Preliminary Study of the Blood Brain Barrier Penetration of Some Organic Compounds and Drugs

Luminita Crisan, Liliana Pacureanu

Department of Computational Chemistry, Institute of Chemistry of Romanian Academy, Timisoara, 24 Mihai Viteazul Avenue, 300223 Timisoara, Romania
e-mail: *lumi_crisan@acad-icht.tm.edu.ro*

Abstract

Partial Least Squares (PLS) regression of blood–brain permeation data (logBB) including 348 diverse organic compounds and drugs was built using 903 Dragon descriptors. The prediction performance of the obtained PLS model is acceptable: the squared correlation coefficient (cumulative sum of squares of all the Y's explained by all extracted components) $R^2_{Y(CUM)}$ = 0.822, the crossvalidated correlation coefficient (cumulative fraction of the total variation of the Y's that can be predicted by all the extracted components) $Q^2_{Y(CUM)}$ = 0.640, the number of independent variables, X=487, for a dataset of 342 compounds (six compounds was outliers). The Y-randomization test demonstrated the absence of chance correlation which is confirmed by the lower values of regression line intercepts for $R^2_{X(CUM)}$ (0.307) and $Q^2_{(CUM)}$ (-0.320). The descriptors such as polar surface area (N,O and N,O,S,P polar contributions), octanol-water partition coefficient (Ghose-Crippen and Moriguchi), hydrophilic factor, complementary information content index and the number of H-bond donor atoms showed the largest Variables Importance in the Projection (VIP) values and can influence the logBB. The values of logBB predicted by our model display lower differences against experimental values of 342 compounds than logBB values predicted by QikProp.

Introduction

The blood–brain barrier (BBB) is a complex system implicated in the normal function of the central nervous system (CNS) through: (i) strictly limiting the passive diffusion of polar substances from the blood to the brain; (ii) mediating the transport of nutrients to the brain and of toxic metabolites and xenobiotics from the brain; (iii) overseeing the migration of circulating immune cells. [1-3] Penetration of blood-brain barrier, represents one of the most important and challenging areas in drug discovery. The presence of the BBB makes difficult the development of new therapies for brain diseases including meningitis, brain abscess, epilepsy, multiple sclerosis, neuromyelitis optica, late-stage neurological trypanosomiasis, Alzheimer's disease, cerebral edema, HIV encephalitis, etc [4]. To measure the drug transport across the blood brain barrier the blood–brain partition coefficient, logBB has been defined, [5] $\log BB = \log(C_{brain}/C_{blood})$, where C_{brain} and C_{blood} are the equilibrium concentrations of the drug in the brain and the blood, respectively.

In vitro experimental determination of BBB permeation is expensive, time consuming and requires compound's stability, purity and assay special conditions, while *in vivo* determinations based on radiolabeled compounds are required in some cases. [6] In 1988 the first theoretical model for a large number of H2 histamine receptor agonists predicting logBB values has been reported. [7] Ever since many attempts to correlate the experimental blood-brain concentration ratio values with physico‑chemical parameters have been reported. [8-24]

In this study the prediction of logBB values based on a larger dataset of compounds belonging to different structural classes collected from literature [12, 22, 23, 25-33] is reported. The aim is to build a comprehensive and general model for the blood brain barrier penetration of different organic compounds and drugs.

Methodology

Dataset. In our study we combined various literature data sets to collect a large-scale logBB dataset comprising 348 experimental logBB values. These dataset are available upon request from the authors and contains compounds that belong to different structural classes: 197 compounds classified as permeable showing positive logBB values, ranging from 0 to 1.64, and 151 compounds classified as non-permeable displaying negative logBB value, ranging from -0.01 to -2.15.

Descriptors. The following classes of descriptors were calculated with the help of Dragon software [34]: of 1D-functional groups, 1D-atom centered fragments, 2D-topological descriptors, 2D walk and path counts, 2D-autocorrelations, 2D-connectivity indices, 2D-information indices, 2D-topological charge indices, 2D-Eigenvalue-based indices, 2D-topological descriptors, 2D-edge adjacency indices, 2D-Burden eigenvalues, molecular properties, 2D-binary fingerprints and 2D-frequency fingerprints starting from the SMILES codes. Molecular descriptors were checked and constant or near-constant variables were excluded. If two descriptors register a correlation coefficient of 0.99 one of them was eliminated. The final set of descriptors used in PLS investigation included 903 molecular descriptors. The complete list of molecular descriptors and their meaning are provided on the Dragon website.[34]

PLS method. PLS analysis is a linear modeling technique [35] aimed at finding the relationship between the independent variable X-matrix (Dragon descriptors) and response Y-matrix (logBB). The information contained in the descriptor X-matrix is projected on a smaller number of latent variables called PLS components, denoted by A. The prediction of Y-values is carried out by extracting a set of 125 orthogonal components from the initial X-matrix, which display the highest predictive power. The number of A factors was determined using the cross-validation method leave seven out, with maximum number of iterations when fitting the model of 200, whereas the confidence level was set at 95%. The VIP reflects the influence of the variables in the PLS model concerning the property Y (i.e., its correlation to all responses), and independent variables X [36]. To evaluate the robustness of the PLS model obtained we used the response permutation method implemented in SIMCA package [36].

Robustness of the QSAR models. Golbraigh demonstrated that the Q^2 is not adequate to assess the predictive ability of the QSAR model. [37] Therefore, Y-randomization test is a widely used technique to evaluate the robustness of a QSAR model. [38] It consists in building a number of QSAR models using the initial descriptor matrix and the randomized Y variable. The plot showing $R^2Y_{(CUM)}$ (cumulative sum of squares of all the Y's explained by all extracted components) and $Q^2_{(CUM)}$ (cumulative fraction of the total variation of the Y's that can be predicted by all the extracted components) for all PLS-DA models (all the Y permuted models, and also the initial model) on the Y-axis and the correlation coefficients between randomized and original response variables on the X-axis was analyzed [37].If the Y-axis intercept of the regression line does not exceed 0.3–0.4 for $R^2_{Y(CUM)}$, and 0.05 for $Q^2_{(CUM)}$, the model is considered free of chance correlation. [38] The selected PLS model was subjected to 999 Y-randomizations.

LogBB prediction by QikProp. The QikProp software [39] developed by Professor William L. Jorgensen [40] fitted to 710 compounds including 500 drugs, one of the state of the art tools in predicting log(BB) was used as reference for our model. In addition to predicting the absorption, distribution, metabolism, and excretion (ADME) physically and pharmaceutically relevant properties of organic molecules or drugs, QikProp provides ranges for comparing a particular molecule properties.

Results and discussions

In order to correlate the experimental logBB values with structural descriptors, the PLS calculations were initiated for 903 descriptors and 348 log BB values [36]. From the 125 principal components resulted, the first 10% of the components already explain 54% of the information content of the X-matrix. The first PLS model was constructed using the initial X matrix, was not satisfactory, therefore we proceed to the improvement of the statistics as follows: (1) the normal probability plot of Y standardized residuals - standard deviation higher than ±3 - was the criterion for gradually eliminating the outliers; (ii) the overfit was reduced by excluding the noise variables (variable coefficient values close to 0). Therefore, six compounds were identified as outliers as their standard deviations exceeded ±3SD (±3.04 to ±4.31) and 416 noise variables were progressively eliminated. The statistical parameters of the final model are suitable for a large dataset of compounds. The cumulative sum of squares (SS) of all the X values explained by all extracted components $R^2_{X(CUM)}$ = 0.559, the cumulative SS of all the Y's explained by all extracted components $R^2_{Y(CUM)}$ = 0.822, and the fraction of the total variation of Y values that can be predicted for all extracted principal components $Q^2_{Y(CUM)}$ = 0.640. The variables which influence markedly our PLS model (VIP > 1.6) include several straightforward descriptors such as polar surface area (PSA - N,O and N,O,S,P polar contributions), octanol-water partition coefficient (Ghose-Crippen and Moriguchi), hydrophilic factor, complementary information content index and the number of H-bond donor atoms. This is in accord with well accepted parameters such as lipophilicity, hydrogen bonding capacity, molecular charge, molecular size, molecular shape, and molecular flexibility which was correlated with log BB. [5] Complementary information content index is an topological index which is calculated based on Shanon information theory [41] Generally speaking, the molecular topology is correlated with a large number of molecular and biological properties. In particular, the topological indices of zero order are of special importance for the suitable description of molar volume of organic compounds which in turn is correlated with logBB [42]. Higher polarity and hydrogen bonding are detrimental for blood-brain penetration, whereas higher molecular volume was positively correlated. [5] PSA is highly correlated with the hydrogen bonding capacity of a compound. [5] Norinder and Haeberlein [43] observed a linear correlation between PSA and the sum of N + O atoms, and concluded that (N + O) $\leq$ 5 is favorable for blood brain penetration. Clark, [44] stated that logP is favorable to get positive values of log BB.

The predictive capacity or validity of a QSAR model is a measure of how accurately the model can predict the biological activity of the set of compounds. The final model was internally validated using, the Y-permutation procedure using 999 randomizations to cover the complete dataset, each time forming a distinct set. The scrambled models were constructed with the same number of latent variables as the final model. The plot displayed in Figure 1 demonstrates that the Y-intercept (logBB-intercept) of the $R^2_{X(CUM)}$ and $Q^2_{(CUM)}$ lines has lower values and indicates no chance correlation for the selected model.

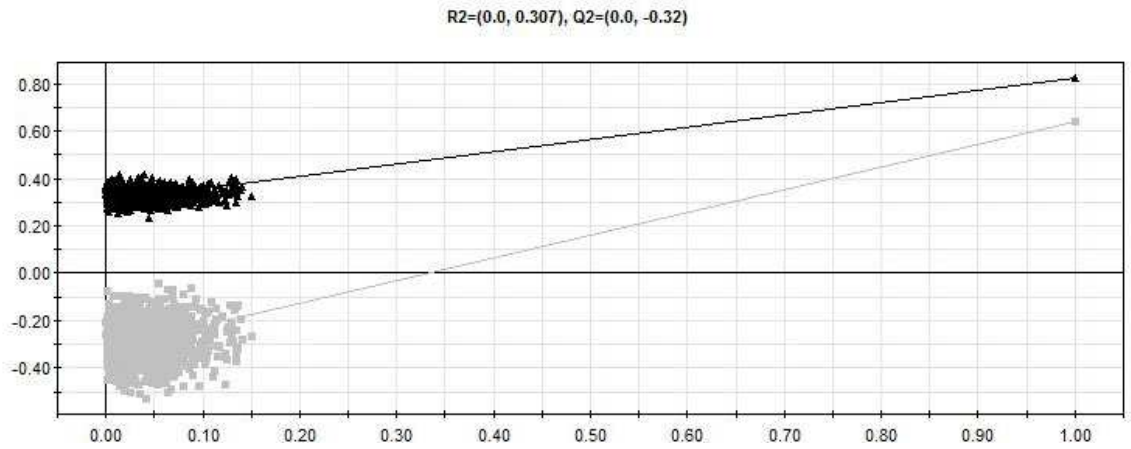

Figure1. Y - Randomization results for the final PLS model. The *x*-axis reports the correlation coefficient between original and permuted response data, while on the *y*-axis are represented R^2 (black triangles) and Q^2 (grey squares) values for the 999 randomized models

Several descriptors displaying higher VIP (Variables Importance in the Projection) values might play a critical role in defining BBB permeability of organic compounds. The top ten descriptors according toVIP magnitudes included in the PLS model are shown in Table 1.

Table 1. The most relevant descriptors of the PLS model

Var ID	*VIP*	*VIPcvSE*	*CoeffCS*	*CoeffCScvSE*	*Descriptor significance*
ALOGP	2.014	0.028	0.035	0.009	Ghose-Crippen octanol-water partition coeff. (logP)
MLOGP	1.926	0.035	0.032	0.010	Moriguchi octanol-water partition coeff. (logP)
BLTD48	1.926	0.035	-0.032	0.009	Verhaar Daphnia base-line toxicity from MLOGP (mmol/l)
TPSA(NO)	1.823	0.021	-0.031	0.007	Topological polar surface area using N,O polar contributions
MLOGP2	1.813	0.029	0.029	0.016	Squared Moriguchi octanol-water partition coeff.
TPSA(Tot)	1.759	0.015	-0.031	0.008	Topological polar surface area using N,O,S,P polar contributions
Hy	1.729	0.049	-0.030	0.022	Hydrophilic factor
ALOGP2	1.722	0.025	0.024	0.009	Squared Ghose-Crippen octanol-water partition coeff.
CIC1	1.636	0.027	0.017	0.011	Complementary Information Content index (neighborhood symmetry of first order)
nHDon	1.629	0.040	-0.033	0.022	Number of donor atoms for H-bonds (N and O)

**VIP* = The influence of every term in the matrix X on all the Y's; *VIPcvSE* = The jack knife standard error of the VIP computed by seven rounds of cross validation; C*oeffCS* = PLS regression coefficients corresponding to centered and scaled X, and scaled (but uncentered) Y; *CoeffCScvSE* = The jack knife standard error of the coefficients CoeffCS computed by seven rounds of cross validation.

For the same dataset of compounds QplogBB (Predicted brain/blood partition coefficient) was calculated with QikProp module from Schrödinger suite. The logBB predicted by our model register lower differences with respect to experimental values than QikProp calculations (see

Figure 2). The highest number of compounds displaying low differences to experimental values (0.05-0.3) is predicted by our PLS model, whereas QikProp predictions exhibit higher differences against experiment.

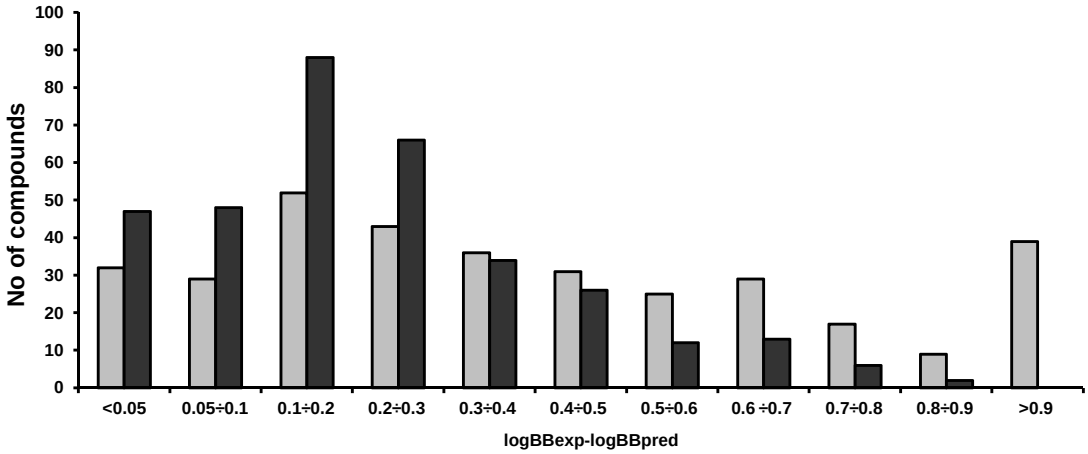

Figure 2. The number of compounds versus logBBexp-logBBpred; black bars render the PLS model and grey bars depict the QikProp prediction.

These results can be explained by the fact that the domain of applicability of the regression equation used by QikProp, is based on N=104 compounds of the molecular weight between 20-525 Da, while the molecular weight for our dataset of N=348 compounds ranges 16-1202 Da.

Conclusions

We have applied a PLS approach to a dataset of 348 compounds with known experimental logBB values, which belong to different structural classes. Some straightforward descriptors such as topological polar surface area, octanol-water partition coefficient and the number of H-bond donor atoms influence the developed PLS model, showing VIP values higher than 1.6. The final PLS model built on a large dataset excluded the risk of arbitrary correlation. Further QSAR experiments using diverse modeling methodologies including 3D descriptors and additional compounds will be pursued.

Acknowledgements

We thank Dr. Erik Johansson (Umetrics, Sweden) for kindly providing the SIMCA P 9.0 program package (L. Kurunczi laboratory) and to Dr. Simona Funar-Timofei for the access to DRAGON software. This project was financially supported by the Project No. 1.2 of the Institute of Chemistry of Romanian Academy, Timisoara.

References

[1] N.J. Abbott, A.A. Patabendige, D.E. Dolman, S.R. Yusof, D.J. Begley, Neurobiol. Dis. 37 (2010) 13.
[2] D.J. Begley, M.W. Brightman, Prog. Drug. Res. 61 (2003) 39.
[3] H. Wolburg, S. Noell, A. Mack, K. Wolburg-Buchholz, P. Fallier-Becker, Cell. Tissue Res. 335 (2009) 75.
[4]S. Paris-Robidas , V. Emond, C. Tremblay, D. Soulet, F. Calon Mol Pharmacol. (2011) 80, 32.
[5] U. Bickel, NeuroRx. 2 (2005) 15.
[6] A. Reichel, D.J. Begley, Pharm. Res. 15 (1998) 1270.

[7] R.C. Young, R.C. Mitchell, T.H. Brown, C.R. Ganellin, R. Griffiths, M. Jones, K.K. Rana, D. Saunders, I.R. Smith, N.E. Sore, T.J. Wilks, J. Med. Chem. 31 (1988) 656.
[8] F. Lombardo, J.F. Blake, W.J. Curatolo, J. Med. Chem. 39 (1996) 4750.
[9] M.H. Abraham, K. Takacs-Novak, R.C. Mitchell, J. Pharma. Sci. 86 (1997) 310.
[10] U. Norinder, P. Sjoberg, T. Osterberg, J. Pharma. Sci. 87 (1998) 952.
[11] D.E. Clark, J. Pharma. Sci. 88 (1999) 815.
[12] J. Kelder, P.D.J. Grootenhuis, D.M. Bayada, L.P.C. Delbressine, J.P. Ploemen, Pharma. Res. 16 (1999) 1514.
[13] P. Crivori, G. Cruciani, P.A. Carrupt, B. Testa, J. Med. Chem. 43 (2000) 2204.
[14] P. Ertl, B. Rohde, P. Selzer, J. Med. Chem. 43 (2000) 3714.
[15] G.M. Keseru, L. Molnar, J. Chem. Inf. Comput. Sci. 41 (2001) 120.
[16] R. Liu, H. Sun, S.S. So, J. Chem. Inf. Comput. Sci. 41 (2001) 1623.
[17] K. Rose, L.H. Hall, L.B. Kier, J. Chem. Inf. Comput. Sci. 42 (2002) 651.
[18] T. Hou, X. Xu, J. Mol. Model. 8 (2002) 337.
[19] M.C. Hutter, J. Comput. Aided Mol. Des. 17 (2003) 415.
[20] M.H. Abraham, Eur. J. Med. Chem. 39 (2004) 235.
[21] A. Yan , H. Liang, Y. Chong, X. Nie, C. Yu, SAR QSAR Environ. Res. 24 (2013) 61.
[22] R. Narayanan, S.B. Gunturi, Bioorg. Med. Chem. 13 (2005) 3017.
[23] L. Zhang, H. Zhu, T.I. Oprea, A. Golbraikh, A. Tropsha, Pharm. Res. 25 (2008) 1902.
[24] M. Muehlbacher, G.M. Spitzer, K.R. Liedl, J. Kornhuber, J. Comput. Aided Mol. Des. 25 (2011) 1095.
[25] S.R. Mente, F. Lombardo, J. Comput. Aided Mol. Des. 19 (2005) 465.
[26] S. Vilar, M. Chakrabarti, S. Costanzi , J. Mol. Graph. Model. 28 (2010) 899.
[27] J.A. Platts, M.H. Abraham, Y.H. Zhao, A. Hersey, L. Ijaz, D. Butina, Eur. J. Med. Chem. 36 (2001) 719.
[28] M.H. Abraham, A. Ibrahim, Y. Zhao, W.E. Acree, J. Pharm. Sci. 95(2006) 2091.
[29] P. Garg, J. Verma, J. Chem. Inf. Model. 46 (2006) 289.
[30] A. Guerra, J.A. Pa´ez, N.E. Campillo, QSAR Comb. Sci. 27 (2008) 586.
[31] K. Rose, L.H. Hall, L.B. Kier, J. Chem. Inf. Comput. Sci. 42 (2002) 651.
[32] D.A. Konovalov, D. Coomans, E. Deconinck, Y. Vander Heyden, J. Chem. Inf. Model. 47 (2007) 1648.
[33] M. Zerara, J. Brickmann, R. Kretschmer, T.E. Exner, J. Comput. Aided Mol. Des. 23 (2009) 105.
[34] Dragon Professional 5.5/2007 is software of Talete S.R.L., Milano, Italy.
[35] S. Wold, M. Sjöström, L. Eriksson, Chemometr. Intell. Lab. 58 (2001)109.
[36] SIMCA-P+ version 9.0, Umetrics AB, Sweden. http://www.umetrics.com.
[37] A. Golbraikh, M. Shen, Z. Xiao, YD Xiao, K.H. Lee, A. Tropsha, J. Comput. Aided Mol. Des. 17 (2003) 241.
[38] L. Eriksson, E. Johansson, N. Kettaneh-Wold, S. Wold, in: Multi- and megavariate data analysis: principles and applications, Umetrics AB, Umea, 2001, pp 92–97, pp 489–491.
[39] QikProp, version 3.9, Schrödinger, LLC, New York, NY, 2014.
[40] W.L. Jorgensen, E.M. Duffy, Bioorg. Med. Chem. Lett. 10 (2000) 1155.
[41] C.E. Shannon, Tech. J. 27 (1948), 379.
[42] A.R. Katritzky, E. Gordeeva, J. Chem. Inf. Comput. Sci. 33 (1993) 835.
[43] U. Norinder, M. Haeberlein, Adv. Drug Deliv. Rev. 54 (2002) 291.
[44] D.E. Clark, J Pharm Sci 88 (1999) 815.