ELSEVIER

Journal of Molecular Graphics and Modelling 21 (2003) 435–448

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

A priori molecular descriptors in QSAR: a case of HIV-1 protease inhibitors I. The chemometric approach[☆]

Rudolf Kiralj, Márcia M.C. Ferreira[*]

Instituto de Química, Universidade Estadual de Campinas, Campinas, SP 13083-970, Brazil

Received 25 May 2001; accepted 14 November 2002

Abstract

A quantitative structure–activity relationship (QSAR) study on 48 peptidic HIV-1 protease inhibitors was performed. Fourteen a priori molecular descriptors were used to build QSAR models. Hierarchical cluster analysis (HCA), principal component analysis (PCA) and partial least squares (PLS) regression were employed. PLS models with 32/16 (model I) and 48/0 (model II) molecules in the training/external validation set were constructed. The a priori molecular descriptors were related to two energetic variables using PLS. HCA and PCA on data from model II classified the inhibitors as slightly, moderately and highly active; three principal components, the chemical nature of which has been highlighted, are enough to describe the enzyme–inhibitor binding. Model I ($r^2 = 0.91$, $q^2 = 0.84$) is comparable to literature models obtained by various QSAR softwares, which justified the use of a priori descriptors.

Keywords: A priori molecular descriptors; QSAR; HIV-1 protease inhibitors; Chemometrics

1. Introduction

Understanding quantitative structure–activity relationships (QSAR) more profoundly includes understanding the difference in dimensionality of molecular representations and of descriptors (i.e. if they are 1D–3D), as is discussed by Van de Waterbeemd and Testa [1]. The benzene molecule (Fig. 1) can be represented with formula (molecular representation) C_6H_6 (1D object), chemical diagram with a regular hexagon (2D object), or structural formula showing molecular planarity (3D object). A 1D formula can give only 1D data (molecular mass, numbers of atoms as in Fig. 1 left, other scalars). A 2D formula (Fig. 1 middle) produces 2D data (2D matrices of topological descriptors, etc.), and 3D representation enables the extraction of 3D data (volume [2] or spatial distribution of some property in the form of 3D matrices, Fig. 1 right). Macroscopic properties are in most cases 1D in form (although describing 3D events), so 2D and 3D data are usually reduced to their 1D forms (Fig. 1 bottom) retaining their 2D and 3D meaning.

QSAR procedures can generate hundreds of 1D–3D, etc. descriptors usually transformed into 1D forms (“classical” QSAR). All these procedures, from the simplest to the most sophisticated, are useful. However, some disadvantages in applying sophisticated QSAR software might be pointed out: (1) treatment of the program as a black box; (2) availability and price; and (3) incompletely interpreted results in publications. For instance, are the descriptors only mathematical concepts or physical properties too complicated to be understood in terms of chemical effects? Ideally, the descriptors should be chosen on the basis of mechanistic considerations or they should be amenable to mechanistic interpretation [3].

Some new trends in QSAR attempt to overcome these difficulties. First, use of extensive or exclusive calculation of descriptors derived only from chemical structures has become standard. Second, “non-empirical structural variables” [4], various 1D–3D descriptors like topological indices [5,6], geometrical or shape descriptors, quantum chemical and others are also now used extensively. Third, exclusion of 3D structural descriptors and use of topological indices (mainly 2D variables) only. These are fast and easy to calculate, encode useful information about various aspects of molecular architecture (size, shape, branching and cyclicity [6]), can be interpreted in terms of quantum mechanics [7].

[☆] The paper was presented in the 13th European Symposium on Quantitative Structure–Activity Relationships: Rational Approaches to Drug Design, Düsseldorf, Germany, 27 August to 1 September 2000. The companion paper, Part II, which interprets a priori molecular descriptors in terms of molecular graphics and modeling, is in this issue of the journal.

[*] Corresponding author. Tel.: +55-19-3788-3102; fax: +55-19-3788-3023.
E-mail address: marcia@iqm.unicamp.br (M.M.C. Ferreira).

1093-3263/02/$ – see front matter © 2002 Elsevier Science Inc. All rights reserved.
PII: S1093-3263(02)00201-2

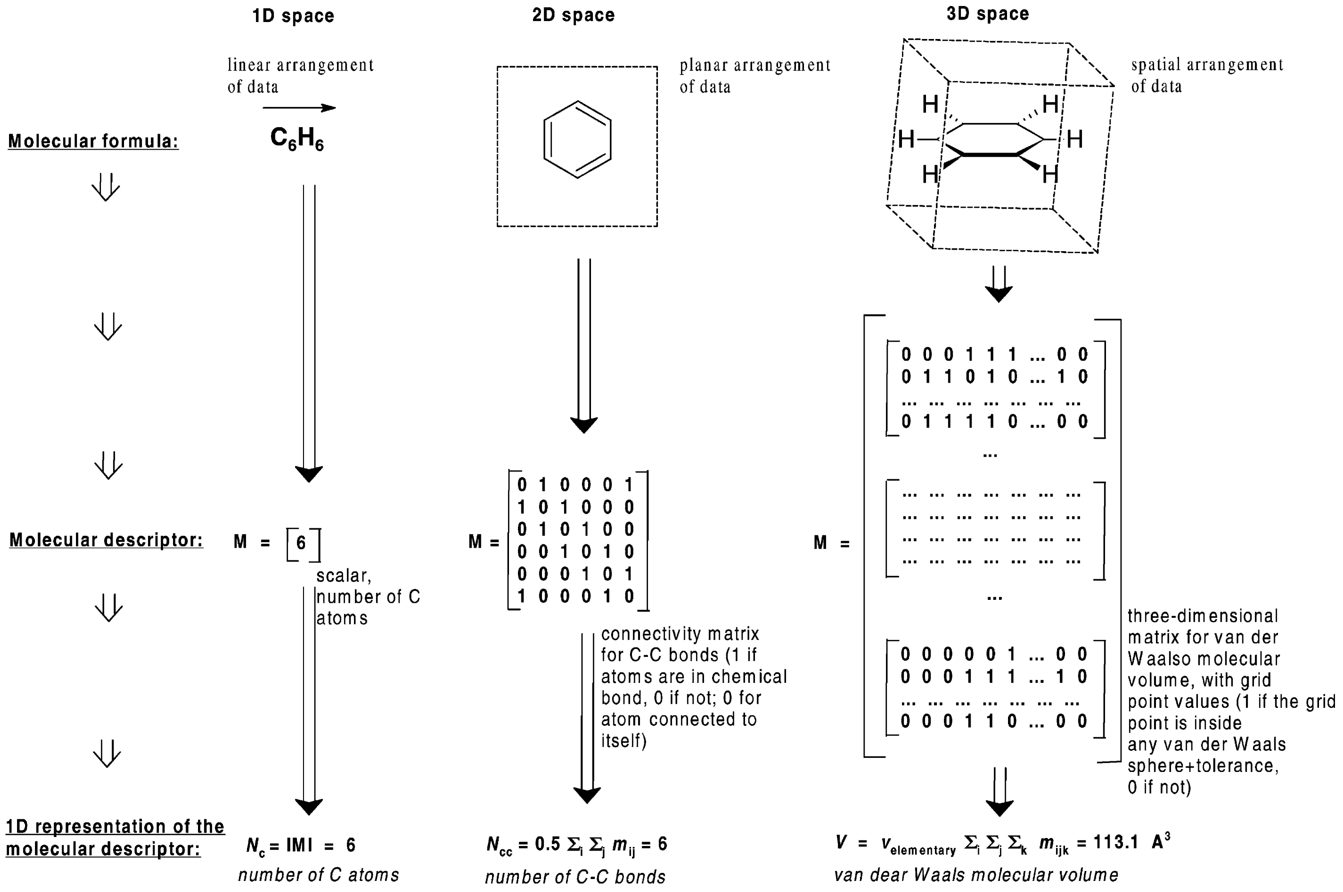
1D space
2D space
3D space
linear arrangement of data
planar arrangement of data
spatial arrangement of data
Molecular formula:
C6H6
H
Molecular descriptor:
M = [6]
scalar, number of C atoms
M =
0 1 0 0 0 1
1 0 1 0 0 0
0 1 0 1 0 0
0 0 1 0 1 0
0 0 0 1 0 1
1 0 0 0 1 0
connectivity matrix for C-C bonds (1 if atoms are in chemical bond, 0 if not; 0 for atom connected to itself)
M =
0 0 0 1 1 1 ... 0 0
0 1 1 0 1 0 ... 1 0
0 1 1 1 1 0 ... 0 0
0 0 0 0 0 1 ... 0 0
0 0 0 1 1 1 ... 1 0
0 0 0 1 1 0 ... 0 0
three-dimensional matrix for van der Waalso molecular volume, with grid point values (1 if the grid point is inside any van der Waals sphere+tolerance, 0 if not)
1D representation of the molecular descriptor:
Nc = IMI = 6
number of C atoms
Ncc = 0.5 Σi Σj mij = 6
number of C-C bonds
V = velementary Σi Σj Σk mijk = 113.1 A3
van dear Waals molecular volume

Fig. 1. Benzene molecule. An illustrative example on dimensions of molecular formulas, derived molecular descriptors, and reduced representation of the descriptors. From the linear, planar and spatial "formula" for the benzene molecule can be derived molecular descriptors with maximum dimension 1D (example: number of carbon atoms), 2D (example: number of carbon–carbon bonds by counting all bounded atomic connections) and 3D (example: molecular volume calculated employing a 3D Riemman summation [2] defined by some set of non-bonding atomic radii), respectively.

Fourth, use of "simple" molecular descriptors [8,9] (1D and 2D), such as number and weight fraction of atomic types, chemical bonds, rings, functional groups, and other indicator variables is a way to simplify and return chemical sense. Topological indices and other 2D descriptors show to be at least as efficient as 3D descriptors in QSAR [4,10–12].

In this work, an a priori approach is introduced, a QSAR methodology where only simple, a priori variables ("known before" any sophisticated, computer-assisted calculation) are employed. A priori variables are generated by hand-count or pocket-calculator using 1D and 2D chemical formulas. The 3D atomic coordinates, structural and extensive databases are not used. The procedure of generation of some a priori variables might appear similar to Hansch and Free–Wilson analyses [13], but there are conceptual differences: (1) the intuitive way of defining descriptors; (2) minimal use of literature data for additive properties; (3) only a few indicator variables used; (4) no exhaustive variable selection required; (5) use of other models besides multiple linear regression (MLR). The results from a sophisticated QSAR methodology, comparative binding energy (COMBINE)-QSAR study on HIV-1 protease inhibitors [14–16] were compared with this a priori study's results.[1] The 48 peptidic inhibitors under study (Figs. 2–4) have four (P_1, P'_1, P_2, P'_2) substituents [17] (scheme in Fig. 4). PCA, hierarchical cluster analysis (HCA) and partial least squares (PLS) results in this work are discussed in terms of the a priori approach and of HIV-1 protease–inhibitor binding. The a priori approach is a helpful tool for QSAR interpretation in terms of basic chemical concepts and can comprise an initial QSAR to be followed by more sophisticated investigation.

2. Methodology

2.1. Calculation of molecular descriptors

QSAR data are in Tables 1–3. Molecular descriptors for 49 HIV-1 protease inhibitors were generated on the basis of 1D and 2D formulas (Figs. 2–4). Only X_6 and X_{12}, were

[1] A new HIV-1 protease inhibitor lopinavir, approved by US FDA, appeared as a pure drug and in combination with ritonavir (a mixture called kaletra) after the submission of this work.

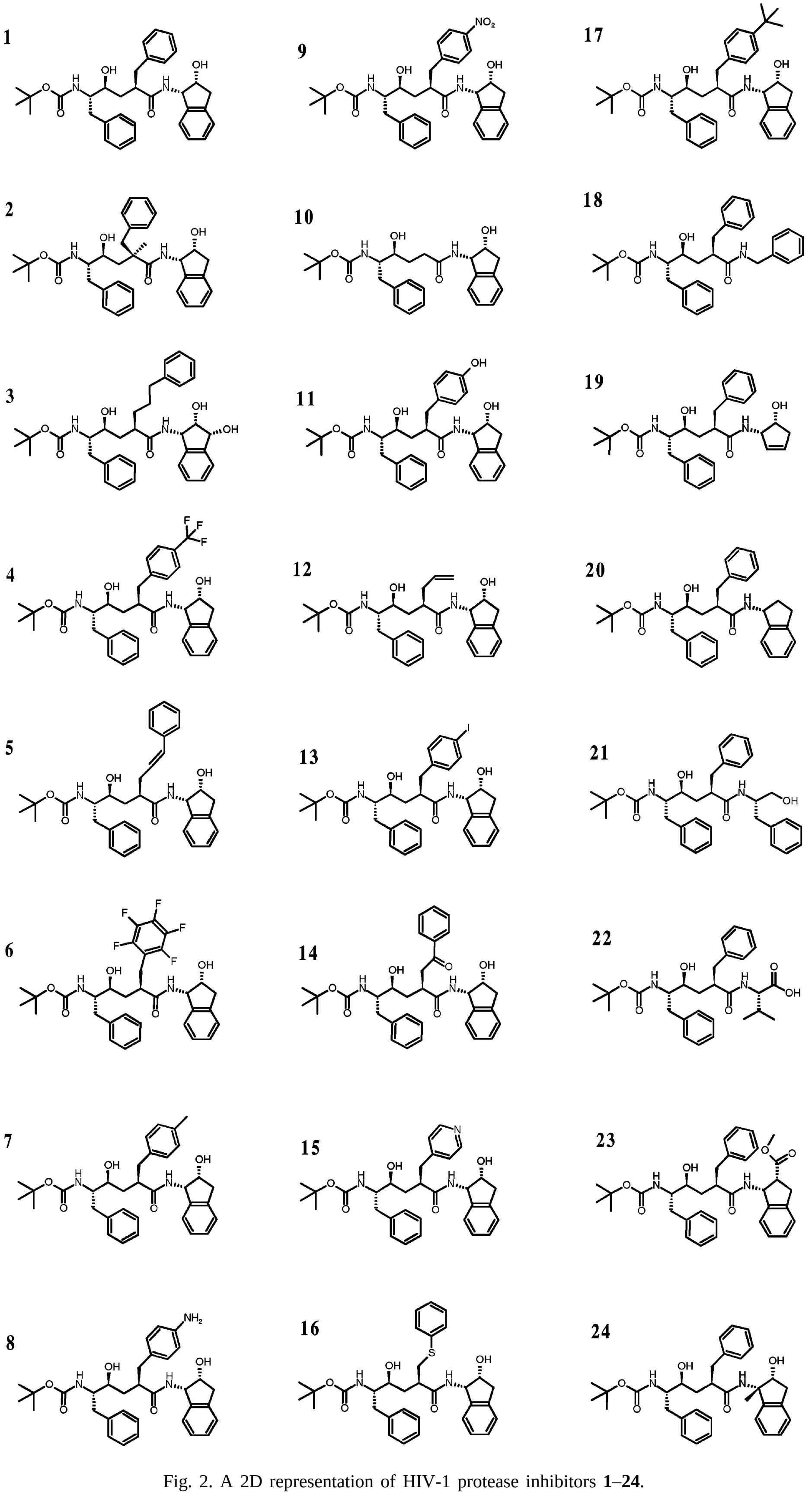

Fig. 2. A 2D representation of HIV-1 protease inhibitors **1**–**24**.

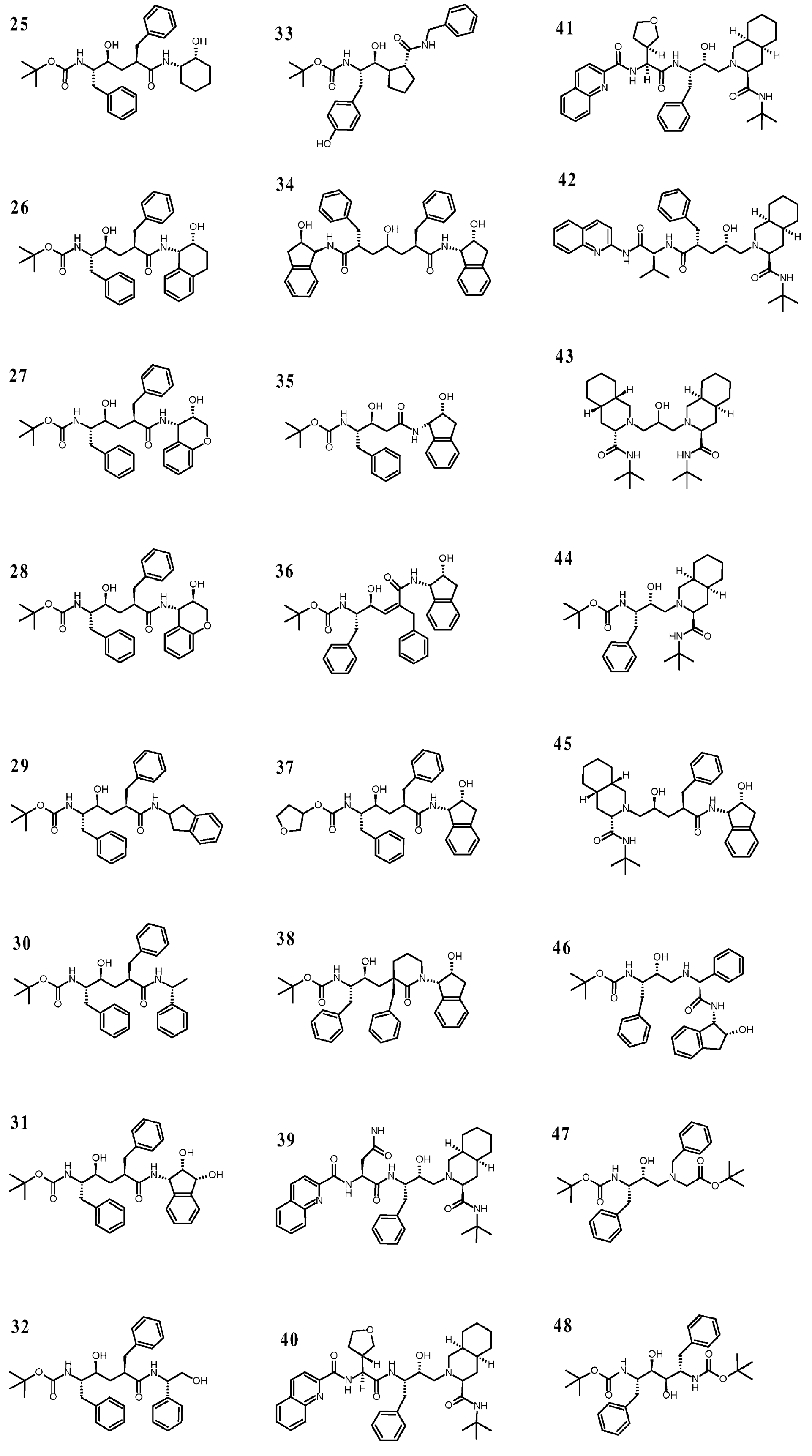

Fig. 3. A 2D representation of HIV-1 protease inhibitors **25**–**48**.

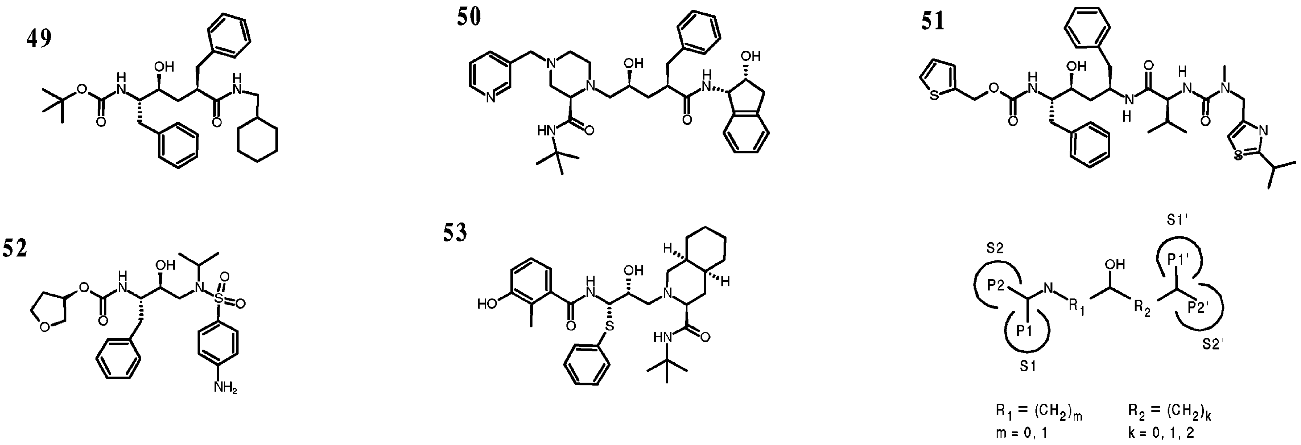

Fig. 4. A 2D representation of HIV-1 protease inhibitors **49**–**60** and schematic representation of inhibitor side chains (substituents) P_1, P'_1, P_2, P'_2 and R_1, R_2 separators in the inhibitors **1**–**60**. Besides saquinavir, **39** four more inhibitors indinavir **50**, ritonavir **51**, amprenavir **52** and nelfinavir **53** are clinically approved in combination with HIV-1 reverse transcriptase inhibitors. The inhibitors **54**–**60** were modeled by modifying the structure of **34**.

Table 1
Definition and description of the variables used in regression models

Symbol	Definition and description
Y	In vitro inhibition activity [14], $\mathrm{pIC}_{50} = -\log \mathrm{IC}_{50}$
X_1 or M_r	Relative molecular mass
X_2	A number of non-σ valence electrons: the number is equal to the count of π-bonds together with the free electron of heteroatoms
X_3	Number of non-hydrogen atoms in planar fragments: this includes aromatic rings and fragments with double bonds
X_4	Number of chemical bonds excluding the bonds with hydrogen
X_5	Number of valence electrons per atom
X_6	Non-σ valence electron surface density X_2/S, where S is van der Waals molecular surface area as a sum of literature surface area increments for atoms and groups [16]
X_7	Number of non-hydrogen atoms in ring systems: this includes both aromatic and aliphatic rings
X_8	Number of groups CX_n, $n = 0$, 1, 2, and 3, where X=H or halogen; C from C=O groups is excluded
X_9	Effective number of substituents based on the following rules: (a) number is 4 for molecule where the substituents are in position with respect to the central chain line as in **1** (standard molecule); (b) if one or two substituents are missing, the number is 3 (**33**, **35**, **44**–**48**) or 2 (**43**), respectively; (c) the number is 3.5 if one of the substituents is smaller (**12**, **18**, **19**, **22**, **25**, **30**, **32**) or in opposite orientation (**28**, **29**, **36**) than in the standard; if the substituent is even smaller, the number is 3.25 (**21** and **42**); (d) the number is 3.5, if one of the substituents is sterically hindered by some little group or atom (by CH_3 in **2**, **23**, **24**; by H in **40**), or via bigger group linked to the main chain (with C=O in **14**; with aliphatic ring in **38**)
X_{10}	Number of potential hydrogen bonds: number of donors (OH, NH, NH_2) + number of acceptors (OH, C=O, –O–)
X_{11}	Effective number of ring substituents (both aromatic and aliphatic) based on the same rules as for X_9: (a) number for molecule **1**, the standard, is 3; (b) number is X_{14}-1 for most of the molecules (**1**–**11**, **13**–**20**, **23**–**33**, **35**, **36**, **38**, **39**, **44**, **46**–**48**) because one substituent is a non-ring system, while for others are special rules as follows; (c) number is 4 when all the substituents are rings (**34**, **41**); (c) number is 3.5 also for some molecules (**37**—a small ring substituent, **40**—sterically hindered ring); (d) number is 3 also for some molecules (**42**—a small non-ring substituent, **45**—one substituent missing); (e) number is 2.5 also for one molecule (**21**—a non-ring and a small ring substituent present in the structure); (f) number is 2 also for some molecules (**12** and **22**—two non-ring substituents present in the structure, **43**—only two substituents present and they are rings)
X_{12} or V_{pol}	The van der Waals volume of polar groups (C=O, $-NH_2$, –NH, –N–, $-CF_3$, –S–, –OH, –O–, $-NO_2$, –I) estimated as van der Waals molecular volume as sum of literature volume increments for atoms and groups [16]
X_{13}	The length of the total “aromatic vector”: number of atoms in localized, delocalized and aromatic π-systems, and the number of atoms with free-electron pairs (N, O, S), and number of C atoms in CH_m groups ($m = 1$, 2, or 3) which can participate in hyperconjugation all these are summed as L_i for some well-defined molecular fragment ($L_i = 1$, if atom is alone); since such fragments are separated with aliphatic groups and are supposed to be independent (orthogonal), they can be understood as aromatic vectors whose summation gives $(\Sigma_i\ L_i{}^2)^{1/2}$ and represents the measure of total (hetero)aromaticity
X_{14}	Similar to X_{13}, the total number of non-σ electrons that can be involved in “aromatic vectors”, what includes: (a) π-electrons of aromatic systems; (b) two electrons for C=C and C=O bonds; (c) two electrons for –N– in aliphatic chains; (d) four electrons for –S–, –O–, –OH; (e) eight electrons for $-NO_2$; (f) two electrons for CH_m ($m = 1$, 2, or 3)
Z_1	Refined AMBER total interaction energy for HIV-1 protease–inhibitor complexes [14]
Z_2	Electrostatic contribution to the free energy of solvation of inhibitor [14]

Table 2
HIV-1 protease inhibitor activity and molecular descriptors X_1–X_8

Number	y^{12}	X_1	X_2	X_3	X_4	X_5	X_6 (Å^{-2})	X_7	X_8
1	9.602	544.694	32	30	43	2.650	0.05395	21	31
2	8.113	558.721	32	30	44	2.627	0.05202	21	32
3	9.721	588.748	34	30	46	2.644	0.05287	21	33
4	9.585	612.693	32	31	47	2.843	0.05099	21	32
5	9.638	570.732	33	32	45	2.643	0.05260	21	33
6	9.222	634.647	32	30	48	3.025	0.05174	21	31
7	9.538	558.721	32	31	44	2.627	0.05225	21	32
8	9.509	559.709	33	31	44	2.659	0.05526	21	31
9	9.569	589.692	38	33	46	2.780	0.06140	21	31
10	5.532	454.569	26	23	37	2.657	0.05283	15	24
11	9.796	560.694	34	31	44	2.691	0.05658	21	31
12	7.561	494.634	33	26	38	2.622	0.06074	15	27
13	9.143	670.591	32	30	44	2.725	0.05104	21	31
14	8.266	572.705	35	32	45	2.707	0.05701	21	31
15	9.276	545.682	33	30	43	2.684	0.05640	21	30
16	9.602	576.760	34	30	44	2.691	0.05525	21	31
17	9.770	600.802	32	31	47	2.565	0.04735	21	35
18	6.943	502.657	30	29	39	2.613	0.05309	18	29
19	8.021	494.634	27	26	38	2.622	0.04923	17	27
20	7.465	528.695	30	30	42	2.608	0.05143	21	31
21	6.161	546.710	32	29	42	2.610	0.05203	18	31
22	6.793	512.649	29	26	38	2.623	0.05023	12	26
23	7.179	574.721	35	34	46	2.667	0.05503	21	32
24	6.673	558.721	32	30	44	2.627	0.05202	21	32
25	6.914	510.677	26	22	39	2.557	0.04526	18	28
26	9.155	558.721	32	30	44	2.627	0.05219	22	32
27	9.745	560.694	34	30	44	2.691	0.05663	22	31
28	7.392	560.694	34	30	44	2.691	0.05663	22	31
29	6.886	544.694	30	30	42	2.608	0.05143	21	31
30	6.836	516.684	30	29	40	2.590	0.05116	18	30
31	10.000	560.694	34	30	44	2.691	0.05639	21	31
32	7.413	532.683	32	29	41	2.633	0.05379	18	30
33	6.230	468.596	26	23	36	2.629	0.05076	17	25
34	9.161	618.777	38	38	51	2.705	0.05843	30	37
35	6.246	440.542	26	23	34	2.688	0.05507	15	23
36	8.886	542.679	33	32	43	2.692	0.05638	21	31
37	10.222	558.678	34	30	45	2.734	0.05902	26	31
38	5.897	584.759	32	30	47	2.621	0.05018	27	34
39	9.638	670.856	37	32	53	2.646	0.05037	26	34
40	8.268	683.896	35	28	55	2.602	0.04634	31	37
41	10.267	683.896	35	28	55	2.602	0.04634	31	37
42	7.277	669.912	33	29	53	2.538	0.04398	26	37
43	5.168	532.814	12	8	52	2.319	0.01914	20	29
44	5.523	501.713	19	15	41	2.434	0.03268	16	27
45	8.116	575.795	25	23	38	2.505	0.03915	25	33
46	6.640	559.709	33	30	44	2.659	0.05477	21	31
47	5.328	484.639	26	22	36	2.560	0.04821	12	26
48	5.862	500.638	28	22	37	2.605	0.04949	12	26
49	4.523	508.705	24	22	40	2.494	0.04105	18	30
50	<8.0	613.804	35	30	49	2.609	0.05521	27	34
51	≈8.9	706.943	39	39	53	2.711	0.05273	22	28
52	≈9.2	491.605	30	22	36	2.776	0.05626	17	23
53	≈8.7	538.749	27	22	41	2.476	0.04192	22	29

calculated as additive properties using fragment increments [18]. X_9–X_{11} were obtained by counting based upon observed logical activity–2D structure rules. In accordance with previous studies [14,15], these descriptors were generated with the assumption that the maximum number of protease subsites (pockets) occupied by inhibitors is four. The 2D formulas [14–19] (Figs. 2–4) contain some stereochemical (3D) information as the drawings are made according to the graphical representation rules recommended by IUPAC [20]. Inhibitor **49**, which was presented but not used in the analysis in previous work [14], is not included in the PCA nor in the external validation set in PLS. However,

Table 3
HIV-1 protease inhibitor molecular descriptors X_9–Z_2 and Y_{pred} activity

Number	X_9	X_{10}	X_{11}	X_{12} (Å^3)	X_{13}	X_{14}	Z_1 (kcal mol^{-1})	Z_2 (kcal mol^{-1})	Y_{pred}
1	4.00	9	3.0	73.4	16.126	48	−80.56	−10.13	9.280
2	3.50	9	2.5	73.4	15.395	46	−76.15	−9.26	7.372
3	4.00	11	3.0	83.8	16.155	52	−84.12	−11.52	9.932
4	4.00	9	3.0	89.0	16.126	48	−82.76	−10.56	9.128
5	4.00	9	3.0	73.4	17.464	50	−82.74	−11.90	9.405
6	4.00	9	3.0	99.4	16.126	48	−79.56	−10.46	9.152
7	4.00	9	3.0	73.4	16.971	50	−81.92	−9.98	9.417
8	4.00	10	3.0	80.9	16.523	50	−81.36	−12.57	9.633
9	4.00	11	3.0	96.9	16.703	56	−84.51	−11.97	9.954
10	3.00	9	2.0	73.4	14.526	40	−67.78	−9.37	5.971
11	4.00	11	3.0	83.8	16.523	52	−81.53	−11.77	9.969
12	3.50	9	2.0	73.4	14.832	44	−74.17	−9.25	6.935
13	4.00	9	3.0	109.1	16.523	48	−83.14	−10.37	9.387
14	3.50	10	2.5	90.9	17.088	50	−81.17	−10.20	7.957
15	4.00	9	3.0	79.6	16.126	48	−81.85	−11.26	9.288
16	4.00	9	3.0	90.9	16.583	52	−80.40	−10.34	9.430
17	4.00	9	3.0	73.4	16.971	50	−85.76	−10.02	9.297
18	3.50	7	2.5	63.0	15.395	44	−73.56	−9.90	6.822
19	3.50	9	2.5	73.4	13.416	44	−75.20	−10.03	7.373
20	4.00	7	3.0	63.0	16.093	44	−77.68	−9.79	8.595
21	3.25	9	2.5	73.4	13.454	46	−70.79	−10.08	6.595
22	3.50	10	2.0	90.9	13.416	42	−69.82	−9.39	6.984
23	3.50	9	2.5	84.2	16.583	54	−75.61	−10.30	7.500
24	3.50	9	2.5	73.4	13.454	46	−78.84	−10.86	7.031
25	3.50	9	2.5	73.4	11.489	40	−74.83	−9.14	7.085
26	4.00	9	3.0	73.4	16.126	48	−81.09	−11.51	9.264
27	4.00	10	3.0	77.1	16.126	52	−82.53	−12.36	9.591
28	3.50	10	2.5	77.1	16.126	52	−76.09	−11.32	7.895
29	3.50	7	2.5	63.0	14.000	42	−76.80	−9.78	6.532
30	3.50	7	2.5	63.0	15.362	42	−75.58	−9.62	6.798
31	4.00	11	3.0	83.8	16.155	52	−82.20	−10.95	9.966
32	3.50	9	2.5	73.4	15.395	46	−74.16	−10.56	7.484
33	3.00	9	2.0	73.4	14.900	38	−65.12	−11.31	6.977
34	4.00	10	4.0	76.6	19.723	60	−88.28	−11.66	11.160
35	3.00	9	2.0	73.4	14.526	40	−61.83	−10.86	6.008
36	3.50	9	2.5	73.4	23.452	48	−79.81	−10.65	8.794
37	4.00	10	3.5	77.1	16.155	52	−83.26	−11.86	9.948
38	3.50	8	2.5	71.1	15.395	46	−66.18	−11.57	7.035
39	4.00	10	3.0	112.6	19.494	52	−86.00	−16.79	9.863
40	3.50	9	3.5	91.3	19.105	50	−81.48	−13.08	9.185
41	4.00	9	4.0	91.3	19.105	50	−91.73	−12.74	10.880
42	3.25	8	3.0	87.6	19.975	42	−80.34	−10.59	7.816
43	2.00	6	2.0	61.6	7.141	20	−73.94	−5.02	1.929
44	3.00	7	2.0	62.4	9.539	28	−70.77	−7.78	4.631
45	3.00	8	3.0	72.6	14.036	38	−80.71	−9.40	6.811
46	3.00	10	2.0	78.6	16.126	50	−72.88	−13.86	6.224
47	3.00	7	2.0	60.9	14.491	38	−68.08	−9.04	5.362
48	3.00	10	2.0	73.6	11.489	40	−66.90	−10.99	5.733
49	3.50	7	2.5	63.0	11.489	34	−72.19	−8.51	6.372
50	4.00	9	3.0	78.2	16.852	50	–	–	–
51	4.00	11	3.5	127.3	20.591	62	–	–	–
52	3.50	10	2.5	82.4	13.675	38	–	–	–
53	3.00	13	3.0	86.6	19.672	38	–	–	–

when literature models take into consideration **49**, this molecule is included in our best PLS for comparative purposes. Furthermore, since there are now five HIV-1 protease inhibitors clinically approved by US FDA Department [19], and one of them is saquinavir **39**, the QSAR variables were derived for all of them (indinavir **50**, ritonavir **51**, amprenavir **52**, nelfinavir **53**). The averages of their experimental activities (IC_{50} values) [19] were expressed as pIC_{50} values and then normalized with respect to pIC_{50} for **39** [14] (Table 1). This procedure, although not entirely accurate, gives approximate and normalized values for pIC_{50}. In vitro pIC_{50} for four of these five inhibitors, measured in the same experimental conditions, range from 7.2 to 8.7 [21].

2.2. Chemometrics

HCA and PCA [22] were carried out using autoscaled data. PLS [22] was performed to build two models: 32/16 (model I, to be comparable to literature models) and 48/0 molecules (model II, to be consistent with the PCA and HCA analysis) in the training/external validation set. The cross-validation strategy in the validation step was leave-two-out. The Pirouette software [23] was used for all chemometrics calculations. Predictions, based on model I, were made also for **49–53**. Two energy variables (AMBER total interaction energy for HIV-1 inhibitor complexes and the electrostatic contribution to the free energy of solvation of substituent (Table 3) from COMBINE-QSAR treatment [16]), were treated as dependent variables and related (through PLS models) to the selected variables for 48 molecules.

3. Results and discussion

Results are presented in Tables 3–7 and Figs. 5–8. HCA plots are in Figs. 5 and 6. PCA plots are in Table 4 and Fig. 7. PLS results are in Tables 3 and 5–7 and Fig. 8.

Table 4
Principal component analysis for 48 samples and 14 variables

PCs	PC1	PC2	PC3
Variance%	56.49	21.86	7.58
Cumulative variance	56.49	78.21	85.79
X_1 or M_r	0.269	0.325	0.234
X_2	0.331	−0.141	−0.086
X_3	0.316	−0.163	−0.260
X_4	0.216	0.405	0.141
X_5	0.224	−0.295	0.244
X_6	0.215	−0.427	−0.163
X_7	0.247	0.352	−0.131
X_8	0.263	0.346	−0.192
X_9	0.292	−0.112	−0.102
X_{10}	0.212	−0.255	0.397
X_{11}	0.285	0.208	−0.130
X_{12} or V_{pol}	0.233	0.016	0.687
X_{13}	0.294	−0.014	−0.188
X_{14}	0.306	−0.224	−0.136

Table 5
Comparison of a priori models with those from literature [14]

Model	Samples	Variables	PCs	r^2	q^2	$SDEP_{cv}$	$SDEP_{ex}$
C_{amber}	32	48	2	0.89	0.70	0.72	0.83
C_{delphi}	32	47	2	0.90	0.73	0.69	0.59
$C_{expanded}$	48	54	2	0.91	0.81	0.66	–
A priori I	32	14	3	0.91	0.85	0.51	1.12
A priori II	48	14	3	0.87	0.77	0.76	–

$SDEP_{cv}$: SDEP (standard error of prediction) of cross-validation, $SDEP_{ex}$: external SDEP.

Table 6
Experimental and predicted activities (pIC_{50} values) for the five clinically approved inhibitors

Sample	Name	Y_{exp}	Y_{pred}
39	Saquinavir	9.638	9.863
50	Indinavir	8.0	9.370
51	Ritonavir	8.9	11.159
52	Amprenavir	9.2	7.741
53	Nelfinavir	8.7	9.234

Table 7
The regression vectors for a priori models I and II

X_i	c_i (model I)	c_i (model II)
X_1 or M_r	−0.0122	−0.0274
X_2	−0.0266	−0.1337
X_3	−0.1156	−0.1014
X_4	−0.0234	−0.0177
X_5	−0.0224	−0.0073
X_6	0.0070	−0.0759
X_7	0.0016	−0.0107
X_8	−0.0068	−0.0021
X_9	0.4682	0.5728
X_{10}	0.2269	0.2309
X_{11}	0.3173	0.4447
X_{12} or V_{pol}	0.0296	0.0103
X_{13}	0.1799	0.1337
X_{14}	0.0997	0.0333

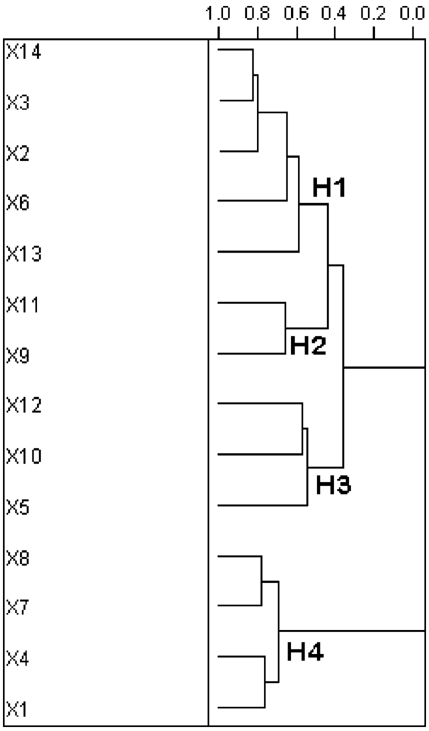

Fig. 5. Hierarchical cluster analysis on 14 variables. **H1**, **H2** and **H3** are sub-clusters of the big cluster, and **H4** is the other, small cluster. This division is based on similarity index 0.50.

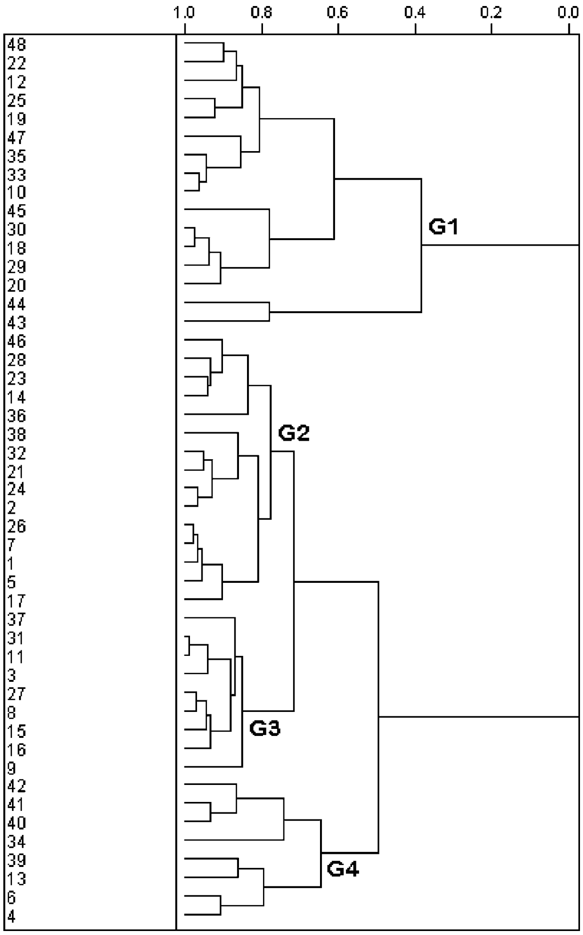

Fig. 6. Hierarchical cluster analysis on 48 samples. **G1** is one cluster, and **G2**, **G3** and **G4** are sub-clusters of the other cluster.

3.1. Biological activity distribution

The biological activity distribution (log units of pIC$_{50}$) reveals three gaps >0.39, and four groups: 5.158–6.246 (**10**, **21**, **33**, **35**, **38**, **43**, **44**, **47**, **48**), 6.640–7.561 (**12**, **18**, **20**, **22–25**, **28–30**, **32**, **42**, **46**), 8.021–8.268 (**2**, **14**, **19**, **40**, **45**), 8.886–10.267 (**1**, **3–9**, **11**, **13**, **15–17**, **26**, **27**, **31**, **34**, **36**, **37**, **39**, **41**). A more regular distribution results when the second and third groups are joined, as has been observed in plots of Y versus X_i. The three groups are characterized by distinctive effective number of substituents X_9 (around 3, 3.5, and 4, respectively). This descriptor has the highest correlation with activity (0.862). In terms of relative activity (IC$_{50\text{rel}}$, with respect to that of **43**, IC$_{50}$ = 6.792 μM), the first group (group I) can be named slightly active (IC$_{50\text{rel}}$ $\approx$ 1–12), the second (group II) moderately active (IC$_{50\text{rel}}$ $\approx$ 30–1300), and the third (group III) highly active (IC$_{50\text{rel}}$ $\approx$ 5200–126 000).

3.2. Classification of the molecular descriptors

A subset of 14 of approximately 30 descriptors were selected. Two descriptors were derived from 1D formula or well-known atomic constants (X_1 and X_5); others were directly counted from 2D formula and atomic valence (X_2 and X_4); the rest were based on 2D formulas and chemical knowledge (stereochemistry from these formulas). The 1D phenomena are described by X_1 and X_5, 2D events are related only to X_2, X_4, X_7, X_8, X_{13}, X_{14}, and the rest are related to 3D events. There are five electronic descriptors (X_2, X_5, X_6, X_{13}, X_{14}), two steric–geometrical (X_9, X_{11}), two electronic–geometrical (X_{10}, X_{12}), one compositional (X_1), one hydrophobic (X_8) and three topological (X_3, X_4, X_7) descriptors. Only X_5 and X_6 are intensive descriptors.

3.3. Hierarchical cluster analysis

The dendogram on variables (Fig. 5) consists of two clusters: a larger (sub-clusters **H1**–**H3**) one and a smaller one (**H4**). The two clusters are distinguished according to the internal structure of the data (behavior around Y versus X_i regression line). **H4** consists of four variables which point out mainly the molecular size (X_1, X_4), shape (X_4, X_7, X_8) and interactions that have no specific direction in space (hydrophobic interactions, X_7, X_8). Pairs of descriptors (X_1, X_4 and X_7, X_8) indicate structural similarity of inhibitors (the same class of peptidic inhibitors), and that most of the rings (substituents P_1, P'_1, P_2, P'_2, especially P_1 and P'_1) are mainly hydrophobic [17,24–26]. Similarly, hydrophobic amino acid residues (Tyr, Pro, Phe, Leu, Ala, Met) of the natural substrates occupy the protease cleavage sites [24]. The descriptors in cluster **H3** (X_5, X_{10}, X_{12}) tend to be more related to electronic properties such as charge distribution, polarity, potential hydrogen bonds. This is in accordance with the fact that electronegative atoms and polar groups in P_2, P'_2, and especially hydrogen bonds are essential for HIV-1 protease–inhibitor binding [17,24–29]. **H2** expresses the complexity of the protease–inhibitor interaction with characteristics like molecular size, topology, steric and conformational properties in terms of two simple variables (X_9, X_{11}). **H1** represents addition details of the electronic distribution, especially the role of non-σ electrons (aromatic, localized, conjugated, free-electron pairs, electrons from CH$_m$ groups in hyperconjugation) responsible for the phenomenon of aromaticity and heteroaromaticity [30,31]. Clustering of X_2, X_3, X_{14} (similarity index 0.8) shows that planar fragments are those contributing mostly to the non-σ electrons. Such (hetero)aromatic and free-electron pair fragments have two important functions. First, they are frequent constituents of compact and/or planar structures (rings) which fit easily into cavities and establish numerous intermolecular interactions. Secondly, having more diffuse electrons, they participate in polar interactions and hydrogen bonds, and also in non-polar interactions (van der Waals and other weak interactions). The cluster analysis on the samples (Fig. 6) shows molecules roughly grouped into two clusters with respect to the activity and molecular size. **G1** (16 samples: **10**, **12**, **18–20**, **22**, **25**, **29**, **30**, **33**, **35**, **43–45**, **47**, **48**) is characterized by low and moderately active compounds.

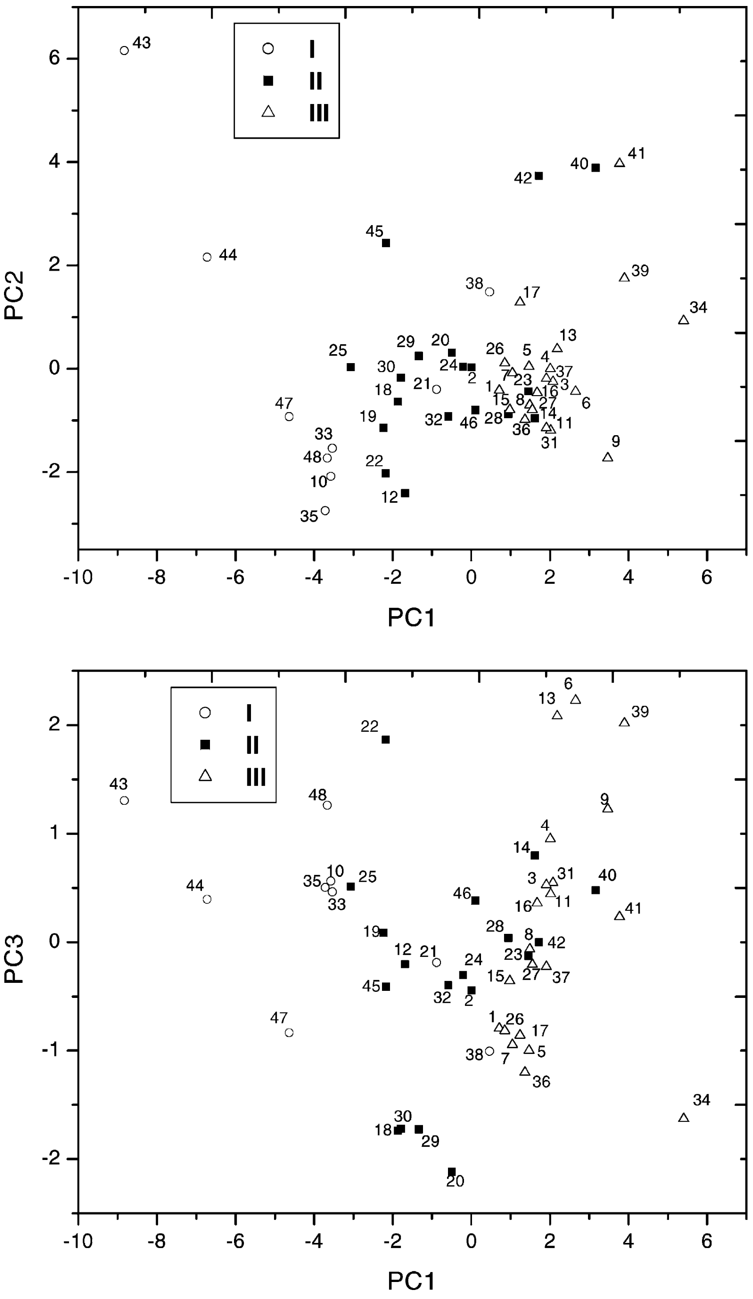

Fig. 7. Principal component analysis on 48 samples using 14 variables. Top: PC2 vs. PC1. Bottom: PC3 vs. PC1. Legend for the activity groups I–III is shown.

Members of this set tend to be the smallest molecules in the set ($M_r = 440$–476), with common structural features including: both the substituents (P_1, P'_1, P_2, P'_2) are small or missing, there is no –OH on P'_2, or P'_2 can be a small ring or a small acyclic systems. This obviously reduces the biological activity. The other, larger cluster consists of three sub-clusters **G2** (15 molecules: **1**, **2**, **5**, **7**, **14**, **17**, **21**, **23**, **24**, **26**, **28**, **32**, **36**, **38**, **46**), **G3** (9 molecules: **3**, **8**, **9**, **11**, **15**, **16**, **27**, **31**, **37**) and **G4** (8 molecules: **4**, **6**, **13**, **34**–**39**, **40**, **41**, **42**). **G2** consists of molecules with medium activity. The molecules can be structurally characterized with respect to molecule **1** as follows: (a) isomers of **1** or close structural analogs (**21**, **32**, **36**, **46**); or (b) having an additional hydrocarbon group (**2**, **5**, **7**, **17**, **24**, **26**, **28**, **38**), or (c) having polar groups at P_1, P_2 which causes sterical hindrance of these substituents (**14**, **23**). **G3** consists of highly active molecules with an electronegative atom more than in P_2, P'_2 of sample **1** (O or N atom). **G4** includes primarily highly active molecules, which are the biggest molecules of the set ($M_\mathrm{r} = 613$–684) and have large P'_1, P_2, P'_2 substituents. Four two-membered sub-clusters (similarity index 0.95) include isomers (**10**, **33**; **2**, **24**; **7**, **26**; **11**, **31**) and three have structurally very similar molecules (**18**, **30**; **1**, **7** or **1**, **26**; **8**, **27**).

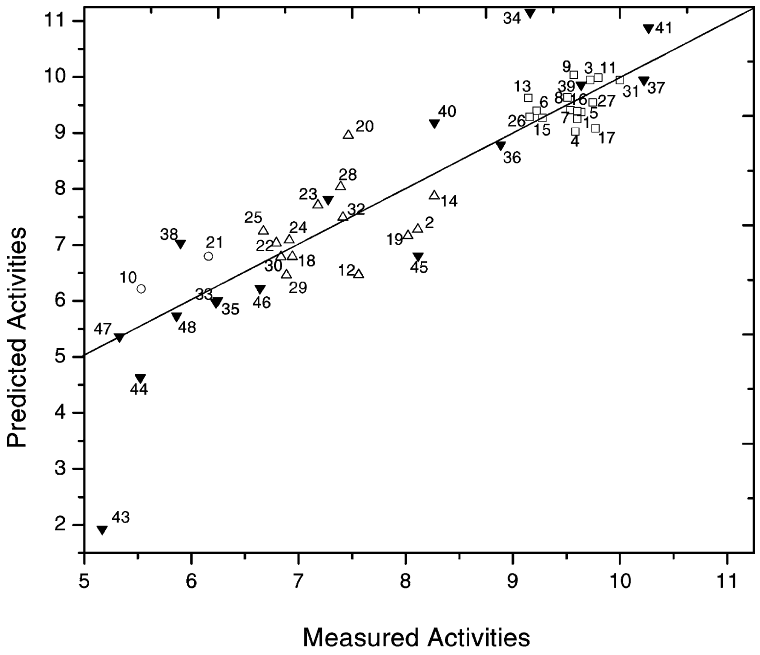

Fig. 8. Correlation between experimental and calculated activities (pIC_{50} values) according to the PLS model I. The training set included 32 molecules, and the external validation set 16 molecules (solid triangles). The activity groups are represented by white circles (group I), triangles (group II) and squares (group III).

3.4. Principal component analysis

The PCA results are listed in Table 4. The first three principal components (PCs) are enough to describe the data set (85.85% of the total variance explained). PC1 roughly separates highly active (group III) inhibitors from slightly active ones (group I), while the moderately active are in the middle (group II) (see Fig. 7). The boundaries of these groups are at PC1 ≈ -3.3 and 0.6, with seven samples being displaced: **21**—isomer of **1** (P_2 is not closed ring); **38**—has a cyclic amide fragment inside the main chain, what causes sterical hindrance to P_2'; **40**—a hydrogen atom in a small P_1 ring disables this ring to be completely exposed to the protease; **23**—there is $–CO_2CH_3$ instead of –OH in P_2', what reduces its hydrogen bonding and the flexibility of P_2'. Steric factors are known to reduce biological activity, which is likely why some molecules are less active than expected (**28**, **14**, **42**). The first two PCs confirm the trends found in HCA. The first three PCs (Table 4) reveal the contribution of molecular descriptors to particular PCs: all the variables are important for PC1 (their coefficients vary in the range 0.21–0.33) with higher contributions from X_2, X_3, X_9, X_{11}, X_{13} and X_{14}, which is in accordance with the HCA results. PC1 is a general PC, closely related to biological activity, and is well expressed in terms of molecular size (cavity, bulk properties) and contents of various types of valence electrons (electronic and hydrophobic properties). The least (**43**, **44**, **47**) and the most active compounds (**34**, **39–41**) in Fig. 7 show that the best inhibitors have maximal effective number of rings and substituents ($X_9 = 4.0$), they are rich in π-electrons and other non-σ electrons from heteroatoms. PC2 mainly includes shape and electronic variables (X_1, X_4, X_6 and X_8 from **H4**) pointing out the complexity of steric and electrotopological properties resulting in activity. PC2 separates more-branched (top: **40–45**) from less-branched molecules (bottom: **9**, **10**, **12**, **35**, **22**, **58**, etc.) (see Fig. 7). Polar groups are most important for PC3 (the most significant variables are X_{10} and X_{12} from **H3**). PC3 separates molecules (relatively to the size of their hydrocarbon parts) rich in electronegative atoms and polar groups (top; **43**, **22**, **48**, **6**, **13**, **39**, etc.) from those having more hydrocarbon (aromatic or aliphatic) fragments (bottom: **20**, **29**, **30**, **18**, **34**, etc.). Thus, PC3 describes the fine (valence electron) distribution of electron density—polarity and hydrogen bond properties. Good inhibitors, besides being electron-rich, have aromatic and hydrophobic fragments. This is in accordance with the fact that hydrophobic interactions are more extensive in number of contacts and contact surface areas than polar interactions. However, the latter are preferable and energetically more favorable [24–26,29,32]. PCA, like HCA, demonstrated that only a few types of molecular properties such as steric/bulk/cavity, electrostatic/electronic/polarity, lipophilic (hydrophobic), hydrogen bonding, electrotopological properties are responsible for drug–receptor interactions, as was well illustrated by Waterbeemd et al. [33].

3.5. Partial least squares regression models

PLS results for models I and II, using 32 and 48 inhibitors in the training set, are in Tables 3 and 5. Both models are comparable with the models of Pérez et al. [16] which used two PCs and obtained better r^2, q^2 and cross-validation standard error of prediction ($SDEP_{cv}$) for 48 samples in the training set than when using 32 samples; their validation was done as random leave-five-out repeated 20 times, which can lead to lower q^2 and higher SDEP than the leave-one-out method [16]. To minimize this difference with respect to our models, a leave-two-out cross-validation algorithm was used. Their C_{delphi} model (32 samples) had six inhibitors (five in the external validation set) with relative error greater than 10% in log units; the average absolute error of prediction was 0.49 (log units) for the validation set and 0.40 for all 48 samples; the outliers with the highest relative error were **33** and **37**, underpredicted by an order of magnitude in IC_{50} units (μM). A priori model I has seven inhibitors (one in the training set) with relative error greater than 10% (Table 3); the average absolute error is 0.77 for the validation set, and 0.46 for 48 samples; the outliers with the greatest relative error (Fig. 8) are **43** and **34** (the former underpredicted by three, and the latter overpredicted by two orders of magnitude in IC_{50} units). Inhibitor **43** has no phenyl groups as substituents (P_1, P'_1), being the smallest and the weakest inhibitor. The **36** is moderately active and electron-rich, but due to an additional double bond in its chain it has reduced flexibility that is required to fit into the protease active site. C_{delphi} predicted activity better than a priori I for 25 molecules. For 22 inhibitors the prediction is reversed. For one molecule the predictions are equal. In the external validation set, seven are predicted better by a priori I and for nine are reversed. The advantages of C_{delphi} model come from incorporated AMBER and electrostatic terms for inhibitor–protease interaction, inhibitor desolvation and solvation [16]. These terms require extensive computer-assisted calculations and so could not be used in an a priori model. OPTIMOL-MM2X model [14] revealed linearity between pIC_{50} and inhibitor–protease interaction energy ($r^2 = 0.78$, $q^2 = 0.76$, $SDEP_{cv} = 0.68$, $SDEP_{ex} = 1.18$) for inhibitors **1**–**32**, **49** in the training set and **33–48** in the external validation set). The equivalent a priori model I (including **49**, $r^2 = 0.90$, $q^2 = 0.81$, $SDEP_{cv} = 0.63$, $SDEP_{ex} = 1.68$) predicted 22 molecules better than OPTIMOL-MM2X; in the validation set, 10 molecules are better predicted by a priori I.

All these comparisons place a priori I model in between C_{delphi} and OPTIMOL-MM2X model. It is worth comparing a priori I to two promotional MLR models obtained by two commercial QSAR software packages of the SciVision company: SCIQSAR3.0 [34] and QSARIS [35]. The same data set was used by both packages as an illustration of their applicability. SCIQSAR3.0 used 30/8 inhibitors in the training/external validation set and five descriptors in their best model ($r^2 = 0.87$, $SDEP_{cv} = 0.50$, no other data available). The best model of QSARIS used 33/15 molecules in the training/validation set and only two descriptors ($r^2 = 0.65$, $q^2 = 0.57$, $SDEP_{cv} = 0.86$, $SDEP_{ex} = 1.49$). Only six inhibitors are predicted better than by the equivalent a priori I in the training set, and five in the validation set.

Hansch and co-workers [36] built a linear regression model with three molecular descriptors and 30 molecules in the training set (the set **1–31**, **49** excluding **24**, **28**). This model ($r^2 = 0.82$, $q^2 = 0.76$, $SDEP_{cv} = 0.69$) is not more quantitatively accurate than a priori model I (with the same molecules in the training set: $r^2 = 0.90$, $q^2 = 0.80$, $SDEP_{cv} = 0.67$). The **49** was predicted by Holloway et al. [14] ($pIC_{50} = 5.532$) better than with a priori I ($pIC_{50} = 6.372$). The **49** is an outlier with the highest residual due to its highly hydrophobic, non-planar cyclohexanyl P'_2. Table 6 shows experimental and predicted activities for the five clinically approved inhibitors **39**, **50–53**. The predictions refer to the group III of highly active inhibitors (with the exception of **52**). Underprediction of amprenavir **52** by more than one, overprediction of indinavir **50** and ritonavir **51** by one to two orders of magnitude in IC_{50} units, can be considered fairly good taking into account the fact that there are no experimental data for all 53 inhibitors measured at the same conditions.

The regression vector coefficients c_i of a priori models I and II are in Table 7. The set **1–32** (used for model I) is structurally more homogeneous than **1–48** (referred to model II). In spite of that, the regression coefficients for both models are quite similar, meaning that a priori descriptors are able to generate a robust model. The highest coefficients refer to X_9 ($c_9 > 0.45$), X_{11} ($c_{11} > 0.30$), X_{10} ($c_{10} > 0.22$), X_{13} ($c_{13} > 0.13$) and X_3 ($c_3 > 0.10$), showing that hydrogen bonds (related to X_{10} and partially to X_{13}) and hydrophobic interactions (described by X_9, X_{11} and partially by X_{13}) are predominant in protease–inhibitor binding. The (hetero)aromaticity variable X_{13} and the number of atoms in planar fragments X_3 take fourth and fifth places. This indicates that new, more efficient HIV-1 protease inhibitors should contain four substituents (mostly ring systems) rich in polar and hydrophobic groups able to participate in electron delocalization. Model II shows the predictional power of the whole set **1–48**, which was extensively described by HCA and PCA.

3.6. Relationships with energetic variables

Energy Z_1 is closely correlated with variables X_4, X_7–X_9 and X_{11} (data: 48 molecules, 14 variables), what is in accordance with high correlation between Z_1 and the anti-viral activity [14,16]. Three PCs are enough to describe the data, and such PLS models are quite sufficient (32/16 molecules in the training/external validation set, 14 variables; $r^2 = 0.88$, $q^2 = 0.76$, $SDEP_{cv} = 2.21\,kcal\,mol^{-1}$ across the range $29.90\,kcal\,mol^{-1}$). According to the regression vector coefficients for X_7–X_{11}, hydrophobic and polar groups, molecular size and shape seem to be significant. Energy Z_2 is correlated with extensive variables X_2, X_3, X_{10} and X_{13} which describe polarity and valence electron distribution (hydrophobic, hydrogen bond properties). This is to be expected due to the nature of inhibitor–polar solvent (water) interactions. PCA with six PCs describes the data well (over

90% of the variance; 48 molecule, 14 variables), and the corresponding PLS model (32 molecules and 14 variables; $q^2 = 0.48$, $r^2 = 0.72$, $\mathrm{SDEP_{cv}} = 0.70\,\mathrm{kcal\,mol^{-1}}$ across a range $8.84\,\mathrm{kcal\,mol^{-1}}$) points out the highest contribution of X_{10}–X_{13} in the regression vector. The electrostatic contribution to the free energy of desolvation of the receptor upon complex formation from the COMBINE-QSAR study [16] shows correlation with variables X_2, X_3, X_{14} (in both scales of desolvation energy). The relationships of a priori descriptors with these energetic variables describing solvation and desolvation phenomena indicates that the a priori variables contain some latent information on interactions including solvent.

4. Conclusion

Fifty-three HIV-1 protease inhibitors, of which 49 were peptidic inhibitors, were described by a priori molecular descriptors, and their anti-viral activities were studied by means of chemometrics, where biological activities for 49 inhibitors having been measured under the same experimental conditions. The chemometric analysis of data for 48 inhibitors demonstrated that the biological activity (more precisely: enzyme–inhibitor binding) is a 3D phenomena in terms of principal components: the first PC is a general PC (bulk, electronic and hydrophobic properties), the second describes stereochemical fit to enzyme (steric and electrotopological properties) and the third is related to distribution of electron density (polarity and hydrogen bonding). The inhibitors are conveniently grouped as slightly, moderately and highly active compounds. In the light of a priori descriptors, a good peptidic inhibitor should have four aromatic and/or ring substituents rich in polar and hydrophobic groups. Fourteen a priori molecular descriptors of various chemical nature (electronic, steric–geometrical, electronic–geometrical, compositional, hydrophobic, topological) well characterized the studied inhibitors and two PLS models were built and successfully compared with those from literature.

Acknowledgements

The authors acknowledge FAPESP for financial support.

References

[1] H. Van de Waterbeemd, B. Testa, The parametrization of lipophilicity and other structural properties in drug design, Adv. Drug Res. 16 (1987) 85–225.

[2] L.M. Rellick, W.J. Becktel, Comparison of van der Waals and semiempirical calculations of the molecular volumes of small molecules and proteins, Biopolymers 42 (1997) 191–202.

[3] M.D. Barratt, J.V. Castell, M. Chamberlain, R.D. Combes, J.C. Dearden, J.H. Fentem, I. Gerner, A. Giuliani, T.J.B. Gray, D.J. Livingstone, W. McLean Provan, F.P. Zbinden, The integrated use of alternative approaches for predicting toxic hazard. The report and recommendations of ECVAM workshop 8, Alternat. Lab. Anim. (ATLA) 23 (1995) 410–429.

[4] S.C. Basak, G.D. Grunwald, G.J. Niemi, Use of graph-theoretic and geometrical molecular descriptors in structure–activity relationships, in: A.T. Balaban (Ed.), From Chemical Topology to Three-Dimensional Geometry, Plenum Press, New York, 1997, pp. 73–116.

[5] N. Trinajstić, Chemical Graph Theory, 2nd ed., CRC Press, Boca Raton, FL, 1992.

[6] S.C. Basak, B.D. Gute, G.D. Grunwald, Graph theory invariants, molecular similarity and QSAR, in: Proceedings of the First Indo-US Workshop on Mathematical Chemistry with Applications in Molecular Design and Hazard Assessment of Chemical, Visva-Bharati University, Santiniketan, West Bengal, India, 9–13 January 1998, Abstract available at http://wyle.nrri.umn.edu/India/Abstract.html [accessed on 5 February 2002].

[7] K. Balasubramanian, Integration of graph theory and quantum chemistry, in: Proceedings of the First Indo-US Workshop on Mathematical Chemistry with Applications in Molecular Design and Hazard Assessment of Chemical, Visva-Bharati University, Santiniketan, West Bengal, India, 9–13 January 1998. Abstract available at http://wyle.nrri.umn.edu/India/Abstract.html [accessed on 5 February 2002].

[8] K.L.E. Kaiser, S.P. Niculescu, Using probabilistic neural networks to model the toxicity of chemicals to the fathead minnow (*Pimephales promelas*): a study based on 865 compounds, Chemosphere 38 (1995) 3237–3245.

[9] R. Todeschini, V. Consoni, Dragon, Version 1.1, Milano Chemometrics and QSAR Research Group, University of Milano, Milan, Italy, 2000.

[10] A. Golbraikh, D. Bonchev, Y.-D. Xiao, A. Tropsha, Novel chiral topological descriptors and their applications to QSAR, in: Proceedings of the 13th European Symposium on Quantitative Structure–Activity Relationships: Rational Approaches to Drug Design, Heinrich-Heine Universität, Düsseldorf, Germany, 27 August to 1 September 2000, Abstract Book, p. 40.

[11] R.D. Brown, Y.C. Martin, The information content of 2D and 3D structural descriptors relevant to ligand–receptor binding, J. Chem. Inf. Comput. Sci. 37 (1997) 1–9.

[12] V.J. Gillet, P. Willett, J. Bradshaw, Reduced graphs as descriptors of bioactivity, in: Proceedings of the Fifth International Conference on Chemical Structures, Noordwijkerhout, The Netherlands, 6–10 June 1999, Plenary lecture 13, Abstract available at http://www.chem-structure.org/ [accessed on 5 February 2002].

[13] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics: Part B, Elsevier, Amsterdam, 1998 (Chapter 37).

[14] M.K. Holloway, J.M. Wai, T.A. Halgren, P.M.D. Fitzgerald, J.P. Vacca, B.D. Dorsey, R.B. Levin, W.J. Thompson, L.J. Chen, S.J. Desolms, N. Gaffin, A.K. Ghosh, E.A. Giuliani, S.L. Graham, J.P. Guare, R.W. Hungate, T.A. Lyle, W.M. Sanders, T.J. Tucker, M. Wiggins, C.M. Wiscount, O.W. Woltersdorf, S.D. Young, P.L. Darke, J.A. Zugay, A priori prediction of activity for HIV-1 protease inhibitors employing energy minimization in the active site, J. Med. Chem. 38 (1995) 305–317.

[15] M. Pastor, C. Pérez, F. Gago, Simulation of alternative binding modes in a structure-binding based QSAR study of HIV-1 protease inhibitors, J. Mol. Graphics Mod. 15 (1997) 363–371.

[16] C. Pérez, M. Pastor, A.R. Ortiz, F. Gago, Comparative binding energy analysis of HIV-1 protease inhibitors: incorporation of solvent effects and validation as a powerful tool in receptor-based drug design, J. Med. Chem. 41 (1998) 836–852.

[17] A. Wlodawer, J. Vondrasek, Inhibitors of HIV-1 protease: a major success of structure-assisted drug design, Annu. Rev. Biophys. Biomol. Struct. 27 (1998) 249–284.

[18] A. Gavezzotti, Molecular packing and correlations between molecular and crystal properties, in: H.-B. Bürgi, J.D. Dunitz (Eds.), Structure Correlation, vol. 2, VCH, Weinheim, 1994, pp. 509–542.

[19] Database for Anti-HIV Compounds. National Institute for Allergy and Infectious Deseases, Bethesda, MD, USA. Data available at http://www.niaid.nih.gov/daids/dtpdb/ [accessed on 5 February 2002].

[20] G.P. Moss, Basic terminology of stereochemistry, Pure Appl. Chem. 68 (1996) 2193–2222.

[21] C. Flexner, HIV-1 protease inhibitors, New Engl. J. Med. 338 (1998) 1281–1292.

[22] K.R. Beebe, R.J. Pell, M.B. Seasholtz, Chemometrics: A Practical Guide, Wiley, New York, 1998.

[23] Pirouette, Version 2.7., Infometrix Inc., Seattle, WA, 2000.

[24] M. Sakurai, S. Higashida, M. Sugano, H. Handa, T. Komai, R. Yagi, T. Nishigaku, Y. Yabe, Studies of human immunodeficiency virus type 1 (HIV-1) protease inhibitors. III. Structure–activity relationship of HIV-1 protease inhibitors containing cyclohexylalanylalanine hydroxyethane dipeptide isostere, Chem. Pharm. Bull. 42 (1994) 534–540.

[25] N. Pattabiram, Occluded molecular surface analysis of ligand–macromolecule contacts: application to HIV-1 protease–inhibitor complexes, J. Med. Chem. 42 (1999) 3821–3834.

[26] A. Velazquez-Campoy, M.J. Todd, E. Freire, HIV-1 protease inhibitors: enthalpic versus entropic optimization of the binding affinity, Biochemistry 39 (2000) 2201–2207.

[27] A.K. Gosh, J.F. Kincaid, D.E. Walters, Y. Chen, N.C. Chaudhuri, W.J. Thompson, C. Culberson, P.M.D. Fitzgerald, H.Y. Lee, S.P. McKee, P.M. Munson, T.T. Duong, P.L. Darke, J.A. Zugay, W.A. Schleif, M.G. Axel, J. Lin, J.R. Huff, Nonpeptidal P_2 ligands for HIV protease inhibitors: structure-based design, synthesis, and biological evaluations, J. Med. Chem. 39 (1996) 3278–3290.

[28] S.S. Abdel-Meguid, B.W. Metcalf, T.J. Carr, P. Demarsh, R.L. DesJarlais, S. Fisher, D.W. Green, L. Ivanoff, D.M. Lambert, K.H.M. Murthy, S.R. Petterway Jr., W.J. Pitts, T.A. Tomaszek Jr., E. Winborne, B. Zhao, G.B. Dreyer, T.D. Meek, An orally bioavailable HIV-1 protease inhibitors containing and imidazole-derived peptide bond replacement: crystallographic and pharmacokinetic analysis, Biochemistry 33 (1994) 11671–11677.

[29] G. Jones, P. Willett, R.C. Glen, A.R. Leach, R. Taylor, Development and validation of a genetic algorithm for flexible docking, J. Mol. Biol. 267 (1997) 727–748.

[30] A.R. Katrizky, M. Karelson, S. Sild, T.M. Krygowsky, K. Jug, Aromaticity as a quantitative concept. 7. Aromaticity reaffirmed as a multidimensional characteristic, J. Org. Chem. 63 (1998) 5228–5231.

[31] C.W. Bird, Heteroaromaticity. 14. The conjugation energies and electronic structures of nonbenzenoid polycyclic aromatic systems, Tetrahedron 54 (1998) 10179–10186.

[32] J.P. Glusker, M. Lewis, M. Rossi, Crystal Structure Analysis for Chemists and Biologists, VCH, New York, 1994, pp. 627–628.

[33] H. Waterbeemd, G. Constantino, S. Clementi, G. Cruciani, R. Valigi, Disjoint principal properties of organic substituents, in: H. Waterbeemd (Ed.), Chemometric Methods in Molecular Design, VCH, Weinheim, Germany, 1995, pp. 103–112.

[34] SCIQSAR, Version 3.0, SciVision, Burlington, MA, 2000. Data available at http://www.scivision.com/gProdPage/tAppNotes/sciQSAR/priori.html [accessed on 5 February 2002].

[35] QSARIS, 2000, SciVision, Burlington, MA. Data available at http://www.scivision.com/gProdPage/tAppNotes/qsarIS/qsarIS_Note3.html [accessed on 5 February 2002].

[36] R. Garg, S.P. Gupta, H. Gao, M.S. Baby, A.K. Debnath, C. Hansch, Comparative quantitative structure–activity relationship studies on anti-HIV drugs, Chem. Rev. 99 (1999) 3526–3601.