Department of Drug Delivery Research, Graduate School of Pharmaceutical Sciences, Kyoto University, Kyoto, Japan

Quantitative structure/property relationship analysis on aqueous solubility using genetic algorithm-combined partial least squares method

S. Wanchana, F. Yamashita and M. Hashida

The present study was initiated to generate a model of predicting aqueous solubility of substances from their molecular structure. For 211 drugs or drug-like compounds, their topological indices were calculated by Molconn-Z software. The optimal subset of the descriptors for the prediction of aqueous solubility was determined by genetic algorithm in combination with partial least squares (PLS) method. Thirty-four descriptors were selected by this method. Using 29 of the descriptors selected, of which the scaled PLS coefficient was significant, the cross-validated predictive q^2 was 0.785 with 19 principal components that was the optimal and the standard error of prediction was 0.676. Thus, it is suggested that the model obtained would exhibit a good performance in predicting the aqueous solubility of compounds.

1. Introduction

Aqueous solubility is one of the most important factors in determining the usefulness of a drug candidate. For example, a poor solubility often hampers the bioavailability and makes the formulation of drugs difficult. The ability to predict the aqueous solubility of the compounds would assist us to remove inappropriate sub-libraries, which might have a tremendous impact both in terms of cost and time required for drug discovery and development.

Theoretical calculation of aqueous solubility is not feasible at present, but various empirical methods have been proposed. In these approaches, molecular descriptors are quantified based on 2-dimensional or 3-dimensional molecular structures, and then correlated with aqueous solubility through multivariate analyses. These molecular representations include fragment descriptors [1–3], topological indices [4–11], and quantum chemical parameters [12, 13]. On the other hand, multiple linear regression [4–7, 12, 13] and artificial neural networks [8–11] are often used to find a quantitative relationship between the descriptors and aqueous solubility.

The goal of the quantitative structure-property relationship (QSPR) is to find a small subset of the large number of calculated descriptors that can effectively predict the property. The exploratory analysis can be automatically achieved by the use of genetic algorithm [14] and simulated annealing [15] that are algorithms to solve combinatorial optimization problems. Both routines are iterative optimization techniques with a small degree of randomness that allow the solution to escape local minima traps and converge to near global conditions. Sutter and Jurs [6] explored multiple linear regression (MLR) models for the prediction of aqueous solubility by the use of both genetic algorithm and simulated annealing and demonstrated that both routines produced very similar models. Taking into account that multicollinearities between the descriptors should be eliminated in MLR analysis, however, combination of the exploratory routines with MLR would be limited for the use. One of the methods to escape the multicollinearity problem is the use of orthogonal transformation in regression analysis, such as principal component regression (PCR) or partial least squares (PLS) methods [16, 17]. Usefulness of the combination of genetic algorithm and partial least squares has been demonstrated by Tropsha et al., who analyzed QSAR modeling of dopamine D1 antagonists and identified the descriptor pharmacophores [18, 19].

In the present study, we explored a QSPR model for predicting aqueous solubility, using a genetic algorithm-combined partial least squares method. Structural descriptors were calculated by Molconn-Z software (Hall Associated Consulting, Quincy, MA) that included connectivity indices, shape indices, electrotopological state (E-state) indices and atom-type E-state indices. A dataset of aqueous solubilities of 211 drugs and drug-like compounds was taken from the literature [9]. We will also compare the ability of our approach with that of the previous method [9].

2. Investigations, results and discussion

To generate a model for predicting aqueous solubility, 220 structural descriptors were obtained by Molconn-Z software. Prior to optimization of a subset of the descriptors by genetic algorithm-based approach, the descriptors showing a heavily skewed distribution (skewness of greater than 3) were removed. The remaining 88 descriptors were subjected to the genetic algorithm-driven subset selection.

Fig. 1 shows the relationship between the average or maximum fitness and generation number. During genetic algorithm-driven optimization, the average of the fitness (See Eq. 5 in Experimental) tends to increase with increasing generation number and reached plateau at around 200 gen-

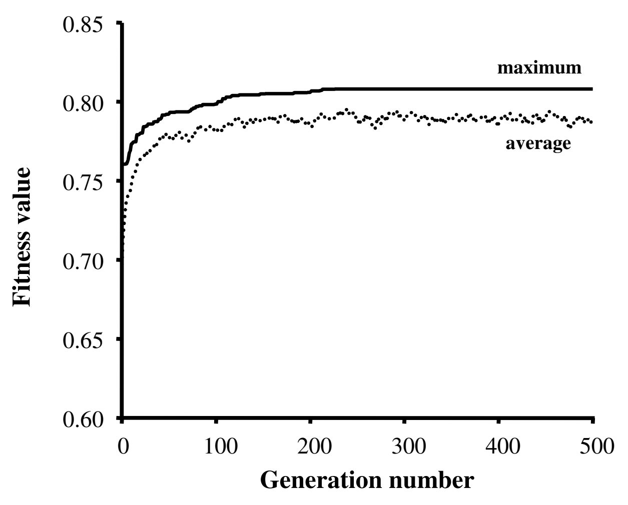

Fig. 1: Relationship between fitness and generation number by genetic algorithm-driven optimization

erations. The best solution at 500th generation was taken for the modeling of aqueous solubility.

The Table summarizes 34 Molconn-Z descriptors selected by genetic algorithm and their scaled PLS regression coefficients. When the descriptors with small scaled PLS regression coefficients were removed, predictability of the model was not so much changed. Therefore, 29 descriptors except $^5\chi_{ch}$, $^6\chi_{ch}$, $^4\chi_{vc}$, $^5\chi_{vch}$, and $^6\chi_{vch}$ were regarded effective for the prediction of aqueous solubility.

Fig. 2 illustrates a typical result of leave-some-out cross-validated prediction. When the leave-some-out prediction was repeated 15 times, the average predictive q^2 was 0.785 ± 0.009 with 19 principal components that was the optimal and the standard error of prediction was 0.676 ± 0.013. Thus, it is suggested that the model obtained would exhibit a good performance in predicting aqueous solubility of compounds.

Table: Scaled PLS regression coefficients of the subset of Molconn-Z descriptors selected by genetic algorithm in combination with PLS regression

Symbol	Description of descriptor	Scaled PLS regression coefficient[a]
fw	Molecular weight	2.8693
$^0\chi$	Path 0 simple connectivity index	7.1452
$^3\chi_p$	Path 3 simple connectivity index	−0.9645
$^5\chi_p$	Path 5 simple connectivity index	−1.2517
$^{10}\chi_p$	Path 10 simple connectivity index	−1.7115
$^0\chi_v$	Path 0 valence connectivity index	−9.0408
$^1\chi_v$	Path 1 valence connectivity index	2.3259
$^2\chi_v$	Path 2 valence connectivity index	0.3538
$^6\chi_{vp}$	Path 6 valence connectivity index	−0.9822
$^7\chi_{vp}$	Path 7 valence connectivity index	1.4442
$^8\chi_{vp}$	Path 8 valence connectivity index	0.2625
$^9\chi_{vp}$	Path 9 valence connectivity index	0.1748
$^5\chi_{ch}$	Chain 5 simple connectivity index	−0.0113
$^6\chi_{ch}$	Chain 6 simple connectivity index	0.0453
$^4\chi_{vc}$	Cluster 4 valence connectivity index	−0.0482
$^5\chi_{vch}$	Chain 5 valence connectivity index	−0.0064
$^6\chi_{vch}$	Chain 6 valence connectivity index	0.0104
totop	Total topological index t	1.7284
sumI	Sum of the intrinsic state values I	−3.8882
SHsOH	The sum of E-state values for hydrogen atom-type (−OH)[b]	0.6656
Hmax	Maximum hydrogen E-state value in molecules	−0.1641
Hmin	Minimum hydrogen E-state value in molecules	−0.5284
SssCH2	Atom type electrotopological state index values for atom types ($-CH_2-$)[b]	−0.7045
SaasC	Atom type electrotopological state index values for atom types (−..C..)[b]	−0.6175
SssNH	Atom type electrotopological state index values for atom types (−NH−)[b]	0.6722
SsssN	Atom type electrotopological state index values for atom types (>N−)[b]	0.5143
numHBa	Number of hydrogen bond acceptors	0.2742
SHCsatu	E-state of C sp^3 bonded to unsaturated C atoms	0.2173
SHHBb	Hydrogen bond donor index	−0.6723
NHBint2	Count of potential internal H-bonds	−0.1745
NHBint3	Count of potential internal H-bonds	0.1302
NHBint7	Count of potential internal H-bonds	0.1024
NHBint10	Count of potential internal H-bonds	0.1878
SHBint2	E-state descriptors of potential internal H-bonds strength	−0.2400

[a] Scaled PLS regression coefficients were calculated by multiplying partial regression coefficient and standard deviation of the descriptors.

[b] The formula of the atom type or group; the bond types between the heavy atoms are s = single (−), d = double (=), and a = aromatic (..)

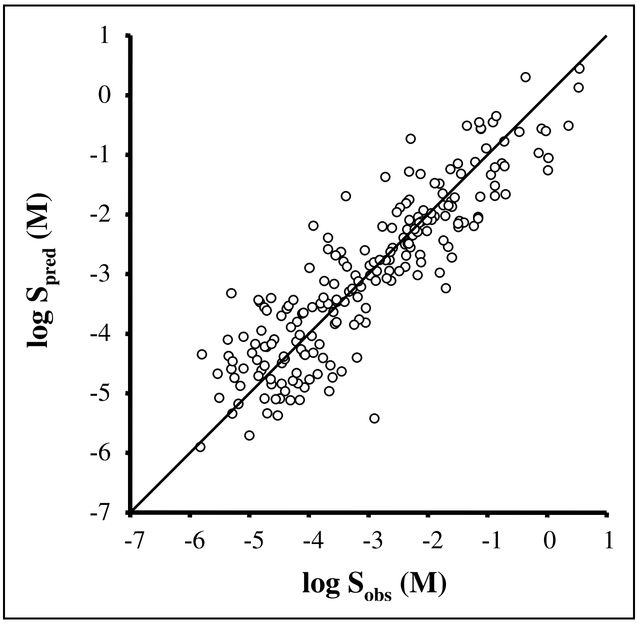

Fig. 2: Leave-some-out cross-validation prediction of aqueous solubility of drugs from the subset of Molconn-Z descriptors selected by genetic algorithm in combination with PLS regression

The PLS regression with 19 principal components was conducted with all data to generate the equation for the prediction of aqueous solubility. When the PLS matrix equation was expanded, the following linear equation was obtained:

$$\begin{aligned}\text{Log S} = {} & 1.0785 + 0.0371\ \text{fw} + 1.8145\ {}^0\chi - 0.3719\ {}^3\chi_p \\ & - 0.5303\ {}^5\chi_p - 2.8378\ {}^{10}\chi_p - 2.5517\ {}^0\chi_v \\ & + 1.0780\ {}^1\chi_v + 0.0900\ {}^2\chi_v - 0.7639\ {}^6\chi_{vp} \\ & + 1.4571\ {}^7\chi_{vp} + 0.3523\ {}^8\chi_{vp} + 0.3515\ {}^9\chi_{vp} \\ & + 0.0282\ \text{totop} - 0.3011\ \text{sumI} + 0.3159\ \text{SHsOH} \\ & - 0.2987\ \text{Hmax} - 1.7444\ \text{Hmin} - 0.2547\ \text{SssCH2} \\ & - 0.3570\ \text{SaasC} + 0.3461\ \text{SssNH} + 0.4580\ \text{SsssN} \\ & + 0.1839\ \text{numHBa} + 0.1776\ \text{SHCsatu} \\ & - 0.3064\ \text{SHHBb} - 0.2469\ \text{NHBint2} \\ & + 0.2265\ \text{NHBint3} + 0.2568\ \text{NHBint7} \\ & + 0.5279\ \text{NHBint10} - 0.0194\ \text{SHBint2} \\ & (r^2 = 0.847,\ s = 0.587,\ n = 211) \qquad (1)\end{aligned}$$

Several researchers demonstrated that artificial neural network is useful for QSPR modeling of aqueous solubility [8−11]. Artificial neural network can perform nonlinear approximations, so that the complicated inter-relationship between explanatory variables and response variable can be simulated. In some cases, however, it is likely that neural network models might be lack of robustness due to a many degree-of-freedom. In the PLS regression, latent factors that account for most of the variation in the response are extracted from the data matrix [16, 17]. The number of latent factors is usually less than that of explanatory variables; in other words, the degree-of-freedom in the PLS regression is less than in the conventional multiple linear regression. In addition, the extracted factors, so-called scores, are orthogonal to one another. Thus, the PLS is generally considered sufficiently robust and useful for the QSPR modeling.

Huuskonen et al. [9] generated a prediction model for aqueous solubility from Molconn-Z parameters using an artificial neural network with a configuration of 23 : 5:1. Their model was able to estimate, with a reasonable degree of accuracy, most of the aqueous solubilities of the training dataset ($r^2 = 0.90$, $s = 0.46$, $n = 160$) and the testing dataset ($r^2 = 0.86$, $s = 0.53$, $n = 51$). When our approach was applied to the same datasets, the PLS regression model consisting of 15 principal components from 53 descriptors was obtained, with a comparable degree of accuracy ($r^2 = 0.85$, $s = 0.57$ for training and $r^2 = 0.89$, $s = 0.47$ for testing). This suggests that selection of the optimal subset of the descriptors might be able to reduce the degree of freedom.

In conclusion, the proposed model, where Molconn-Z descriptors were used as molecular descriptors, was able to predict aqueous solubility with a reasonable degree of accuracy. The descriptors were derived from 2-D molecular graphs without the need for complicated 3-D conformational analysis. Since the calculation of Molconn-Z descriptors is easy and fast, our prediction approach would be useful for the screening of large virtual combinatorial libraries. The present study also revealed that genetic algorithm-combined partial least squares method is useful for optimization of the QSPR model and would be widely applicable to QSPR modeling of various physicochemical and biological activities.

3. Experimental

3.1. Dataset

The aqueous solubilities of 211 drugs were taken from the literature [9]. The solubility were expressed as log units of molar solubility (M) and ranged from 0.545 to −5.824.

3.2. Calculated molecular descriptors

The topological descriptors were calculated by Molconn-Z software (Hall Associated Consulting, Quincy MA) on the basis of 2-dimensional structures. A total of 220 connectivity, shape and atom-type E-state indices were calculated from the two-dimensional geometry. Prior to application of genetic algorithm, the descriptors representing heavily skewed distribution with a skewness of greater than 3 were removed. This method reduced the entire descriptor pools to 88 members.

3.3. Genetic algorithm-driven optimization

A population of 100 random subsets of the structural descriptors was generated. Each subset was encoded as a binary string of digits, so-called chromosome. The value of "1" implied that the descriptor was regarded as of importance; while the value of "0" implied that the descriptor was disregarded. The length of each string should be equal to the total number of descriptors. Considering both goodness-of-fit to the training data and predictability of the testing data, the following value was basically defined as the fitness in genetic algorithm optimization:

$$\text{fitness} = \frac{2}{\frac{1}{R^{*2}} + \frac{1}{\text{predictive } q^2}} \tag{2}$$

where,

$$R^{*2} = 1 - \frac{n-1}{n-c-1}\,\frac{\sum (y_i - y_{cal})^2}{\sum (y_i - \bar{y})^2} \quad \text{for training dataset} \tag{3}$$

$$\text{Predictive } q^2 = 1 - \frac{\sum (y_i - y_{pred})^2}{\sum (y_i - \bar{y})^2} \quad \text{for testing dataset} \tag{4}$$

where n was the number of compounds; c was the optimal number of components; $\bar{y}$ was an average of y_i values; and y_{cal} and y_{pred} were the theoretical values for training and testing datasets, respectively. The regression coefficient corrected by the degree of freedom (R^{*2}) was used to avoid over-fitting to the training data. Please note that the fitness is the maximal when R^{*2} and predictive q^2 are equal assuming that their total is constant. Leave-some-out cross-validation was made by randomly dividing the original dataset into 7 groups. The calculation was repeated three times and the fitness was then as follows:

$$\text{fitness} = \frac{6}{\frac{3}{R^{*2}} + \sum^{3} \frac{1}{\text{predictive } q^2}} \tag{5}$$

In genetic algorithm, two strings were selected randomly by a roulette wheel selection method according to fitness values. Two-point crossover of the "parent" strings was performed at a predefined probability (P) of 0.8. One of the new strings was taken, subjected to random mutation ($P = 0.01$), and stored in the next generation. Top 5 strings with high fitness values were kept for the next generation (Elite = 5). In each generation, a series of these steps were repeated until the predefined number (100) of population was obtained. After 500 generations, the best string that was generated by genetic algorithm-driven optimization was taken. The standard error (s) and correlation coefficient (r) for prediction were evaluated as follows:

$$s = \sqrt{\frac{\sum (y_i - y_{pred})^2}{n}} \tag{6}$$

$$r = \sqrt{\frac{\sum (y_i - \bar{y})(y_{i,pred} - \bar{y}_{pred})}{n}} \tag{7}$$

This research paper was presented during the 3rd Conference on Retrometabolism Based Drug Design and Targeting, May 13–16, 2001, Amelia Island. Florida, USA.

References

1 Klopman, G.; Wang, S.; Balthasar, D. M.: J. Chem. Inf. Comp. Sci. **32**, 474 (1992)
2 Kühne, R.; Ebert, R.-U.; Kleint, F.; Schmidt, G.; Schüürmann, G.: Chemosphere **30**, 2061 (1995)
3 Lee, Y.-H.; Myrdal, P. B.; Yalkowsky, S. H.: Chemosphere **33**, 2129 (1996)
4 Patil, G. S.: J. Hazard. Mater. **36**, 35 (1994)
5 Nelson, T. M.; Jurs, P. C.: J. Chem. Inf. Comput. Sci. **34**, 601 (1994)
6 Sutter, J. M.; Jurs, P. C.: J. Chem. Inf. Comp. Sci. **36**, 100 (1996)
7 Huuskonen, J.: Environ. Toxicol. Chem. **20**, 491 (2001)
8 Huuskonen, J.; Salo, M.; Taskinen, J.: J. Pharm. Sci. **86**, 450 (1997)
9 Huuskonen, J.; Salo. M.; Taskinen, J.: J. Chem. Inf. Comp. Sci. **38**, 450 (1998)
10 Mitchell, B. E.; Jurs, P. C.: J. Chem. Inf. Comp. Sci. **38**, 489 (1998)
11 Huuskonen, J.; Rantanen, J.; Livingstone, D.: Eur. J. Med. Chem. **35**, 1081 (2000)
12 Bodor, N.; Huang, M. -J.: J. Pharm. Sci. **81**, 954 (1992)
13 Huibers, P. D. T.; Katritzky, A. R.: J. Chem. Inf. Comp. Sci. **38**, 283 (1998)
14 Luke, B. T.: J. Chem. Inf. Comp. Sci. **34**, 1279 (1994)
15 McClelland, H. E.; Jurs, P. C.: J. Chem. Inf. Comp. Sci. **40**, 967 (2000)
16 Geladi, P.; Kowalski, B. R.: Anal. Chim. Acta **185**, 1 (1986)
17 Geladi, P.; Kowalski, B. R.: Anal. Chim. Acta **185**, 19 (1986)
18 Hoffman, B.; Cho, S. J.; Zheng, W.; Wyrick, S.; Nichols, D. E.; Mailman, R. B.; Tropsha, A.: J. Med. Chem. **42**, 3217 (1999)
19 Tropsha, A.; Zheng, W.: Curr. Pharm. Des. **7**, 599 (2001)

Prof. Dr. Mitsuru Hashida
Department of Drug Delivery Research
Graduate School of Pharmaceutical Sciences
Kyoto University
Sakyo-ku, Kyoto 606−8501
Japan