Available online at www.sciencedirect.com
SCIENCE DIRECT®

Journal of Pharmaceutical and Biomedical Analysis 37 (2005) 411–415

JOURNAL OF PHARMACEUTICAL AND BIOMEDICAL ANALYSIS

www.elsevier.com/locate/jpba

Short communication

Prediction of the aqueous solubility of benzylamine salts using QSPR model

Vimon Tantishaiyakul*

Department of Pharmaceutical Chemistry, Faculty of Pharmaceutical Sciences, Prince of Songkla University, Hat-Yai, Songkhla 90112, Thailand

Received 29 April 2004; received in revised form 1 November 2004; accepted 1 November 2004
Available online 9 December 2004

Abstract

Models predicting aqueous solubility of benzylamine salts were developed using multivariate partial least squares (PLS) and artificial neural network (ANN). Molecular descriptors, including binding energy (BE) and surface area of salts (SA), were calculated by the use of Hyperchem and ChemPlus QSAR programs for Windows. Other physicochemical properties, such as hydrogen acceptor for oxygen atoms, hydrogen acceptor for nitrogen atoms, hydrogen bond donors, hydrogen bond forming ability, molecular weight (MW), and calculated log partition coefficient (clog *P*) of *p*-substituted benzoic acids, were also used as descriptors. In this study, the predictive ability of ANN, especially multilayer perceptron (MLP) architecture networks, was founded to be superior to PLS models. The best ANN model derived, a 6-1-1 architecture, had an overall R^2 of 0.850 and root mean square error (RMSE) for cross-verification and test set of 0.189 and 0.185 log units, respectively. Since all the utilized descriptors are readily obtained from calculation, these derived models offer the advantage of not requiring the experimental determination of some descriptors.

Keywords: Salt; Solubility; Benzylamine; Binding energy

1. Introduction

Aqueous solubility is one of the most important physicochemical properties that plays a significant role in various physical and biological processes and has a marked impact on the design and pharmaceutical formulation development. For weak electrolyte drugs, salt formation is a common approach to improve its solubility, since it is a much simpler method than complex molecular modifications. Using different counter-ions can result in salts with difference in physicochemical properties. Until now, various organic and inorganic salts of acidic and basic drugs have been prepared and their physicochemical properties were subsequently determined in order to aid the selection of the most suitable salt for drug development [1,2].

Numerous in silico based methods for prediction of solubility of organic compounds have been developed [3–9]. Nevertheless, the quantitative structure–property relationship (QSPR) methods enabling prediction of solubility of salts are less well investigated. There was a study concerning QSPR of salt solubility indicating that no correlation was found between diclofenac salt solubility and any one parameter of either pK_a, hydrophilicity, or melting point of counter-ions [1]. Parshad et al. have recently reported models for predicting aqueous solubility of benzylamine salts, and their best models using different set of descriptors gave R^2 of 0.82 and Q^2 of 0.72 for training set, and R^2 of 0.74 and Q^2 of 0.72 for test set [10]. Intrinsic dissolution rate, intrinsic solubility of unionized acids, Charton's steric parameter, Hansch hydrophobic parameter, and molecular weight (MW) were reported as important descriptors for their derived models. Nevertheless, some of the utilized descriptors are based on experimental measurement.

* Tel.: +66 74 288864; fax: +66 74 428239.
E-mail address: tvimon@ratree.psu.ac.th.

0731-7085/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpba.2004.11.005

Salt solubility is a complex process, and one important factor that might govern aqueous solubility of salts is the electrostatic interaction between cationic and anionic species of the ion pair. The models for predicting aqueous solubility of diclofenac salts based on calculated structural descriptors and binding energy (BE) have previously been developed using partial least squares (PLS) regression [11]. PLS is a linear technique that can determine the relative importance of descriptors. However, some nonlinear relationships may involve in salt solubility. Artificial neural network (ANN), which is capable to recognizing nonlinear relationships, is usually used to generate QSPR models. Over the past few years, ANN models have been successfully employed in various aqueous solubility prediction studies [12–15]. The objective of this investigation is to model experimentally determined solubility of salts from computationally derived molecular descriptors, and to compare the predictive performance of PLS and ANN methods. The same set of salts from Parshad et al. was employed in this study [10].

2. Methods

2.1. Data set

Aqueous solubility data for the 22 benzylamine salts were taken from Parshad et al. [10]. These values were converted from mM to logarithm of salt solubility (log S). The solubilities of these benzylamine salts of p-substituted benzoic acids are listed in Table 1.

2.2. Molecular modeling and descriptor calculation

Molecular modeling calculations were performed using HyperChem 5.1 for Windows (Hypercube, FL, USA). The MM+ molecular mechanics force field was first run to get close to the optimized geometry. The conformation obtained from molecular mechanics was subjected to a refined geometry optimization using the PM3 semiempirical quantum chemistry.

Binding energy was calculated as previously described [11,16,17]. In brief, the salt/ion pair constituted by the benzylamine cation and the negatively charged p-substituted benzoic acid was calculated to obtain the total energy of ion pair ($\mathrm{TE}_{\text{ion-pair}}$). The interaction energy of the ion pair ($E_{\text{interaction}}$) was computed as the difference between the total energy of the ion pair and the sum of the energy of benzylamine ($E_{\text{benzylamine}}$) and organic acid (E_{acid}). The negative of the interaction energy is termed as binding energy:

$$E_{\text{interaction}} = \mathrm{TE}_{\text{ion}} - [E_{\text{benzylamine}} + E_{\text{acid}}]$$

$$\mathrm{BE} = -E_{\text{interaction}}$$

The ChemPlus QSAR Properties 1.5 (Hypercube, FL, USA) was employed for further calculation of the surface area of the salts. Hydrogen bond-forming ability of the acid (Hb) is the sum of hydrogen bond numbers of various groups including oxygen hydrogen bonding acceptor (HaO), nitrogen hydrogen bonding acceptor (HaN), and hydrogen bonding donor (Hd) was calculated as described by Xia et al. [18]. The log P value of acid was calculated directly from the molecular structure using the ClogP program (Biobyte, CA, USA). The values of these descriptors are presented in Table 1.

2.3. Statistical analysis

The software package used for conducting PLS analysis was Unscrambler 6.01 (Computer-Aided Modelling A/S, Trondheim, Norway). PLS is a bilinear modeling technique where information in the descriptor matrix **X** is projected onto a small number of underlying (“latent”) variables called PLS components, referred to as PCs. The matrix **Y** is simultaneously used in estimating the “latent” variables in **X** that will be most relevant for predicting the **Y** variables. All descriptor variables were preprocessed by autoscaling, using weights based on the variables' standard deviation and the data were mean-centered prior to PLS processing. The number of significant PCs for the PLS algorithm was determined using the cross-validation method. With cross-validation, some samples were kept out of the calibration and used for prediction. The process was repeated so that each of the samples was kept out once. The predicted values of left-out samples were then compared to the observed values using prediction error sum of squares (PRESS). The PRESS obtained in the cross-validation was calculated each time that a new PC was added to the model. The optimum number of PCs was concluded as the first local minimum in the PRESS versus PC plot. PRESS is defined as

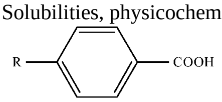
$$\mathrm{PRESS} = \sum_{i=1}^{n} (\hat{y} - y)^2$$

where $\hat{y}$ is the estimated value of the ith object and y is the corresponding reference value of this object.

ANN was also used to estimate the functional relation between the molecular descriptors and the solubility. The advantage of ANN is the inclusion of nonlinear relations in the model. In this study, ANN calculations were performed with Statistica 6.1 (Stat Soft, OK, USA). This program can search automatically for the optimal type/architecture of ANN such as multilayer perceptrons (MLP) and radial basis function (RBF). It can also select the number of input variables and hidden units, and the settings of various control parameters in the training algorithms that can affect the final performance of the network to fit a particular data set.

This study, the training algorithms used to optimize the network included back propagation and conjugate gradient for MLP, and K-means, K-nearest neighbor as well as pseudo-invert for RBF. The data set was divided into three subsets: the training, selection, and test sets. The number of compounds in the training, selection and test sets was 12, 5, and

Table 1
Solubilities, physicochemical, and molecular parameters of *p*-substituted benzoic acids and their benzylamine salts



p-substituted benzoic acid

Compound no.	*R*	*S* (mM)[a]	log *S*[b]	HaO[c]	HaN[d]	Hd[e]	Hb[f]	clog *P*[g]	BE[h]	SA[i]	MW[j]
1	H	839	2.924	4	0	1	5	1.885	120.93	476.92	122.04
2	CH_3	145	2.161	4	0	1	5	2.384	122.99	504.69	136.15
3	C_2H_5	98.2	1.992	4	0	1	5	2.913	108.77	508.56	150.17
4	C_3H_7	59	1.771	4	0	1	5	3.442	124.21	562.92	164.08
5	*i*-C_3H_7	97	1.987	4	0	1	5	3.312	122.85	554.01	164.08
6	C_4H_9	38.9	1.590	4	0	1	5	3.971	124.17	594.31	178.10
7	*t*-C_4H_9	35	1.544	4	0	1	5	3.711	123.92	571.26	178.10
8	C_6H_5	12.8	1.107	4	0	1	5	3.773	121.14	584.67	198.07
9	Cl	120	2.079	4	0	1	5	2.696	120.49	505.61	156.00
10	Br	85.4	1.931	4	0	1	5	2.846	119.74	514.23	199.95
11	I	33.9	1.530	4	0	1	5	3.106	119.98	518.99	247.93
12	OCH_3	306	2.486	6	0	1	7	2.023	122.80	518.13	152.05
13	OH	138	2.140	6	0	2	8	1.557	122.61	488.38	138.03
14	NO_2	77.6	1.890	8	0	1	9	1.838	113.68	513.84	167.02
15	CH_2OH	964	2.984	6	0	2	8	0.847	123.52	519.70	152.05
16	CN	256	2.408	4	1	1	6	1.545	117.07	509.36	147.03
17	NH_2	272	2.435	4	1	3	8	1.042	123.24	491.97	137.05
18	$NH(CH_3)$	231	2.364	4	1	2	7	1.730	123.15	520.52	151.06
19	$N(CH_3)_2$	62.6	1.797	4	1	1	6	2.275	122.15	549.55	165.08
20	CF_3	29.3	1.467	4	0	1	5	2.940	117.03	520.30	190.02
21	SO_2NH_2	110	2.041	8	1	3	12	0.452	115.93	550.38	201.01
22	$CONH_2$	118	2.072	6	1	3	10	0.702	118.51	528.32	165.04

[a] Salt solubility (mM).
[b] Logarithm of salt solubility.
[c] Number of hydrogen bond acceptor oxygen atoms of *p*-benzoic acid derivatives.
[d] Number of hydrogen bond acceptor nitrogen atoms of *p*-benzoic acid derivatives.
[e] Number of hydrogen bond donor atoms of *p*-benzoic acid derivatives.
[f] Hydrogen bond formation ability of *p*-benzoic acid derivatives.
[g] Calculated log *P* of *p*-benzoic acid derivatives.
[h] Binding energy of salt.
[i] Surface area of salts (A^2).
[j] Molecular weight of *p*-benzoic acid derivatives.

5, respectively, and the compound for each set was randomly selected. The neural networks were trained using the training subset only. The selection subset was used to keep an independent check on the performance of the networks during training, with deterioration in the selection error indicating over-learning. If over-learning occurs, the network will stop training the network and restore it to the state with minimum selection error. The test set was purely used to check that the selection error was not artificial. The network model will generalize if the selection and test errors are close together. The goodness of fit was evaluated by root mean square error (RMSE) which is defined as

$$\mathrm{RMSE} = \sqrt{\frac{\mathrm{PRESS}}{n}}$$

where n is number of compounds. Initially, all descriptors (Table 1) were used as input variables, the number of hidden units varied from 1 to 4, and a single output was log *S*.

3. Results and discussion

For predicting aqueous solubility of salt, it is important that the descriptor set should describe both the interaction between the ion pair species and the factors that influence the solubility of each species. The selected descriptors that might be of importance in modeling in this study are listed in Table 1. These descriptors include binding energy of salt, hydrogen bonding parameters of acid, lipophilicity (clog *P*) of the acid, surface area of the salt, and molecular weight of the acid.

The relationship between these calculated descriptors and log *S* of 22 salts has been analyzed. The predictive model-building abilities of two methods, PLS and ANN, were compared.

The PLS model including all descriptors and 22 salts resulted in the optimal number of seven PCs. The models showed a high squared correlation coefficient with R^2 of 0.858 (RMSE of 0.168), but a low squared correlation coefficient for cross-validation with Q^2 of 0.665 (RMSE of 0.270). This indicates a fairly good model, but probably low predic-

Table 2
Prediction profiles and statistical data for certain ANN models

Model no.	MLP configuration[a]	Descriptors in model[b]	Overall R^2	RMSE[c]		
				Train	Select	Test
A	6-2-1	HaO, Hb, clog P, BE, SA, MW	0.868	0.166	0.160	0.172
B	6-1-1	HaO, Hb, clog P, BE, SA, MW	0.850	0.164	0.189	0.185
C	6-1-1	HaN, Hb, clog P, BE, SA, MW	0.831	0.196	0.169	0.174
D	4-1-1	HaO, Hb, clog P, MW	0.829	0.196	0.186	0.181
E	5-1-1	HaO, Hb, clog P, SA, MW	0.827	0.161	0.225	0.205

[a] Input–hidden–output nodes.
[b] Abbreviations for descriptors are listed in Table 1.
[c] Number of compounds for training set, selection set, and test set is 12, 5, and 5, respectively.

tive ability for new data. The predictive ability of this PLS model is slightly lower than the model reported by Parshad et al. [10].

Although PLS has been achieved in predicting aqueous solubility of diclofenac salts [11], there may be nonlinear dependencies of log S of benzylamine salts and this set of descriptors. A QSPR model based on ANN was thus investigated. A preliminary analysis using all available descriptors was done to determine the optimal type, the input variables, and the number of the neurons in the hidden layer. Among the trained networks, the performance of MLP is better than RBF, and the resulting MLP models are listed in Table 2.

Compared to PLS analysis, improved predictive performance was observed by ANN approaches (models A, B, and C). The overall R^2 are fairly high for all ANN models ranging from 0.831 to 0.868 and RMSEs for training sets of 0.172 to 0.185. In addition, RMSEs for selection sets which were used to cross-verify the performances of training algorithms are low (0.160–0.189). Furthermore, the differences in RMSEs for selection and test sets are small, reflecting the generalization performance and high predictive ability of the networks.

The ratio (ρ) of the number of data points in the training set to the number of connections in the ANN has been proposed as a criterion for a network size. This ratio should be in the range of 1.8–2, as when the ratio approaches to 1, the network possibly has overfitting problems [19]. Subsequently, it has been demonstrated that overfitting may not be a problem if overtraining is avoided by cross-validation [20]. The ρ values in models A, B, and C vary from 0.7 (of the 6-2-1 architecture) to 1.3 (of the 6-1-1 configuration). In spite of the generalization performance and the low errors of cross-verify set (selection set) of model A, models B and C are more preferable based on simplicity principle. Due to higher overall R^2 and better generalization, model B was chosen as the final model, and the plot of actual versus predicted log salts solubility of the model is presented in Fig. 1.

The optimal number of descriptors chosen for the ANN models (A, B, and C) is six. All of these models use Hb, clog P, BE, SA, and MW as independent variables. HaO and HaN are also used by some models. Lipinski et al. have suggested that four parameters including molecular weight, log P, number of hydrogen bond donors, and the number of hydrogen bond acceptors are globally associated with aqueous solubility of compounds [21]. In addition to these four parameters, binding energy between each ion could play an important role in aqueous salt solubility. By adding BE, ANN models with good overall R^2 and predictive ability can be developed (models A, B and C). Without BE, the overall performance decreases as shown in models D and E (Table 2).

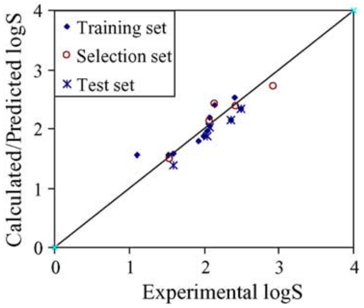


Fig. 1. Plot of calculated/predicted vs. experimental log S for the training, selection, and test set compounds used to develop model B.

In conclusion, ANN models with good predictive performance for estimating aqueous solubility of salt of the same basic compound have been developed. The ANN models were found to be more successful than PLS analysis, reflecting that the relationship between descriptors and solubility of benzylamine salts is nonlinear. These ANN models include only the calculated molecular descriptors, which make them suitable for use in salt designing.

Acknowledgement

This work was partially supported by Prince of Songkla University.

References

[1] K.M. O'Connor, O.I. Corrigan, Int. J. Pharm. 226 (2001) 163–179.
[2] R.T. Forbes, P. York, J.R. Davidson, Int. J. Pharm. 126 (1995) 199–208.
[3] D. Yaffe, Y. Cohen, G. Epinosa, A. Arenas, F. Giralt, J. Chem. Inf. Comput. Sci. 41 (2001) 1177–1207.
[4] X. Chen, S.J. Cho, S. Venkatesh, J. Pharm. Sci. 91 (2002) 1838–1852.

[5] H. Gao, V. Shanmugasundaram, P. Lee, Pharm. Res. 19 (2002) 497–503.

[6] W.L. Jorgensen, E.M. Duffy, Adv. Drug. Deliv. Rev. 54 (2002) 355–366.

[7] I.V. Tetko, V.Y. Tanchuk, T.N. Kasheva, A.E.P. Villa, J. Chem. Inf. Comput. Sci. 41 (2001) 1488–1493.

[8] P. Bruneau, J. Chem. Inf. Comput. Sci. 41 (2001) 1605–1616.

[9] R. Liu, H. Sun, S.S. So, J. Chem. Inf. Comput. Sci. 41 (2001) 1633–1639.

[10] H. Parshad, K. Frydenvang, T. Liljefors, C.S. Larsen, Int. J. Pharm. 237 (2002) 193–207.

[11] V. Tantishaiyakul, Int. J. Pharm. 275 (2004) 133–139.

[12] J. Huuskonen, J. Chem. Inf. Comput. Sci. 40 (2000) 773–777.

[13] O. Engkvist, P. Wrede, J. Chem. Inf. Comput. Sci. 42 (2002) 1247–1249.

[14] P.D.T. Huibers, A.R. Katritzky, J. Chem. Inf. Comput. Sci. 38 (1998) 283–292.

[15] J. Huuskonen, M. Salo, J. Taskinen, J. Pharm. Sci. 86 (1997) 450–454.

[16] C. Aleman, D. Zanuy, Chem. Phys. Lett. 319 (2000) 318–326.

[17] B. Madhan, P. Thanikaivelan, V. Subramanian, J.R. Rao, B.U. Nair, T. Ramasami, Chem. Phys. Lett. 346 (2001) 334–340.

[18] C.Q. Xia, J.J. Yang, S. Ren, E.J. Lien, J. Drug Target. 6 (1998) 65–77.

[19] T.A. Andrea, H. Kalayeh, J. Med. Chem. 34 (1991) 2824–2836.

[20] I.V. Tetko, D.J. Livingstone, A.I. Luik, J. Chem. Inf. Comput. Sci. 35 (1995) 826–833.

[21] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, Adv. Drug Deliv. Rev. 23 (1997) 3–25.