Prediction of Polar Surface Area and Drug Transport Processes Using Simple Parameters and PLS Statistics

Thomas Österberg*,† and Ulf Norinder‡

Pharmaceutical Sciences R&D, Pharmacia Corporation, SE-112 87 Stockholm, Sweden, and AstraZeneca R&D Södertälje, SE-151 85 Södertälje, Sweden

Received June 17, 2000

Modeling of the calculated polar surface area of drugs with rapidly derived descriptors (i.e., the number of hydrogen bonds accepting oxygen and nitrogen atoms and the number of hydrogen atoms bonded to these) using partial least squares projection to latent structures (PLS) analysis is described. The statistical analysis showed strong relationships between the hydrogen-bonding descriptors and the calculated polar surface area of five chemically diverse sets of drugs ($R^2 > 0.93$ and $Q^2 > 0.69$, $n = 11$, 20, 45, 70, and 74, respectively). The statistical models (using H-bonding descriptors and log *P*) of transport across Caco-2 cells ($n = 11$), brain−blood partitioning (two data sets, $n = 45$ and 70) and percent intestinal absorption ($n = 20$) showed $R^2 = 0.92$, 0.72, 0.76, and 0.81 and $Q^2 = 0.74$, 0.75, 0.71, and 0.73, respectively. The inclusion of log *P* improved two models, had no effect on one model, and had a slightly negative impact on one model. The combination of H-bonding descriptors with log *P* is similar to the Lipinski "rule-of-five" mnemonic. However, by using a multivariate statistical method (e.g., PLS), the prediction becomes quantitative instead of qualitative. Good statistical models were derived which permit fast computational screening and prioritization of virtual compound libraries.

1. INTRODUCTION

Several authors have advocated using the calculated polar surface area (PSA) for the prediction of various drug transport processes.[1−7] The PSA is generally defined as the area associated with nitrogen and oxygen atoms and the hydrogen atoms bonded to these. Sulfur atoms and hydrogen atoms attached to sulfur may also be included in the PSA. The PSA is dependent on the conformation and possible internal hydrogen bonding. The simplest method utilizes a single low-energy conformer of the molecule to calculate the PSA. The dynamic PSA is a more sophisticated method taking a Boltzmann weighted average of all conformers within 2.5 kcal mol^{-1} of the lowest energy conformer found during a conformational search.[1] However, since the PSA calculation utilizes the molecular area occupied by the nitrogen, oxygen, and hydrogen attached to these atoms, it is interesting to study how well the PSA could be modeled by the number of these atoms in a molecule. The main aim of the study was to model PSA using hydrogen-bonding descriptors (i.e., the number of hydrogen bond accepting oxygen and nitrogen atoms and the number of hydrogen atoms bonded to these) using partial least squares projection to latent structures (PLS) analysis. A secondary aim was to study whether the prediction of drug transport processes could be improved by the combination of the hydrogen-bonding descriptors and calculated log *P*.

2. METHOD OF CALCULATION

2.1. Human Intestinal Absorption Data. The experimental values for the dynamic PSA and the human intestinal absorption data (data set 1) for 20 drugs were taken from Palm et al.[1] The experimental values for the static PSA (data set 2) for 74 drugs were taken from Clark.[4] The intestinal absorption values (% Abs) were converted with the logit transformation: logit Abs = ln[(Abs − c1)/(c2 − Abs)], where c1 = 0.2 (lower asymptotic value) and c2 = 101 (higher asymptotic limit).

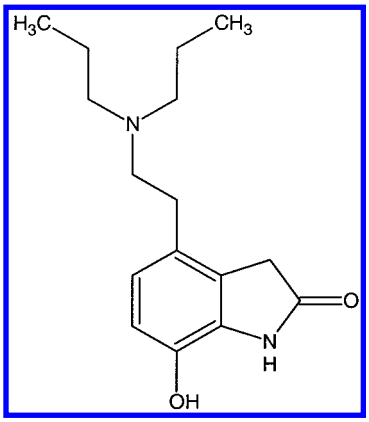

Figure 1. Chemical structure of SKF 89 124.

2.2. Caco-2 Cell Permeability Data. The calculated polar water accessible surface area (PWASA) and the experimental apparent permeability values (log P_{app}) for 11 β-adrenoceptor blockers were taken from Krarup et al.[6]

2.3. Brain−Blood Partitioning Data. The calculated static PSA and the experimental values for the ratio of the brain−blood partition coefficient (BB) for 70 compounds (data set 1) were taken from Clark.[3] The calculated dynamic PSA and the experimental values for the ratio of the BB for 45 compounds (data set 2) were taken from Kelder et al.[5] BB is defined as (C_{brain}/C_{blood}).

* Corresponding author. Phone: + 46 8-695 42 63. FAX: + 46 8-695 41 29. E-mail: thomas.osterberg@eu.pnu.com.
† Pharmacia Corporation.
‡ AstraZeneca.

10.1021/ci000065l CCC: $19.00
Published on Web 10/12/2000

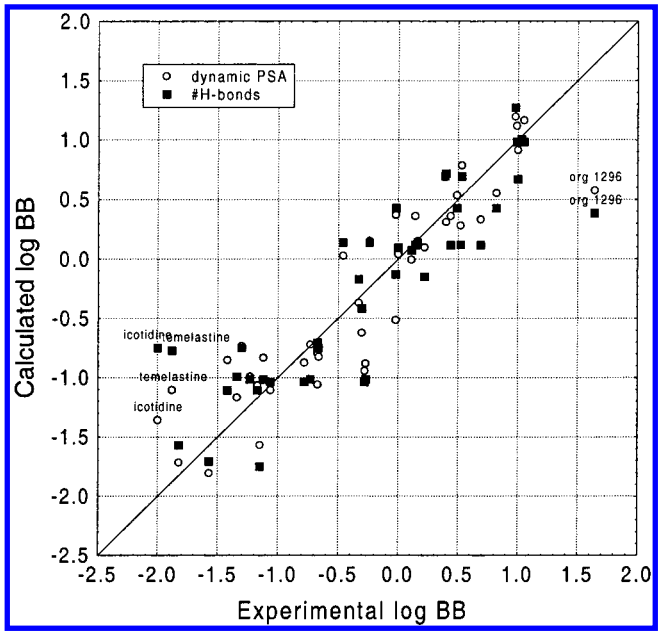

Figure 2. Comparisons between predictions of log BB by the dynamic PSA method and PLS model using H-bonding descriptors. Note that both methods underestimate log BB of compound org 1296 and overestimate log BB of icotidine and temelastine.

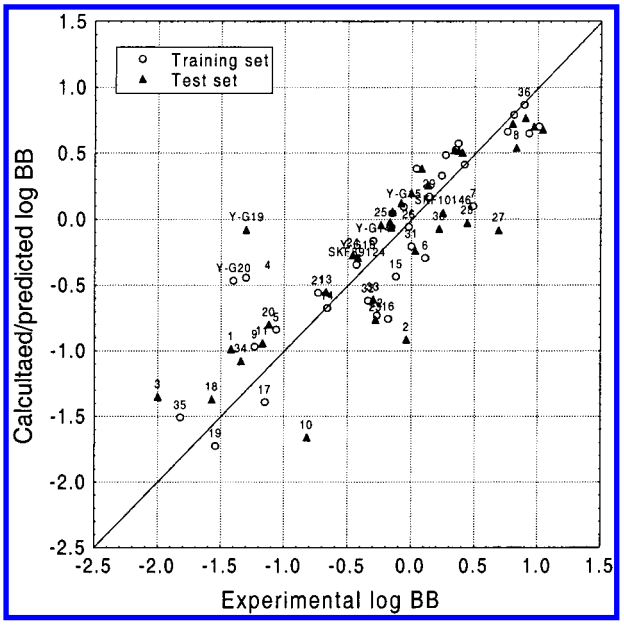

Figure 3. Experimental versus calculated/predicted log BB. Compounds numbered as in ref 3.

2.4. Designation of Hydrogen-Bonding Atoms and Calculation of log *P*. The number of nitrogen and oxygen atoms in the chemical formula was taken as the number of acceptor nitrogen (#HBAn) and oxygen atoms (#HBAo), respectively. The number of hydrogen bond donor atoms was the sum of hydrogen atoms bound to oxygen and to nitrogen atoms (#HBD). As an example, SKF 89 124, $C_{16}H_{24}N_2O_2$ (Figure 1), included in BB data set 1, has two hydrogen bond accepting oxygens, two hydrogen bond accepting nitrogens, and two hydrogen bond donors.

The software ACD/logP[8] was used to calculate the octanol/water partitioning coefficient (log *P*).

2.5. Statistical Analysis. The relationships between the response variables (e.g., PSA, brain−blood partitioning) and the calculated descriptors were determined using the PLS method. The PLS analyses were performed with SIMCA-P, 7.0, Umetrics, Sweden, using default settings. Four models were derived: (I) calculated PSA using the hydrogen-bonding parameters (II) experimental transport data (e.g., brain−blood partitioning) using the hydrogen-bonding parameters, and (III) experimental transport data using the hydrogen-bonding parameters and calculated log *P*. To assess the external predictive ability of the method, two data sets (BB data sets 1 and 2) were divided into a training set and a test set (model IV). The compounds were arranged in increasing order of the response variable (log BB), and every second compound was selected as the training set, the remaining compounds being used as the test set.

3. RESULTS AND DISCUSSION

Strong linear relationships between the calculated static PSA and the calculated dynamic polar surface have been reported,[4,5] ($R^2 = 0.96$, $n = 45$ and $R^2 = 0.99$, $n = 20$). On the basis of this fact, there seems to be little reason to use the computationally intensive calculation of the dynamic PSA in comparison to the static PSA. However, we found that further simplification was possible by using only the number of hydrogen bond forming atoms and PLS statistics. The PLS analysis (see Table 2) showed that the PSA values of the five data sets could be modeled surprisingly well ($R^2 > 0.93$ and $Q^2 > 0.69$).

The results from statistical modeling of the transport of β-adrenoceptor blockers across Caco-2 cell monolayers with the hydrogen-bonding parameters and log *P* were very good ($R^2 = 0.92$ and $Q^2 = 0.74$) and similar to the relationship obtained with the PWASA parameter ($R^2 = 0.98$).[6]

The statistical modeling of the brain−blood partitioning (log BB) with the hydrogen-bonding parameters only showed fair results, with $R^2 = 0.65$ and $Q^2 = 0.64$ (data set 1) and $R^2 = 0.76$ and $Q^2 = 0.75$ (data set 2). The relationships between the experimental log BB versus the calculated log BB obtained with the dynamic PSA values or the PLS model using hydrogen-bonding parameters are shown in Figure 2. It can be readily observed that both methods underestimate the brain−blood partitioning of the compound org 1296 and overestimate the partitioning of icotidine and temelastine. However, the model based on the hydrogen-bonding parameters shows a much higher overestimate of temelastine and icotidine compared to the PSA method. A possible explanation for the overestimate of icotidine and temelastine is that these compounds may be subjected to an efflux mechanism (e.g., P-glycoprotein) in the blood−brain barrier. The combination of the hydrogen-bonding parameters and log *P* improved the statistics of BB data set 1 ($R^2 = 0.76$ and $Q^2 = 0.75$), but not of BB data set 2 ($R^2 = 0.72$ and $Q^2 = 0.71$). This is probably explained by the fact that data set 1 contains many small organic solutes, which generally show very strong linear relationships with log *P*.

To estimate the external predictive ability of the present method (hydrogen-bonding descriptors and log *P*), two data sets (BB data sets 1 and 2) were divided into a training set and test set. The results of the statistical analyses (see Table 2) gave RMSEP values of 0.37 and 0.50, respectively. A plot of experimental versus calculated/predicted brain−blood partitioning coefficients (data set 2) is shown in Figure 3. It may be concluded that the present method seem to be

Table 1. Calculated log *P* (ACD/log *P*) and Hydrogen-Bonding Descriptors

no.	descriptor	designation
1	calculated octanol/water partition coefficient	log *P*
2	number of hydrogen bond acceptor nitrogen atoms	#HBAn
3	number of hydrogen bond acceptor oxygen atoms	#HBAo
4	number of hydrogen bond donor atoms on nitrogen and oxygen atoms	#HBD

Table 2. PLS Statistics of the Derived Models[a]

model	N^{tr}	N^{te}	*Npc*	R^2	Q^2	$RMSE^{tr}$	$RMSEP^{te}$	constant	#HBAo[b]	#HBAn[b]	#HBD[b]	log *P*[b]
Caco-2 cells												
PWASA (I)	11	–	2	0.936	0.694	7.115	–	2.358	0.455	0.096	0.948	–
log P_{app} (II)	11	–	2	0.920	0.656	0.124	–	−10.716	−0.509	−0.024	−0.945	–
log P_{app} (III)	11	–	2	0.920	0.742	0.124	–	−10.716	−0.515	0.033	−0.971	−0.027
BB (data set 1)												
PSA (I)	70		1	0.963	0.958	7.539	–	1.157	0.261	0.489	0.441	–
log BB (II)	69[c]	–	1	0.649	0.640	0.453	–	−0.235	−0.238	−0.382	−0.365	–
log BB (III)	69[c]	–	1	0.756	0.746	0.378	–	−0.235	−0.231	−0.370	−0.354	0.253
log BB (IV)	35	34[c]	1	0.714	0.682	0.423	0.370	−0.228	−0.210	−0.363	−0.362	0.252
BB (data set 2)												
PSA dynamic (I)	45			0.940	0.936	6.414	–	1.889	0.340	0.541	0.468	–
log BB (II)	45		1	0.758	0.747	0.452	–	−0.267	−0.310	−0.487	−0.417	–
log BB (III)	45		1	0.720	0.707	0.486	–	−0.267	−0.226	−0.354	−0.304	0.286
log BB (IV)	23	22	1	0.721	0.684	0.518	0.502	−0.252	−0.218	−0.318	−0.356	0.247
intestinal absorption (data set 1)												
PSA dynamic (I)	20		2	0.936	0.901	14.39	–	1.675	0.580	0.259	0.505	–
logit Abs (%) (II)	20		2	0.806	0.734	1.497	–	0.145	−0.532	−0.214	−0.466	–
logit Abs (%) (III)	20		2	0.811	0.732	1.478	–	0.145	−0.521	−0.237	−0.447	0.069
intestinal absorption (data set 2)												
PSA (I)	74		2	0.931	0.904	12.32	–	2.029	0.731	0.340	0.298	–

[a] N^{tr}, number of compounds in the training set; N^{te}, number of compounds in the test set; Npc, number of PLS components; R^2, ordinary correlation coefficient; Q^2, cross-validated correlation coefficient; $RMSE^{tr}$, root-mean-square error for the dependent variable of the training set; $RMSEP^{te}$, root-mean-square error for the dependent variable of the test set. [b] Scaled and mean centered regression coefficient. [c] Compound **12** excluded.

predictive from both an internal (cross-validation) and an external point of view with respect to the predictions of the test sets.

The modeling of intestinal absorption (Abs%) was performed with a logit transformation of the response variable, as described previously.[9] In brief, the variable (Abs%) has a closed scale (0−100% absorption), and sigmoidal curves often result from analysis of absorption data. By using a logit transformation (see 2.1), the nonlinear behavior of the Abs% variable is reduced. The results of modeling intestinal absorption (data set 1) showed a fair linear relationship ($R^2 = 0.81$ and $Q^2 = 0.73$). It is noteworthy that a model without log *P* showed a similar relationship ($R^2 = 0.81$). Data set 2 contains many compounds that are subjected to active uptake or efflux, and the Abs% was therefore not modeled. However, since the modeling of PSA for this data set was good ($R^2 = 0.93$ and $Q^2 = 0.90$), the modeling of intestinal absorption would probably be similar to the results reported by Clark.[4]

Inspection of the R^2 and Q^2 values (Table 2) for models with and without log *P* show that log *P* had a positive effect on the Caco-2 and BB data set 1, no effect on intestinal absorption (data set 1), and a slightly negative effect on BB data set 2. Inclusion of the descriptor molecular mass improved the statistics of the Caco-2 and log BB models, but worsened the model of intestinal absorption (data not shown).

It has been shown[10] that log *P* values should be considered to be composed of two factors, namely a steric and a polar contribution:

$$\log P = aV - \Lambda$$

where *V* is the molar volume and Λ accounts for the polarity of the molecule including H-bonding capacity. We have previously reported that models of drug transport using "noncomposite" lipophilicity terms, i.e., polarizability and/or polarity terms, a surface term encompassing information regarding the size of the molecule, and various hydrogen-bonding terms, were among the best models with respect to both internal and external predictive ability.[11] However, since log *P* has a strong tradition in the modeling of drug transport processes, it was of interest to study its importance in the modeling of the data sets included in this study.

The Lipinski[12] rule-of-five concept has been implemented by many drug companies as an early alert system for compounds with absorption problems.

The rule-of-five mnemonic states that poor absorption or permeation is more likely when the molecular mass is over 500, there are more than five H-bond donors, the sum of N's and O's exceeds 10, and the MLOGP is over 4.15 (or CLOGP is over 5). Substrates for biological transporters are exceptions.

If two of the conditions specified above are met, the compound is flagged. The present method combines some of these parameters, but by utilizing PLS analysis a qualitative prediction is obtained.

What are the advantages of models based on the simple hydrogen-bonding parameters and calculated log *P* over models based on a static or dynamic PSA? To answer this question, one may, in our opinion, divide the purpose of deriving mathematical models for drug transport processes into two categories: (i) more advanced purposes aimed at understanding the mechanisms of a particular process; (ii)

pragmatic electronic MTS or HTS screening methods aimed at prioritizing medium to large virtual libraries.

For the first case, more advanced computational methods taking the dynamic behavior of the structures into account may be needed, while for the second case speed and ease of computation are of primary importance. Thus, for the second case, the combination of the hydrogen-bonding parameters and calculated log *P* represents descriptors that are simple and fast to compute as well as being interpretable from a chemical point of view. Furthermore, the descriptors are derived from two-dimensional molecular graphs without the necessity of three-dimensional geometry generation and associated complications with conformational analysis.

We believe that these simple parameters, in combination with PLS statistics, are useful for deriving predictive quantitative structure−property relationships for drug transport processes and that these models permit electronic screening of large collections of virtual structures in order to detect compounds with potential drug transport problems at an early stage of the preclinical phases of drug development.

REFERENCES AND NOTES

(1) Palm, K.; Luthman, K.; Ungell, A. L.; Strandlund, G.; Artursson, P. Correlation of drug absorption with molecular surface properties. *J. Pharm. Sci.* **1996**, *85*, 32−39.

(2) Palm, K.; Stenberg, P.; Luthman, K.; Artursson, P. Polar molecular surface properties predict the intestinal absorption of drugs in humans. *Pharm. Res.* **1997**, *14*, 568−571.

(3) Clark, D. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 2. Prediction of blood-brain barrier penetration. *J. Pharm. Sci.* **1999**, *88*, 815−821.

(4) Clark, D. Rapid calculation of polar surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption. *J. Pharm. Sci.* **1999**, *88*, 807−814.

(5) Kelder, J.; Grootenhuis, P.; Bayada, D.; Delbressine, L.; Ploemen, J.-P. Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs. *Pharm. Res.* **1999**, *16*, 1514−1519.

(6) Krarup, L. H.; Christensen, I. T.; Hovgaard, L.; Frokjaer, S. Predicting drug absorption from molecular surface properties based on molecular dynamics simulations. *Pharm. Res.* **1998**, *15*, 972−978.

(7) van de Waterbeemd, H.; Camenisch, G.; Folkers, G.; Raevsky, O. A. Estimation of Caco-2 cell permeability using calculated molecular descriptors. *Quant. Struct.-Act. Relat.* **1996**, *15*, 480−490.

(8) *ACD/logP Batch*, version 3.50; A.C.D. Inc., 33 Richmond St. West, Suite 605, Toronto, ON MSH 2L3, Canada.

(9) Norinder, U.; Österberg, T.; Artursson, P. Theoretical calculation and prediction of intestinal absorption of drugs using Molsurf parametrization and PLS statistics. *Eur. J. Pharm. Sci.* **1999**, *8*, 49−56.

(10) van de Waterbeemd, H.; Testa, B. The parametrization of lipophilicity and other structural properties in drug design. *Adv. Drug Res.* **1987**, *16*, 85−225.

(11) Norinder, U.; Österberg, T. The applicability of computational chemistry in the evaluation and prediction of drug transport properties. *Perspect. Drug Discovery Des.* **2000**, *19*, 1−18.

(12) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

CI000065L