Available online at www.sciencedirect.com
SCIENCE DIRECT®
Bioorganic & Medicinal Chemistry Letters
ELSEVIER
Bioorganic & Medicinal Chemistry Letters 14 (2004) 851–853

A neural network based prediction of octanol–water partition coefficients using atomic5 fragmental descriptors

László Molnár,[a] György M. Keserű,[a,*] Ákos Papp,[b] Zsolt Gulyás[c] and Ferenc Darvas[b]

[a]*Department of Chemical Information Technology, Budapest University of Technology and Economics, Szent Gellért tér 4., H-1111 Budapest, Hungary*
[b]*ComGenex, Inc., Bem rkp. 33–34, H-1027 Budapest, Hungary*
[c]*ComGRID, Ltd, Stáhly u. 13, H-1085 Budapest, Hungary*

Received 31 July 2003; revised 19 November 2003; accepted 4 December 2003

Abstract—An artificial neural network based approach using Atomic5 fragmental descriptors has been developed to predict the octanol–water partition coefficient (log*P*). We used a pre-selected set of organic molecules from PHYSPROP database as training and test sets for a feedforward neural network. Results demonstrate the superiority of our non-linear model over the traditional linear method.

Lipophilicity plays a key role in rational drug design since it is of primary importance in drug absorption and distribution. The 1-octanol–water partition coefficient (log*P*) is widely used in QSAR/QSPR[1] approaches after its first usage as a measure of lipophilicity by Hansch and Leo.[2] Since the measurement of log*P* is still not a high-throughput process, it is not convenient for the measurement of combinatorial libraries, and calculation methods are used instead. Virtual libraries are also used in drug discovery and for them computation is the only way to obtain log*P* values.

There are many methods for the prediction of log*P*,[3–6] but basically, it can be calculated in four ways. It can be built up from hydrophobic fragmental values, it can be derived from a knowledge of other physico-chemical or electronic parameters and their relationship to log*P*, or it can be predicted using molecular descriptors combining them by linear or non-linear (e.g., linear regression or neural network) algorithms. Traditionally, the most frequently used methods were based on the first approach applying linear models, profiting from the additive-constitutive nature of log*P*. Earlier approaches broke down the molecule to larger fragments,[7–9] while newer methods are based on atomic contributions,[10–12] but since the relationship between the structure and log*P* is non-linear, all of them have to use correction terms. The newest prediction programs use neural networks and differ from each other in the molecular descriptors they use as input.[13–15] Our present method belongs to this last group, but is unique in a way that its input are atomic fragmental descriptors. The fragment definitions are the same as in the knowledge base of the Atomic5 method.

Atomic5 is the name of a linear log*P* calculation method implemented in Pallas PrologP[16] program. It is based on Ghose–Crippen[11] fragmentation, but uses modified atomic contributions and additional correction terms.

In this work we report an artificial neural network based non-linear approach, which is able to predict log*P* at higher accuracy compared to the traditional linear models. We used a pre-selected set of organic molecules from PHYSPROP[17] database. Preselection process included the elimination of compounds with erroneous chemical structure, inorganic molecules and organometallics. 12729 Compounds with experimental log*P* data were divided into three sets. The largest set, containing 8729 molecules, was used as a Training set. The remaining 4000 molecules were divided evenly into two sets, containing 2000 member of each, namely Test1 and Test2. Division of the whole data set into the three sets

Keywords: Octanol–water partition; Log*P*; Preduction; Neural network.
* Corresponding author. Fax: +36-1-4633953; e-mail: gykeseru@eik.bme.hu

0960-894X/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmcl.2003.12.024

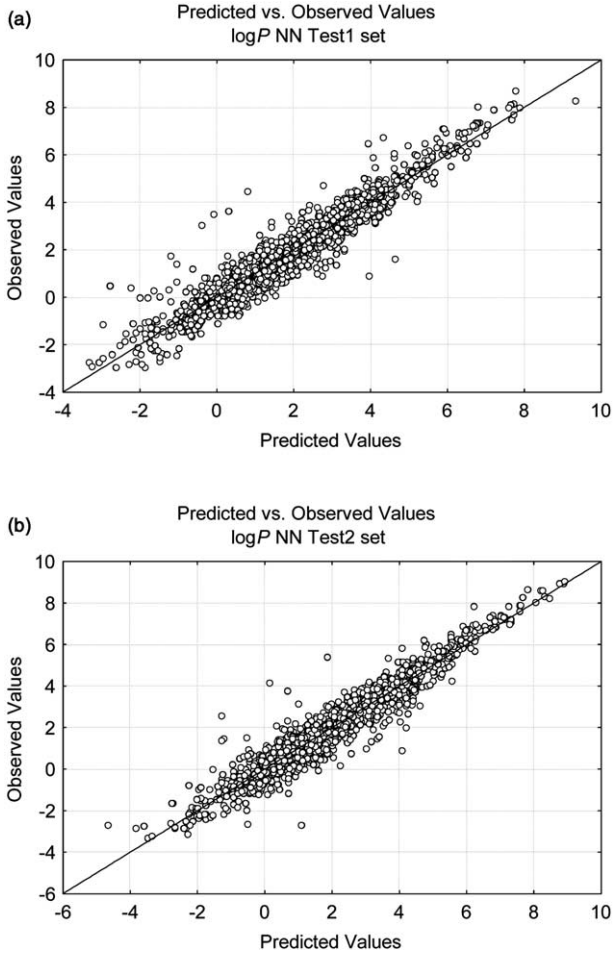

Figure 1. Correlation between predicted and observed log*P* values in Test1 (a) and Test2 (b) sets.

was carried out by a random ordering algorithm. Distribution profiles of measured log*P* values were virtually identical in the whole dataset and all of the subsets.

The neural network was designed to contain 130 input neurons, 12 hidden neurons and one output neuron. The training of the neural network was carried out by the insertion of the 130 Atomic5 descriptors of the Training set onto the input layer and the experimental log*P* value onto the output layer of the neural network. The Test1 set was used to monitor the quality of generalisation ability of the neural network at each learning cycle. The best-trained neural network (which has the lowest sum of square errors /SSE/ on Test1 set) was saved. At the end of training, Test2 set was used as an external validation set. All neural network operations were carried out using Stuttgart Neural Network Simulator (SNNS).[18]

Table 1. Correlation between predicted and observed log*P* values for Test1 and Test2 sets

	R^2	F	df	No. of compounds	Std. Error	*t*
Training set	0.94	148645	1.87	8729	0.01	17.58
Test1 set	0.91	19789	1.20	2000	0.02	9.15
Test2 set	0.92	22638	1.20	2000	0.02	6.55

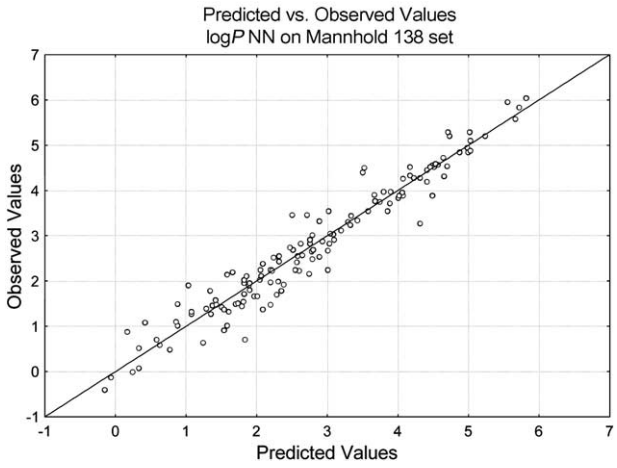

Figure 2. Correlation between predicted and observed log*P* values for Mannhold set.

Figure 1 shows correlations between predicted and observed log*P* values in Test1 (graph a) and Test2 (graph b) sets.

Table 1 reports the performance of our trained net on the training set as well as on Test1 and Test2 sets

After the validation of the trained neural network by the Test1 and Test2 set, we used an additional external validation set which is a log*P* evaluation set compiled by R. Mannhold and K. Dross, containing 138 diverse compounds[19] (Fig. 2).

We also compared our neural network based approach to the linear log*P* model of Atomic5 descriptor set. The comparison clearly shows the superiority of the non-linear approach (Table 2).

Finally the predictive power of our neural net based approach has been compared to other methods reported in ref 19. Table 3 shows the performance of five different log*P* prediction tools all of which were able to calculate log*P* for the total Mannhold set.

Table 2. Statistical description of correlation between predicted and observed log*P* values for the Mannhold set

Prediction	R^2	F	df	No. of compounds	Std. Error	*t*
Atomic5 Linear	0.89	1075	1.14	138	0.09	1.38
NNLogP	0.94	2044	1.14	138	0.07	0.09

Table 3. Comparison with other methods on the Mannhold set[19]

	f-SYBYL	KOWWIN	CHEMICALC-2
Acceptable	81.9	90.6	68.8
Disputable	13.8	5.8	17.4
Unacceptable	4.3	3.6	13.8
	MOLCAD	Tsar 2.2	NNlogP
Acceptable	68.1	68.1	84.1
Disputable	20.3	30.3	14.5
unacceptable	11.6	11.6	1.4

Predictions for this comparison have been classified as acceptable ($\log P_{pred}$–$\log P_{exp} < \pm 0.5$), disputable ($\pm 0.5 \leq \log P_{pred}$–$\log P_{exp} \leq \pm 1.0$) and unacceptable ($\log P_{pred}$–$\log P_{exp} > \pm 1.0$). This comparison revealed that our NNlog*P* outperforms most of the methods by predicting acceptable log*P* for 84.1% of the set. Although KOWWIN showed higher rate of acceptable predictions, this method generated unacceptable results for a significantly larger subgroup of compounds than NNlog*P*. Therefore we concluded that NNlog*P* predicts log*P* with reasonable accuracy and its application prevents unacceptable predictions almost completely.

Although it is early to generalize our results, we do believe that the developed method is forming the first member of a new class of pseudo-linear algorithms, where the precision of the non-linear approaches is combined with the transparency of the early linear methods.

References and notes

1. Clark, D. E.; Pickett, S. D. *Drug Discovery Today* **2000**, *5*, 49.
2. Leo, A.; Hansch, C.; Elkins, D. *Chem. Rev.* **1971**, *71*, 525.
3. Van de Waterbeemd, H. *Hydrophobicity of Organic Compounds. How to Calculate It by PC's*, in Booksoft Series. Darvas, F. (Ed.), CompuDrug International, Vienna, 1986.
4. Rekker, R. F.; Mannhold, R. *Calculation of Drug Lipophilicity; The Hydrophobic Fragmental Constant Approach*; VCH: Weinheim, 1992.
5. Leo, A. *Chem. Rev.* **1993**, *93*, 1281.
6. Martin, Y. C., Duban, M. E., Bures, M. G., DeLazzer, J. In Pharmacokinetic Optimization in Drug Research, Official publication of the logP2000 Symposium, VHCA-VCH, Zürich, 2000, 485.
7. Rekker, R. F. *The Hydrophobic Fragmental Constant. Its Derivation and Application. A Means of Characterizing Membrane Systems*; Elsevier: Amsterdam, 1977.
8. Rekker, R. F.; De Kort, H. M. *Eur. J. Med. Chem.* **1979**, *14*, 479.
9. Hansch, C.; Leo, A. *Substituent Constants for Correlation Analysis in Chemistry and Biology*; Wiley: New York, 1979.
10. Broto, P.; Moreau, G.; Vandycke, C. *Eur. J. Med. Chem.—Chim. Ther.* **1984**, *19*, 71.
11. Ghose, A. K.; Crippen, G. M. *J. Comput. Chem.* **1986**, *7*, 565.
12. Viswandadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163.
13. Duprat, A. F.; Huynh, T.; Dreyfus, G. *J. Chem. Inf. Comput. Sci.* **1998**, *4*, 586.
14. Tetko, I. V.; Tanchuk, V. Y.; Villa, A. E. *J. Chem. Inf. Comput. Sci.* **2001**, *5*, 1407.
15. Wegner, J. K.; Zell, A. *J. Chem. Inf. Comput. Sci.* **2003**, *3*, 1077.
16. The Pallas PrologP program is a Trademark of CompuDrug, Inc., www.compudrug.com.
17. The Physical Properties Database (PHYSPROP) is a trademark of Syracuse Research Corporation, www.syrres.com.
18. SNNS: Stuttgart Neural Network Simulator, University of Stuttgart, 1995.
19. Mannhold, R.; Dross, K. *Quant. Struct.-Act. Relat* **1996**, *15*, 403.