Journal of Computer-Aided Molecular Design, 8 (1994) 449–456
ESCOM

J-CAMD 252

Chemometric QSAR studies of antifungal azoxy compounds

Kiyoshi Hasegawa[a], Takeo Deushi[a], Hiroshi Yoshida[b], Yoshikastu Miyashita[b,*] and Shin-ichi Sasaki[b]

[a]*Tokyo Research Laboratories, Kowa Co. Ltd., 17-43-2, Noguchi-cho Higashimurayama, Tokyo 189, Japan*
[b]*Department of Knowledge-Based Information Engineering, Toyohashi University of Technology, Tempaku-cho, Toyohashi 441, Japan*

Received 21 December 1993
Accepted 14 March 1994

Key words: PLS; QSAR; Factorial design; Chemometrics; Antifungal activity; Azoxy compounds

SUMMARY

Quantitative structure–activity relationships (QSARs) for 16 azoxy compounds with antifungal activity have been studied by the combined approach of a partial least-squares method and factorial design. The PLS model equation suggested the structural requirements of two substituents, R_1 and R_2, for the antifungal activity. The sterically bulky and hydrophobic R_1 substituents and electron-withdrawing R_2 substituents are favorable for the activity. We propose candidate compounds which are more potent than the compounds based on QSAR data. In this study, we show that the chemometric approach is a powerful tool for QSAR studies and drug design.

INTRODUCTION

In medicinal chemistry, effective modelling of activity profiles is an important task in the development of novel therapeutic drugs. Quantitative structure–activity relationships (QSARs) have been widely studied for this purpose. QSAR leads to a mathematical model, relating the variations in chemical structures to those in biological activities. Conventionally, the Hansch approach using classical multiple linear regression (MLR) [1] has been used in QSAR studies. However, the MLR method has some disadvantages (collinearity and unreliable prediction problems) and is not always applicable to QSAR modelling [2].

Recently, chemometrics has been developed from analytical chemistry [3,4]. It is a new discipline which applies mathematical and statistical tools in chemistry. Some chemometric methods (SIMCA, PLS, etc.) are useful for solving QSAR problems [5] which were intractable before.

*To whom correspondence should be addressed.

Abbreviations: PLS, partial least squares; FD, factorial design; MLR, multiple linear regression; PPs, principal properties.

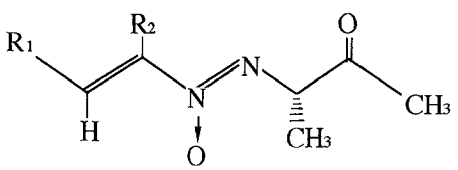
0920-654X/$ 10.00 © 1994 ESCOM Science Publishers B.V.

Among the various kinds of chemometric methods, the partial least-squares (PLS) method plays an important role in analytical chemistry [6] and QSAR analysis [7–9].

Hellberg et al. [10] have proposed principal properties (PPs) of natural amino acids, derived from 29 physicochemical parameters for each amino acid. They have applied the PLS method to several peptide QSAR problems using these PPs as structural descriptors and obtained excellent QSAR models. Skagerberg et al. [11] have proposed the PPs of 100 aromatic substituents, which are available from nine physicochemical parameters. Recently, De Meo et al. [12] have carried out QSAR studies for antibacterial and antimycotic activities of benzofused heteroaromatic derivatives using aromatic PPs. Thus, PPs derived from various physicochemical parameters possess significant information and can be used as useful structural descriptors.

In this paper, we have performed the QSAR analysis of antifungal azoxy compounds using PLS modelling techniques. The aromatic PPs proposed by Skagerberg et al. [11] were used as structural descriptors of the azoxy compounds for the PLS analysis. The aromatic PPs of some substituents in the data set compounds that are not listed in Ref. 11 were calculated by the original physicochemical parameters and principal component loadings. After PLS analysis, the PLS model equation was used to interpret the structural requirements for antifungal activity. The structural requirement information was combined with factorial designs (FD) [13] and the PLS model equation was used to predict candidate compounds with higher antifungal activity.

TABLE 1
CHEMICAL STRUCTURE AND OBSERVED AND CALCULATED ANTIFUNGAL ACTIVITY OF AZOXY COMPOUNDS



No.	R_1	R_2	Obsd[a]	Calcd[b]
1	CH_3	H	4.10	4.55
2	C_2H_5	H	5.04	4.86
3	n-C_3H_7	H	5.06	5.14
4	n-C_4H_9	H	5.39	5.60
5	n-C_5H_{11}	H	6.00	5.92
6	$CH{=}CH_2$	H	4.12	4.00
7	C_6H_5	H	4.85	4.57
8	$CH_2C_6H_5$	H	5.77	5.57
9	n-C_4H_9	CH_3	6.30	6.02
10	n-C_4H_9	n-C_5H_{11}	4.93	5.46
11	H	C_6H_5	4.54	4.66
12	SCH_3	C_6H_5	4.91	4.58
13	4-NO_2-C_6H_4	H	3.72	3.73
14	$(CH_2)_2C_6H_5$	H	5.20	5.32
15	$(CH_2)_2C_6H_5$	CH_3	5.51	5.74
16	n-C_4H_9	4-Cl-C_6H_4	6.22	5.93

[a] Observed antifungal activity (log(1/MIC)).
[b] Antifungal activity, calculated by means of Eq. 8.

MATERIALS AND METHODS

Data set for QSAR study

Sixteen azoxy compounds with antifungal activity were used for the QSAR studies. The antifungal activity was expressed as the logarithm of the reciprocal of the MIC (Minimum Inhibitory Concentration) against *Trichophyton rubrum*. These compounds were synthesized and tested in our group [14]. The chemical structures and the antifungal activity are listed in Table 1.

Calculation of principal properties

The structural descriptor used in the PLS analysis is a six-dimensional vector **x**:

$$\mathbf{x} = (Z_1(R_1), Z_2(R_1), Z_3(R_1), Z_1(R_2), Z_2(R_2), Z_3(R_2)) \quad (1)$$

The elements of the vector **x** are three PPs ($Z_1(R_1)$, $Z_2(R_1)$, $Z_3(R_1)$) for substituent R_1 and three PPs ($Z_1(R_2)$, $Z_2(R_2)$, $Z_3(R_2)$) for substituent R_2. These PPs were taken from the work of Skagerberg et al. [11]. From principal component loadings, it is seen that the first PP mainly represents steric bulkiness and hydrophobicity. The second and third PP mainly represent the electronic properties and hydrophobicity, respectively.

Skagerberg et al. [11] used nine physicochemical parameters to estimate the PPs of the substituents. Each substituent is expressed by a vector **u** as follows:

$$\mathbf{u} = (\pi, MR, \sigma_m, \sigma_p, L, B_i, B_{ii}, B_{iii}, B_{iv}) \quad (2)$$

where π is a hydrophobic substituent constant and MR is the molar refractivity. σ_m and σ_p are the Hammett sigma constant in the meta and para positions, respectively. L, B_i, B_{ii}, B_{iii} and B_{iv} are modified Verloop steric parameters. B_i is the smallest width and B_{ii} is the opposite of B_i. B_{iii} is orthogonal to B_i and smaller than B_{iv}.

		$\mathbf{u} =$ (π,	MR,	σ_m,	σ_p,	L,	B_i,	B_{ii},	B_{iii},	B_{iv})
(1)	4-NO_2-C_6H_4	1.70	31.69	*	0.23	7.66	1.70	1.70	3.11	3.11
(2)	$(CH_2)_2C_6H_5$	2.66	34.65	*	0.12	8.33	1.66	4.50	1.90	1.90
(3)	4-Cl-C_6H_4	2.67	30.36	*	-0.08	7.74	1.70	1.70	3.11	3.11

Data Reduction

$$\mathbf{z}^T = (\mathbf{P}^T\mathbf{P})^{-1}\mathbf{P}^T\mathbf{u}^T$$

	Z_1	Z_2	Z_3
(1)	5.614	-2.216	0.560
(2)	6.367	-1.273	0.597
(3)	5.913	-1.404	0.957

Fig. 1. A diagram for the procedure of calculating principal properties. An asterisk represents a missing value.

Twelve compounds (1–12) have substituents with known PPs; the remaining four compounds (13–16) have substituents with unknown PPs. These are 4-NO_2-C_6H_4, $(CH_2)_2C_6H_5$ and 4-Cl-C_6H_4. In order to develop a QSAR model for all 16 compounds, it is necessary to calculate three PPs for the unknown substituents by the following procedures (see Fig. 1). The π and MR values were estimated by the empirical model, using an additive rule [15]. The σ_p value was referred from the Hansch–Leo data [15]. The Verloop steric parameters were estimated using the Sterimol program [16]. Only the σ_m value is not available from the Hansch–Leo data. Hence, this property is missing in the nine physicochemical parameters of the substituents.

Wold et al. [17] proposed the NIPALS algorithm to calculate principal component scores for an incomplete data matrix during the training phase. When predicting the PPs vector, this is done by means of multiple regression where the new physicochemical data **u** is regressed against the loading matrix **P**. The regression is performed on nonmissing data in **u** and corresponding nonmissing variables in **P**. Three PPs for the substituents are calculated using the following equation:

$$\mathbf{z}^T = (\mathbf{P}^T\mathbf{P})^{-1}\mathbf{P}^T\mathbf{u}^T \tag{3}$$

where $\mathbf{P}^T$ is the transpose of the principal component loading matrix reported by Skagerberg et al. [11]. Table 2 lists PPs of all substituents, including the computed ones used in the QSAR analysis and prediction.

Statistical method

The PLS method was employed to obtain the structural requirements for antifungal activity. PLS is a novel multivariate statistical method for QSAR studies. For a detailed explanation of PLS, see Refs. 6, 7 and 18–20. A brief introduction of this method will be given below.

TABLE 2
PRINCIPAL PROPERTIES (Z_1, Z_2, Z_3) USED AS DESCRIPTORS FOR PLS ANALYSIS AND PREDICTION OF ACTIVITY

Substituent	Z_1	Z_2	Z_3
H	0.000	0.000	0.000
CH_3	1.220	−0.213	0.816
C_2H_5	2.275	−0.326	0.796
n-C_3H_7	3.203	−0.425	0.817
n-C_4H_9	4.298	−0.452	0.846
n-C_5H_{11}	5.231	−0.526	0.871
$CH{=}CH_2$	2.367	−0.944	0.593
C_6H_5	4.572	−1.398	1.099
$CH_2C_6H_5$	3.021	−0.552	2.501
SCH_3	2.381	−1.066	0.706
4-NO_2-C_6H_4	5.614	−2.216	0.560
$(CH_2)_2C_6H_5$	6.367	−1.274	0.597
4-Cl-C_6H_4	5.913	−1.404	0.957
NHC_6H_5	3.780	−0.244	1.997
Cyclohexyl	4.876	−0.855	1.869
SO_2CF_3	1.989	−3.731	1.281
SO_2F	1.243	−3.610	0.969
SF_5	2.410	−3.285	1.410

PLS relates the variance in a data set **y** to the variance in independent parameters **X**, using latent variables **t**. The **X** and **y** blocks can be represented in a principal component-like expression:

$$\mathbf{X} = \mathbf{X}_0 + \sum_{a=1}^{A} \mathbf{t}_a \mathbf{p}_a^T + \mathbf{E} \tag{4}$$

$$\mathbf{y} = \mathbf{y}_0 + \sum_{a=1}^{A} \mathbf{t}_a q_a + \mathbf{f} \tag{5}$$

where $\mathbf{X}_0$ and $\mathbf{y}_0$ are the corresponding mean value matrices, $\mathbf{t}_a$, $\mathbf{p}_a$ and q_a are the *a*th latent variable and loadings for the **X** and **y** block, respectively, **E** and **f** are residuals, A is the number of PLS components determined by cross-validation [21] and **t** is a linear combination of independent parameters **X** and PLS weight **w**, Eq. 6.

$$\mathbf{t} = \mathbf{X}\,\mathbf{w} \tag{6}$$

Two types of PLS analyses, PLS1 and PLS2, exist. PLS1 relates one **y** variable to **X** and PLS2 relates several variables in **Y** to **X**. In the present QSAR study, we employ PLS1. **y** is the antifungal activity of 16 azoxy compounds and **X** is 6 PPs for the substituents R_1 and R_2. The PLS analysis was carried out using the Unscrambler software package developed by Martens and Næs [6] on an IBM PS/2 microcomputer.

RESULTS AND DISCUSSION

PLS analysis of antifungal activity data

The data set (1–12) in which the PPs of the substituents for 12 compounds are known was studied first. The antifungal activity and six PPs were mean centered. A five-component PLS model was derived by the cross-validation procedure [21]. Converting the latent variables **t** to the original six PPs, the following QSAR equation (MLR-like equation) was obtained.

$$\begin{aligned}\log(1/\mathrm{MIC}) = {} & \underset{(\pm 0.016)}{0.460 Z_1(R_1)} + \underset{(\pm 0.044)}{0.980 Z_2(R_1)} + \underset{(\pm 0.075)}{0.523 Z_3(R_1)} - \underset{(\pm 0.034)}{0.447 Z_1(R_2)} - \underset{(\pm 0.160)}{1.109 Z_2(R_2)} + \\ & \underset{(\pm 0.285)}{1.297 Z_3(R_2)} + \underset{(\pm 0.097)}{3.583} \\ n = 12,\ A = 5,\ s = 0.187,\ r^2 = 0.960,\ q^2 = {} & 0.336\end{aligned} \tag{7}$$

where n, A, s, r^2 and q^2 are the number of compounds, the number of PLS components, the standard error, the squared conventional correlation coefficient and the cross-validated r-squared value by the leave-one-out procedure, respectively. The figures in parentheses are the standard deviations of the coefficients determined by the bootstrap analysis. It appears that structural information for the R_2 substituent is not reliable, because data variation of the R_2 substituent in the data set is not very large.

For the dataset (1–16) in which some PPs of the substituents were estimated, the activity and PPs were mean centered. Similar to the first study, a four-component PLS model was obtained by cross-validation and transformed to an MLR-like equation:

$$\log(1/MIC) = \underset{(\pm 0.024)}{0.438Z_1(R_1)} + \underset{(\pm 0.038)}{1.323Z_2(R_1)} + \underset{(\pm 0.038)}{0.400Z_3(R_1)} - \underset{(\pm 0.042)}{0.199Z_1(R_2)} - \underset{(\pm 0.130)}{0.633Z_2(R_2)} + \underset{(\pm 0.145)}{0.650Z_3(R_2)} + \underset{(\pm 0.085)}{3.974} \quad (8)$$

$n = 16$, $A = 4$, $s = 0.310$, $r^2 = 0.876$, $q^2 = 0.530$

QSAR Eq. 8 is significantly improved with respect to the validation step, compared with QSAR Eq. 7. The cross-validated r-squared value is not very high, but it is still interesting to predict the antifungal activity. The antifungal activity calculated by Eq. 8 is shown in Table 1. Figure 2 shows the plot of the calculated versus observed antifungal activity.

Note that QSAR Eqs. 7 and 8 have coefficients with the same signs for R_1 and R_2. From the QSAR equation, potent azoxy compounds can be found by increasing the steric bulkiness and hydrophobicity of the R_1 substituent and increasing the electron-withdrawing ability of the R_2 substituent.

Candidate compounds and factorial designs

Skagerberg et al. [11] have classified 100 aromatic substituents into eight groups by two-level (+/–) factorial design (FD) in three PPs of the substituents. Two-level FD in three PPs is geometrically represented by a cube in which eight vertices have combinatorial coordinates (+,+,+), (+,+,–), (+,–,+), (+,–,–) (–,+,+), (–,+,–), (–,–,+), (–,–,–) [13]. Each vertex corresponds to one substituent group. The substituents belonging to the same group are then considered to have similar physicochemical properties. We can easily propose candidate compounds with higher activity according to this idea.

The MLR-like QSAR equation indicates that the R_1 substituent should be selected from the (+,+,+) group in FD. Similarly, the R_2 substituent should be selected from the (–,–,+) group. The (+,+,+) group has the following seven substituents: *s*-C_4H_9, *n*-C_4H_9, *n*-C_5H_{11}, C_6H_5, OC_6H_5, NHC_6H_5 and cyclohexyl. The (–,–,+) group has the following 10 substituents: Br, SO_2F, SF_5, I,

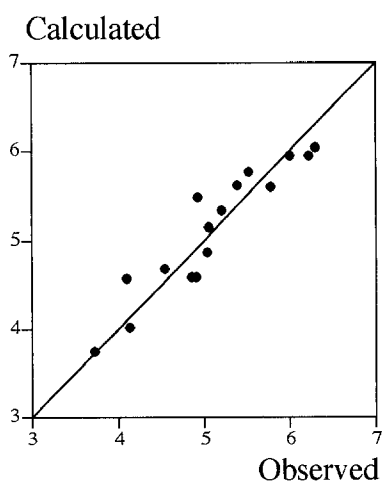


Fig. 2. Plot of calculated against observed antifungal activity of the azoxy compounds under investigation.

TABLE 3
CANDIDATE COMPOUNDS WITH INCREASED ANTIFUNGAL ACTIVITY

No.	R_1	R_2	Predicted[a]	Std[b]
1	NHC_6H_5	SO_2CF_3	8.86	0.506
2	NHC_6H_5	SO_2F	8.73	0.491
3	NHC_6H_5	SF_5	8.58	0.446
4	n-C_5H_{11}	SO_2CF_3	8.72	0.518
5	n-C_5H_{11}	SO_2F	8.58	0.504
6	n-C_5H_{11}	SF_5	8.44	0.459
7	Cyclohexyl	SO_2CF_3	8.53	0.508
8	Cyclohexyl	SO_2F	8.40	0.492
9	Cyclohexyl	SF_5	8.24	0.447

[a] Antifungal activity, predicted by means of Eq. 8.
[b] Prediction standard error.

CF_3, SCF_3, SO_2CF_3, CF_2CF_3, PMe_2 and COC_3H_7. From QSAR Eq. 8, the candidate compounds with the best three substituent types for R_1 (NHC_6H_5, n-C_5H_{11}, cyclohexyl) and with the best three substituent types for R_2 (SO_2CF_3, SO_2F, SF_5) can be proposed. The predicted antifungal activities from Eq. 8 and the prediction errors are listed in Table 3. The best candidate compound ($R_1 = NHC_6H_5$, $R_2 = SO_2CF_3$) has an antifungal activity of 8.86, which is about two orders larger than that of the QSAR data set compounds.

There is no way of testing the real reliability, except for synthesizing and testing the candidate compounds. However, it is important to assess the reliability of interpolated and extrapolated activities with statistical methods. It is possible to evaluate a measure of reliability for the predicted activity by calculating the sample RSD (Residual Standard Deviation) and comparing it with the RSD value of the calibration set [7,22]. If a sample RSD is smaller than the RSD value of the calibration set, the predicted activity is more reliable. In Table 2 it is easily seen that the R_1 substituents in the calibration set span the PPs space for R_1 fairly well and the candidate substituents (NHC_6H_5, n-C_5H_{11}, cyclohexyl) are located inside the PPs space of the calibration set. Therefore, the activity contributions of these three R_1 substituents to overall activity are reliable. On the other hand, the R_2 substituents in the calibration set span the PPs space for R_2 in a biased way and the candidate substituents of R_2 (SO_2CF_3, SO_2F, SF_5) are located outside the PPs space of the calibration set. The activity contributions of these three R_2 substituents are not very reliable. We should also consider the nonlinear behavior of the activity. However, it should be noted that the novel R_2 substituents of the candidate compounds were derived from the QSAR model equation. It is valuable to synthesize such compounds and to test their antifungal activity.

CONCLUSIONS

This paper demonstrates that the combined approach of PLS analysis and FD is useful for QSAR studies. From the PLS analysis, we have obtained the structural requirement information for antifungal activity. This information was used to predict candidate compounds with increased antifungal activity.

The aromatic PPs were employed as structural descriptors for the QSAR studies. PPs are considered to have three major advantages. First, they have a lower dimension than the original physicochemical parameters and are orthogonal. Thus, the choice of suitable structural descriptors from

many physicochemical parameters is not necessary, which is desirable and straightforward. Second, PPs can be grouped in factorial and fractional factorial designs. Aromatic substituents [11] or amino acids [10] belonging to a group are assumed to have similar physicochemical properties. The selection of substituents in the desirable direction is easy. In drug optimization, this is very important and quite meaningful information. Third, PPs can be easily calculated from the original physicochemical parameters. Even though they contain missing values due to experimental or computational difficulties, the estimated PPs are satisfactory. This easily leads to the development of compounds with novel substituents or peptides including nonnatural amino acids for drug optimization.

The PPs were proven to be useful structural descriptors for QSAR modelling. The PLS method has a high ability for finding the regularities between biological data and chemical structures. In the future, PLS modelling and FD using PPs as structural descriptors will be applied to many QSAR problems.

ACKNOWLEDGEMENTS

This work was supported by the Ministry of Education, Science and Culture under a Grant-in-Aid for Scientific Research and the Ishida Foundation. We thank the Computer Center of the Institute for Molecular Science for providing their computation facilities.

REFERENCES

1 Hansch, C. and Fujita, T., J. Am. Chem. Soc., 86 (1964) 1616.
2 Dunn III, W.J., Chemometrics Intelligent Lab. Syst., 6 (1989) 181.
3 Kowalski, B.R., Chemometrics; Theory and Application, American Chemical Society, Washington, DC, 1977.
4 Massart, D.L., Vandeginste, B.G.M., Deming, S.N., Michotte, Y. and Kaufman, L., Chemometrics: A Textbook, Elsevier, Amsterdam, 1988.
5 Miyashita, Y., Li, Z. and Sasaki, S., Trends Anal. Chem., 12 (1993) 50.
6 Martens, H. and Næs, T., Multivariate Calibration, Wiley, Chichester, 1989.
7 Dunn III, W.J., Wold, S., Edlund, U., Hellberg, S. and Gasteiger, J., Quant. Struct.–Act. Relatsh., 3 (1984) 131.
8 Miyashita, Y., Ohsako, H., Takayama, C. and Sasaki, S., Quant. Struct.–Act. Relatsh., 11 (1992) 17.
9 Hasegawa, K., Miyashita, Y., Sasaki, S., Sonoki, H. and Shigyou, H., Chemometrics Intelligent Lab. Syst., 16 (1992) 69.
10 Hellberg, S., Sjöström, M., Skagerberg, B. and Wold, S., J. Med. Chem., 30 (1987) 1126.
11 Skagerberg, B., Bonelli, D., Clementi, S., Cruciani, G. and Ebert, C., Quant. Struct.–Act. Relatsh., 8 (1989) 32.
12 De Meo, G., Pedini, M. and Ricci, A., Il Farmaco, 45 (1990) 313.
13 Box, G.E.P., Hunter, W.G. and Hunter, J.S., Statistics for Experiments, Wiley, New York, NY, 1978.
14 Deushi, T. et al., manuscript in preparation.
15 Hansch, C. and Leo, A., Substituent Constants for Correlation Analysis in Chemistry and Biology, Wiley, New York, NY, 1979.
16 Verloop, A., Hoogenstraaten, W. and Tipker, J., In Äriense, E.J. (Ed.) Drug Design, Vol. 7, Academic Press, New York, NY, 1976, pp. 165–207.
17 Wold, S., Albano, C., Dunn III, W.J., Esbensen, K., Hellberg, S., Johansson, E. and Sjöström, M., In Martens, H. and Russwurm Jr., H. (Eds.) Food Research and Data Analysis, Applied Science Publishers, London, 1983, pp. 147–188.
18 Geladi, P. and Kowalski, B.R., Anal. Chim. Acta, 185 (1986) 19.
19 Höskludsson, A., J. Chemometrics, 2 (1988) 221.
20 Glen, W.G., Dunn III, W.J. and Scott, D.R., Tetrahedron Comput. Methodol., 2 (1989) 349.
21 Wold, S., Technometrics, 20 (1978) 397.
22 Okuyama, T., Miyashita, Y., Kanaya, S., Katsumi, H., Sasaki, S. and Randić, M., J. Comput. Chem., 9 (1988) 636.