 cohesive interactions in the crystal such that a higher MP leads to lower solubility.

log S = 0.8 – log P –0.01(MP – 25) **(2)**

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. 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. Augmentation of eq. (1) with this term led to an equation that yields an r^2 of 0.82 and a rms error of 0.88. However, 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 x HBAC was tried in place of ESXC/SASA, but it did not improve the correlation. However, adjusting this for size with HBDN x 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 x HBDN^{½}/SASA as a reasonably simple and effective cohesive index, and the best five-descriptor equation that could be found is eq. (3). The correction for carboxylic acids is no longer significant and has been dropped.

log S = 0.3158•ESXL + 0.6498•HBAC +2.192•#amine – 1.759•#nitro – 161.6•HBAC• HBDN^{½}/SASA + 1.181 **(3)**

Eq. (3) 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]

QikProp Results

With QikProp, the same descriptors are found to be the most significant as from the Monte Carlo simulations. However, the solute-water Coulomb and Lennard-Jones energies are no longer available, and it is often found that a somewhat larger number of descriptors, ca. 8, are found to be fully significant from the F ratios. For log P, the regression equation for 270 compounds yields an r^2 of 0.92 and rms error of 0.55. For log S, the corresponding figures for 190 compounds are 0.88 and 0.69. And, for log BB, eq. (4) has an r^2 of 0.84 and rms error of 0.31 with the dataset of 61 compounds.

$$\textbf{log BB} = 0.001300 \bullet \text{FOSA} - 0.004332 \bullet \text{FISA} + 0.6337 \bullet \#\text{amine} - 0.0751 \bullet \mu - 0.1369 \bullet \#\text{rotor} + 0.04192 \quad (4)$$

There are only five significant descriptors; hydrophobic surface area and non-conjugated amines increase the brain concentration, while increased polarity, as reflected in the hydrophilic surface area and dipole moment, and flexibility increase the concentration of the compound in blood. The results are illustrated in Figure 1.

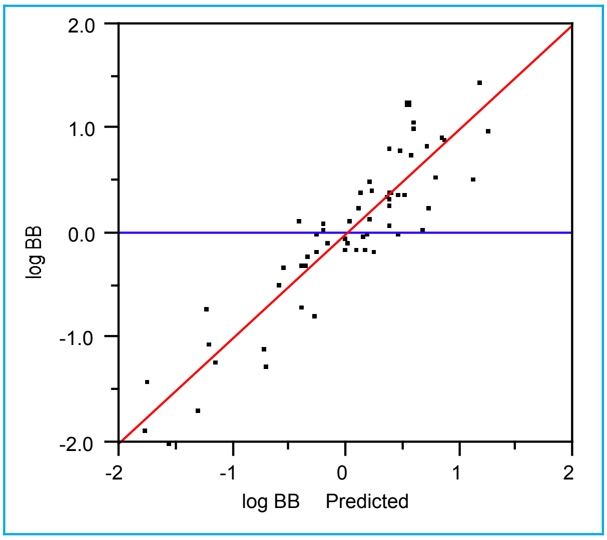

Figure 1: QikProp vs experimental results for log BB.

Conclusion

In summary, log P, log S, and log BB can be predicted well using regression equations with only 4-8 descriptors. The descriptors correspond to easily interpreted quantities. They suggest changes that can be made in a structure to guide an analog's properties into a desired range. The current methods are 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.

Acknowldegments

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, 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*, 353-380.

[4] Yalkowsky, S. H. "Solubility and Solubilization in Aqueous Media", Oxford University Press: Oxford, 1999.

[5] Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1-18.

[6] Huuskonen, J.; Salo, M.; Taskinen, J. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 450-456.

[7] Mitchell, B. E.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 489-496.

[8] Abraham, M. H.; Le, J. *J. Pharm. Sci.* **1999**, *89*, 868-880.

[9] (a) Duffy, E. M.; Jorgensen, W. L. *J. Am. Chem. Soc.* **2000**, *122*, 2878-2888. (b) Jorgensen, W. L.; Duffy, E. M. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 1175-1178.

[10] (a) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926-935. (b) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225-11236.

[11] Storer, J. W.; Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. *J. Comput. Aided Mol. Design* **1995**, *9*, 87-110.

[12] Jorgensen, W. L. *BOSS Version 4.2*; Yale University: New Haven, CT, 2000.

[13] (a) Luco, J. M. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 396-404. (b) Kelder, J.; Grootenhuis, P. D. J.; Bayada, D. M.; Delbressine, L. P. C.; Ploemerr, J.-P. *Pharm.*

Res. **1999**, *16*, 1514-1519.

[14] *JMP Version 3*; SAS Institute Inc., Cary, NC, 1995.