QSAR of Progestogens: Use of *a Priori* and Computed Molecular Descriptors and Molecular Graphics

Rudolf Kiralj, Yuji Takahata and **Márcia M. C. Ferreira***

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

Abstract

Quantitative Structure-Activity Relationship (QSAR) study of two sets of oral progestogens was carried out by using Principal Component Analysis (PCA), Hierarchical Cluster Analysis (HCA) and Partial Least Squares (PLS). *A priori*, computed (at DFT 6-31G** level) and molecular graphics and modeling descriptors were employed. Molecular graphics and modeling studies of crystal structures of complexes progesterone receptor (PR)-progesterone, Fab'-progesterone and PR-metribolone have been performed. QSAR of progestogens is a three-dimensional phenomenon (over 96% of information is explained by the first three Principal Components), which can be, although it exhibits significant non-linearity, treated well with linear methods such as PLS. Progestogen activity depends primarily on double bond contents and resonance effects which define the skeletal conformation, and also on substituent characteristics (size, conformational and electronic properties). Sterical relationships between a substituent at $C6(sp^2)$ or $C6(sp^3)$-α and sulfur atom from Met 801 residue of PR are important for progesterone binding to the protein and can be quantified. Basically the same was observed for substituents at β-C10 with respect to residue Met759.

---

* To receive all correspondence

Key words: progesterone, chemometrics, principal component analysis, hierarchical cluster analysis, partial least squares

Abbreviations:
PR = progesterone receptor
HCA = Hierarchical Cluster Analysis
MMFF = molecular mechanics force field
VWN = Vosco-Wilk-Nussair
ROP = relative oral progestatinal activity
AOP = acetoxyprogesterone
X = the substituent atom covalently bound to C_6
X' = any other substituent atom except X

List of Symbols
pIC — activity defined as $\log(1/IC)$
pIC_{exp} — experimental activity
pIC_{pre} — calculated activity
m — the number of multiple bonds counted as the number of π-electron pairs (for set I)
n — the weighted number of non-σ valence electrons of the multiple bonds and heteroatoms
k — the number of changes in substitutents with respect to the most active molecule **VI** in set I
V — van der Waals volume of substituent at $C6(sp^2)$ or α-substituent at $C6(sp^3)$
N_6 — the number of electrons from double bonds, free pairs of O and of halogen of substituent at C6
L_j — frontier orbital density at j-th skeleton atom
L_{13} — frontier orbital density at C13
L_{14} — frontier orbital density at C14
M_4 — unweighted 3D-Morse descriptor (signals 4)
M_{11} — unweighted 3D-Morse descriptor (signals 11)
R_0 — Rte GETAWAY descriptor
S_6 — the projected area of the substituent at C6 with hydrogens
S_6' — the projected area of the substituent at C6 without hydrogens

S — the projected area of H6β and substituents at C1, C2, C6-β, C9−C13, C21
P_1 — the cumulative descriptor $P_1 = S + S_6'$
D_{XS} — the X S distance
D_{CX} — the X−C6 bond length
D'_{XS} — the X' S distance
D'_{XX} — the X−X' bond length or any other bond length in the substituent
R_S — van der Waals radius of methionine sulfur
R_X — van der Waals radius of atom X
R'_X — van der Waals radius of atom X'
Δ — parameter calculated as $\Delta = D_{XS} - (R_S + R_X)$
P_6 — molecular modeling descriptor including geometry around X, defined as $P_6 = S(R_S - |\Delta|)/D_{XS}$.
y — dependent variable in parabolic fit $y = a + b\,x + c\,x^2$
x — independent variable in $y = a + b\,x + c\,x^2$
a, b, c — the regression coefficients in $y = a + b\,x + c\,x^2$
$\sigma(a), \sigma(b), \sigma(c)$ — the statistical errors of the coefficients a, b, c,
n' — descriptor defined as $n' = n + (c/b)n^2$
m' — descriptor defined as $m' = m + (c/b)m^2$
L_{13}' — descriptor defined as $L_{13}' = L_{13} + (c/b)L_{13}^2$
D_{5-10} — the length of C5−C10 bond
D_{9-10} — the length of C9−C10 bond
R — the correlation coefficient from prediction
Q — the correlation coefficient from validation
D_H — the minimum distance of PR and water atoms (including hydrogens) from a particular progesterone hydrogen atom
D — the minimum distance of PR and water atoms (excluding hydrogens) from a particular progesterone hydrogen atom
η_H — the sum of valence electrons of PR and water atoms (including hydrogens) inside 5.5 Å cut-off distance sphere around a particular progesterone hydrogen atom
η — the sum of valence electrons of PR and water atoms (excluding hydrogens) inside 5.5 Å cut-off distance sphere around a particular progesterone hydrogen atom

Table 1. The IUPAC names of progestogens (AOP = acetoxyprogesterone)*

Mol.	Name*	Mol.	Name*
I	progesterone	9	6α-bromo-17α-AOP
II	norethisterone (norethindrone)	10	6α-methyl-17α-AOP
III	norgestimate	11	6α-choro-17α-AOP
IV	levonorgestrel	12	6α-bromo-1-dehydro-17α-AOP
V	desogestrel	13	6α-fluoro-1-dehydro-17α-AOP
VI	gestodene	14	6-fluoro-1,6-bisdehydro-17α-AOP
VII	17-deacetylnorgestimate	15	6α-methyl-1-dehydro-17α-AOP
VIII	5-α-dihydrotestosterone	16	6α-chloro-1-dehydro-17α-AOP
1	norethisterone (norethindrone)	17	6-methyl-1,6-bisdehydro-17α-AOP
2	17α-AOP	18	6-methyl-6-dehydro-17α-AOP
3	17-α-ethinyltestosterone	19	6-fluoro-6-dehydro-17α-AOP
4	21-chloro-1,6-bisdehydro-17α-AOP	20	6-chloro-1,6-bisdehydro-17α-AOP
5	6α-nitro-17α-AOP	21	6-chloro-6-dehydro-17α-AOP
6	6β-chloro-17α-AOP	22	17α-hydroxy-17β-ethynil-gona-4-ene-3-one
7	6α-fluoro-17α-AOP	23	6-iodo-6-dehydro-17α-AOP
8	21-fluoro-1,6-bisdehydro-17α-AOP	24	6-ethyl-6-dehydro-17α-AOP

* Set I: molecules **I – VIII**; set II: molecules **1 – 24.** Note that there is only one molecule in common for both sets: **II** and **1.**

1 Introduction

Progestogens are widely known as oral contraceptives, but the current literature reveals that health research [1 – 4] (hormone replacement and various anti-cancer therapies, gynecological disorders etc.) and veterinary science [5, 6] are today the two most promising areas of their application. The lack of large and homogeneous progestogen activity data is the major limitation to have an entirely clear picture of the progestogen behavior at molecular level. This class of compounds has been target of various (Quantitative) Structure-Activity Relationship (Q)SAR studies in the last four decades [7]. These studies were confronting the difficulty in describing well the molecular properties of progestogens without having the 3D receptor-drug complex structure, and in treating the non-linearity of steroid QSAR [8]. Recently, the crystal structure of progesterone receptor (PR)-progesterone complex [9] made it possible to explain mutations at atomic level [10] and perform more promising drug design. In previous work [11], SAR studies of oral contraceptive activity (OCA) of set I (**I – VIII**) of progesterones, and oral progestational activities (OPA) of 17α-acetoxyprogesterones in set II (**1 – 24**), presented in Figure 1 and Table 1, were performed. The authors employed molecular descriptors calculated by semi-empirical molecular orbital methods. The aim of the present work is to apply different methods and descriptors to extend the study to QSAR level in order to deepen the knowledge on the PR-progestogen binding. *A priori* descriptors [12, 13], computed molecular descriptors at *ab initio* level, and molecular graphics and modeling descriptors are used in this work. The relationships among the descriptors and among the samples are studied using Principal Component (PCA) and Hierarchical Cluster (HCA) Analyses [14]. Partial Least Squares (PLS) regression [14] is used to build models for both progestogen data sets and to predict activities of proposed progestogen derivatives. Molecular graphics is employed at

qualitative (3D visualization of molecular and crystal structure of PR-progesterone complex) and quantitative level (generation of molecular graphics and modeling descriptors). Structural study of PR-progesterone and related complexes Fab'-progesterone and PR-metribolone serves as a supporting tool for the chemometric analysis.

2 Methodology

2.1 Modeling and Geometry Optimization of Progestogens

Molecular structure of progesterone was built, optimized with MMFF94 force field [15] (Monte Carlo conformational search), semi-empirical PM3 [16] and finally with *ab initio* density-functional method, DFT, (6-31G**, VWN functional [17]) incorporated in Titan [18]. Structures of progestogens were built by modifying the structure of progesterone optimized with PM3, and then they were optimized at the same DFT level. Molecular properties for all the molecules were calculated at this DFT level.

2.2 Generation of Molecular Descriptors

The oral contraceptive activities for set I are from literature [19]. The original data are given as inhibition of ovulation (mg/d), were converted arbitrarily to IC (the molar concentration of drug daily necessary to inhibit ovulation [11]) and further to the form pIC = log(1/IC). The activities for set II are relative oral progestatinal activity (ROPA) (Clauberg assay) taken from literature [20], where the reference compound is norethindrone (**1**).

There are many ways to generate descriptors in (Q)SAR studies. Information that can be obtained from (Q)SAR depends very much on the type of descriptors. Whole molecular, substituent, atomic and bond descriptors for
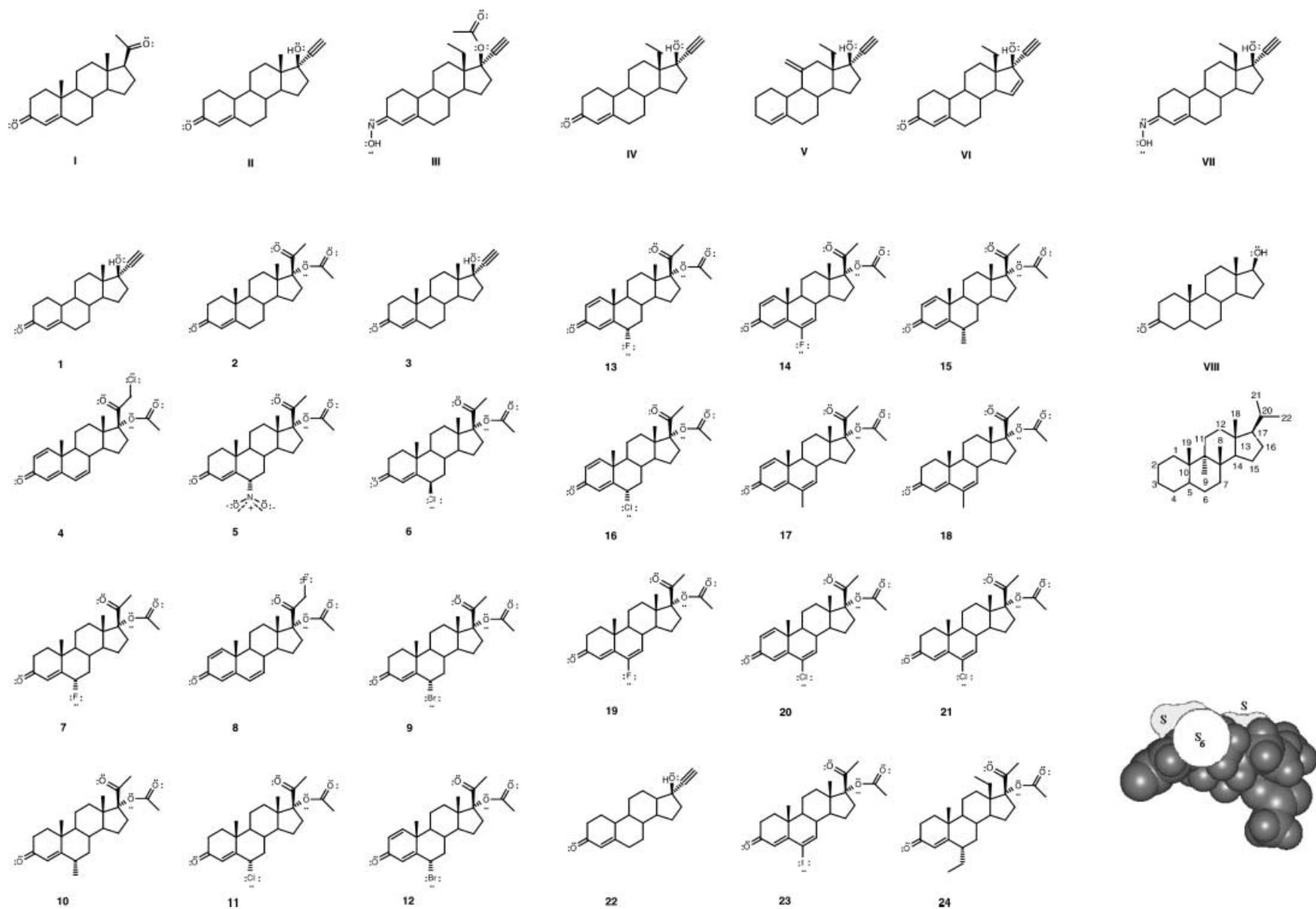
Figure 1. Schematic representation of set I (top and top right, **I – VIII**) and set II (**1 – 24**) of progestogens including multiple bonds, free electron pairs and the nomenclature system (right). Training sets: **I – VI**, **1 – 21**. Prediction sets: **VII** and **VIII**, **22 – 24**. Surface areas S and S_6 for set II are also shown (right bottom).

both sets of progestogens were generated in three different ways as follows.

a) *A priori* descriptors [12, 13], "known before" any sophisticated computer-assisted calculation. They are simple descriptors based on 2D chemical structures and chemical intuition, sometimes using well-known tabulated data, and can be calculated without computer aid. They are defined in this study in the following way:

m – the number of multiple chemical bonds counted as the number of π-electron pairs (*i.e.* those pairs that eventually can participate in π...π intermolecular interactions), as follows: 1 for double bond, 2 for triple bond;

n – the weighted number of non-σ valence electrons of the multiple bonds and heteroatoms (which can participate in electron delocalization and π...π interactions), counted in the following way: 2 for double bond, 2 for triple bond (overall population 0.5 for each electron since the two π-orbitals are mutually perpendicular and subject to free rotation), 2 for O in $-$OH and $-$O$-$, and 3 for halogens (overall population 0.5 for each electron due to free rotation) with added electrons from $-$CH$-$, $-$CH$_2-$and $-$CH$_3$ aliphatic groups (*i.e.* from C-H groups involved in hyperconjugation) in the neighborhood of multiple bonds (effective population 0.577 per electron for fixed regular tetrahedral groups and 0.5 for free rotation groups);

k – the number of changes in substituents with respect to the most active molecule **VI** in set I, counted in a way that each presence/absence of substituent at some position, as well as change of the substituent at the same position, counts 1;

V – van der Waals volume of substituent at C6(sp^2) or α-substituent at C6(sp^3) ($-$CH$_3$, $-$Cl, $-$F, $-$H, $-$Br, $-$I, $-$CH$_2$CH$_3$, $-$NO$_2$), as the sum of volume increments [21];

N_6 – the number of electrons from double bonds, free pairs of O and of halogen (if halogen is conjugated with a double bond) of C6 substituent.

b) Structural and electronic descriptors from DFT calculations.

L_j – frontier orbital density [22] at j-th skeleton atom as $L_j = \Sigma_i \, c_{ij}^2$ where c_{ij} is an atomic orbital coefficient at j-th atom in i-th molecular orbital (LUMO);

M_4, M_{11} – unweighted 3D-Morse descriptors (signals 4, 11) [23];

R_0 – Rte GETAWAY descriptor, and other molecular descriptors using Dragon [24];

other descriptors calculated by Titan – partial atomic charges, total and frontier orbital energies, dipole moments, moments of inertia, lipophilicity, ovality, bond lengths in the steroid skeleton etc.

c) Molecular graphics-based descriptors called molecular graphics and modeling descriptors [13, 25 – 27], calculated for set II.

These descriptors can be calculated by measuring surface areas of atoms, molecules or functional groups projected orthogonally onto the plane of projection (the paper of the picture printout) [13]. Additionally measured structural parameters like non-bonding distances, can be useful in combination with these surface areas. In the present study, the progestogen molecules were aligned in the view along C6(sp^2)-substituent or C6(sp^3)-α-substituent bond (viewing from the substituent towards the progesterone skeleton) by using PLATON [28], and a few areas were measured by empirical method [13] (Figure 1 right down). These areas were used as molecular graphics descriptors:

S_6, S_6' – the projected area of the substituent at C6 with and without hydrogens, respectively. S_6' was set to zero for the C$_6$-αH atom;

S – the projected area of H6β and substituents at C1, C2, C6-β, C9$-$C13, C21

P_1 – the cumulative descriptor $P_1 = S + S_6'$

Substitution effects include changes in molecular size and partially in its shape, in the double bond contents in rings A-C, and conformational changes of the rings A-C. All these structural variations of the set are reflected well in the measured areas S_6 and S, and this phenomenon seems to be in accordance with the induced fit model for steroids [7], where the PR molecule adopts an appropriate conformation to maximize the binding with the steroid molecule. This was the reason to calculate molecular graphics-based descriptors as chemically understandable concepts, easy to derive.

Molecular modeling descriptors incorporated parameters of progesterone-methionine 801 interaction geometry based on the crystal structure of PR-progesterone complex [9] in the following way. Molecules of various derivatives of progesterone were modeled by changing substituents at C6 (H, CH$_3$, CH$_2$CH$_3$, NO$_2$, F, Cl, Br, I) in PR-progesterone complex, and some PR-progestogen geometry parameters were measured by Titan. Van der Waals radii by Bondi [29] were used, and Δ parameter was calculated as $\delta = D_{XS} - (R_S + R_X)$ where D_{XS} is the X...S distance (X is the substituent atom covalently bound to C6, and S is sulfur atom from the methionine residue), R_S, R_X are van der Waals radii of methionine sulfur (1.80 Å and of X. It can happen that X sterically hinders the progestogen binding to PR. This justifies the generation of some molecular graphics and modeling descriptors taking into account the geometry around X, such as descriptor P_6 defined as $P_6 = S\,(R_S - |\Delta|\,/\,D_{XS}$ where $R_S - |\Delta|$ is normalized with respect to the X...S distance. Information on substitutions at other positions is introduced to P_6 by surface area S.

2.3 Chemometrics

Study of nonlinear relationships between pIC and molecular descriptors in set I, and between ROPA and (V, N_6) in set II, was performed. Molecules with similar structural characteristics were grouped and linear $y = a + b\,x$ and parabolic relationships $y = a + b\,x + c\,x^2$ ($y =$ pIC or ROPA, $x = V$, N_6 or descriptors in set I) were established using Matlab 5.4 [30]. Statistical t-test on regression coefficients was performed in order to find if linearity or simple non-linearity is relevant for these steric and electronic descriptors.

The generated knowledge-based *a priori* and molecular graphics and modeling descriptors showed moderate to high

correlation with the two biological activities (correlation coefficients above 0.5). In order to have more complete description on progestogens, variable selection for molecular descriptors generated by other methods was performed based on the correlation matrix. Only a few steric and electronic descriptors from DFT were selected (cut-off for correlation coefficient 0.5). In order to find a suitable electrotopological or similar descriptors, the variable selection for descriptors generated by Dragon was carried out in the same way. Only a few electrotopological descriptors have been selected. After the variable selection, the obtained data sets were autoscaled prior to QSAR analysis. PCA and HCA were applied on the training (set I: **I** – **VI**; set II: **1** – **21**) and training + prediction set (set I: **I** – **VIII**; set II: **1** – **24**). Incremental linkage method was used in HCA.

PLS models were built and validated by leave-one-out cross-validation using the descriptors previously selected for PCA and HCA. Biological activity for five molecules (**VII**, **VIII** from set I, **22** – **24** from set II) were predicted. The PLS models were compared to analogous PCR (Principal Component Regression) models.

All PCA, HCA and PLS/PCR calculations were performed by using Piroutte software [31].

2.4 Molecular Graphics and Modeling Studies on PR-Progesterone Complex

Progesterone hydrogens in PR-progesterone complex [9] were generated by Titan. The neighborhood (enzyme and water) space inside 5.5 Å cut-off distance sphere (determined in previous study as the most suitable [13]) around each progesterone hydrogen atom was described by four steric and electronic descriptors, as has been shown to be an efficient approach for HIV-1 peptidic inhibitors [13]. First, the minimum intermolecular distance between a particular progesterone hydrogen atom and surrounding neighborhood atoms was measured: as D excluding the hydrogen atoms from the neighborhood, and as D_H including them. In the next step, the neighborhood atoms inside the cut-off sphere were counted, and the corresponding numbers of valence electrons were summed: η – the sum of valence electrons from the neighborhood non-hydrogen atoms, η_H – the analogous sum including the neighborhood hydrogens. Local routines [32] were used to perform these calculations. The four descriptors were autoscaled and treated by PCA and HCA in a standard procedure. The most appropriate linkage method in HCA was selected on the basis of its interpretability in terms of 3D progesterone and PR-progesterone complex structures *i.e.* to recognize the clusters of hydrogens in the PR-progesterone crystal structure [9]. This way, the complete linkage method showed to be superior to the single and incremental methods. The objective of these PCA and HCA analyses was to find out which progesterone hydrogens are suitable for substitution with respect to the free space around them in the PR pockets.

Steric, hydrophobic and electronic PR-progesterone complementarity, geometry of specific intermolecular interactions and search for free space for new substituents was performed by the molecular graphics software WebLab ViewerLite [33]. Fixed PR and progesterone in PR-progesterone complex were studied by modifying progesterone into progestogens: substituents from sets I and II were generated in empty PR pockets by using Titan.

Such visualization of steric, hydrophobic and electronic PR-progesterone relationships helps to understand progesterone activity as a phenomenon determined by these steroid properties. Even though geometry of such complexes have not been optimized, they can aid in better understanding of drug-receptor interaction. Besides, PR flexibility is limited and so large amino-acid displacement caused by insertion of a big substituent is not expected. This approach, based on unoptimized geometry of a drug-receptor complex, can be called *a priori* modeling, as has been shown for peptidic HIV-1 protease inhibitors [13]. It is

Table 2. Structure-activity data for set I*

Ds.[a]\Mol	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**
pIC^b_{exp}	3.02	5.77	6.17	6.62	6.71	6.89	(active)[c]	(inact.)[c]
pIC^b_{pre}	3.04	5.77	6.30	6.43	6.60	7.04	6.08	0.23
m	3	4	5	4	4	5	4	1
n	10.462	11.462	15.462	13.462	15.770	16.6161	13.462	6.308
k	4	1	2	0	2	0	1	3
D_{5-10}/Å	1.510	1.500	1.501	1.500	1.506	1.501	1.502	1.511
D_{1-10}/Å	1.528	1.520	1.522	1.520	1.522	1.519	1.521	1.507
D_{9-10}/Å	1.547	1.523	1.533	1.533	1.529	1.532	1.532	1.548
L_{13}	0.0433	0.0676	0.0704	0.0669	0.0573	0.0668	0.0740	0.0429
L_{14}	0.0378	0.0293	0.0214	0.0292	0.0143	0.0289	0.0226	0.0371
R_0	32	29	30	29	30	28	29	32
m'	2.011	2.242	2.254	2.253	2.242	2.242	2.242	0.890
n'	6.773	7.035	7.405	7.355	7.389	7.312	7.355	4.967
L_{13}'	0.0281	0.0305	0.0301	0.0305	0.0306	0.0306	0.0295	0.0280

[a] Molecular descriptors. D_{5-10}, D_{1-10} and D_{9-10} are bond C5−C10, C1−C10 and C9−C10 lengths, respectively. See text for explanation for the definition of other descriptors. [b] The experimental and predicted (PLS) biological activity as −logIC. [c] Predicted in literature as active/inactive [11].

assumed that there would not be more than approximately 1 Å displacement between the raw and optimized complex geometries at the active site hole.

Visualization of two related complexes, PR-R1881 (R1881 or metribolone; PDB: 1E3K) [10] and Fab'-progesterone (PDB: 1DBB) [34] was performed by Titan and WebLab ViewerLite and compared to PR-progesterone complex at qualitative level.

3 Results and Discussion

3.1 Set I of Progestogens
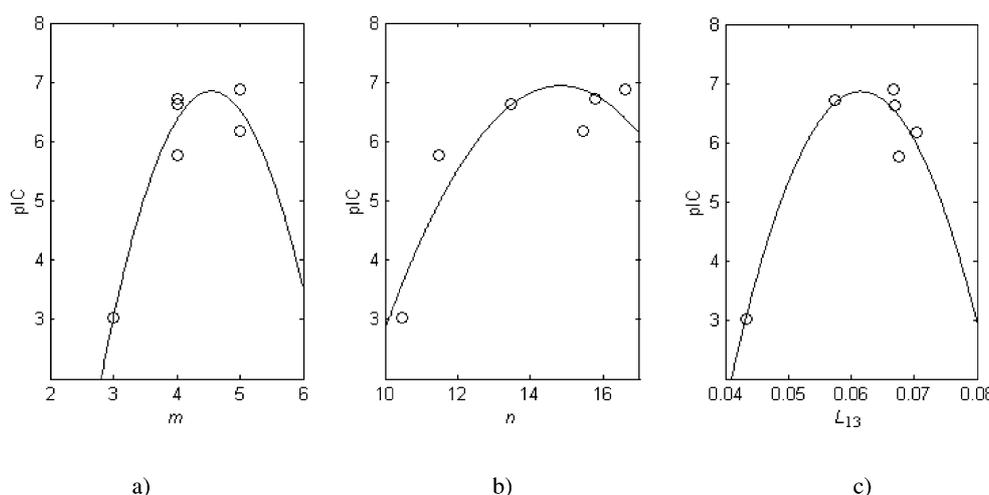
3.1.1 Non-linear Relationships in Set I

Twelve selected molecular descriptors are listed in Table 2. L_{13} and L_{14} are LUMO electron densities at C13 and C14, respectively. Non-linearity of pIC-x relationship (x is any molecular descriptor) was studied by comparing the correlation coefficients r between experimental (pIC$_{exp}$) and calculated (pIC$_{pre}$) activities, the mean absolute deviation δ of pIC$_{pre}$ from pIC$_{exp}$, the ratios $a/\sigma(a)$, $b/\sigma(b)$, $c/\sigma(c)$ from linear pIC $= a + b \, x$ and parabolic pIC $= a + b \, x + c \, x^2$ fits ($\sigma(a)$, $\sigma(b)$, $\sigma(c)$ are statistical errors of the coefficients a, b, c, respectively), and the corresponding t-values. Visual inspection of pIC-x plot was also performed. In general, parabolic fits have much higher coefficients r and low δ than linear fits for some descriptors. Visual inspection and the ratios $a/\sigma(a)$, $b/\sigma(b)$, $c/\sigma(c)$ showed that parabolic regression can be used only in the case of molecular descriptors n, m, L_{13}. The Student test maximum probability for linear/parabolic fits are 0.82/0.16, 0.92/0.04, and 0.65/0.03 for n, m and L_{13}, respectively. Thus new descriptors, including non-linear terms from the parabolic fits, were calculated: n', m', L_{13}' of the form $x' = x + (c/b) \, x^2$ (Table 2). Figure 2 illustrates non-linear relationships pIC-x, where x is n, m and L_{13}. The

non-linearity is quite pronounced, thus a simple linear approximation would not be the most appropriate. What would be implication of the non-linearity? pIC reaches its maximum with respect to the variation of any of the three descriptors, n, m and L_{13}. The physical meaning of the two descriptors, n and m, is somewhat similar. Roughly speaking, they represent the total number of loosely bound electrons in the molecule. L_{13} is electron density of LUMO at C13. n, m and L_{13} must take certain optimum values (in a range being not too small nor too large), in order to attain the highest activity (pIC). This is quite reasonable, due to the specific nature of the receptor (PR). The number, position, size, shape, electronic properties of the substituents which are appropriate for hydrogen bonding, hydrophobic, polar and other intermolecular interactions can exhibit non-linearity when the enzyme pockets are limited in size and flexibility, predominantly hydrophobic or polar. The values of descriptors such as n, m and L_{13} must be also limited.
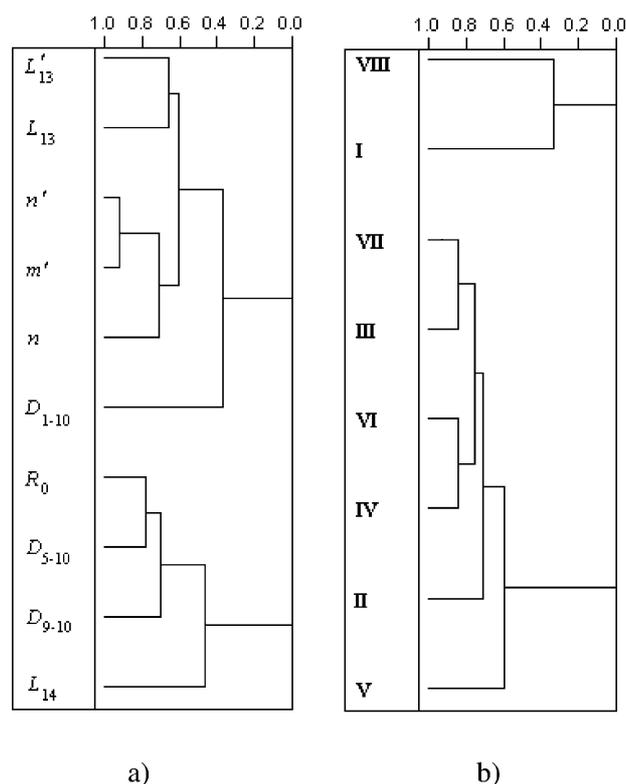
3.1.2 Hierarchical Cluster Analysis

The results of HCA analysis are shown in Figure 3. The analysis included all the samples and 10 variables; m and k were excluded in variable selection for PLS, and for the sake of the same consistence were excluded in HCA. Dendogram of variables (Figure 3a) exhibits two well defined clusters. The four-membered one contains R_0, D_{5-10}, D_{9-10}, L_{14} which decrease as pIC increases. The variables in the other cluster n, n', m', L_{13}, L_{13}' (with exception of D_{1-10}) increase as pIC increases. These observations can be checked using the data in Table 2. It is interesting to note that the similarity index between L_{13} and L_{13}' is only 0.66, and between n and n' is not greater than 0.71. These facts indicate that non-linear terms included in molecular descriptors n' and L_{13}' contain some useful information for (Q)SAR study of set I.

Dendogram of samples (Figure 3b) consists of two separated clusters: the small one (low active or inactive **I**

Figure 2. Non-linear activity-molecular descriptor relationships in set I: a) pIC-m; b) pIC-n; c) pIC-L_{13}, where m is the number of multiple bonds, n is the weighted number of non-σ valence electrons, and L_{13} is the frontier orbital density at C13.

Figure 3. The results of Hierarchical Cluster Analysis for set I: a) dendogram of variables exhibiting two clusters, depending on if variables increase (smaller cluster) or decrease (bigger cluster) as pIC increases; b) dendogram of samples consisting of two progestogen classes: low active/inactive (**I**, **VIII**) and moderately/highly active (**II** – **VII**).

and **VIII**) and the big one (moderately and highly active **II** – **VII**). The position of samples **VII** and **VIII** in the dendogram is, in general, in accordance with previous predictions of their progestational activities [11], that is, **VII** is active and **VIII** is low active. Besides, the structural characteristics (see Figure 1) determine the clustering in Figure 3b. **I** and **VIII** differ only in the presence of Me group at position 21 and in hybridization around C4—C5 bond, what makes them to be the little cluster. Molecules **II** – **VII** belong to other class of progesterone derivatives, 17-α-acethylene progesterones. Molecules **IV** and **VI** differ only in the formal multiplicity of the bond C15—C16 (similarity index 0.84), and **III** and **VII** in the presence of MeCO-group at position 17-β (similarity index 0.85). These two sub-clusters are clearly visible in Figure 3b. Furthermore, **II** differs from **IV** only in the presence of Me instead of Et at position 18. Molecule **V** has no C=O group at position 3 but has ethylene group at C11, what makes this molecule to be the most isolated in the big cluster.
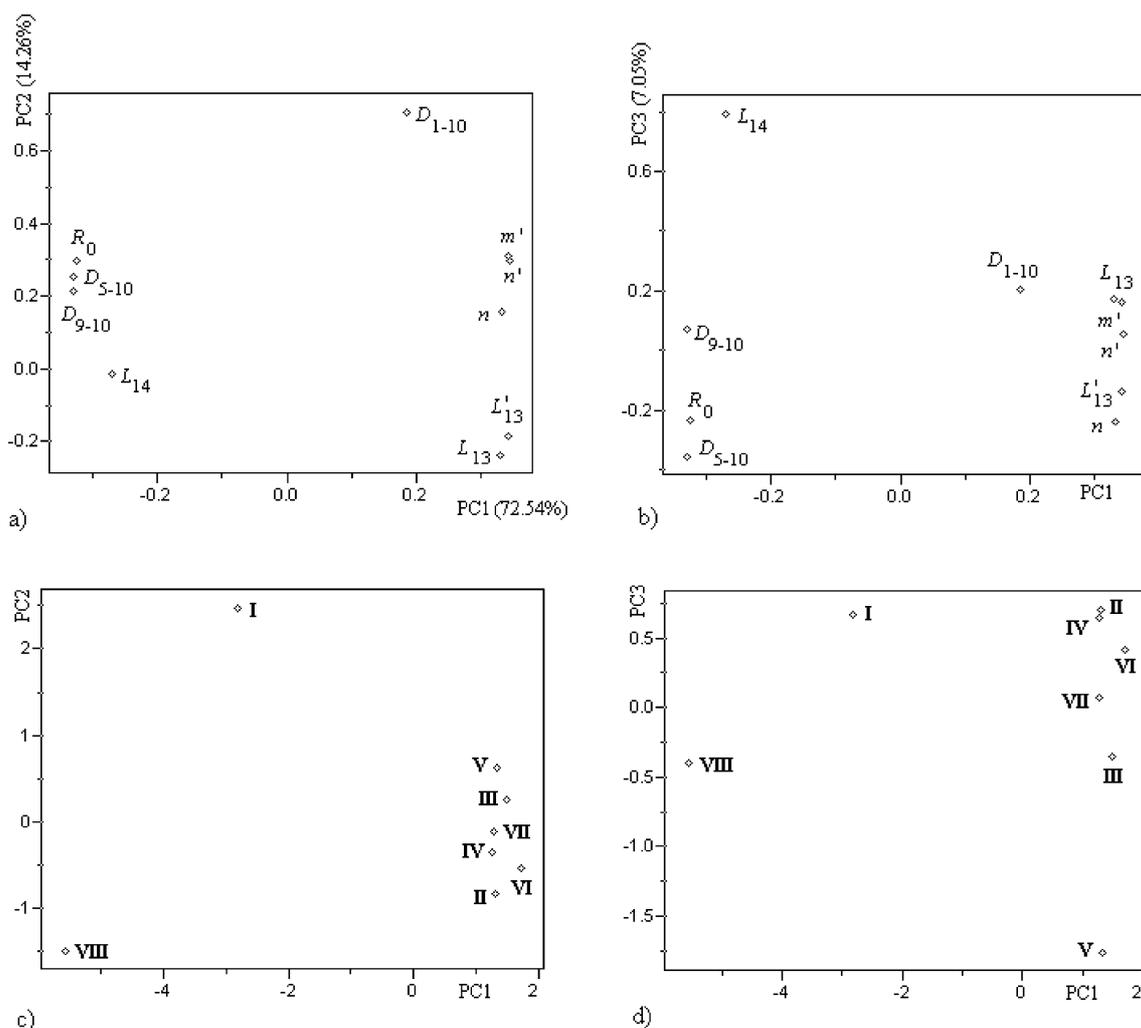
3.1.3 Principal Component Analysis

The first three principal components (PCs) describe 93.86% information of the original data set. The loadings plots

(Figure 4a, b) reveal two well separated groups with respect to PC1. These groups are equivalent to the two clusters in HCA (Figure 3b). PC3 shows that L_{13} and L_{13}', as well as n and n' do not contain the same information. The distance between the variables in each pair is pronounced, being more than 25% of the maximum distance between D_{5-10} and L_{14} along PC3. Furthermore, there are more pairs of other ("linear") variables characterized by $\leq 25\%$ of the maximum distance in PC3.

The scores plots (Figure 4c, d) exhibit a cluster containing molecules **II** – **VII**, while **I** is far and **VIII** even farther from the cluster. This arrangement of the samples reminds on dendogram of samples (Figure 3b). Low active **I** and **VIII** are characterized by long bonds D_{5-10} and D_{9-10} (compare the loadings and the scores plots in Figure 4); molecules **II** – **VII** have shorter bonds D_{5-10} and D_{9-10}, the lengths of which are in narrow intervals $1.500-1.506$ and $1.523-1.533$ Å, respectively (Table 2). By other words, the degree of electron delocalization *via* mechanisms of conjugation and hyperconjugation is less pronounced in skeleton of **I** and **VIII** than that of **II** – **VII**. High contribution of descriptors n, n', m' to PC2 (Figure 4a, b) *i. e.* higher contents of multiple bonds and non-σ valence electrons, is another aspect confirming the fact on electron delocalization in progestogens skeleton. Moderately and highly active compounds have also high L_{13}, L_{13}' and low L_{14}, what could reflect the importance of hydrophobic substituent at position 13 (Me, Et) in intermolecular interaction with hydrophobic residues of PR. It can be said that, in general, the PCA results agree with the HCA results.

3.1.4 Partial Least Squares

The best PLS model (training set: **I** – **VI**, prediction set: **VII**, **VIII**) included 10 variables and 3 PCs reaching $Q^2 = 0.739$, $R^2 = 0.992$. This model is better than the analogous PCR model ($Q^2 = 0.669$, $R^2 = 0.978$). It can be noticed that the non-linearity in variables n' and L_{13}' is crucial for these models. By excluding the two variables, it becomes impossible to construct good models (PLS with 3 PCs: $Q^2 = 0.117$, $R^2 = 0.975$; PCR with 2 PCs: $Q^2 = 0.000$, $R^2 = 0.949$, while with 3 PCs is even worse). The predicted activities by the PLS model with linear descriptors (Table 2) do not deviate from experimental activities more than 0.19 in log units (pIC). Compound **VII** is predicted active (Table 2), almost as **III**. The dendogram (Figure 3b) and the scores plots (Figure 4c, d) show significant similarity between these two compounds, not being a surprize that they behave equally in biological sense. Compound **VIII** is predicted inactive (Table 2), what is in accord to HCA (Figure 3b) and PCA (Figure 4c, d) studies. Besides, the previous work [11] confirms the predictions for **VII** and **VIII**. The regression vector (Table 3) can lead to some conclusions about non-linearity in progestogen QSAR. The coefficient for L_{13}' is greater than that of L_{13} (their absolute values are considered). Besides, the regression coefficients for n and n' are practically equal, while the coefficient for m' is greater than

Figure 4. The results of Principal Component Analysis for set I: the loadings plots – a) PC1 *vs.* PC2, and b) PC1 *vs.* PC3; the scores plots – c) PC1 *vs.* PC2; d) PC1 *vs.* PC3.

those for other four descriptors (D_{5-10}, D_{9-10}, L_{13}, L_{14}). These comparisons reconfirm the importance of introducing non-linear terms into molecular descriptors for progestogens.

3.2 Set II of Progestogens

3.2.1 Non-linear Relationships in Set II

Activity (ROPA, Table 5 depends parabolically on van der Waals volume (V) of substituents at C6 or C21 in sets (**12**, **13**, **15**, **16**) and (**2**, **7**, **9 – 11**), as is presented in Figure 5. The corresponding maximum linear/parabolic fit probabilities from the Student test are 0.73/0.11 and 0.22/0.16, respectively. Sample **5** was excluded from further analysis due to relatively high deviation. Analytically obtained parabola for set (**2**, **4**, **8**) is very much similar to these two parabolas (practically the same shape, see Figure 5a). Each parabola means that the activity changes with the size of substituent at C6 or C21. The differences among the parabolas in position with respect to the origin confirm that significant contribu-

Table 3. The PLS regression vector for set I

Descriptor	Coefficient
n	0.20
D_{5-10}	0.0016
D_{1-10}	− 0.18
D_{9-10}	0.0034
L_{13}	− 0.043
L_{14}	− 0.066
R_0	− 0.21
m'	0.11
n'	0.21
L_{13}'	0.19

tion to activity comes from the skeleton conformation *i.e.* the number and distribution of double bonds in it. The activities for other two sets (**14**, **17**, **20**) and (**18**, **19**, **21**), depend parabolically on V also (the two parabolas are of the same shape, Figure 5b). The maximum Student probabilities for linear fit for these two parabolas and for that for (**2**, **4**, **8**)

Table 4. Measured projected areas S_6 and S for set II

No	$S_6/\text{Å}^2$*	$S/\text{Å}^2$*	C6**	others#	Mol	$S_6/\text{Å}^2$*	$S/\text{Å}^2$*	C6**	others#
1	4.44(4)	6.59(4)	H	7H, 2CH$_3$	**13**	6.70(4)	6.84(4)	F	3H, 2CH$_3$
2	4.77(4)	9.48(5)	H	6H, 2CH$_3$	**14**	6.77(4)	9.11(5)	F	2H, 2CH$_3$
3	4.76(4)	7.91(4)	H	6H, 2CH$_3$	**15**	11.89(5)	6.45(4)	CH$_3$	3H, 2CH$_3$
4	4.45(4)	9.88(5)	H	2H, 2CH$_3$, Cl	**16**	9.67(4)	7.78(4)	Cl	3H, 2CH$_3$
5	13.87(6)	5.88(4)	NO$_2$	5H, 2CH$_3$	**17**	12.10(5)	8.94(5)	CH$_3$	2H, 2CH$_3$
6	4.30(4)	9.91(5)	H	5H, 2CH$_3$	**18**	12.17(6)	10.01(5)	CH$_3$	4H, 2CH$_3$
7	6.59(4)	9.00(5)	F	5H, 2CH$_3$	**19**	6.70(4)	10.12(5)	F	4H, 2CH$_3$
8	4.46(4)	9.65(5)	H	2H, 2CH$_3$, F	**20**	9.65(5)	9.41(5)	Cl	2H, 2CH$_3$
9	10.86(5)	8.97(5)	Br	5H, 2CH$_3$	**21**	9.55(5)	10.78(5)	Cl	4H, 2CH$_3$
10	12.09(6)	8.48(5)	CH$_3$	5H, 2CH$_3$	**22**	4.41(3)	3.24(3)	H	7H
11	9.49(5)	8.18(5)	Cl	5H, 2CH$_3$	**23**	12.37(6)	10.18(5)	I	4H, 2CH$_3$
12	10.80(5)	7.41(4)	Br	3H, 2CH$_3$	**24**	17.94(7)	8.03(5)	Et	4H, CH$_3$, Et

* Experimental errors in brackets. ** S_6 is measured for C6 substituent. # S is measured for other substituents (see text).

Table 5. Some parameters of the C6–S(Met 801) geometry

| C6	$D_{CX}/\text{Å}^a$	$D_{XS}/\text{Å}^b$	$D'_{XS}/\text{Å}^c$	$R_X/\text{Å}^d$	$R'_X/\text{Å}^e$	$D'_{XX}/\text{Å}^f$	$\Delta/\text{Å}^g$
H	1.096	4.290	3.637	1.70	1.20	–	0.790
CH$_3$	1.531	3.441	2.947, 2.978, 4.013	1.70	1.20	1.096	-0.059
NO$_2$	1.471	3.466	3.527, 3.321	1.46	1.52	1.210	0.206
F	1.383	3.503	–	1.47	–	–	0.233
Cl	1.772	3.352	–	1.75	–	–	-0.198
Br	1.933	3.301	–	1.85	–	–	-0.349
I	2.132	3.247	–	1.98	–	–	-0.533
Et	1.531	3.441	C: 4.677, H: 2.947, 2.978, 4.602, 5.326, 5.343	1.70	C: 1.70, H: 1.20	C: 1.531, H: 1.096	-0.059

^{a} X–C6 bond length (X – the substituent atom covalently bounded to C6). ^{b} X...S distance (S from Met 801). ^{c} X'...S distance (X' is any other substituent atom except X). ^{d} van der Waals radius of X. ^{e} van der Waals radius of X'. ^{f} X–X' bond length as well as other bond lengths in the substituent. ^{g} Δ parameter (see text).

confirm that linearity as not being the relevant relation with activity for variable V (0.67, 0.64, 0.91, respectively). These molecules are characterized by relatively flat region between C3 and C7. Activity is also parabolically dependent on the number of partially delocalized electrons N_6 in sets with the same substituent at C6: CH$_3$ (**10**, **15**, **17**, **18**), Cl (**11**, **16**, **20**, **21**) and F (**7**, **13**, **14**, **19**), as shown in Figure 5c. The maximum Student probabilities for linear/parabolic fits are 0.68/0.04, 0.40/0.13 and 0.61/0.11, respectively. The number of double bonds in steroid rings also shows to be non-linearly related with the activity, as is in Figure 5d. The Student maximum probability for the linear/parabolic fit is 0.24/0.004. Dispersion around the parabola is pronounced, indicating that molecules with common properties should be treated as separated sets, as was the case in Figures 5b and 5c. These non-linear trends in progesterone QSAR can be explained as follows. ROPA has its maximum when the value of V, N_6 and the number of double bonds in steroid rings takes optimum value. The value of V, for instance, must not be too small neither too large in order for a steroid to attain high biological activity (ROPA). Similar interpretation to that for V and N_6 applies to the number of double bonds in steroid rings. All these observations on non-linearity in set II are analogous to those in set I.

3.2.2 Hierarchical Cluster Analysis

Table 4 contains measured projected surface areas S_6 and S. Geometrical parameters of C6-S(Met801) are in Table 5. These data were used to calculate molecular graphics and modeling descriptors (see Methodology). Five molecular descriptors were chosen in the variable selection step (Table 6). Dendogram of variables (Figure 6a) shows that M_{11} is isolated and M_4 is relatively distant from the molecular graphics descriptors (S_6', P_1, P_6). The similarity index between P_1 and S_6' is 0.80, what is in accord with the definition of P_1 (see the methodology section). Surprisingly, the similarity index between P_6 and P_1 is only 0.62, meaning that introducing parameters of PR-progesterone geometry (Δ, D_{XS}, Table 5) into molecular graphics descriptors brings some new information about progestogen binding to the PR.

Dendogram of samples (Figure 6b) exhibits two distinct clusters; the smaller one contains samples with low biological activity (see Table 6). The big cluster consists of two sub-clusters; most of molecules in one of them are moderately active, and in the other most of molecules are highly active. Compounds **22**, **23** and **24** are predicted as low, moderately and highly active, respectively. One would expect such a trend due to molecular structure of these molecules.

a)

b)

c)

d)
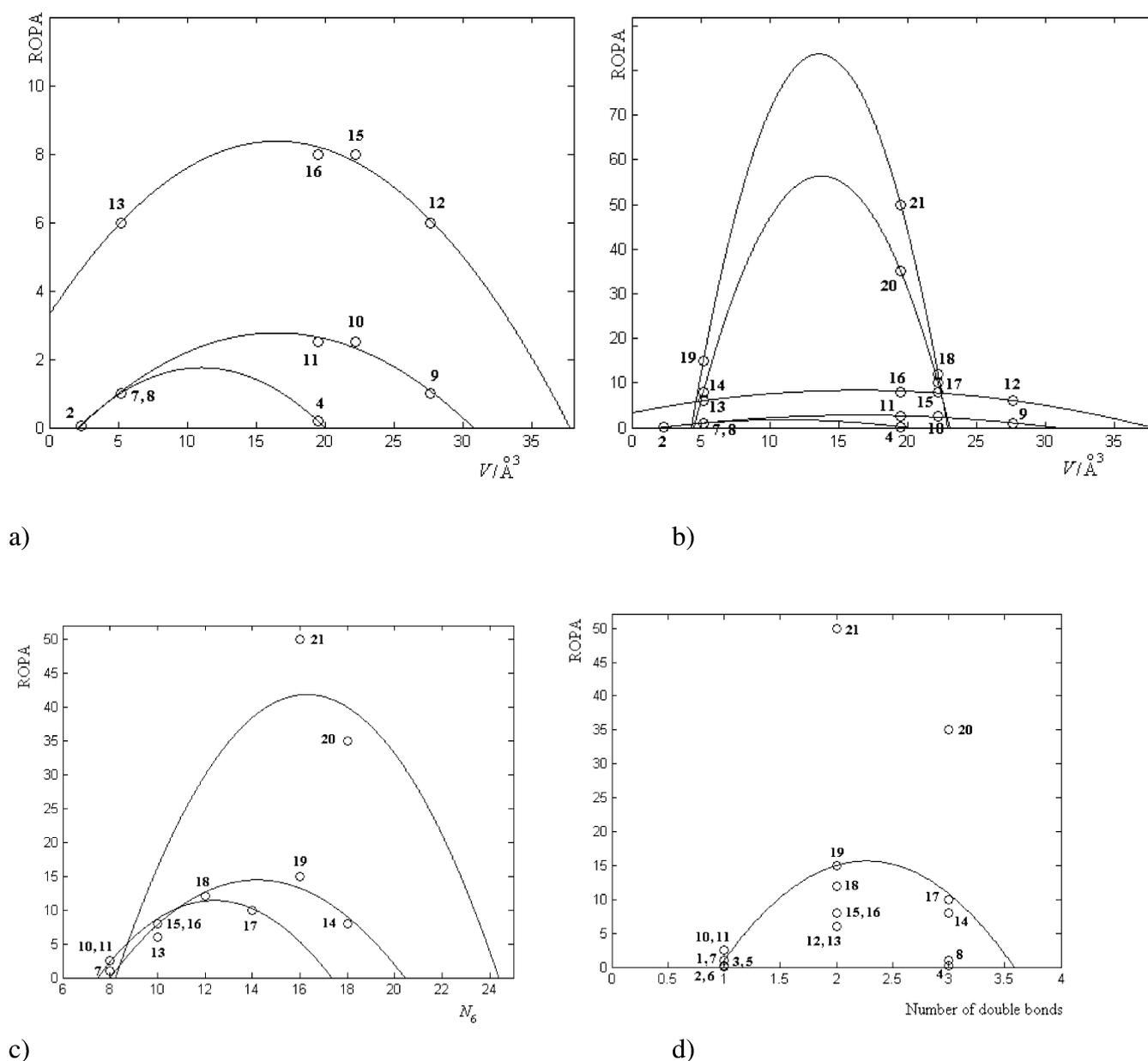
Figure 5. Non-linear activity (ROPA) – molecular descriptor (x) relationships in set II: a) for $x = V$, represented by three parabolas with the same shape; b) for $x = V$, for all the five parabolas; c) for $= N_6$ represented by three parabolas; d) for $x =$ the number of double bonds in steroid rings, visualized by one parabola. V is the van der Waals volume of substituent at C6(sp²) or a-substituent at C6(sp³). N_6 is the number of electrons from double bonds and lone pairs of oxygen or halogen atom at C6.

3.2.3 Principal Component Analysis

Set II can be well described by three PCs (96.04% of total original information). Variables are grouped basically in the same way as in HCA. Molecular graphics descriptors have predominant contribution in PC1 (Table 7), while M_4 is the most important for PC2. PC3 is mostly related to M_{11}.

Clusters of low, moderately and highly active samples are observed in PC1-PC2 scores plot (Figure 7a) in the same way as in HCA (Figure 6b). Highly active compounds are positively related to molecular graphics-based descriptors

and negatively to M_{11}. The opposite is for low active compounds, while moderately active stand between these two tendencies. Molecular-graphics based descriptors contain information on more factors determining PR-progestogen binding: progestogen conformation (especially of rings A and B), content of π- and lone pair electrons, substituent position, and some substituent properties (mainly steric). By other words, these descriptors correspond to enhanced progestogen fitting to the PR active site hole, what is the reason why the quadrant PC1 > 0, PC2 < 0 is occupied mosty by highly active molecules (Figure 7a). In PC1-PC3

Table 6. Structure-activity data for set II. Experimental [19] and predicted (PLS) Relative Oral Progestational Activities (ROPA) are listed in the last two columns.

No	M_4	M_{11}	$S_6'/\text{Å}^2$	$P_1/\text{Å}^2$	$P_6/\text{Å}^2$	IC_{exp}	IC_{pred}
1	10	0	0	6.59	1.55	1	0.23
2	11	−1	0	9.48	2.23	0.07	0.17
3	11	0	0	7.91	1.86	0.20	0.15
4	9	−1	0	9.88	2.33	0.20	0.71
5	12	−2	7.49	13.37	2.70	0.25	0.43
6	10	−1	0	9.91	2.33	0.50	0.37
7	11	−1	6.59	15.59	4.03	1[a]	1.85
8	8	−1	0	9.65	2.27	1	1.31
9	11	−1	10.86	19.83	3.94	1	4.89
10	12	−1	9.08	17.56	4.29	2.50[b]	1.87
11	11	−1	9.49	17.67	3.91	2.50[b]	3.27
12	10	−2	10.80	18.21	3.26	6	5.17
13	9	−1	6.70	13.54	3.06	6	3.85
14	8	−1	6.77	15.88	4.08	8	14.66
15	11	−2	9.08	15.53	3.26	8	1.64
16	10	−1	9.67	17.45	3.72	8	5.88
17	9	−2	9.08	18.02	4.52	10	13.58
18	10	−2	9.08	19.09	5.06	12	9.79
19	9	−2	6.70	16.82	4.52	15	8.47
20	8	−1	9.65	19.06	4.50	35	36.21
21	9	−2	9.55	20.33	5.15	50	23.58
22	9	−2	0	3.24	0.76	–	0.17
23	13	−2	12.37	22.55	3.97	–	1.77
24	9	−1	14.14	22.17	4.06	–	39.54

[a] Reference [35]. [b] Reference [36].

Table 7. PCA results for set II.

	PC1	PC2	PC3
M_4	0.091	0.978	0.003
M_{11}	−0.318	0.036	0.945
S_6'	0.547	0.060	0.121
P_1	0.560	−0.012	0.220
P_6	0.526	−0.198	0.210

Table 8. The PLS regression vector for set II

	Y1
M_4	−0.46
M_{11}	0.053
S_6'	0.37
P_1	0.19
P_6	0.28

plot (Figure 7b) the samples are arranged in three lines: (**1**, **3**), (**2**, **4**, **6 – 11**, **13**, **14**, **16**, **20**, **24**) and (**5**, **12**, **15**, **17 – 19**, **21 – 23**) corresponding to three distinct values of M_{11} (0, − 1, − 2, respectively). Activity for **22 – 24** is predicted as in HCA (compare Figures 6b and 7a).

3.2.4 Partial Least Squares

The best PLS model (Figure 8) with three PCs reaches $Q^2 = 0.692$, $R^2 = 0.811$. The PCR model ($Q^2 = 0.693$, $R^2 = 0.805$) is similar to the PLS model. Significant deviations of predicted activities from experimental (greater than 4 in IC units) are obtained for samples **9**, **14**, **15**, **17**, **19**, **21** (Table 5). Activities for samples **22 – 24** are predicted as expected. **22** without CH_3 groups at positions 10 and 13 should be low active (too small inhibitor, doesn't interact with S from Met801); **23** with large monoatomic C6-substitutent (iodine atom), due to sterical hindrance to S(Met 801) could be low to

moderately active; **24** with CH_3CH_2 at position 6 could be highly active (it fits into the pocket better than any monoatomic or spherical polyatomic substituent). Geometry parameters in Table 5 illustrate well such behavior. The C6−X bond length (see Methodology section) and the molecular volume vary as the substituent at 6-α position changes; this results in more or less preferable interaction with S(Met 801). In the case of CH_3 and CH_3CH_2, there is a C...S and two H...S interactions at the distance of sums of van der Waals radii (Table 5). In other words, situation in which Δ parameter is close to zero represents the best choice for substituents with respect to Met 801 residue. The regression vector (Table 8) exhibits important contribution of all molecular graphics descriptors. ROPA increases as S_6', P_1 and P_6 increase. Highly active compounds are characterized by high S_6', P_1, P_6 and low M_4, M_{11} in PCA also (see Tables 7 and 8).
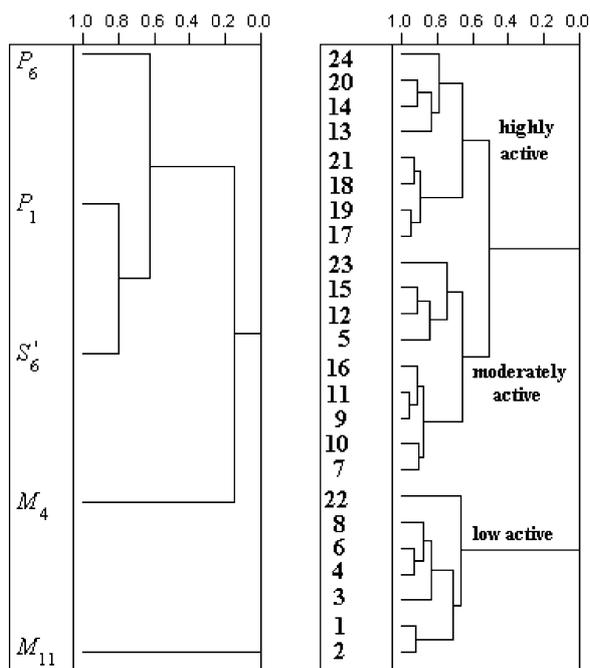
Figure 6. The results of Hierarchical Cluster Analysis for set II: a) dendogram of variables showing separation of electrotopological from molecular graphics variables; b) dendogram of samples exhibiting three distinct clusters belonging to highly, moderately and low active compounds.

3.3 PR-Progesterone Complex

3.3.1 PCA and HCA on Progesterone Hydrogens

Accessibility of progesterone hydrogens to PR amino acid residues was considered as the free space around the hydrogens suitable for substitution (Table 9). HCA (Figure 9) reveals that thirty hydrogens are grouped into a small 5-membered cluster and big cluster consisting of 12- and 13-membered sub-clusters. These three types of hydrogens are

called *central*, *free* and *occupied*, respectively (Figure 10a). The central hydrogens (colored pink in Figure 10a) are located in the central rings. Furthermore, these hydrogens are "hidden" from PR residues, but they are in the neighboring pockets so that substituting them by bigger groups could enhance PR-progesterone binding. The free hydrogens (colored yellow in Figure 10a) are located between the central rings and the molecular terminal polar groups. There is free space in the neighboring pockets, and so substitution of the hydrogens by appropriate groups would positively affect the PR-progestogen binding. The occupied hydrogens (colored blue in Figure 10a) are those that are exposed to PR residues more than any others, and are situated mostly around the progesterone terminal polar groups. In general, there is no free space in the neighboring pockets that would support the substitution of these hydrogens by any group. These observations on substitution of progesterone hydrogens were confirmed using molecular graphics of PR active site-progesterone complex with the same hydrogen classification. The activities of sets I and II confirm these findings: substitution at 21 (**4**, **8**) does not increase the activity significantly, while at 17α (most of the molecules), 11β (**V**), 18a (Et in **IV–VII**), 6 (many molecules in set II) increases the activity significantly. As the progesterone hydrogen descriptors (Table 9) describe isotropic distribution of the free space with relatively small cut-off distance (up to 5.5 Å), a special anisotropic distribution of the free space could not be detected (as for example, the free space for positioning the long pocket at 17α-H). But our results agree with the most known substitutions [7] which work well due to PR-progesterone molecular complementarity or are used as protection against reduction of terminal carboxyl groups: 6, 7α, 11β, 14α, 16α, 17α, 21. The first three PCs describe the data well (97.99% of total variance). The three groups of hydrogens from HCA can be clearly observed along PC1 (Figure 9b). Other sub-clusters from HCA are also visible using the first three PCs. Both HCA and PCA suggest additional substitutions at 7β, 8, 9, 12α, 15β, 18a and 19b.
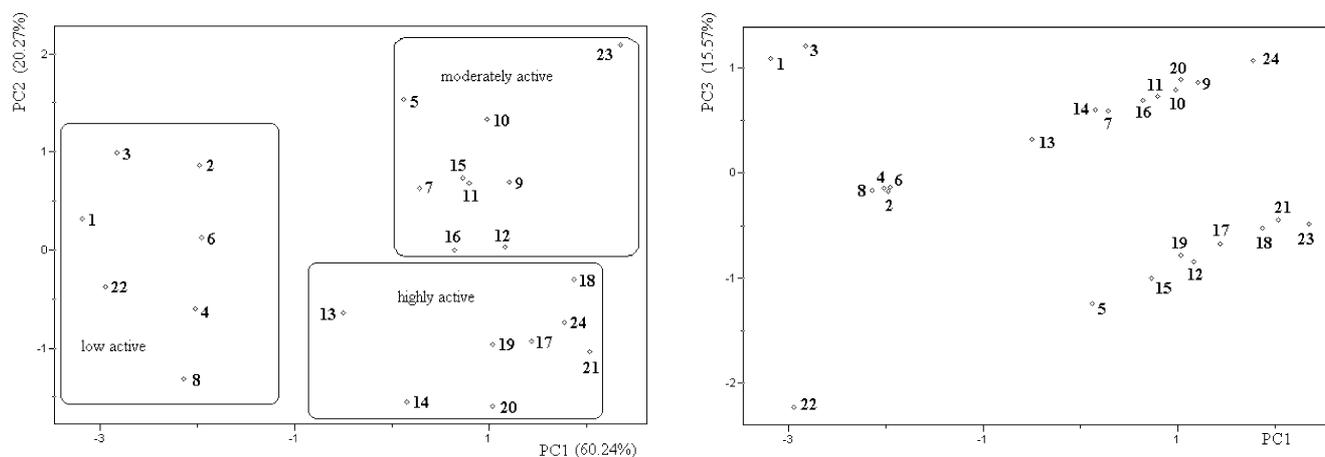

Figure 7. PCA scores plots for set II. a) PC1 *vs.* PC2 with well pronounced separation of low active from moderately and highly active progestogens; b) PC1 *vs.* PC3 data arranged in parallel lines due to descriptor M_{11}.

Table 9. Progesterone hydrogen descriptors derived from experimentally determined PR-progesterone complex.

H[a]	D_H/Å[b]	D/Å[b]	η[c]	η_H^c	H[a]	D_H/Å[b]	D/Å[b]	η[c]	η_H^c
H4	2.5	3.3	55	143	H21a	2.2	3.0	62	147
H6-β	3.1	3.5	52	139	H12-α	2.4	3.0	44	116
H6-α	2.9	3.6	53	135	H12-β	2.3	2.3	43	118
H7-β	2.7	3.4	43	94	H11-β	2.5	3.3	32	98
H7-α	2.8	3.2	37	79	H11-α	2.5	2.5	46	126
H8	3.0	3.0	29	56	H1-α	2.9	3.4	47	161
H9	3.0	3.9	30	68	H1-β	2.8	2.8	56	161
H14	3.5	4.4	32	74	H2-β	2.4	2.5	55	154
H15-β	3.1	3.6	47	120	H19a	2.4	2.6	49	146
H15-α	2.4	3.0	51	119	H18b	2.4	3.1	36	97
H16-β	2.4	3.1	50	130	H18a	2.7	3.2	29	82
H16-α	2.5	2.8	53	132	H18c	2.6	3.1	41	113
H17	2.7	3.5	43	95	H19b	2.8	3.4	43	117
H21c	2.5	3.3	62	168	H2-α	2.3	3.0	69	184
H21-β	2.2	2.2	58	139	H19c	2.7	3.5	41	110

[a] Hydrogen names as in Figure 10. [b] Minimal distance of PR atoms (D_H: including hydrogens, D: non-hydrogen atoms) from each progesterone hydrogen. [c] Sum of valence electrons of PR atoms (η_H: including hydrogens, η: non-hydrogen atoms) inside 5.5 Å cut-off distance sphere around each progesterone hydrogen.
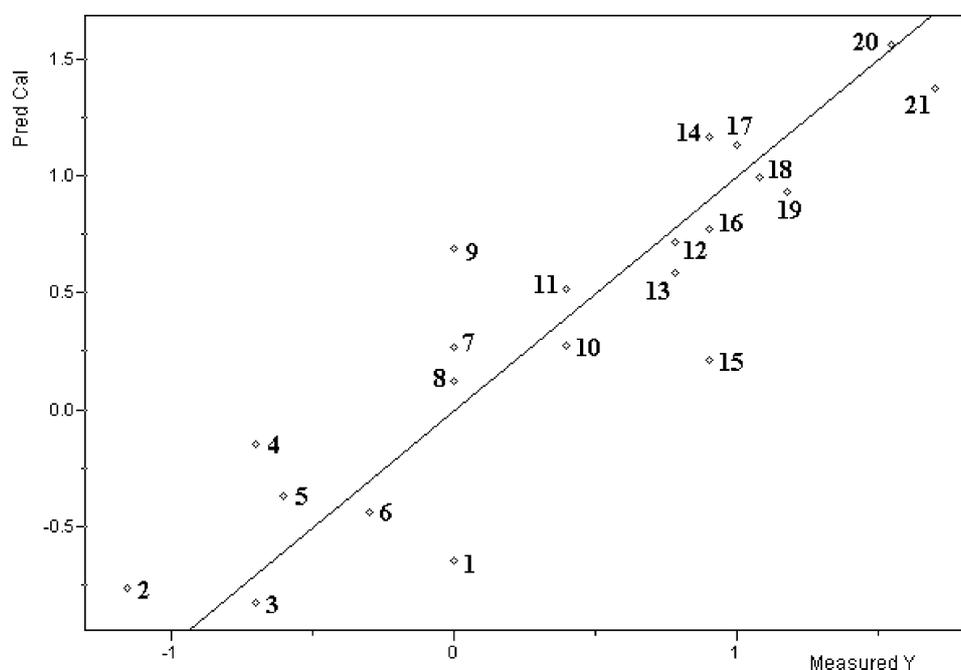

Figure 8. PLS model for set II using three Principal Components. Predicted *vs.* measured activities (ROPA) are presented.

3.3.2 Molecular Graphics of PR Pockets Around the Progesterone

Van der Waals model of the complex progesterone-PR active site hole (consisting of 20 amino-acids and a water molecule [9]) clearly shows four different pockets where the progesterone molecule is not buried by its receptor (Figures 10b–e). In Figure 10b, six progesterone atoms in the proximity of sulfur from methionine 801, (O3, H4, H6-α, H6-β, H7-α, H7-β) are exposed to the pocket and they are visible. There is free space available for substitution at these atomic positions. In a view perpendicular to the steroid rings, the two Me groups (at C10 and C13) are in an empty pocket (Figure 10c). In Figure 10d (in the view opposite to that in Figure 10c), H-17α and other central hydrogens are clearly visible. In fact, molecular graphics nicely illustrates geometrical relationships between the three types of progesterone hydrogens and PR amino-acid residues. The fourth orientation is towards the β side chain at C17, with the smallest free pocket (Figure 10e). As the pictures show, sterical complementarity, polar-polar and hydrogen bonding (involving progesterone keto groups), and hydrophobic-hydrophobic (around the steroid ring) interactions are the dominant binding forces.

a)   b)
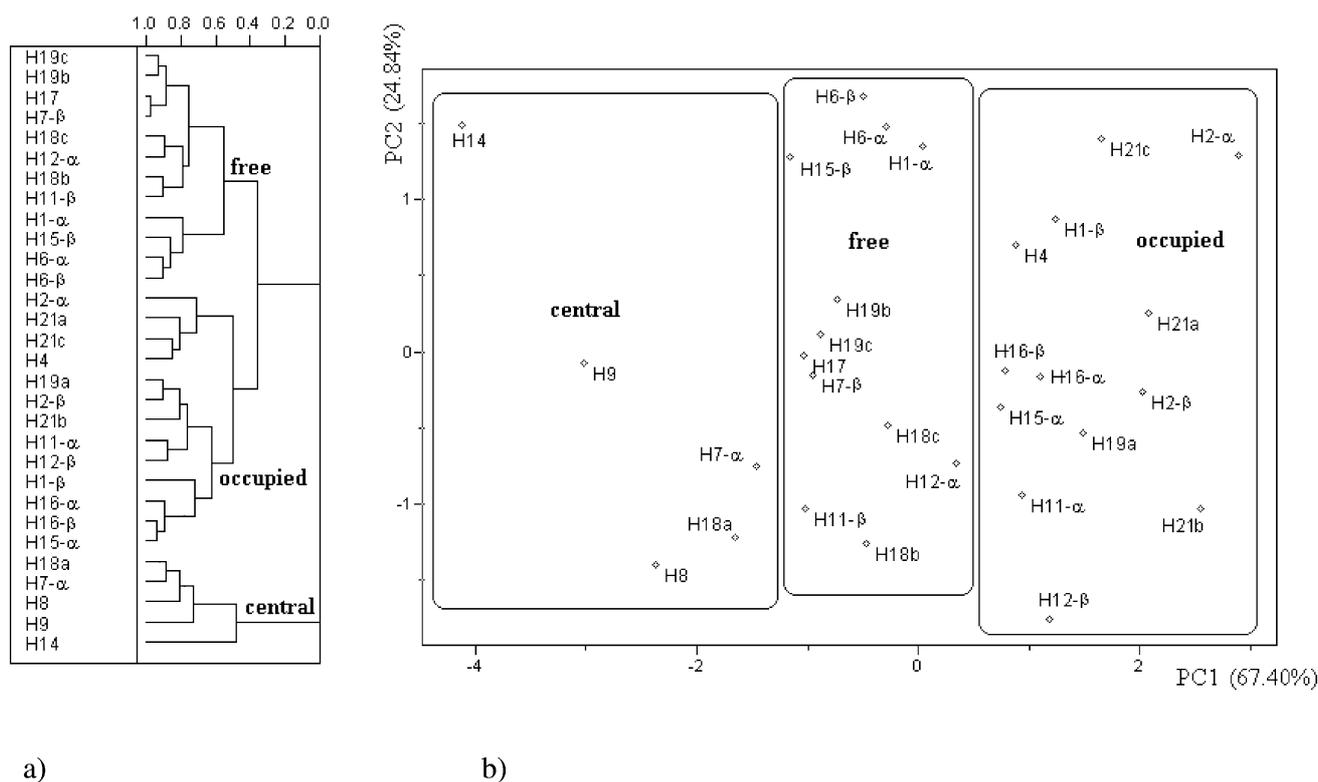
Figure 9. Results of a) HCA and b) PCA for progesterone hydrogen atoms showing the three groups of hydrogen atoms (central, free and occupied).

Sets I and II include progestogens with substituents at positions 3, 6, 11, 15 – 21. Molecular graphics shows that substituents at all these positions fit into the four pockets (even keeping the PR geometry frozen), as already discussed in the previous section (Figures 10b – e). Figure 11 illustrates a few cases of the substitution. Certain hydrogen atoms from the progesterone molecule, which are exposed to one of the four pockets (Figure 10), were substituted by an atom or a group of atoms. Cl-6α substituent is in contact with S(Met 801), occupying the central part of the pocket (Figure 11a). This may be an explanation why **11**, which has Cl-6α substituent, shows higher activity than **6** (which has Cl-6β substituent). If H6-α is substituted by a smaller (as F in **7**) or greater atom (as Br in **9**) than Cl (as in **11**), the PR-progestogen binding is destabilized. The compounds **7** and **9** show lower ROPA than **11** (Table 6). Et-6α (Figure 11b), besides making attractive contacts with S(Met 801), fills the pocket farther left to the sulfur, something that a monoatomic, methyl or hydroxyl group cannot do. Substitution of H18a by Me enhances the progestational activity; HS(Cys 891) group interacts favorably with this Me (Figure 11c). Activity of **III** – **VI** is greater than that of **I** and **II** (Table 4). In fact, Et can also be placed at C10 and C13, and Me or similar substituent (as $CH_2=$) at C11 [7]. Molecules with Et (**III** – **VII**) are more active than those without it (**I**, **II**, **VIII**). Small substituents like halogens, are acceptable at C21 (**4**, **8**) although this substitution does not increase the activity significantly (Figure 11d). There is a free space around O3,

and that is the reason why it can be substituted (like in **III**, **VII**, Figure 11e). C17-α substituents (**II** – **VII**, set I) with short chains (2 – 3 atoms, like in HC≡C−C17α) fit nicely the free space available (Figure 11f). Such substitution must result in substantial stabilization of the complex. The compounds **II** – **VII** are more active than **I** which does not have C17-α substituent. It is obvious that C17-α and C6-α (**5-7**, **9** – **21**) substitutions are common in both sets, I and II. Most of substitutions at these positions enhance the progestational activity substantially. This is in accord with rational progestogen design [7].

3.3.3 Fab′-Progesterone and PR-Metribolone Complex

Fab′ is another protein which acts as a receptor of progesterone. The crystal structure of its complex with progesterone is known [34]. Although the progesterone molecule is practically at the surface of the protein, a part of the molecule is well buried (Figure 12a). There is some free space around Me at C18. What would happen if Me is substituted with Et? This substitution could significantly enhance the Fab′-progesterone binding, as could be observed in set I. Another similarity with PR-progesterone complex is the free space around C21 (Figure 12b). Metribolone (R1881) in PR is another example how some basic properties of progestogens determine their binding affinity to receptor. R1881 has only 17β-OH and 17α-Me groups, and conjugated system of four double bonds (O3=C3,
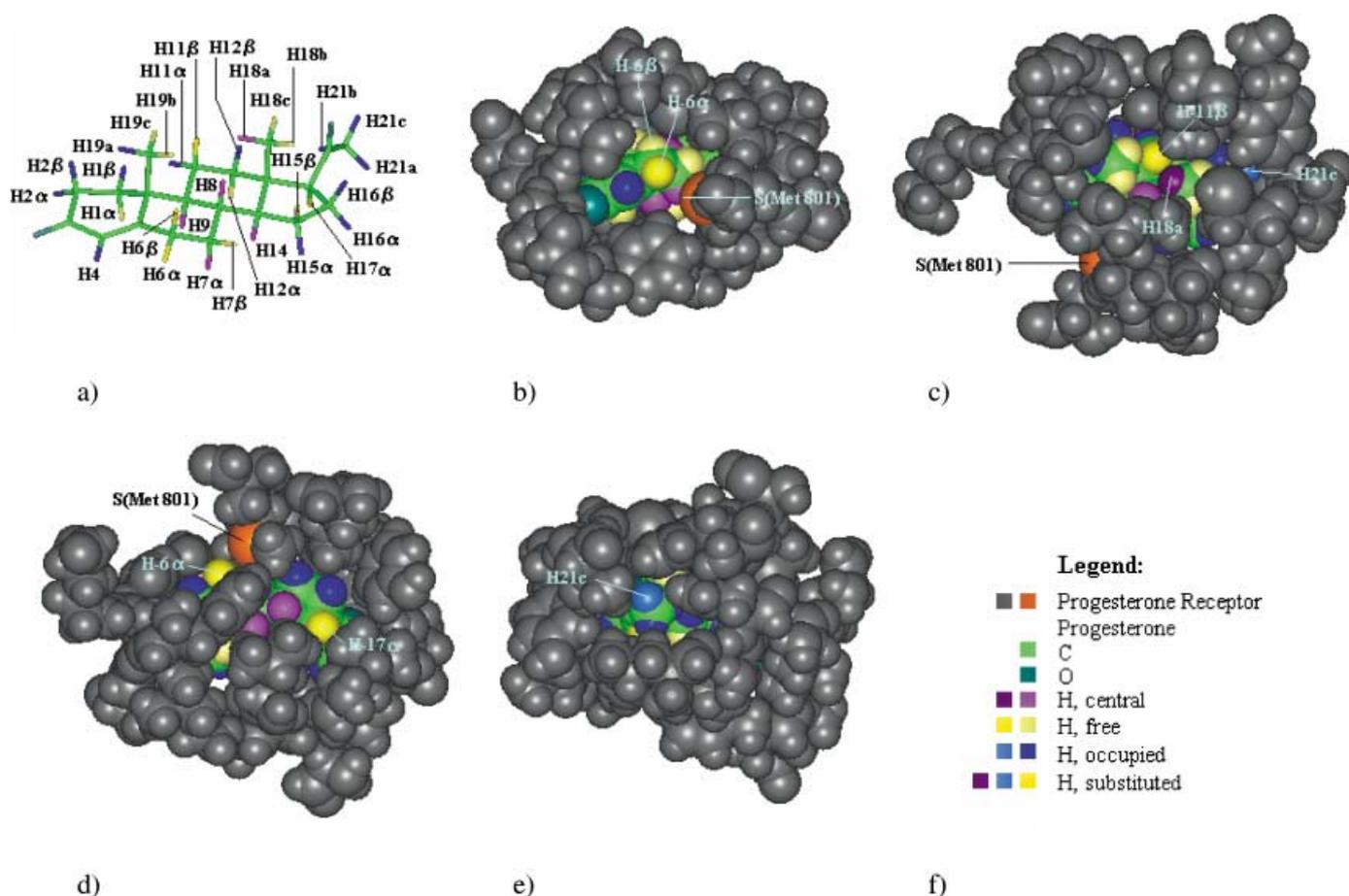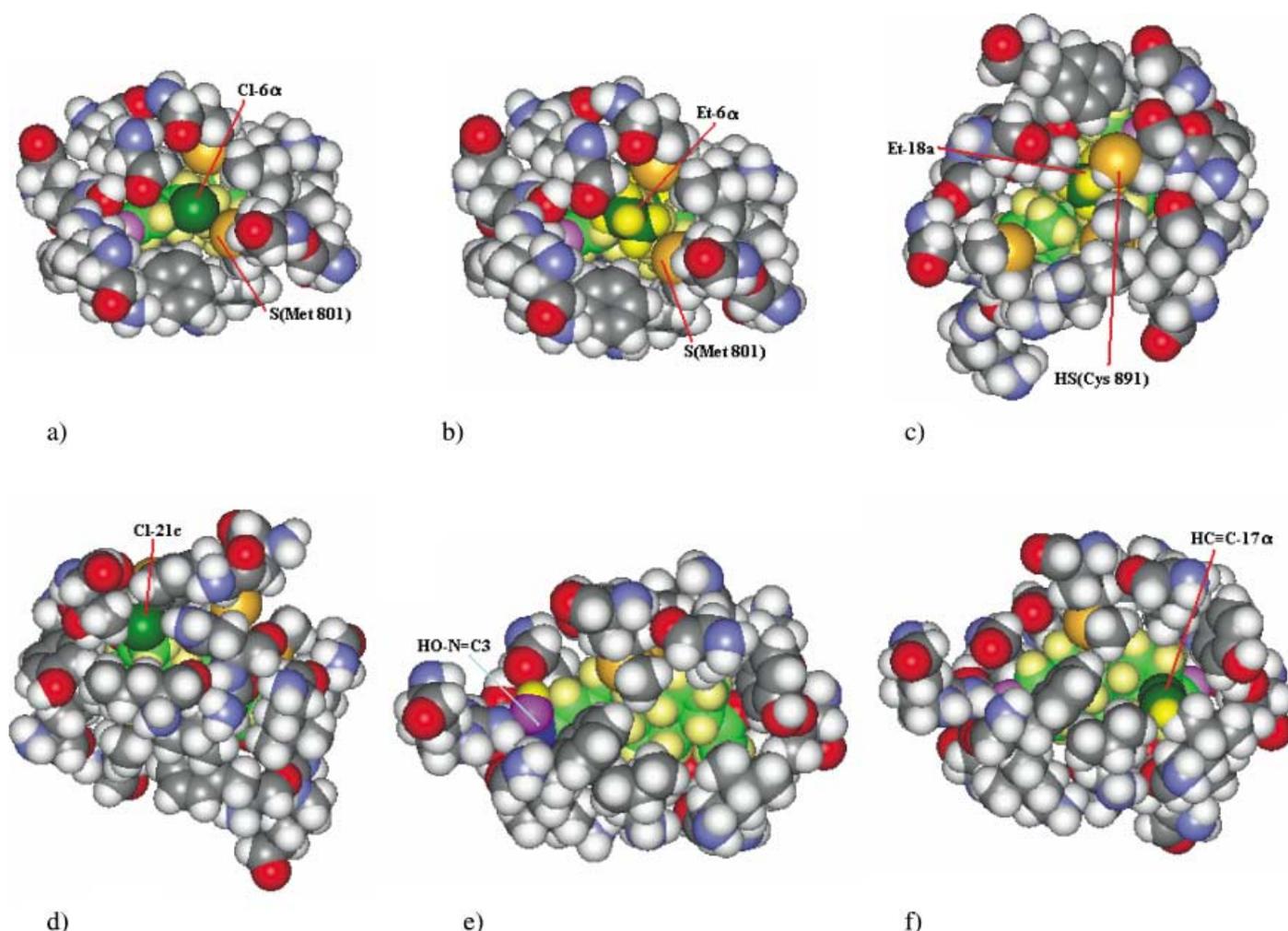
Figure 10. Molecular graphics of PR active site-progesterone complex. a) Central (pink), free (yellow) and occupied (blue) progesterone hydrogens. b) – e) The four PR pockets pronounce free space around the progesterone molecule. Hydrogens substituted in sets I and II are marked also. f) Legend. PR active site is gray except important sulfur atoms (orange). Progesterone coloring from a) is applied to other pictures. Hydrogen atoms which are found to be substituted in sets I and II are colored darker (H6-α, H11-β, H17-α, H18a) or lighter (H21c) than other hydrogens from the same group.

C4=C5, C9=C10, C11=C12, Figure 12c). R1881 strongly binds to PR [10] *via* hydrogen bonds through both keto and hydroxyl groups, and participates in numerous hydrophobic interactions. The molecule is more flat than progesterone due to the four conjugated double bonds in rings A-C. The conjugated bonds are rich in π-electrons because about 2/3 of molecular skeleton is involved in conjugation and hyperconjugation. These resonance effects were found to be important even for progesterone, determining the differences between its keto groups in hydrogen bond formation with PR [37]. Thus some molecular descriptors in this work can be called (hetero)aromaticity descriptors (m, n, k, D_{5-10}, S_6', P_1, P_6). Molecular graphics search for pockets around R1881 in PR revealed only one (Figure 12d): O3, H4, H6-α, H6-β, H7-α, H7-β are not burried completely. This is in accord with the fact that R1881 binding affinity to PR is 1.8 times greater than that for progesterone [10]. It is reconfirmed that double bonds play an important role as electronic, structural and conformational parameter.

3.3 Unifying the Molecular Graphics and Chemometric Results

Thus far, set I and set II of progestogens were studied applying both chemometrics and molecular graphics techniques separately. The two different techniques gave different aspects of the same phenomenon. Now it is possible to unite the information obtained from these two techniques. Regarding to set I, one can observe that from 10 descriptors used in chemometric analysis (Table 2), three descriptors D_{1-10}, D_{5-10}, D_{9-10} involve C10 (interatomic distances between C10 and C1, C5 and C9, respectively). This can be understood as a consequence of the fact that all the highly active molecules **II** – **VI** do not have Me at C10, while the low active **I** has. The presence or absence of Me at C10 affects sensibly those values of D_{1-10}, D_{5-10}, D_{9-10}. Molecular graphics study of PR-progesterone complex revealed that there is no large free space available that permits substitution at β-C10. The presence of Me at this position seems to prevent the optimum fit of ring A of progesterone to the

Figure 11. Molecular graphics of PR active site-progesterone complexes. The amino-acids are colored conventionally (N: blue, C: gray, H: white, O: red, S: gold) and progestogens unconventionally (N: dark blue, C: green, H: yellow, O: pink, substituent atoms are darker) for the sake of clarity.

counterpart of PR. The two descriptors L_{13} and L_{14} are electron density of LUMO at C13 and C14, and they were also employed in chemometric analysis. All the highly active molecules **II – VI** have Et at C13, while the low active **I** does not have. The magnitude of L_{13} and L_{14} must depend closely on the presence or absence of Et at C13. The molecular graphics analysis showed that the presence of the Et at C13 stabilizes PR-progesterone complex, because it fits well to the pocket on the β side of C13. Et group has just appropriate size to occupy the free pocket. In this way it can be understood the reason why some of the descriptors were selected.

It had been predicted previously that a progestin receptor site established an intimate specific contact with ring A, but far less specific contact with the reminder of steroid [38]. This situation was clearly observed when Fab′-progesterone interaction in Figures 12a and 12b was discussed. In previous work [11], it was found that the calculated atomic charge at C10 was intimately related with oral contraceptive activity

of set I. This is in accord with the present results. It has been shown that the ring A plays important role on the oral contraceptive activity of the compounds. However, the role of the substituents at other positions, especially at C13 and C17, was not clear in the previous works. The present molecular graphics analysis has given an explanation to the role of the substituents at these positions. Appropriate substituents at C13 and C17 fit well in the existing pockets (see Figures 11c, 11d and 11f) enhancing ligand-PR binding energy. This eventually increases oral contraceptive activity.

In the case of set II, the five descriptors, P_1, P_6, S'_6, M_4 and M_{11} were selected and used in chemometric analysis. The first three descriptors (P_1, P_6, S'_6) are closely related to the shape and size of substituent at C6. There is a free pocket around α-C6. The majority of compounds in set II are 6-α substituted steroids. These 6-α substituents occupy the free pocket. When a 6-α substituent has optimal size and shape, it fits nicely into the pocket increasing the biological activity.

a)   b)   c)

d)
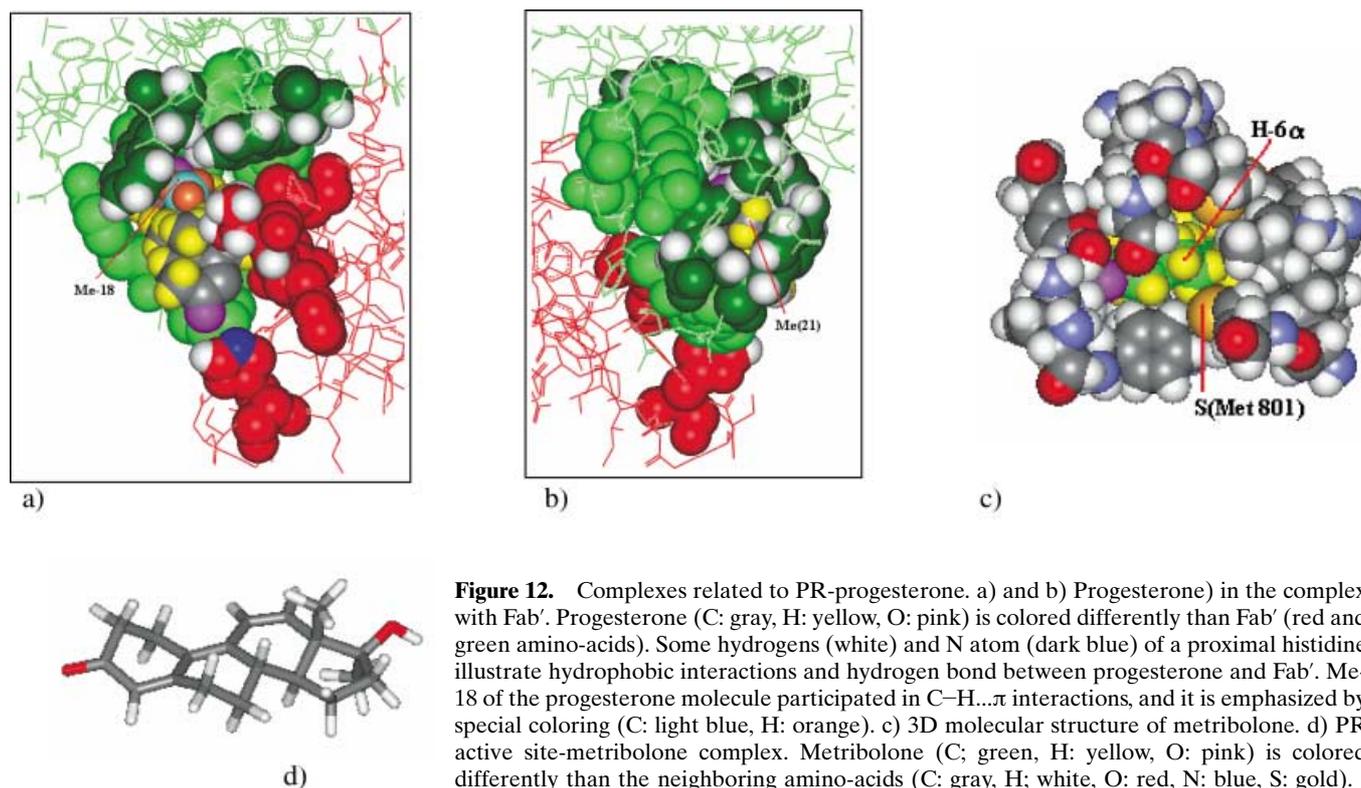
Figure 12. Complexes related to PR-progesterone. a) and b) Progesterone) in the complex with Fab′. Progesterone (C: gray, H: yellow, O: pink) is colored differently than Fab′ (red and green amino-acids). Some hydrogens (white) and N atom (dark blue) of a proximal histidine illustrate hydrophobic interactions and hydrogen bond between progesterone and Fab′. Me-18 of the progesterone molecule participated in C−H...π interactions, and it is emphasized by special coloring (C: light blue, H: orange). c) 3D molecular structure of metribolone. d) PR active site-metribolone complex. Metribolone (C; green, H: yellow, O: pink) is colored differently than the neighboring amino-acids (C: gray, H; white, O: red, N: blue, S: gold).

The three descriptors P_1, P_6, and S'_6 are intimately related with the shape and size of 6-α substituents.

Comparison between set I and set II reveals that there is no compound in set I which has a substituent at C6, while the majority of compounds in set II have such an substituent. Set I concerns oral contraceptive activity, whereas set II oral progestational activity.

Finally, it is interesting to compare the differences in PR-steroid binding for progesterone and metribolone, taking into consideration positions of amino-acid residues which are in close contact with the steroid molecule. Crystal structures of PR-progesterone and PR-metribolone complexes reveal that, due to the differences in molecular structure of progesterone and metribolone, residue Met759 shows the greatest positional variation. In the case of PR-progesterone, S from this residue interacts with Me at C10, being at distance 3.47 Å from C(Me), and far from steroidal C5, C9 and C10 (4.72, 6.05 and 5.16 Å, respectively). As a consequence, there is great PR pocket extending up to C13, capable to accept Et at this position (as has been already shown in Figs. 10c and 11c). Met759-steroid contact is quite different in PR-metribolone complex. The metribolone rings A-C act as a boat which can accept relatively large atom as S from Met759. Thus S is placed almost equally close to C5, C9 and C10 (4.11, 4.01, 3.90 Å, respectively). Also, there is no empty PR pocket around C10−C13 (see Fig. 12c and the corresponding previous discussion). Thus, Me at β-C10 sterically hinders S from Met 759 to bind the steroid rings A-C more efficiently. If Me at this position would be

replaced by H, one would expect stronger Met759-steroid interaction. Set I, as was already discussed, shows that **II**−**VII** (with H at C10) are more active than **I** and **VIII** (with Me at C10).

It is interesting to note that two methionine residues, Met801 and Met 756, play an important role in PR-steroid binding through specific $S \cdots X$ (X=C, H, O, F, Cl, Br, I) interactions, including $S \cdots C(\pi)$ interactions. Both, statistical analysis [39] and study of particular crystal structures [40] of sulfur containing proteins revealed this type of interaction as being responsible for protein and enzyme-substrate complex stabilization. In a mechanistic sense, one can have better idea on sulfur role in PR-steroid interaction, as has been already noticed in chemometric and molecular graphics analyses in this work.

Now it is possible to summarize that hybridization, substitution, conformational and other changes at C10 and its close surroundings in the steroid ring determine the strength of PR-steroid binding significantly. On the other side, these variations are well described by molecular descriptors including atom C10 and its surroundings, or even by global molecular descriptors, as was already discussed above.

4 Conclusion

Combined chemometric, molecular graphics and modeling, quantum mechanics and structural studies on the two sets of

progestogens gave more insight into the behavior of these compounds at molecular level, rather than any technique in particular. Chemometrics aided to select significant molecular descriptors that are related with biological activities. Molecular graphics studies on PR-ligand and Fab'-ligand complexes provided useful information helping to understand why and how some of the substituents are related to the activity. The information obtained with the application of molecular graphics helped to understand and interpret the nature of the molecular descriptors that were selected and used in chemometrics. Thus the two approaches are complementary. Previous findings on steric, electronic and hydrophobic PR-progesterone relationships reported in literature are confirmed. The number of electrons involved in resonance effects (heteroaromaticity: conjugation, hyperconjugation) is primarily important (both as electronic and conformational property of progestogens) for activity. Substitution effects strongly depend on substitution position (determined by both progesterone and PR structure), substituent size, conformational and electronic properties. Substituents at $C6(sp^2)$ or $C6(sp^3)$-α can enhance or weaken the binding due to sterical interactions with S(Met 801). Substituents at C10 can sterically hinder S(Met759). QSAR of progestogens can be described by three PCs in PCA. PLS and PCR regression models are satisfactorily good with 3 PCs. Progestogen activity exhibits high degree of nonlinear (probably parabolic) functional dependence on some of the molecular descriptors. Linear methods like PLS can treat progestogen QSAR well by introducing non-linear descriptors. We believe that our findings can be used in further studies directed towards progestogen application in health and veterinary science.

Acknowledgement

References

[1] Gambrell , R. D., Jr., Hormone replacement therapy and breast cancer risk, *Arch. Family Med. 5*, 341 – 348 (1996).

[2] Gambrell, R. D., Jr., Use of progestogens in postmenopausal women, *Int. J. Fertility 34*, 315 – 321 (1989).

[3] Desport, J. C., Blanc-Vincent, M. P., Gory-Delabaere, G., Bachmann, P., Beal, J., Benamouzig, R., Colomb, V., Kere, D., Melchior, J. C., Nitenberg, G., Raynard, B., Scheiner, S., and Senesse, P., Standards, Options and Recommendations (SOR) for the use of appetite stimulants in oncology, Work group, Federation of the French Cancer Centres (FNCLCC), *Bull. Cancer 87*, 315 – 328 & 490 (2000).

[4] Baziad, A., and Pache, T., Cardiovascular disease and hormone replacement therapy, *First Consensus Meeting on Menopause in the East Asian Region*, Ratnam, S. S., and Campana, A., (Eds.), Geneva, 26 – 30 May 1997, http://matweb.hcuge.ch/matweb/bookmp/

[5] Kustritz, M. V. R., Use of Supplemental Progesterone in Management of Canine Pregnancy, in: *Recent Advances in Small Animal Reproduction*, Concannon, P. W., England, G., and Verstegen, J. (Eds.), International Veterinary Information Service, Ithaca, New York, USA, http://www.ivis.org

[6] Löfstedt, R., *Progesteron & Estradiol (P & E) for the Control of Equine Estrous Cycles,* VETGATE, The UK's gateway to high quality Internet resources in Animal Health, http://www.upei.ca/ ~ lofstedt/opence/pande.html

[7] Bursi, R., and Groen, M. B., Application of (quantitative) structure-activity relationships to progestogens: from serendipity to structure-based design, *Eur. J. Med. Chem. 35*, 787 – 796 (2000).

[8] So, S.-S., van Helden, S. P., van Geerestein, J. V., and Karplus, M., Quantitative Structure-Activity Relationship Studies of Progesterone Receptor Binding Steroids, *J. Chem. Inf. Comput. Sci. 40*, 762 – 770 (2000).

[9] Williams, S. P., and Sigler, P. B., Atomic structure of progesterone complexed with its receptor, *Nature 393*, 392 – 396 (1998).

[10] Matias, P. M., Donner, P., Coelho, T., Thomaz, M., Peixoto, C., Macedo, S., Otto, N., Joschko, S., Scholz P., Wegg, A., Bäslersm S., Schäfer, M., Egner, U., and Carrondo, M. A., Structural Evidence for Ligand Specificity in the Binding Domain of the Human AB Androgen receptor, *J. Biol. Chem. 275*, 26164 – 26171 (2000).

[11] Vendrame, R., Ferreira, M. M. C., Collins, C. H., and Takahata, Y., Structure-Activity Relationship (SAR) of Contraceptive Progestogens Studied with Pattern Recognition Methods using Calculated Physicochemical parameters. *J. Mol. Graph. Mod. 20*, 345 – 358 (2002).

[12] Kiralj, R., and Ferreira, M. M. C. *A priori* descriptors in QSAR: A Case of HIV-1 Protease Inhibitors. I. The Chemometric Approach, *J. Mol. Graph. Mod. 21*, 435 – 448 (2003).

[13] Kiralj, R., and Ferreira, M. M. C., *A priori* descriptors in QSAR: A Case of HIV-1 Protease Inhibitors. II. Molecular graphics and modeling, *J. Mol. Graph. Mod.*, in press.

[14] Beebe, K. R., Pell, R. J., and Seasholtz, M. B., *Chemometrics: A Pratical Guide*, Wiley, New York 1998.

[15] Halgren, T. A., Merck molecular force field 1. Basis, form, scope, parameterization, and performance of MMFF94, *J. Comp. Chem. 17*, 490 – 519 (1996) .

[16] Stewart, J. J. P., Optimization of Parameters For Semiempirical Methods 1. Method, *J. Comput. Chem. 10*, 209 – 220 (1989).

[17] Vosko, S. H., Wilk, L., and Nusair, M., Accurate Spin-Dependent Electron Liquid Correlation Energies For Local Spin-Density Calculations – A Critical Analysis, *Can. J. Phys. 51*, 1200 – 1211 (1980).

[18] Titan, version 1, 2000. Wavefunction, Inc., Irvine, CA, USA.

[19] Rebar, R. W., and Zeserson, K., Characteristics of the new progestogens in combination oral-contraceptives, *Contraception 44*, 1 – 11 (1991) .

[20] Shoppe, C. W., *Chemistry of the steroids*, Butterworth, 2nd ed., London 1964, p. 185.

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

[22] Fukui, K., Theory of orientation and stereo selection, in: Hafner, K., Lehn, J.-M., Rees, C. W., Schleyer, P. R., Trost, B. M., and Zahradnik, R. (Eds.), *Reactivity and Structure Concepts in Organic Chemistry*, Springer-Verlag, Berlin 1975, Vol. 2, p. 39.

[23] Schuur, J., and Gasteiger, J., Infrared spectra simulation of substituted benzene derivatives on the basis of a 3D structure representation, *Anal. Chem. 69*, 2398 – 2405 (1997).

[24] Todeschini, R., and Consonni, V., Dragon, version 1.1. Milano, Italy, 2000.

[25] Kiralj, R., and Ferreira, M. M. C., Molecular Graphics-Structural and Molecular Graphics Descriptors in a QSAR Study of 17-α-Acetoxyprogesterones, *J. Braz. Chem. Soc. 14*, 20 – 26 (2003).

[26] Kiralj, R., Kojic-Prodic, B., Piantanida, I., and Zinic, M., Crystal and molecular structures of diazapyrenes and a study of π π interactions, *Acta Cryst. B55*, 55 – 69 (1999).

[27] Ferreira, M. M. C., and Kiralj, R., On Physico-Chemical Properties of PAHs, 11th Brazilian Symposium on Theoretical Chemistry, Caxamby, MG, Brazil, November 18 – 21, 2001.

[28] Spek, A. L., PLATON, A Multipurpose Crystallographic Tool, version 31000, Utrecht University, Utrecht, The Netherlands, 2000.

[29] Bondi, A., van der Waals Volumes and Radii, *J. Phys. Chem. 68*, 441 – 451 (1964) .

[30] Matlab 5.4, The MathWorks Inc., Natick, MA, USA, 2001.

[31] Pirouette 3.01, Infometrix Inc., Seattle, WA, USA, 2001.

[32] Kiralj, R., and Ferreira, M. M. C., F77 routines for study of inhibitor-enzyme interactions, Instituto de Química, Universidade Estadual de Campinas, Campinas, SP, Brazil, 2000.

[33] WebLab ViewerLite, version 4.0., Molecular Simulations Inc., San Diego, CA, USA, 2000.

[34] Arevalo, J. H., Stura, E. A., Taussig, M. J., and Wilson, I. A., Three-dimensional structure of an anti-steroid Fab' and progesterone-Fab' complex,. *J. Mol. Biol. 231*, 103 – 118 (1993).

[35] Bowers, A., Ibanez, L. C., and Ringold, H., Steroids 128, Synthesis of Halogenated Steroid Hormones – 6-α-Fluoro-17-α-Acetoxyprogesterone and some Unsaturated Analogs – A New Class of Highly Active Oral Progestational Hormones, *J. Amer. Chem. Soc. 81*, 5991 – 5993 (1959).

[36] Ringold, H. J., Batres, E., Bowers, A., Edwards, J., and Zderic, J., Steroids 127, 6-Halo Progestational Agents, *J. Amer. Chem. Soc. 81*, 3485 – 3486 (1959) .

[37] Le Quest, J.-Y., Boquet, G., Bethelot, M., and Laurence, C., Hydrogen Bonding of Progesterone: a Combined Theoretical, Spectroscopic, Thermodynamic, and Crystallographic Database Study, *J. Phys. Chem. B104*, 11816 – 11823 (2000) .

[38] Duax, W. L., Grifin, J. F., and Weeks, C. M., in: Sluyser, M. (Ed.), *Interaction of Steroid Hormone Receptors with DNA*, Ellis Horwood, Chichester 1985, p. 83.

[39] Iwaoka, M., Takemoto, S., Okada, M., and Tomoda, S., Weak Nonbonded S X (X=O, N, and S) Interactions in Proteins. Statistical and Theoretical Studies, *Bull. Chem. Soc. Jpn. 75*, 1611 – 1625 (2002).

[40] Druhan, L. J., and Swenson, R. P., Role of Methionine 56 in the Control of the Oxidation-Reduction Potentials of the Clostridium *beijerinskii* Flavodoxin: Effects of Substitutions by Aliphatic Amino Acids and Evidence for a Role of Sulfur-Flavin Interactions, *Biochemistry 37*, 9668 – 9678 (1998).