

# Application of the Electrotopological State Index to QSAR Analysis of Flavone Derivatives as HIV-1 Integrase Inhibitors

John K. Buolamwini,<sup>1,3</sup> Krishnamchari Raghavan,<sup>2</sup> Mark R. Fesen,<sup>2</sup> Yves Pommier,<sup>2</sup> Kurt W. Kohn,<sup>2</sup> and John N. Weinstein<sup>2</sup>

Received August 21, 1996; accepted October 8, 1996

**Purpose.** A QSAR study based on electrotopological state (E-state) indices was conducted for a series of flavone HIV-1 integrase inhibitors to guide drug design.

**Methods.** E-state indices formulated to encode electronic and topological information for each skeletal atom in a molecule (Kier and Hall *Pharm. Res.* 7:801–807 (1990)) were calculated using the Molconn-X program, and partial least squares (PLS) multivariate regression was used to derive QSAR models.

**Results.** Predictive models with correlation coefficients ( $r^2$ ) of 0.98 (3 PLS components) and 0.99 (5 PLS components) and corresponding cross-validated correlation coefficients (c.v.  $r^2$ ) of 0.51 and 0.73, were obtained for inhibition of cleavage and integration, respectively, with one molecule omitted from the analysis.

**Conclusions.** E-state indices at C6, C3', C5', C5, and O4 were found to be more important for prediction of activity than those for any of the other 12 flavone skeletal atoms that are common to the molecules in the data set.

**KEY WORDS:** Electrotopological state index; quantitative structure-activity relationships; HIV-1 integrase inhibitors; flavones; partial least squares.

## INTRODUCTION

Current therapy for AIDS (acquired immunodeficiency syndrome) is inadequate, hence the continuing efforts to find new treatment strategies (1,2). Human immunodeficiency virus (HIV) integrase is one of the most recent additions to the list of potential molecular targets for anti-HIV therapy (1). The integrase is essential to the life cycle of HIV, being responsible for incorporation of reverse-transcribed proviral DNA into the host genome (3–5). The incorporation occurs by two tandem steps: i) 3' processing (cleavage) that removes two nucleotides from the 3' end of the proviral DNA; ii) integration of the processed DNA into the host genome.

*In vitro* assays for separate measurement of cleavage and integration by HIV-1 integrase have been developed (6) and used to identify several families of compounds with potent HIV-1 integrase inhibitory activity (7). Quercetin, representing the flavone family, was one of these compounds. Subsequently,

a series of other flavones, mainly polyhydroxylated and glycosylated derivatives, were tested (8). Here, we use electrotopological state (E-state) chemical descriptors (9) in quantitative structure-activity relationship (QSAR) studies of these flavones as inhibitors of HIV-1 integrase *in vitro*.

E-state indices were recently introduced by Kier, Hall and their associates as nonempirical atomic level structural descriptors suitable for QSAR studies (9,10). E-state indices encode electronic and topological information in a single number for each skeletal atom in the hydrogen-suppressed graph of a molecule. The E-state index value ( $S$ ) of an atom  $i$  is given by the sum of the intrinsic state value ( $I$ ) and perturbations ( $\Delta I$ ) by the fields of all other atoms in the chemical graph, including substituent atoms. Thus,  $S$  is defined by the following set of equations:

$$S = I + \Delta I, \quad (1)$$

$$I = ((2/N)^2 \delta^v + 1) / \delta, \quad (2)$$

$$\delta^v = \sigma + \pi + n - h, \quad (3)$$

$$\delta = \sigma - h, \quad (4)$$

$$\Delta I_i = \sum (I_j - I_i) / r_{ij}^2, \quad (5)$$

where  $N$  is the principal quantum number (i.e., row in which the atom occurs in the periodic table),  $\sigma$  the number of sigma electrons contributed by the atom,  $\pi$  the number of pi electrons contributed by the atom,  $h$  the number of hydrogen atoms attached to the atom,  $n$  the number of lone-pair electrons on the atom, and  $r$  the number of atoms in the shortest graph path connecting atom  $i$  and any other atom  $j$  in the chemical graph of the whole molecule.

Our application of E-state QSAR to flavone HIV-1 integrase inhibitors will aid in the identification of submolecular loci important for biological activity. It may also provide further insights for lead optimization.

## METHODS

Two dimensional structures of the molecules under study were constructed using the Chem-X molecular modeling program (Chemical Design Ltd., Oxon., England) running on an Indigo Elan workstation (Silicon Graphics Inc., Mountain View, CA). Their connection tables were written in CSSR format and imported into the Molconn-X program (version 2.0, 1993, Hall Associates Consulting, Quincy, MA) for calculation of E-state indices.

QSAR models were derived by the partial least squares (PLS) statistical technique, as implemented in the SYBYL/QSAR program package (Tripos Associates Inc., St. Louis, MO), on an Indigo 2 workstation (Silicon Graphics Inc., Mountain View, CA). PLS was introduced in the early 1980's for multivariate regression (11). It is a variant of principal component regression in that the original variables are replaced by a smaller set of linear combinations. Unlike the case in principal component analysis, however, the dimensionally-reduced orthogonal set of variables is constrained to maximize the communality of predictor and response variable blocks.

<sup>1</sup> Department of Medicinal Chemistry and National Center for the Development of Natural Products, School of Pharmacy, University of Mississippi, University, Mississippi 38677.

<sup>2</sup> Laboratory of Molecular Pharmacology, Division of Basic Sciences, National Cancer Institute, Bldg. 37, Room 5C25, National Institutes of Health, Bethesda, Maryland 20892.

<sup>3</sup> To whom correspondence should be addressed.

Table I (modified from Fesen *et al.* (8)) lists a set of flavones active against HIV-1 integrase. Shown are the chemical structures, together with negative  $\log(\text{IC}_{50})$  values for the cleavage and integration steps. These two quantities were used as the dependent ( $y$ ) variables in our QSAR study. Predictive models were derived using E-state indices for the 17 skeletal atoms common to all molecules in the set as multivariate descriptors. The leave-one-out crossvalidation method was used to test the predictive abilities of the models. The crossvalidation coefficient c.v.  $r^2$  is given by Eq. 6:

$$\text{c.v. } r^2 = 1.0 - \frac{\sum (y_{\text{pred}} - y_{\text{actual}})^2}{\sum (y_{\text{actual}} - y_{\text{mean}})^2} \quad (6)$$

The number of PLS components maintained in the model was that which minimized the cross-validated standard error of the estimate ( $s$ ), as defined in Eq. 7.

$$s = \sqrt{\frac{\text{PRESS}}{n - c - 1}} \quad (7)$$

where PRESS is the predictive error sum of squares (i.e., the

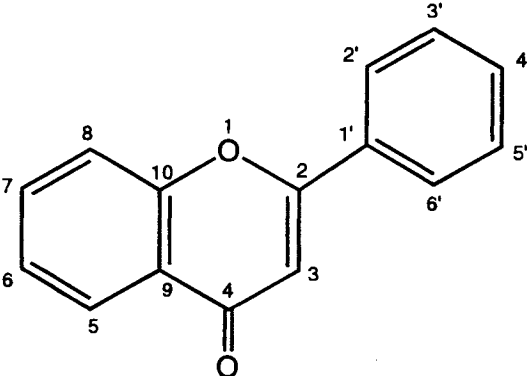
numerator  $\sum (y_{\text{pred}} - y_{\text{actual}})^2$  in Eq. 6),  $n$  the number of compounds, and  $c$  the number of components.

## RESULTS

Calculations based on all 15 compounds yielded low c.v.  $r^2$  values for both cleavage and integration, and examination of the PLS results showed that 6-methoxyluteolin (Compound 14) was an "outlier" in both the cleavage and integration models based on the residuals. Consequently, Compound 14 was eliminated, and PLS models with good predictive abilities were obtained for the rest of the compounds in the data set.

Randomization control studies involving the permuted activity data (not shown) supported the outlier status of Compound 14. The statistical effects of omitting one compound were also examined by calculating Bonferroni corrections (12), as described in the subsection on PLS equations. In this set of molecules, besides Compound 14, there are only two other 6-position-substituted compounds (baicalein, 1 and quercetagenin, 2), the substituent being a hydroxyl group in both cases. These compounds (1 and 2) were also the two most potent in the

Table I Structures and HIV Integrase Inhibitory Activity of Flavones<sup>a</sup>



Flavones	-Log(IC <sub>50</sub> ) <sup>b</sup> values		Ring Substituents								
	Cleavage	Integration	3	5	6	7	8	2'	3'	4'	5'
1 quercetagenin	6.10	7.00		OH	OH	OH	OH			OH	OH
2 baicalein	5.92	5.37		OH	OH	OH	OH				
3 robinetin	5.23	5.80	OH			OH			OH	OH	OH
4 myricetin	5.12	5.60	OH	OH		OH			OH	OH	OH
5 quercetin	4.63	4.87	OH	OH		OH			OH	OH	
6 fisetin	4.55	5.07	OH			OH			OH	OH	
7 luteolin	4.48	4.60		OH		OH			OH	OH	
8 myricetrin	4.40	4.99	RH <sup>c</sup>	OH		OH			OH	OH	OH
9 quercetrin	4.22	4.41	RH	OH		OH			OH	OH	
10 rhamnetin	4.21	4.54	OH	OH		OMe			OH	OH	
11 avicularin	4.18	4.60	AR <sup>d</sup>	OH		OH			OH	OH	
12 gossypin	4.16	4.64	OH	OH		OH	GL <sup>e</sup>		OH	OH	
13 morin	4.12	4.50	OH	OH		OH		OH		OH	
14 6-methoxyluteolin	4.03	4.41		OH	OMe	OH			OH	OH	
15 kaempferol	4.01	4.19	OH	OH		OH				OH	

<sup>a</sup> Modified from Fesen *et al.* (8).

<sup>b</sup> -Log(IC<sub>50</sub>) is negative log<sub>10</sub> of the molar concentration of compound that caused 50% inhibition of HIV-1 integrase activity.

<sup>c</sup> RH = rhamnose.

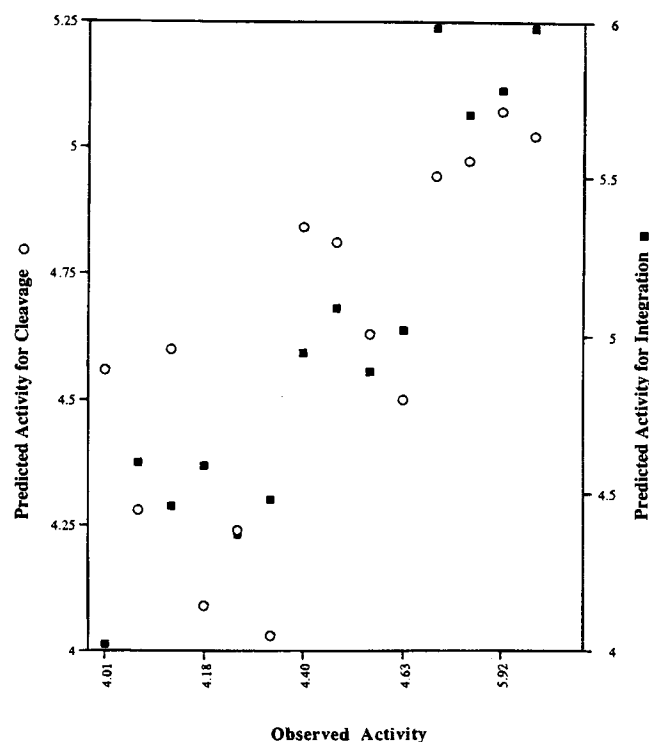
<sup>d</sup> AR = arabinose.

<sup>e</sup> GL = glucose.

assays (Table I). The algorithm appeared not to distinguish effectively between the 6-methoxy substituent in Compound **14** and the activity-enhancing 6-hydroxy substituent in Compounds **1** and **2**. Interestingly, the same molecule was also an outlier in parallel studies by Raghavan *et al.* (13) using a very different QSAR method, comparative molecular field analysis (CoMFA).

PLS was combined with leave-one-out crossvalidation to determine the optimum number of components to use for deriving final QSAR models with 14 compounds. The number of components corresponding to the lowest cross-validated standard error of estimate (*s*) was used to derive the final non-crossvalidated PLS model, the predictive ability of which is the c.v.  $r^2$  corresponding to the lowest *s*. The *s* stopping criterion is more conservative than the SYBYL/QSAR default stopping criterion (14), imposing a penalty for addition of components (and consequent reduction in degrees of freedom).

The final PLS models included 3 and 5 components for the cleavage and integration data, respectively. In Figure 1, the leave-one-out crossvalidated predictions obtained are plotted against experimental values. The c.v.  $r^2$  values indicate that the models are reasonably predictive. Randomization controls were performed to show the robustness of the predictive models obtained with only 14 compounds (data not shown). The PLS regression equations obtained are given as Eqs. 8 and 9 below. The subscript refers to the flavone core position. Thus  $S_{O1}$  is the E-state value of the oxygen atom in position 1. Regression equations with high correlation coefficients ( $r^2$ ) and small standard errors of estimate (sd) were obtained.



**Fig. 1.** Cross-validated Predictions of Cleavage (c.v.  $r^2 = 0.513$ ) and Integration (c.v.  $r^2 = 0.734$ ) PLS models.  $\circ$ , predicted activity for cleavage;  $\blacksquare$ , predicted activity for integration.

**Table II.** Effect of Leaving out Each Atom E-state (S) Index on Predictive Ability (c. v.  $r^2$ ) of QSAR Models

E-state index left out	Leave-one-out c. v. $r^2$	
	Cleavage	Integration
None	0.513	0.734
$S_{O1}$	0.521	0.729
$S_{C2}$	0.514	0.737
$S_{C3}$	0.632	0.676
$S_{C4}$	0.517	0.736
$S_{C5}$	0.506	0.607
$S_{C6}$	-1.076	-0.569
$S_{C7}$	0.511	0.730
$S_{C8}$	0.569	0.735
$S_{C9}$	0.521	0.736
$S_{C10}$	0.519	0.735
$S_{C1'}$	0.515	0.735
$S_{C2'}$	0.540	0.738
$S_{C3'}$	0.550	0.489
$S_{C4'}$	0.580	0.889
$S_{C5'}$	0.353	0.693
$S_{C6'}$	0.513	0.732
$S_{=O}$	0.471	0.510

### PLS Regression Equations

The following equations were obtained from the standard output of the SYBYL/QSAR package:

#### Inhibition of Cleavage

$$\begin{aligned}
 -\text{Log}(\text{IC}_{50}) = & 11.475 - 0.149(S_{O1}) + 0.017(S_{C2} + 0.005(S_{C3}) \\
 & + 0.015(S_{C4}) + 0.250(S_{C5}) - 0.888(S_{C6}) \\
 & - 0.123(S_{C7}) + 0.138(S_{C8}) + 0.027(S_{C9}) \\
 & + 0.043(S_{C10}) + 0.014(S_{C1'}) + 0.107(S_{C2'}) \\
 & - 0.209(S_{C3'}) + 0.184(S_{C4'}) - 0.290(S_{C5'}) \\
 & - 0.024(S_{C6'}) - 0.421(S_{O4}) \quad (8)
 \end{aligned}$$

$r^2 = 0.975$ ,  $sd = 0.121$ ,  $F = 131$ ,  $n = 14$ , c.v.  $r^2 = 0.513$ , (Probability that c.v.  $r^2 = 0$  is  $< 0.0001$ ), number of components = 3, with one compound (**14**) omitted.

#### Inhibition of Integration

$$\begin{aligned}
 -\text{Log}(\text{IC}_{50}) = & 12.700 - 0.189(S_{O1}) - 0.023(S_{C2}) \\
 & - 0.140(S_{C3}) - 0.015(S_{C4}) + 0.309(S_{C5}) \\
 & - 1.110(S_{C6}) - 0.155(S_{C7}) + 0.215(S_{C8}) \\
 & + 0.016(S_{C9}) + 0.043(S_{C10}) - 0.031(S_{C1'}) \\
 & - 0.035(S_{C2'}) - 0.380(S_{C3'}) - 0.048(S_{C4'}) \\
 & - 0.236(S_{C5'}) - 0.052(S_{C6'}) - 0.470(S_{O4}) \quad (9)
 \end{aligned}$$

$r^2 = 0.994$ ,  $sd = 0.093$ ,  $F = 255$ ,  $n = 14$ , c.v.  $r^2 = 0.734$ , (Probability that c.v.  $r^2 = 0$  is  $< 0.0001$ ), number of components = 5, with one compound (**14**) omitted.

If a Bonferroni correction (12) is applied for omission of Compound **14** as an outlier, the critical P-value is reduced by

a factor of  $15!(14!1!) = 15$ . Therefore a two-sided 5% critical p-value would be  $0.025/15 = 0.00167$ . The p-values obtained from Eqs. 8 and 9 still show statistical significance even after this highly conservative correction.

The relative importance of the different atom E-state indices to the predictive models, was determined by eliminating each E-state variable in turn and performing cross-validated PLS runs using the rest of the variables (Table II). Comparatively, the E-state value at C6 was the single most important variable contributing to changes in activity for both the cleavage and integration models. Its elimination resulted in very poor c.v.  $r^2$  values,  $-1.076$ , and  $-0.569$  for cleavage and integration, respectively (Table II). The other E-state values of importance (albeit much less) for prediction of both cleavage and integration inhibitory activity, were  $S_{C3'}$ ,  $S_{C5}$ ,  $S_{C5'}$ , and  $S_{O4}$ . The order of importance of E-state values with respect to predictive ability was as follows:

Cleavage:  $S_{C6} \gg S_{C5'} > S_{O4} > S_{C5}$

Integration:  $S_{C6} \gg S_{C3'} > S_{O4} = S_{C5} > S_{C3}$

## CONCLUSIONS

The catalytic core of HIV-1 integrase has been crystallized and analyzed by X-ray diffraction (15), but no such data are yet available for a complex of HIV integrase with any of its substrates or inhibitors. Therefore, structure-based design of integrase inhibitors must rely heavily, for the time being on QSAR, and on database searching methods. Atomic level chemical descriptors such as the E-state indices used in this QSAR study have the advantage over global (whole-molecule) descriptors in that they can potentially give more insight into submolecular regions associated with or influencing interaction between ligand and receptor.

Changes in the vicinities of flavone skeletal atoms C6, C5, C3', C5', and O4 were shown to be important to the prediction of HIV integrase inhibitory activity. Changes that result in a decrement in the E-state values at C6, C3', C5' and/or O4 are predicted to enhance HIV-1 integrase inhibitory activity. Polyhydroxylation appears to be required for potency (7, 16); our study suggests that substitution of an electronegative group, such as a hydroxyl at C6 (causing a decrease in the E-state value) will enhance activity. A parallel 3D QSAR study (13) by CoMFA on the same data set indicated similar regions of the molecules to be important for HIV-1 integrase inhibitory activity. This general concordance of very different QSAR approaches could be coincidental, but it probably indicates that the results obtained are relatively robust. CoMFA is based on 3D

electrostatic and steric interaction energies generated between aligned molecules and a probe atom (17), whereas the E-state approach uses 2D atom-level topological connectivity information, as well as atom type and electronegativity, to encode invariant molecular structure descriptors. The observation that the E-state approach performs comparably to CoMFA for this data set is interesting in that the latter is more computationally intensive than the former. The E-state method may prove especially useful for identification of pharmacophores that can be used to search large 2D molecular databases.

## ACKNOWLEDGMENTS

This work was supported in part by start-up funds from the University of Mississippi School of Pharmacy and the NIH Intramural AIDS Targeted Antiviral Program. We also thank Dr. Lawrence V. Rubinstein and Dr. Barry Bunow for reviewing the statistics.

## REFERENCES

1. M. I. Johnston, and D. F. Hoth. *Science* **267**:1286-1293 (1993).
2. P. S. Anderson, G. L. Kenyon and G. R. Marshall (eds.), *Perspectives in Drug Discovery and Design 1: Supplement to Computer-Aided Molecular Design*, ESCOM Science Publishers, 1993.
3. H. E. Varmus and P. O. Brown. In D. E. Berg and M. M. Howe (eds.), *Mobile DNA*, American Society for Microbiology, Washington DC, 1989, pp. 53-108.
4. R. L. LaFemina, C. L. Schneider, H. L. Robbins, P. L. Callahan, K. Legrow, E. Roth, W. A. Schleif, and E. A. Emini. *J. Virol.* **66**:7414-7419 (1992).
5. F. D. Bushman, and R. Craigie. *Proc. Natl. Acad. Sci. USA* **88**:1339-1343 (1991).
6. A. Engelman, K. Mizuuchi, and R. Craigie. *Cell* **67**:1211-122 (1991).
7. M. R. Fesen, K. W. Kohn, F. Leuteurtre, and Y. Pommier. *Proc. Natl. Acad. Sci. U.S.A.* **90**:2399-2403 (1993).
8. M. R. Fesen, Y. Pommier, F. Leuteurtre, S. Hiroguchi, J. Yung, and K. W. Kohn. *Biochem. Pharmacol.* **48**:595-608 (1994).
9. L. B. Kier, and L. H. Hall. *Pharm. Res.* **7**:801-807 (1990).
10. L. B. Kier, and L. H. Hall. *Adv. Drug Res.* **22**: 1-138 (1992).
11. H. Wold. In *Systems under Indirect Observation*, Part II, Joreskog, K. G. and Wold, H., Eds., North Holland, Amsterdam, 1982, pp. 1-54.
12. Kleinbaum, D. G.; Kupper, L. L. and Muller, K. E. *Applied Regression Analysis and Other Multivariate Methods*, PWS-KENT Publishing Company, Boston, 1988.
13. K. Raghavan, J. K. Buolamwini, M. R. Fesen, Y. Pommier, K. W. Kohn, and J. N. Weinstein. *J. Med. Chem.* **38**:890-897 (1995).
14. *Theory Manual of SYBYL Molecular Modeling Software, Version 6.2*, Tripos Associates, Inc., St. Louis, MO, 1995.
15. F. Dyda, A. B. Hickman, T. M. Jenkins, A. Engelman, R. Craigie, and D. R. Davies. *Science* **266**:1981-1986 (1994).
16. T. R. Burke, Jr., M. R. Fesen, A. Mazumder, J. Wang, A. M. Carothers, D. Grumberger, J. Driscoll, K. Kohn, and Y. Pommier. *J. Med. Chem.* **38**:4174-4178 (1995).
17. R. D. Cramer III, D. E. Paterson, and J. D. Bunce. *J. Am. Chem. Soc.* **110**:5959-5967 (1988).