ELSEVIER

Journal of Molecular Structure (Theochem) 541 (2001) 81–88

THEO CHEM

www.elsevier.nl/locate/theochem

A multiple linear regression and partial least squares study of flavonoid compounds with anti-HIV activity

C.N. Alves[a,b], J.C. Pinheiro[a], A.J. Camargo[b], M.M.C. Ferreira[c], R.A.F. Romero[d], A.B.F. da Silva[b,*]

[a]*Departamento de Química, Centro de Ciências Exatas e Naturais, Universidade Federal do Pará, CP 11101, 66075-110, Belém, PA, Brazil*
[b]*Departamento de Química e Física Molecular, Instituto de Química de São Carlos, Universidade de São Paulo, CP 780, 13560-970, São Carlos, SP, Brazil*
[c]*Departamento de Físico-Química, Instituto de Quimica, Universidade Estadual de Campinas, Campinas, SP, 13081-970, Brazil*
[d]*Departamento de Ciência da Computação e Estatística, Instituto de Ciência da Computação e Estatística, Universidade de São Paulo, CP 668, 13560-970, São Carlos, SP, Brazil*

Received 20 April 2000; revised 6 September 2000; accepted 6 September 2000

Abstract

The molecular orbital semi-empirical method PM3 was employed to calculate a set of molecular properties (variables or descriptors) of 21 flavonoid compounds with anti-HIV activity. The correlation between biological activity and structural properties was obtained by using the multiple linear regression and partial least squares methods. The model obtained showed not only statistical significance but also predictive ability. The significant molecular descriptors related to the compounds with anti-HIV activity were: electronegativity (χ) and the charges on atoms C3 and C7 (Q_3 and Q_7, respectively). These variables led to a physical explanation of electronic molecular property contributions to HIV inhibitory potency.

Keywords: Flavonoids; Multiple linear regression; Partial least squares; PM3; Human immunodeficiency virus; QSAR

1. Introduction

Human immunodeficiency virus type 1 (HIV-1) [1,2] is the causative agent of acquired immunodeficiency syndrome (AIDS), which is characterized as a systemic and fatal disorder. Recently, a series of flavonoid compounds with active anti-HIV principle have been isolated and synthesized [3–5]. Previous works have also reported a series of other activities related to flavonoid compounds [6–10].

Extensive synthetic and structure–activity studies have been carried out on anti-HIV compounds [11–13]. The overall picture which emerges from these studies shows that the hydrophobic, electrostatic, and steric characteristics of substituents have a predominant role in the anti-HIV activity [14–15].

In a previous work [16], we made use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) to investigate which descriptors would be more effective for classifying 21 flavonoid compounds according to their degree of anti-HIV activity. In that work we found that only five variables, namely LUMO (the energy of the lowest unoccupied molecular orbital), χ (electronegativity),

* Corresponding author. Tel.: +55-16-273-9975; fax: +55-16-273-9975.
E-mail address: alberico@iqsc.sc.usp.br (A.B.F. da Silva).

0166-1280/01/$ - see front matter
PII: S0166-1280(00)00755-7

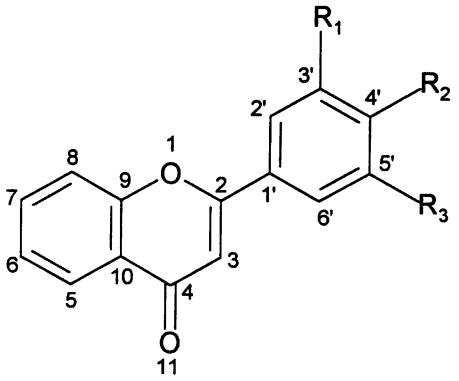

Fig. 1. Structural skeleton and numbering of the flavonoid compounds studied.

and Q_2, Q_3 and Q_7 (charges on atoms 2, 3 and 7, respectively), were responsible for classifying the 21 flavonoid in more active and less active compounds.

In the present work, the multiple linear regression (MLR) and partial least squares (PLS) methods [17–18] are employed with the aim to obtain a correlation between the five previously calculated descriptors and the anti-HIV activity of the 21 flavonoid compounds [16]. The main structure and numbering adopted are shown in Fig. 1, and their structures are shown in Fig. 2.

2. Calculation

The biological evaluation of the flavonoids was done by using the log of the numerical indicator for activity, EC_{50}, that indicates pharmacological potency (concentration which inhibits virus replication by 50%) [3]. The respective log EC_{50} values for all the 21 flavonoid compounds studied are shown in Table 1.

The five variables previously selected by PCA and HCA analyses [16] were used initially for the construction of the MLR and PLS models for the data matrix described in Table 1. The PIROUETTE and MATLAB softwares [19–20] were used in our statistical analyses.

The descriptor matrix $\mathbf{X} = (n,m)$ with $n = 21$ rows (n being the number of samples) and $m = 5$ columns was autoscaled as pretreatment, i.e. was meancentered (each element of data matrix was subtracted by its mean column) and scaled to variance of one (each element divided by the standard deviation of its column) before the MLR and PLS analyses.

The multivariate technique of partial least squares (PLS) regression was used for modelling [18,21]. In the PLS method, the latent variables are computed taking into account both the descriptors and activity values (y), making use of more information when building the model. Contrary to MLR regression method, PLS allows the simultaneous use of strongly intercorrelated descriptors by focusing the systematic covariances in the **X** block in a few latent variables. In modelling it is essential to determine its complexity to avoid overfitting. The predictive capability of the resulting model depends on the quality of the data (the more and better the data available, the more accurate prediction is possible), and the number k of significant latent variables necessary. Cross-validation is a practical and reliable method for testing this significance. In principle, the so-called 'leave-one-out' approach consists in developing a number of models with one sample omitted at the time. After developing each model, the omitted data is predicted and the differences between actual and predicted y(log EC_{50}) values are calculated. The sum of squares of these differences is computed and finally the performance of the model (its predictive ability) can be given by the standard error of prediction (SEP) defined as

$$\mathrm{SEP} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}$$

where y is the experimental log EC_{50}, $\hat{y}$ is the predicted value and n is the number of samples used for model building.

The predictive ability of the model was also quantified in terms of the Q^2 which is defined as:

$$Q^2 = 1.0 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \quad \text{where } \bar{y} = y_{\mathrm{mean}}$$

3. Results and discussion

3.1. Multiple linear regression (MLR) analysis

From the data matrix displayed in Table 1, the

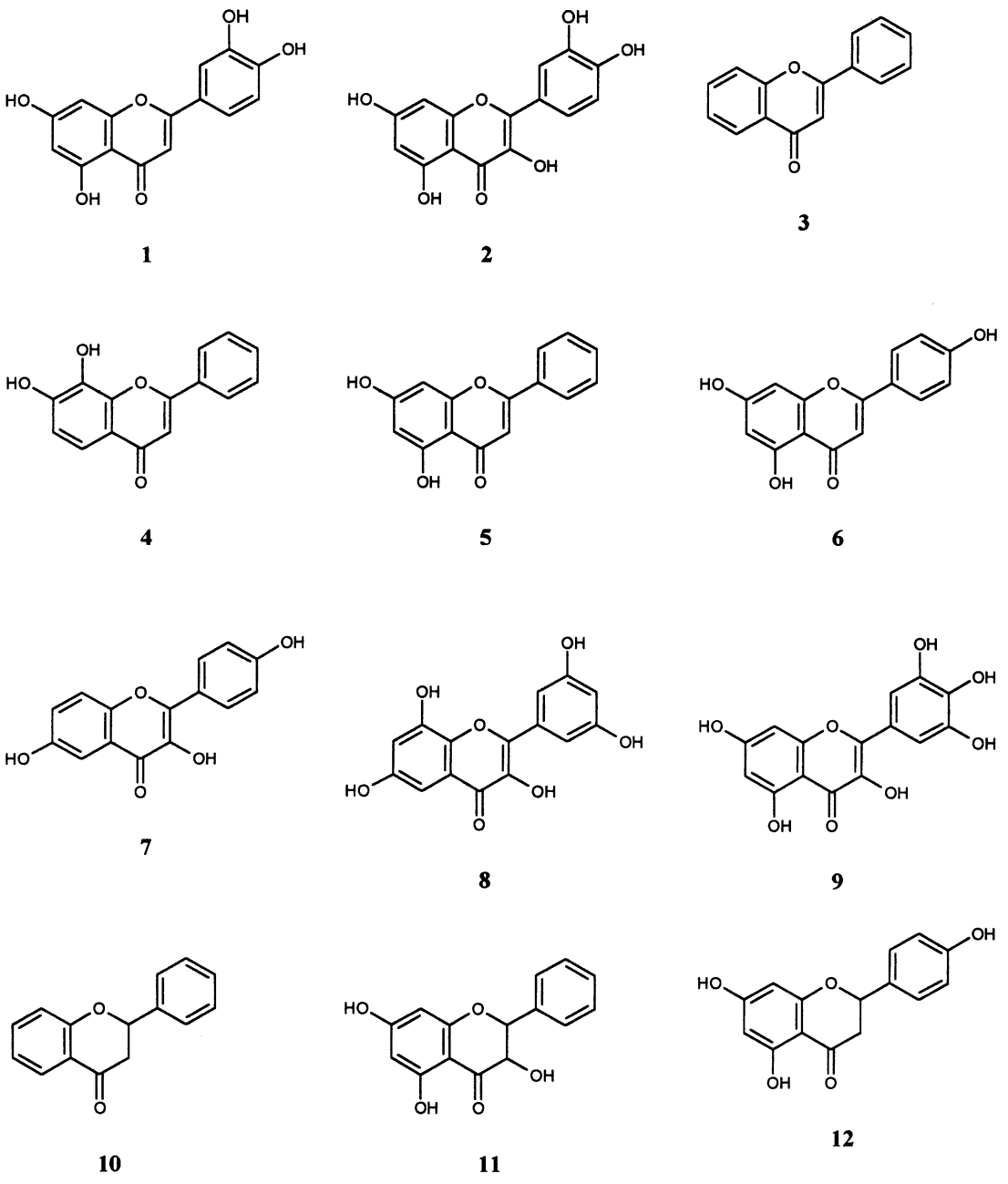

Fig. 2. Structure of the 21 flavonoids studied.

following regression equation was obtained:

$$\log EC_{50} = 15.98 + 0.12\text{LUMO} - 2.81\chi - 1.80Q_2 + 0.92Q_3 - 0.94Q_7 \quad (1)$$

$$R^2 = 0.98\ Q^2 = 0.74\ \text{SEP} = 0.29\ F = 10.57$$

where R^2 is the general correlation coefficient, and F is the Fisher test for significance of the equation.

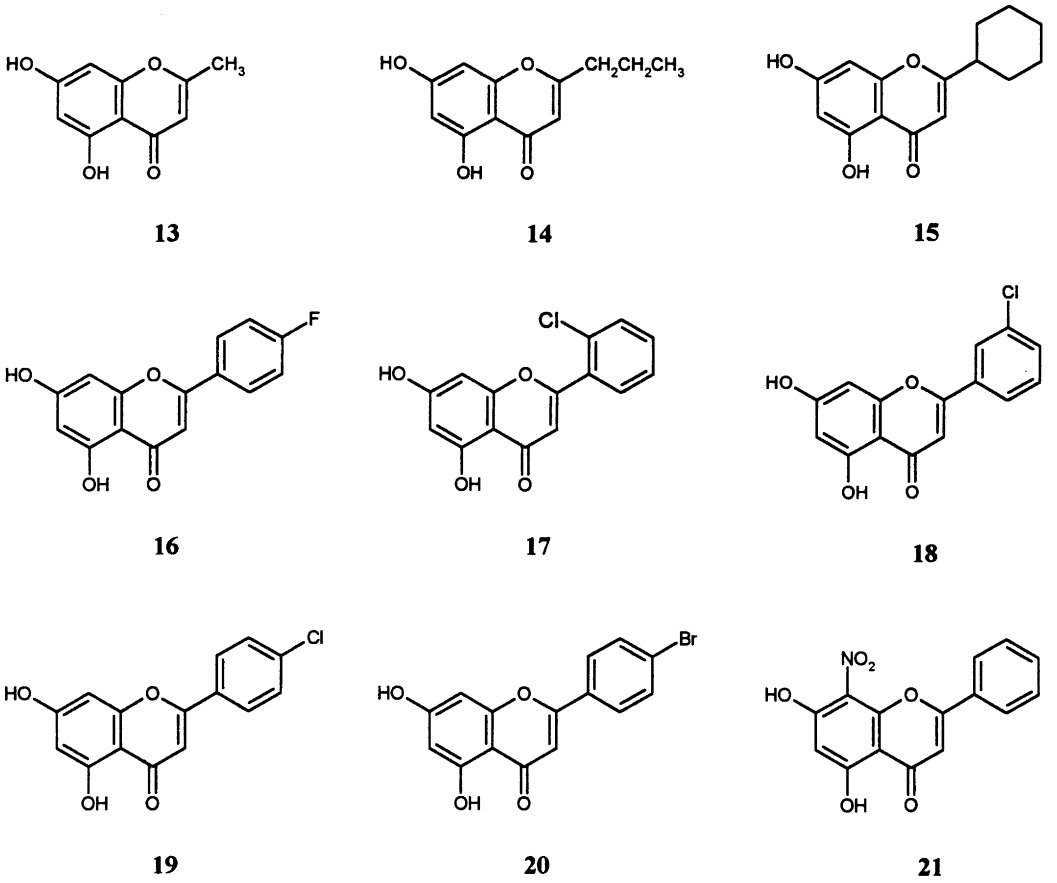

Fig. 2. (*continued*)

The statistical quality of Eq. (1) is good and accounts for 98% of the variance in $\log EC_{50}$. Compound 21 was excluded, since it was considered as an outlier due to the fact that none of the descriptors used in this study describe the presence of the group NO_2 in carbon number 8 (C8). It is interesting to notice that only compound 21 contains a NO_2 group in C8 (see Fig. 2).

The evaluation of the regression coefficients in the model — Eq. (1) — shows that χ e Q_2 have the highest contributions to the variation in $\log EC_{50}$. In a following step, the so-called statistical method of stepwise multiple regression procedure, based on the forward-selection and back ward-elimination methods, was used for variable selection with the aim to obtain the best regression equation (in such a way that variables that swallow little increment or are redundant in the explanation of the dependent variable ($\log EC_{50}$) were not included). In order to avoid overfitting or difficulties in interpretation of the resulting models, pairs of variables with $r \geq 0.7$ were classified as intercorrelated ones, and only one of the variables was included in the model. After some consideration, the following equation was obtained:

$$\log EC_{50} = 16.39 - 2.90\chi + 1.18Q_3 - 0.98Q_7$$
$$R^2 = 0.90\ Q^2 = 0.80\ \text{SEP} = 0.25\ F = 24.77 \quad (2)$$

There is some general improvement in the statistics of Eq. (2) as compared to Eq. (1). The F test as well as the predictive capability expressed by Q^2 and SEP, are slightly better in Eq. (2) than in Eq. (1) since Eq. (2) has smaller residuals for higher activity.

Table 1
Values of the five most important properties (descriptors) that classify the 21 flavonoid compounds studied and their respective values of log EC_{50} (units: Lumo and χ are in eV)

Compound	LUMO	χ	Atomic charges			log EC_{50} [3]
			Q_2	Q_3	Q_7	
1	1.064	5.089	0.1670	−0.3072	0.2021	1.00
2	1.082	4.904	0.1015	−0.0978	0.2115	2.12
3	0.904	5.101	0.1563	−0.3130	−0.0442	1.70
4	0.978	5.151	0.1545	−0.3043	0.1350	1.00
5	1.064	5.155	0.1663	−0.3079	0.2027	0.70
6	1.030	5.093	0.1762	−0.3155	0.2020	0.95
7	0.975	4.798	0.1091	−0.1197	−0.0750	2.09
8	1.080	4.943	0.0894	−0.0977	−0.1372	2.52
9	1.171	5.057	0.0875	−0.0870	−0.2132	1.54
10	0.500	4.956	0.0646	−0.1909	−0.0372	1.76
11	0.770	5.067	0.1217	0.0160	0.2113	1.45
12	0.591	4.923	0.1382	−0.1848	−0.2078	1.96
13	0.656	4.953	0.1118	−0.3041	0.2050	1.69
14	0.652	4.953	0.1093	−0.3043	0.2050	1.43
15	0.647	4.949	0.1132	−0.3066	0.2052	1.28
16	1.225	5.265	0.1669	−0.3087	0.2051	0.60
17	1.147	5.164	0.1697	−0.3075	0.2047	0.70
18	1.167	5.219	0.1639	−0.3064	0.2062	0.60
19	1.186	5.215	0.1657	−0.3084	0.2044	0.60
20	1.217	5.255	0.1621	−0.3040	0.2056	0.70
21	1.413	5.589	0.1846	−0.3069	0.2727	1.08

From Table 2, one can see the agreement between the observed and calculated values are satisfactory.

3.2. Partial least squares (PLS) analysis

The same five selected descriptors from previous work [16] were also used here as independent variables, and log EC_{50} as the dependent variable. Using one latent variable, the conventional R^2, SEP and Q^2 values were 0.83, 0.25 and 0.81, respectively. The regression equation was shown to be the following:

$$\log EC_{50} = -0.47\,\mathrm{LUMO} - 0.87\chi - 0.78Q_2 + 0.67Q_3 - 0.66Q_7 \quad (3)$$

$$R^2 = 0.83 \;\; Q^2 = 0.81 \;\; \mathrm{SEP} = 0.25 \;\; F = 12.17$$

The second model was built by using the 20 compounds and the same variables as used in Eq. (2). The PLS analysis resulted as a significant component model with the following statistics:

$$\log EC_{50} = -0.88\chi + 0.67Q_3 - 0.66Q_7 \quad (4)$$

$$R^2 = 0.85 \; Q^2 = 0.82 \; \mathrm{SEP} = 0.24 \; F = 24.86$$

The quality of both PLS models (Eqs. (3) and (4)) may be demonstrated by the comparison between the observed and calculated activities given in Table 3, but the statistical quality of Eq. (4) is slightly better than Eq. (3). Here it is interesting to remember that the PLS method, differently from the MLR method, works always with linearly independent variables [18,21].

In fact both PLS models, the first one with five descriptors (LUMO, χ, Q_2, Q_3 and Q_7) and the second one with three descriptors (χ, Q_3 and Q_7), can be used in future studies, but we believe the second one is more appropriate to work with due to the least number of descriptors and also because the best regression equation was obtained for both MLR and PLS methodologies when only the three descriptors χ, Q_3 and Q_7 were used.

It can be observed that for the set of compounds studied in this work, higher values for the variable χ combined with high positive charges on atom 7 (C7)

Table 2
The Anti-HIV activity calculated by using MLR

Compound	log EC_{50}				
	Observed value	Calculated value[a]	Residue	Calculated value[b]	Residue
1	1.00	1.0430	−0.0430	1.0621	−0.0621
2	2.12	1.6882	0.4318	1.7265	0.3935
3	1.70	1.1255	0.5745	1.1814	0.5186
4	1.00	0.9407	0.0593	0.9472	0.0528
5	0.70	0.8615	−0.1615	0.8733	−0.1733
6	0.95	1.0076	−0.0576	1.0422	−0.0922
7	2.09	2.5854	−0.4954	2.4781	−0.3881
8	2.52	1.9617	0.5583	1.9812	0.5388
9	1.54	2.1599	−0.6199	1.9488	−0.4088
10	1.76	1.9487	−0.1887	1.8200	−0.0600
11	1.45	1.3591	0.0909	1.5612	−0.1112
12	1.96	2.0511	−0.0911	2.1246	−0.1646
13	1.69	1.4225	0.2675	1.4160	0.2740
14	1.43	1.4899	−0.0599	1.4598	−0.0298
15	1.28	1.5272	−0.2472	1.4977	−0.2177
16	0.60	0.5531	0.0469	0.5363	0.0637
17	0.70	0.8396	−0.1396	0.8452	−0.1452
18	0.60	0.6963	−0.0963	0.6861	−0.0861
19	0.60	0.7072	−0.1072	0.6977	−0.0977
20	0.70	0.5837	0.1163	0.5627	0.1373
21	1.08	–	–	–	–

[a] From the regression equation (1).
[b] From the regression equation (2).

and high negative charges on atom 3 (C3) lead to an increasing of the anti-HIV activity. Here it is also important to mention that the descriptor χ is related to the strength of molecular association by charge transfer and the atomic charges (Q_3 and Q_7) to the electrostatic interaction between the drug and a center of opposite charge on the receptor.

A performance comparison between MLR and PLS models showed that the PLS model have substantially better predictive capability (higher Q^2 and smaller SEP) than the MLR model. However, as expressed by the correlation coefficients, the data-fitting ability for both MLR and PLS models seems to be similar.

4. Conclusions

Significant regression equations were obtained by multiple linear regression and partial least squares methods for 20 flavonoid compounds according to their anti-HIV activity. The best regression equation obtained was based on the following descriptors: electronegativity (χ) and atomic charges on atoms C3 and C7 (Q_3 and Q_7, respectively). The model obtained showed not only statistical significance but also predictive ability and revealed that higher values for χ combined with high positive charges on C7 and high negative charges on C3 lead to an increasing of the anti-HIV activity. These variables allowed a physical explanation of electronic molecular properties contributing to HIV inhibitory potency as the electronic character relates directly to the electron distribution of interacting molecules at the active site.

A comparison of the performance between the MLR and PLS models showed the PLS have substantially better predictive capability than the MLR model, even thought their correlation coefficients are comparable. This indicates clearly that the correlation coefficient by itself is not a good parameter for testing the model performance. Also, it has been shown that

Table 3
The Anti-HIV activity calculated by using PLS

Compound	log EC_{50}				
	Observed value	Calculated value[a]	Residue	Calculated value[b]	Residue
1	1.00	0.9193	0.0807	1.0037	−0.0037
2	2.12	1.7041	0.4159	1.7560	0.3640
3	1.70	1.2167	0.4833	1.2827	0.4173
4	1.00	0.9910	0.0090	0.9696	0.0304
5	0.70	0.8423	−0.1423	0.8751	−0.1751
6	0.95	0.8748	0.0752	0.9802	−0.0302
7	2.09	2.0745	0.0155	2.3510	−0.2610
8	2.52	1.9646	0.5554	2.0844	0.4356
9	1.54	1.9371	−0.3971	2.0718	−0.5318
10	1.76	2.1905	−0.4305	1.8037	−0.0437
11	1.45	1.6804	−0.2304	1.6665	−0.2165
12	1.96	1.9603	−0.0003	2.1146	−0.1546
13	1.69	1.5165	0.1735	1.2793	0.4107
14	1.43	1.5313	−0.1013	1.2790	0.1510
15	1.28	1.5185	−0.2385	1.2824	−0.0024
16	0.60	0.6231	−0.0231	0.6467	−0.0467
17	0.70	0.7788	−0.0788	0.8550	−0.1550
18	0.60	0.7290	−0.1290	0.7474	−0.1474
19	0.60	0.7169	−0.1169	0.7540	−0.1540
20	0.70	0.6620	0.0380	0.6692	0.0308
21	1.08	–	–	–	–

[a] From the regression equation (3).
[b] From the regression equation (4).

PLS is an excellent tool for those cases where the descriptors are by any means correlated.

Acknowledgements

The authors would like to thank CAPES and FINEP (Brazilian Agencies) for the financial support in this work.

References

[1] F. Barré-Sinoussi, J.C. Chermann, F. Rey, M.T. Nugeyre, S. Chamaret, J. Gruest, C. Dauguest, C. Axler-Blin, F. Vézinet-Brun, C. Rouzioux, W. Rozenbaum, L. Montagnier, Science 220 (1983) 868.
[2] R.C. Gallo, S.Z. Salahuddin, M. Popovic, G.M. Shearer, M. Kaplan, B.F. Haynes, T.J. Palker, R. Redfeild, J. Oleske, B. Safai, G. White, P. Foster, P.D. Markhan, Science 224 (1984) 500.
[3] C. Hu, K. Chen, Q. Shi, R.E. Kilkuskie, Y. Cheng, Kuo-Hsiung Lee, J. Nat. Prod. 57 (1995) 42.
[4] Hui-Kang Wang, Yi Xia, Zheng-Yu Yang, S.L.M. Natschke, Kuo-Hsiung Lee, Adv. Exp. Med. Biol. 439 (1998) 191.
[5] P.J. Houghton, Stud. Nat. Prod. Chem. 21 (2000) 123.
[6] H. Ishitsuka, C. Ohsawa, T. Ohiwa, I. Umeda, Y. Suhara, Agents Chemother. 22 (1982) 611.
[7] T.N. Kaul, E. Middletown Jr., P.L. Ogra, J. Med. Virol. 15 (1985) 71.
[8] R. Vrijsen, L. Everaert, L.M. Van Hoof, M.A.J. Vlietinck, D.A. Berghe, A. Boeye, Antiviral Res. 7 (1987) 35.
[9] T. Nagai, Y. Miyaichi, T. Tomimori, Y. Suzuki, H. Yamada, Antiviral Res. 19 (1992) 207.
[10] J.A. Beutler, J.H. Cardellina, G.M. Cragg, J. Nat. Prod. 55 (1992) 207.
[11] H. Tanaka, H. Takashima, M. Ubasawa, K. Sekiya, I. Nitta, M. Baba, S. Shigeta, R.T. Walker, E. Clercq, T. Miyasaka, J. Med. Chem. 35 (1992) 337.
[12] S. Hannongbua, L. Lawtrakul, C.A. Sotriffer, B.M. Rode, Quantum Struct. Act. Relat. 15 (1996) 389.
[13] H. Ishitsuka, C. Ohsawa, T. Ohiwa, I. Umeda, Y. Suhara, Agents Chemother. 22 (1982) 611.
[14] S. Hannongbua, L. Lawtrakul, J. Limtrakul, J. Comput.-Aided Mol. Design 10 (1996) 145.

[15] J.M. Luco, F.H. Ferretti, J. Chem. Inf. Comput. Sci. 37 (1997) 392.
[16] C.N. Alves, J.C. Pinheiro, A.J. Camargo, A.J. de Souza, R.B. Carvalho, A.B.F. da Silva, J. Mol. Struct. (Theochem) 491 (1999) 123.
[17] B.R. Kowalski, Chemometrics, Mathematics and Statistics in Chemistry, Reidel, Dordrecht, 1984.
[18] P. Geladi, B.R. Kowalski, Anal. Chim. Acta 185 (1986) 1.
[19] Pirouette Multivariate Data Analysis for IBM PC Systems, Version 2.0. Informetrix: Seattle, WA 1996.
[20] MATLAB, Mathworks Inc., Natick, MA, 1998.
[21] R.D. Cramer III, J.D. Bunce, E.D. Patterson, Quantum Struct.–Act. Relat. 7 (1988) 18.