Predicting Blood-Brain Transport of Drugs: A Computational Approach

Subhash C. Basak,[1,3] Brian D. Gute,[1] and Lester R. Drewes[2]

Received November 17, 1995; accepted February 16, 1996

Purpose. This study was conducted to determine the efficacy of using nonempirical parameters in the estimation of blood-brain transport, inferred from central nervous system (CNS) activity, for a set of twenty-eight compounds.

Methods. A discriminant function analysis was used to construct three distinct models based on topological indices, a hydrogen-bonding parameter, and logP.

Results. These models correctly predict the CNS activity of twenty-seven of the twenty-eight compounds.

Conclusions. Nonempirical parameters may be used effectively in the estimation the cerebrovascular penetration for known and newly designed drugs.

KEY WORDS: central nervous system; brain uptake; hydrogen bonding; lipophilicity; molecular similarity; nonempirical parameters.

INTRODUCTION

Blood-brain (BB) transport of central nervous system (CNS) active molecules is crucial to their effectiveness as therapeutic agents. Transendothelial transport of chemicals depends on multiple factors which are reported to include binding to plasma constituents, ionization at physiological pH, time-dependent plasma-concentration profile (which is dependent upon the distribution, metabolism and elimination processes), cerebral blood flow (which determines the access of drugs to the central vasculature) and lipophilicity of the molecules (1). All important determinants of BB entry, except cerebral flow, can be manipulated through structural modification of the drug molecule.

Various authors have attempted to predict BB transport of molecules from their partition coefficients (P), determined experimentally with different solvent pairs, e.g., octanol/water, cyclohexane/water. However, in many cases the correlations between partition coefficient and BB transport are not very good (1). Also, solvatochromic parameters, such as dipolarity/polarizability (π^H), hydrogen-bond acceptor basicity (β^H), hydrogen-bond donor acidity (α^H), and intrinsic volume (Vx) have been used to estimate the passage of molecules across the vascular endothelium into the brain (1).

Many of the parameters used to predict BB transport are either experimental properties or parameters based on experimental data. In drug design, one has to evaluate a very large number of chemicals in order to decide which handful of molecular structures should be tested extensively for the optimum discovery of novel therapeutic agents. In such cases, one is often confronted with candidate molecules for which there is little or no experimental data or that have not yet been synthesized. Evaluation schemes based solely on experimental parameters will be of limited value in such situations.

We have been involved in the development of models for selecting bioactive analogs of chemicals, as well as for predicting physicochemical, biochemical and toxic properties of different sets of molecules, using quantitative structure-activity relationships (QSAR) and quantitative molecular similarity analysis (QMSA) methods. Our predictive models are based on parametric values which can be calculated directly from molecular structure, and include topological indices (TIs), substructures, geometrical parameters, and quantum-chemical descriptors (2–6).

Seelig et al. (7) attempted to establish a correlation between the BB entry and the surface activity of twenty-eight, diverse compounds. A lipophilicity analysis was run side-by-side with their study to show the problems of using partitioning coefficients for predicting BB transport as inferred from central nervous system (CNS) activity. A critical analysis of their results shows that, while their predictions were reasonable for compounds that exhibited low values of both critical micelle concentrations (CMC) and concentration for onset of surface-activity (C_0), the method was inaccurate for compounds exhibiting higher CMC and C_0 values. Therefore, it was of interest to see if we could develop a more accurate predictive model for this set of compounds using nonempirical parameters; computed algorithmically from molecular structures.

TIs have been used to develop QSAR and QMSA models pertinent to pharmacokinetics, pharmacodynamics, and toxicology (2–6, 8, 9). In the present paper, we have carried out a discriminant function analysis (DFA) of the twenty-eight compounds analyzed by Seelig et al. (7) in an effort to predict their BB transport from theoretical parameters of chemical structure, specifically: a set of 102 TIs, a computed hydrogen-bonding parameter, HB_1, (10) and calculated logP (11).

MATERIALS AND METHODS

Database

The twenty-eight compounds analyzed in this study, along with their observed CNS activity levels, were those presented by Seelig et al. (7), as shown qualitatively in Table I.

Calculation of Parameters (TIs, HB_1 and logP)

TIs were calculated using the POLLY software developed by Basak et al. (12). The program calculates 102 TIs; including the Wiener index, Randić's connectivity index, higher order connectivity indices, information-theoretic indices defined on a distance matrix, indices of neighborhood complexity and Balaban's J indices; from the SMILES line-notation entry of molecular structure. These indices have been described in detail previously (4) and brief definitions are included in Table II. The hydrogen bonding parameter, HB_1, was calculated by the software, H-BOND, developed by Basak (10) and based on the work of Ou et al. (13). LogP values for the compounds were

[1] Natural Resources Research institute, University of Minnesota, Duluth, 5013 Miller Trunk Highway, Duluth, Minnesota 55811.
[2] Department of Biochemistry and Molecular Biology, School of Medicine, University of Minnesota, Duluth, Duluth, Minnesota 55812.
[3] To whom correspondence should be addressed.

0724-8741/96/0500-0775$09.50/0

Table I. CNS Activity Levels, Reported and Predicted (from Jack-knifing), for Twenty-eight Compounds: Reported Activity (A_R), Predicted Activity for Model 1 (A_{P1}), Model 2 (A_{P2}), and Model 3 (A_{P3}). Compounds with High CNS Activity Are Designated by + and Compounds with Low or No Activity Are Designated by −

No.	Compound	A_R	A_{P1}	A_{P2}	A_{P3}
1	(R)-apomorphine·HCl	+	+	+	+
2	Chlorpromazine·HCl	+	+	+	+
3	Clonidine·HCl	+	+	+	+
4	Desipramine·HCl	+	+	+	+
5	Doxylamine succinate	+	+	+	+
6	*cis*-Flupentixol·2HCl	+	+	+	+
7	Haloperidol·HCl	+	+	+	+
8	Imipramine·HCl	+	+	+	+
9	Naltrexone·HCl	+	+	+	+
10	Perphenazine·2HCl	+	+	+	+
11	Promazine·HCl	+	+	+	+
12	Promethazine·HCl	+	+	+	+
13	Roxindole methane sulfonate	+	+	+	+
14	Tamitinol·2HCl	+	+	+	+
15	Thiopental sodium	+	+	+	+
16	Thioridazine·HCl	+	+	+	+
17	Astemizole·2HCl	−	−	−	−
18	Carebastine	−	−	−	−
19	Domperidone·HCl	−	−	−	−
20	Ebastine methane sulfonate	−	−	−	−
21	Loperamide·HCl	−	−	−	−
22	Terfenadine	−	−	−	−
23	Atenolol	−	−	−	−
24[a]	Mequitazine·HCl	−	+	+	+
25	Salbutamol hemisulfate	−	−	−	−
26	Carmoxirol·HCl	−	−	−	−
27	Furosemide	−	−	−	−
28	Pirenzepine·HCl	−	−	−	−

[a] Mequitazine is the only compound which was misclassified by all three models as a CNS^+ compound, instead of a CNS^- compound.

calculated by the CLOGP software, version 3.2 (11). The symbols and non-transformed values for the TIs used in our models, as well as HB_1 and logP are presented in Table III.

Model Development/Statistical Analysis

All TI values were transformed by the natural logarithm of the index plus one, because some TIs may be several orders of magnitude greater than others. After this step, the TIs were examined and completely correlated indices (those with r = 1.0) were removed, leaving 94 TIs.

Since such a large set of values, 96 including HB_1 and logP, could lead to spurious results, it was necessary to reduce the amount of data which would be used for model construction. Because many of the TIs are highly intercorrelated, the 94-dimensional space can be represented by a subspace without significant loss of information. The statistical package, SAS (14), was used to perform a principal component analysis (PCA) on the set of 94 TIs, combining them into a set of principal components (PCs). Principal components explain the maximum variance while maintaining maximal orthogonality between PCs. From the PCA analysis, seven PCs were retained, all of which had eigenvalues greater than 1.0.

Next, subsets of TIs were selected from the seven PCs based on the correlation of the TIs to the PCs. The first subset used the two most-highly correlated TIs for each of the seven PCs, while the second subset selected TIs from the PCs based proportionately on the eigenvalues of the PCs. From the PCs with eigenvalues close to 60, the 6 most-highly correlated TIs were selected; for PCs with eigenvalues between 10 and 20, two TIs were used, and PCs with eigenvalues between 1 and 10, the one TI with the highest correlation was chosen.

Three separate data sets were used in the DFA model construction. One set contained all 94 TIs, while the other two sets used the subsets described previously. HB_1 and logP were included in all three sets. Selection of indices for the final models was done using the SAS procedure, STEPDISC, and the models were crossvalidated using the DISCRIM procedure (14). Indices were examined using a stepwise selection. This process begins with no variables in the model, and adds the one variable which contributes the most to the model. In a stepwise fashion, variables continue to be added and removed from the model based on their ability to meet the criteria, first, for inclusion in the model, and second, for remaining in the model. In this fashion, the model selects specific variables as it is constructed and eliminates any which become irrelevant. Validation of the models was conducted through jack-knifing.

RESULTS

Model 1 was constructed from five parameters: CIC_4, CIC_5, $^2\chi^v$, $^6\chi^v_C$, and HB_1; model 2 used six parameters: CIC_4, CIC_5, P_2, $^6\chi_C$, $^5\chi^v_{Ch}$, and HB_1; and model 3 also used six parameters: CIC_4, CIC_5, M_1, $^6\chi_C$, $^5\chi^v_{Ch}$, and HB_1. Although the logP variable was available for selection during model construction, it was not incorporated into any of the models.

Each of the three models performed well, having 100% accuracy for classifying CNS^+ compounds and 91.75% accuracy for CNS^- compounds. This was true in the initial model construction and after the jack-knifing procedure. The final results from the jack-knifing procedure are shown in Table I. The scoring results used by the DFA procedure for classification of the compounds are presented in Figure 1.

DISCUSSION

The purpose of this study was to evaluate the utility of nonempirical parameters for estimating transendothelial transport of chemicals. To this end, we used a DFA procedure to incorporate nonempirical parameters into three separate models for predicting the BB transport for a set of well-documented CNS-active chemicals (7). Each model correctly predicted BB transport for twenty-seven of the twenty-eight compound set.

One of the most interesting findings was the reliance of each model on a hydrogen-bonding parameter (HB_1) and the exclusion of the logP parameter. Analysis of the step-wise discrimination procedure, used to select features for model inclusion, reveals that logP was not statistically significant in any step of model construction. This finding reinforces the conclusion of Seelig et

Table II. TIs[a], HB_I(10), and log P(11) Values Used for Each of the 28 Compounds

N	CIC_4	CIC_5	$^2X^v$	$^6X^v_C$	P_2	^6X_C	$^5X^v_{Ch}$	M_1	$^5X^b_{Ch}$	HB_I	log P
1	0.291	0.291	5.796	0.000	35	0.000	0.000	116	0.000	5	3.010
2	0.588	0.588	7.011	0.000	32	0.000	0.000	110	0.000	2	5.200
3	0.957	0.957	3.699	0.000	20	0.000	0.062	70	0.125	5	2.340
4	0.970	0.970	5.366	0.000	29	0.000	0.000	102	0.000	3	4.390
5	0.816	0.816	5.266	0.030	28	0.118	0.000	98	0.000	3	1.850
6	0.741	0.596	8.704	0.004	47	0.144	0.000	160	0.000	4	5.900
7	0.857	0.857	7.350	0.019	39	0.072	0.000	134	0.000	4	3.520
8	1.189	1.189	6.066	0.000	31	0.000	0.000	108	0.000	2	4.710
9	0.468	0.468	8.557	0.051	47	0.078	0.121	152	0.169	5	1.820
10	0.717	0.566	8.614	0.000	41	0.000	0.000	142	0.000	5	5.570
11	1.088	1.088	6.430	0.000	30	0.000	0.000	104	0.000	2	4.280
12	1.107	1.107	6.545	0.000	31	0.000	0.000	106	0.000	2	4.650
13	0.431	0.431	7.088	0.000	39	0.000	0.032	136	0.037	5	4.990
14	0.614	0.614	4.275	0.000	19	0.000	0.000	68	0.000	5	1.900
15	0.831	0.831	4.660	0.057	23	0.144	0.000	78	0.000	6	2.980
16	0.422	0.422	8,937	0.000	39	0.000	0.000	134	0.000	2	6.420
17	0.719	0.719	8.905	0.000	52	0.000	0.023	180	0.032	6	6.060
18	1.210	1.210	10.413	0.064	55	0.250	0.000	190	0.000	6	1.860
19	0.556	0.481	8.423	0.000	49	0.000	0.056	166	0.102	9	9.450
20	1.423	1.423	10.733	0.083	52	0.144	0.000	180	0.000	3	6.160
21	1.306	1.306	9.904	0.056	53	0.179	0.000	180	0.000	5	3.900
22	1.385	1.385	11.144	0.121	54	0.289	0.000	184	0.000	5	6.240
23	0.817	0.827	4.815	0.000	24	0.000	0.000	86	0.000	9	−0.109
24	0.978	0.978	8.130	0.000	39	0.000	0.000	132	0.000	2	5.120
25	0.981	0.981	5.193	0.000	24	0.000	0.000	82	0.000	8	0.111
26	0.415	0.415	7.424	0.000	42	0.000	0.032	146	0.037	6	5.530
27	0.188	0.188	6.010	0.017	32	0.118	0.036	108	0.056	11	2.040
28	0.569	0.569	6.654	0.000	41	0.000	0.000	140	0.000	8	1.710

[a] The original, non-log scaled values for the TIs (CIC_4, CIC_5, $^2X^v$, $^6X^v_C$, P_2, ^6X_C, $^5X^v_{Ch}$, M_1, $^5X^b_{Ch}$) are presented. TIs were transformed by the natural logarithm of the TI plus one for model creation and validation.

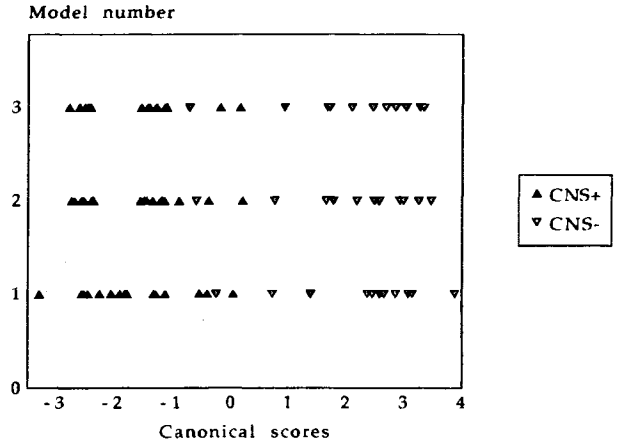

Fig. 1. Canonical scoring produced by the DFA procedure for classification of CNS activity.

al. (7), that lipophilicity is not a reliable parameter for classifying the CNS activity of this particular set of twenty-eight compounds. The two models that initially incorporated only 14 or 15 parameters are the most reliable. The model developed from the set of all TIs, HB_I, and logP, while giving the same results as the other two models, risks spurious results through the possible inclusion of extraneous parameters or the deletion of relevant parameters due to the large number of calculations required.

Earlier studies have shown that molecular weight, which roughly correlates with molecular size, is inversely related to transendothelial transport (15). Our study has concurred with this finding, in that each model contains one descriptor which is correlated with molecular size, P_2, M_1, and to a lesser extent, $^2\chi^v$.

In an earlier study, Basak and Grunwald (9) developed a model aimed at predicting mutagenicity/non-mutagenicity for a set of 463 chemicals. The DFA model using TIs correctly predicted the activity for 75% of the chemicals. Addition of quantum chemical parameters, viz., dipole moment (μ), HOMO energy (E_{HOMO}), LUMO energy (E_{LUMO}), and heat of formation (H_F), did not make a significant improvement to the predictive accuracy of the model. The relative success of the current study is all the more encouraging, because a high level of classification accuracy was achieved using only structurally-based parameters, plus a generalized quantifier of hydrogen-bonding capacity.

These results are also encouraging in their implications for development of new therapeutic agents that are active in the central nervous system. The ability to predict CNS activity from parameters that can be calculated directly from chemical structure would help to speed the development of new drugs, allowing drug design based on predicted activity—even the potential of compounds that have yet to be synthesized could be evaluated for development as therapeutic agents.

Table III. Topological Index Symbols and Definitions

I_D^W	Information index for the magnitudes of distances between all possible pairs of vertices of a graph
$\overline{I_D^W}$	Mean information index for the magnitude of distance
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
I^D	Degree complexity
H^V	Graph vertex complexity
H^D	Graph distance complexity
$\overline{IC}$	Information content of the distance matrix partitioned by frequency of occurrences of distance h
O	Order of neighborhood when IC, reaches its maximum value for the hydrogen-filled graph
I_{ORB}	Information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
O_{ORB}	Maximum order of neighborhood of vertices for I_{ORB} within the hydrogen-suppressed graph
M_1	A Zagreb group parameter = sum of square of degree over all vertices
M_2	A Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
IC_r	Mean information content or complexity of a graph based on the r^{th} (r = 0–5) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	Structural information content for r_{th} (r = 0–5) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	Complementary information content for r_{th} (r = 0–5) order neighborhood of vertices in a hydrogen-filled graph
$^h\chi$	Path connectivity index of order h = 0–6
$^h\chi_C$	Cluster connectivity index of order h = 5–6
$^h\chi_{PC}$	Path-cluster connectivity index of order h = 4–6
$^h\chi^V$	Valence path connectivity index of order h = 0–6
$^h\chi_C^V$	Valence cluster connectivity index of order h = 5–6
$^h\chi_{PC}^V$	Valence path-cluster connectivity index of order h = 4–6
$^h\chi^b$	Valence path connectivity index of order h = 0–6
$^h\chi_C^b$	Valence cluster connectivity index of order h = 5–6
$^h\chi_{PC}^b$	Valence path-cluster connectivity index of order h = 4–6
P_h	Number of paths of length h = 0–9
J	Balaban's J index based on distance
J^B	Balaban's J index based on multigraph bond orders
J^X	Balaban's J index based on relative electronegativities
J^Y	Balaban's J index based on relative covalent radii

ACKNOWLEDGMENTS

This paper is contribution number 176 from the Center for Water and the Environment (CWE) of the Natural Resources Research Institute. Research reported in this paper was supported, in part, by grant F49620-94-1-0401 from the United States Air Force, a grant from Exxon Biomedical Sciences, Inc., and the Structure-Activity Consortium (SARCON) of the Natural Resources Research Institute of the University of Minnesota. The authors would also like to extend their thanks to Greg Grunwald (CWE) and Douglas Bristol (Laboratory of Environmental Carcinogenesis and Mutagenesis, NIEHS, NIH) for helpful discussions.

REFERENCES

1. H. Chadha, M. H. Abraham, R. C. Mitchell. Correlation and prediction of blood-brain distribution, *Proceedings of an International Conference on the New Concept of a Blood-Brain Barrier,* London, 4–5 July 1994.
2. S. C. Basak, G. D. Grunwald. Estimation of lipophilicity from structural similarity, *New J. Chem.* **19**:231–237 (1995).
3. S. C. Basak, G. D. Grunwald. Molecular similarity and estimation of molecular properties, *J. Chem. Inf. Comput. Sci.* **35**:366–372 (1995).
4. S. C. Basak, G. D. Grunwald. Use of graph theoretic parameters in risk assessment of chemicals, *Toxicol. Lett.* **79**:239–250 (1995).
5. S. C. Basak, B. D. Gute. Use of graph-theoretic parameters in predicting inhibition of microsomal p-hydroxylation of aniline by alcohols: a molecular similarity approach, *Proceedings of the Hazardous Waste and Public Health: International Congress on the Health Effects of Hazardous Waste,* Princeton Scientific Publishing Co., Inc., in press (1996).
6. S. C. Basak, B. D. Gute, G. D. Grunwald. Estimation of normal boiling points of haloalkanes using molecular similarity, *Croat. Chem. Acta* in press (1996).
7. A. Seelig, R. Gottschlich, R. M. Devant. A method to determine the ability of drugs to diffuse through the blood-brain barrier, *Proc. Nat. Acad. Sci., USA* **91**:68–72 (1994).
8. S. C. Basak. Use of molecular complexity indices in predictive pharmacology and toxicology: a QSAR approach, *Med. Sci. Res.* **15**:605–609 (1987).
9. S. C. Basak, G. D. Grunwald. Predicting genotoxicity of chemicals using nonempirical parameters, *Proceedings of the International Cancer Congress,* New Delhi (India), 30 Oct.-5 Nov., 1994, pp. 413–416.
10. S. C. Basak. H-BOND: A program for calculating hydrogen-bonding parameter. University of Minnesota, Duluth, 1988.
11. A. Leo, D. Weininger. CLOGP, version 3.2 user refernce manual, Pomona Medicinal Chemistry Project, Pomona College, Claremont, CA. Distributed by Daylight Chemical Information Systems, Inc., 1984.
12. S. C. Basak, D. K. Harriss, V. R. Magnuson. POLLY 2.3: Copyright of the University of Minnesota, 1988.
13. X. Ou, Y. Ouyang, E. J. Lien. Examination of quantitative relationship of partition coefficient (LogP) and molecular weight, dipole moment and hydrogen bond capability of miscellaneous compounds, *J. Mol. Sci.* **4**:89–95 (1986).
14. SAS Institute, Inc. In SAS/STAT User's Guide, Release 6.03 ed., SAS Institute Inc., Cary, NY, 1988, chapters 16, 31, pp. 359–448, 909–922.
15. E. J. Lien. Structure-activity relationships and drug disposition, *Ann. Rev. Pharmacol. Toxicol.* **21**:31–61 (1981).