Development of a Chirality-Sensitive Flexibility Descriptor for 3+3D-QSAR

Máté Dervarics,[†] Ferenc Ötvös,[‡] and Tamás A. Martinek*[,†]

Institute of Pharmaceutical Chemistry, University of Szeged, H-6701 Szeged, POB 121, Hungary, and
Institute of Biochemistry, Biological Research Center of the Hungarian Academy of Sciences,
H-6726 Szeged, Temesvári krt 62, Hungary

Received December 29, 2005

Multidimensional QSAR methodologies can be used to predict the active conformation encoded in the conformational preferences of molecules in an active series by utilizing conformational sampling. In the 3+3D-QSAR approach, the conformational free energy loss is modeled with internal coordinate-based flexibility descriptors. While the pharmacophore point pair distance descriptors introduced earlier proved useful in the construction of QSAR models and in the prediction of important features of the active conformation, they are inherently incapable of describing the chiral arrangement of the pharmacophores. As an improvement, a chirality-sensitive flexibility (CSF) descriptor is now introduced, which is based on the distance between a pharmacophore point and a plane defined by three pharmacophore points. The performance of the CSF descriptor was tested on two active series: 37 endomorphin analogues with opiate activity and 38 PGF2α analogues with antinidatory activity. The newly devised descriptor resulted in improved QSAR models in terms of both prediction accuracy and precision of the chiral geometric features of the predicted active conformations.

INTRODUCTION

With a view to the design of more effective and less adverse drugs, 3D-QSAR methods are of crucial importance. The evolution of descriptors with spatial attributes has furnished a great wealth of direct information which can assist medicinal chemists with their decisions. Highly creditable 3D-QSAR methods have emerged from the approach of the grid-based comparative analysis of the molecular interaction field (MIF).[1−5] Unfortunately, grid-based methods are dependent on the molecular alignment: an improper alignment can easily lead to poor results. Even worse is the fact that the success of an MIF-based 3D-QSAR calculation is not generally dependent on finding the real receptor-bound conformation but rather on a highly self-consistent alignment[6] that sometimes creates merely an illusion of 3D-QSAR.[7] The efforts made to alleviate the alignment problem have resulted in the alignment-free 3D-QSAR methods (e.g., GRIND,[8] EVA,[9] MaP,[10] MoRSE,[11] MS-WHIM,[12] DiP,[13] CoMMA,[14] and IDA[15]) and in the chirality-sensitive alignment-free descriptors.[16] These techniques successfully generate indirect spatial descriptors independent of translational and orientational state of the molecules, but the assumed active conformation is still required. The bioactive series used to benchmark these algorithms lacked extensive flexibility, and hence the question of finding active conformation was not critical. Highly flexible and pharmacologically interesting molecules initiated the studies where conformational preferences were correlated with the biological activity.[17,18] In the class of multidimensional QSAR methodologies, the conformational sampling is incorporated into the algorithm. The 4D-QSAR of Hopfinger et al. utilizes a multiple alignment approach,[19−21] while the multidimensional QSAR family of Vedani et al. invokes a receptor surrogate model allowing an estimation of the effects of the conformational flexibility, the induced fit, and the desolvation upon binding.[22−24] The 3+3D-QSAR separates the effect of MIF from the free energy change encountered during the transformation of a conformational ensemble of the ligand to an active conformation on a theoretical background.[25] This approach first requires the a flexibility 3D-QSAR model building to predict the active conformation and its quantitative effect on the activity and a subsequent traditional 3D-QSAR if the full 3+3D-QSAR description is desired. The flexibility descriptors are generated by using a simplified pharmacophore-point-pair (PPP) distance representation, which offers an easy mode of interpretation with respect to the predicted active conformation. The PPP-based 3+3D-QSAR flexibility descriptors, however, have a serious blind spot: they are insensitive to the chirality. A set of descriptors utilized to predict the active conformation cannot distinguish two enantiomers, despite the fact that there may be a difference of several orders of magnitude in their biological activity. As biological systems are able to recognize the difference,[26] we have made an attempt to improve our original idea, and we introduce here a newly devised chirality-sensitive flexibility (CSF) descriptor for 3+3D-QSAR purposes. We now present a definition for the CSF descriptor, which is tested on two active series that were studied earlier in 4D- and 3+3D-QSAR works.

METHODS

The general principles of flexibility descriptor calculation in 3+3D-QSAR were described earlier,[25] and we merely give a brief summary here. The computation is based on the theorem (eq 1) that the conformational free energy loss

* Corresponding author e-mail: martinek@pharm.u-szeged.hu.
[†] University of Szeged.
[‡] Biological Research Center of the Hungarian Academy of Sciences.

10.1021/ci0505574 CCC: $33.50 © 2006 American Chemical Society
Published on Web 03/24/2006

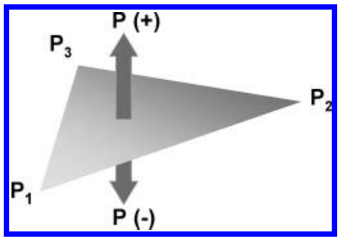


Figure 1. Schematic definition of the point-plane distance for the chirality-sensitive flexibility (CSF) descriptor.

encountered during the transformation of the conformational ensemble of a ligand into the active conformation prior to binding ($\Delta G^0{}_{\mathrm{conf}}$) is exclusively dependent on the population of the active conformation within the conformational ensemble in a target molecule-free environment (p_{a}):

$$\Delta G_{\mathrm{conf}}{}^{0} = -RT\ln p_{\mathrm{a}} \qquad (1)$$

According to eq 1, the natural logarithm of the populations of the hypothetical active conformation along the active series is proportional to the conformational free energy loss, and thus it can make a significant contribution to the free energy of receptor binding. To identify the phase cell corresponding to the active conformation, the conformational space of the ligands must be assigned in a common reference frame. This can be achieved by defining a conformational subspace on the basis of the distances between the assignable pharmacophore points. The flexibility descriptor proposed earlier is derived from the PPP distance distribution over a conformational ensemble.

For stereochemical pattern recognition, a chiral object is necessary as probe. The most obvious choice would be the dihedral angle of an ordered atom sequence. Although there are several examples of the successful description of the peptides with dihedral angles, serious difficulties can arise with the assignment of dihedral angles when there are different types of residues (e.g. α- and *β*-amino acids) or nonpeptide molecules in the examined active series. We therefore propose an ordered point quartet-based (PPQ) descriptor, where the distance is measured between the first pharmacophore point (P) and the plane defined by the three remaining pharmacophore points (P_1, P_2, P_3; Figure 1). This definition gives a distance-type descriptor (as the original PPP distance) but with positive and negative values depending on the position of P relative to the direction of the normal vector of the reference plane, thereby affording the ability to recognize stereochemical differences. The point-plane distance (d) can be calculated in a straightforward way by using the determinant in eq 2 from analytical geometry

$$\mathbf{D} = \begin{vmatrix} x - x_1 & y - y_1 & z - z_1 \\ x_2 - x_1 & y_2 - y_1 & z_2 - z_1 \\ x_3 - x_1 & y_3 - y_1 & z_3 - z_1 \end{vmatrix} \qquad (2a)$$

$$d = \frac{\mathbf{D}}{\sqrt{\mathbf{A}_{11}{}^2\mathbf{A}_{12}{}^2\mathbf{A}_{13}{}^2}} \qquad (2b)$$

where x, y, and z designate the usual Cartesian coordinates of points $P(x,y,z)$, $P_1(x_1,y_1,z_1)$, $P_2(x_2,y_2,z_2)$, and $P_3(x_3,y_3,z_3)$, and $\mathbf{A}_{ij}$ is the corresponding adjoint subdeterminant of **D**. If n pharmacophore points are assigned in the active series, $\binom{n-1}{3}$ planes can be defined for each point P, which finally leads to $n\binom{n-1}{3}$ distances. With the help of the above definition, the point-plane distance distributions can be calculated over the conformational ensemble.

The conformational sampling was performed by using a hybrid molecular dynamics − Monte Carlo (MC-MD) approach implemented in the Chemical Computing Group's Molecular Operating Environment (MOE), with which 50 000 conformations were generated for each molecule with the help of molecular mechanics force field MMFF94.[27,28] The starting geometries for endomorphins were constructed by using the Protein Builder module in MOE, and the backbone dihedrals were set to an extended conformation. For the prostaglandins, the Molecule Builder module was utilized, and all the flexible bonds in the side chains were oriented in an antiperiplanar position. The molecules were minimized before the conformational sampling procedure. The shielding effect of the aqueous solvent was taken into account by using a distance-dependent dielectric value set to $\epsilon = 4.5 * r$ (r is the interatomic distance), this setup having proved efficient in reproducing experimentally determined conformational behavior in water.[29] Every MC step was followed by four 1 fs velocity-Verlet MD steps, and the temperature was set to 300 K. The simulation comprised an initial 1 ns equilibration phase and a subsequent 5 ns sampling stage. The structures were saved every 0.1 ps. This conformational sampling protocol proved to be sufficiently independent of the starting geometry and resulted in acceptably low errors in the derived descriptors.

The binning procedure and the descriptor calculation were carried out with SVL scripts coded in our laboratory, using MOE. The SVL routine library is available from the authors upon request. The bin size Δd for the discretization of the population distribution function can be regarded as an adjustable parameter, but in this work $\Delta d = 0.1$ Å was chosen as default. The distance range for the binning was set to $-20 - +20$ Å. The correlations of the individual descriptors were analyzed via the Pearson correlation coefficient, which measures the collinearity with the biological activity. The logarithm function cannot be defined on distance bins with zero population; these were therefore excluded from the correlation analysis. The Pearson correlation coefficients were set to zero in these cases. By using the Pearson correlograms obtained for each PPQ, the descriptors in the neighborhood of the maximum positive correlation coefficients were systematically selected. No descriptor was selected for PPQs where the maximum correlation did not exceed the limit of 0.3, and the correlogram did not exhibit a single dominant positive lobe or peak. The peaks/lobes were detected by finding the largest correlation value on the curve, and the peak width was determined by following the curve in both directions until the threshold limit (0.3) was reached. If a second peak was detected outside the width of the highest peak with a local maximum correlation higher than (rejection factor)*(largest value), then the PPQ was excluded from further analysis as an uncertain descriptor. As we did not have a priori quantitative criteria for omitting the multiple peak PPQs, the QSAR models were optimized with respect to the rejection factor. The LMO q^2 values exhibited a maximum at around the rejection factors of 0.7−0.8, indicating the noise content of the PPQs with multiple comparable peaks. This protocol furnishes a considerably reduced descriptor space, which may help to avoid the overfitting and chance correlation and to remove the noisy variables. To decrease the fluctuation, the descriptors can be calculated by averaging parameters in the neighborhood of the maximum Pearson correlation PPQ distance. In this work, descriptors exhibiting a correlation higher than 90% of the maximum value were averaged, and thus the lower and upper boundaries for the distance attributes of the descriptors were assigned accordingly.

As regards the QSAR model building approaches, we utilized the partial least squares (PLS) NIPALS algorithm devised by Wold et al.[30] The QSAR models were constructed by using the code in Fortran77 by Ponder et al.[31,32] The combination of PLS with the stepwise Monte Carlo simulated annealing variable selection (MCA) algorithm was tested too.[33] The software was supplemented in our laboratory with all the additional functionalities necessary for descriptor scaling, leave-multiple-out, and scrambling validation. The MCA-PLS algorithm was compiled by the authors on the basis of the original program QSAR. The modified code is available upon request. The parameters were left unscaled. The goodness and the statistical relevance of the linear models were assessed carefully.[34] We adopted the tough validation protocol described by Baumann et al.[10] with minor modifications. It contains leave-one-out and leave-multiple-out cross-validations, where 50% (18 and 19 for endomorphins and prostaglandins, respectively) of the compounds in the original active series are randomly excluded from the training set, and the prediction is performed for the excluded prediction sets. To obtain a statistically reliable estimation for the predictive power, the cross-validation was performed with a great number of steps, always using a newly selected training and prediction set. During the model optimization stages, the leave-50%-out (L50%O) cross-validations were run with 300 cross-validation steps to keep the computational demand at an acceptable level, and the final models were evaluated for the predictive power with 3000 cross-validation steps. The protocol also involves the sampling of the probability distribution function of the L50%O q^2 for a chance correlation by calculating 3D-QSAR models with randomly scrambled target biological activities. The random biological data were generated 1000 times, and the full model optimization protocol (300 L50%O cross-validation steps) was carried out on each of them. To assess the confidence levels of the models, the L50%O q^2 values were compared with the highest and the 99% percentile values obtained from the chance correlation distribution.

For back-projection purposes, the relative descriptor importance was estimated with the product of the PLS coefficients of the final 3D-QSAR model and the standard deviation of the descriptors over the active series (coeff*stddev). The labels of the descriptors indicate the PPQ and the specific distance, and these 3D characteristics of the most important descriptors can therefore be used to filter out the structures meeting the criteria from the original conformational sampling databases. These conformers were used to represent the predicted conformation of the active state.

RESULTS AND DISCUSSION

Table 1. Training Set of the Endomorphin Analogues with Biological Activities Determined in in Vitro Competitive Inhibition Receptor Assays[a]

no.	compound	K_i,[b] nmol	ln(1/K_i) uncorrected[b]	ln(1/K_i) corrected[b]
1	H-Tyr-Pro-Phe-Phe-NH$_2$ (EM2)	9.530	−2.254	−2.254
2	H-Tyr-Pro-Phe-Trp-NH$_2$ (EM1)	4.210	−1.437	−1.437
3	H-Tyr-Pro-Phe-Pro-NH$_2$[c]	80.000	−4.382	−4.382
4	[(2S,3S)-β-Me-Phe^4]EM2[c]	1.667	−0.511	−0.511
5	[(2S,3R)-β-Me-Phe^4]EM2	69.500	−4.241	−4.241
6	[(2R,3R)-β-Me-Phe^4]EM2[c]	250.000	−5.521	−5.521
7	[(2R,3S)-β-Me-Phe^4]EM2	104.000	−4.644	−4.644
8	[D-Tic^2]EM2	166.667	−5.116	−5.116
9	[D-Pro^2]EM2	1610.000	−7.384	−7.384
10	[D-Ala^2]EM2	0.310	1.172	1.172
11	[Dmt^1]EM2	0.210	1.561	1.561
12	[Dmt^1][(2S,3S)-β-Me-Phe^4]EM2[c]	0.350	1.050	1.050
13	[Dmt^1]EM1[c]	0.014	4.269	4.269
14	[(2S,3S)-β-Me-Phe^3]EM2	45.300	−3.813	−3.813
15	[(2S,3R)-β-Me-Phe^3]EM2	106.000	−4.663	−4.663
16	[(2R,3R)-β-Me-Phe^3]EM2[c]	7090.000	−8.866	−8.866
17	[(2R,3S)-β-Me-Phe^3]EM2	4910.000	−8.499	−8.499
18	[(2S,3S)-β-Me-Phe^4]EM1[c]	0.800	0.223	0.223
19	[(2S,3R)-β-Me-Phe^4]EM1[c]	26.300	−3.270	−3.270
20	[(2R,3R)-β-Me-Phe^4]EM1	45.300	−3.813	−3.813
21	[(2R,3S)-β-Me-Phe^4]EM1	107.300	−4.676	−4.676
22	[(1R,2R)-Aic^3]EM2	298.000	−5.697	−6.508
23	[(1S,2S)-Aic^3]EM2	128.000	−4.852	−5.274
24	[(1R,2R)-Aic^2]EM2[c]	1000.000	−6.908	−8.275
25	[(1S,2S)-Aic^2]EM2	241.000	−5.485	−6.198
26	[(1R,2S)-Cip^2]EM2	674.000	−6.513	−7.699
27	[(1S,2R)-Cip^2]EM2[c]	14.600	−2.681	−2.104
28	[(1R,2S)-Cip^2]EM1	7205.000	−8.883	−11.158
29	[(1S,2R)-Cip^2]EM1	2.500	−0.916	0.472
30	[Dmt^1][(1S,2R)-Cip^2]EM2	1.200	−0.182	1.544
31	[Dmt^1][(1R,2S)-Cip^2]EM2	28.000	−3.332	−3.055
32	[Dmt^1][(1S,2R)-Achc^2]EM2[c]	2.700	−0.993	0.360
33	[Dmt^1][(1R,2S)-Achc^2]EM2	980.000	−6.888	−8.246
34	[(1S,2R)-Achc^2]EM2[c]	0.500	0.693	−0.388
35	[(1R,2S)-Achc^2]EM2	263.000	−5.572	−6.278
36	[(1S,2R)-Achc^2]EM1	1.500	−0.405	−1.421
37	[(1R,2S)-Achc^2]EM1[c]	467.000	−6.146	−6.818

[a] For the structures of the unnatural β-amino acid monomers, see Chart 1. [b] For the details of the experimental conditions and the correction protocol, see ref 25. [c] The marked compounds were included in the external test set.

The feasibility testing of the CSF descriptor was first performed on the series of endomorphin analogue tetrapeptides with μ-opiate activity (Table 1).[25,35−37] The sources of the studied biological data were the K_i values from μ-opiate competitive inhibition receptor assays, determined in three different setups: **1**−**21** were measured against Tyr-D-Ala-Gly-NMe-Phe-Gly-ol (DAMGO) as radioligand in the rat brain, **22** and **33** were measured against DAMGO as radioligand in the mouse brain, and **34**−**37** were measured against endomorphin-2 as radioligand. The K_i values are independent of the K_d values of the radioligands and are directly comparable. Since the level of structural homology between receptor molecules from different sources is very high, the biological data were scaled by a method described in ref 25 to account for the systematic differences due to the incidentally different receptor preparations and experimental conditions.

The first step was the definition of the pharmacophore points (PPs) which are important for binding to the μ-opioid receptor and characteristic to describe flexibility. The following assignment rules were defined for the active series (Chart 2): the N-terminal NH$_2$ group (**a**); the phenolic OH group in the Tyr1 residue (**b**); the C═O oxygen of the first amide group (**c**); a distant side-chain carbon atom on the second residue (**d**); the C═O oxygen of the second amide

Chart 1. Chemical Structures of the Non-natural β-Amino Acids Built into the Molecules

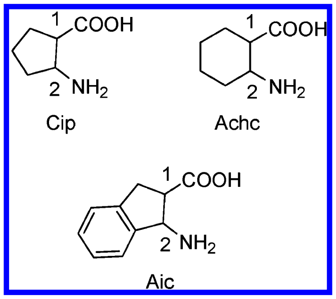
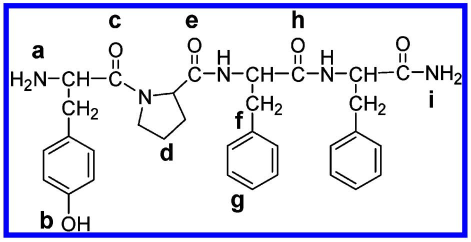


Chart 2. Applied Pharmacophore Point Assignment for the Endogenous Ligand Endomorphin-2



group (**e**); to account for the orientation of the side-chain, two atoms were selected as PPs from the aromatic group (**f** and **g**); the C=O oxygen of the third amide group (**h**); and the C-terminal NH_2 group (**i**). Experimental and modeling results to date rationalize our selection.[38−41]

After the PP assignment, the descriptors were computed for all 504 constituent PPQs in the range −20.0 to 20.0 Å with a resolution of 0.1 Å. The Pearson correlation coefficients with the biological activity for each descriptor were calculated. It is clearly seen from the diagrams that the variation in the correlation coefficients is deterministic as a function of the corresponding PPQ distance (Figure 2). In most cases, the correlograms give acceptably high correlation values. It may be concluded that the descriptors in the neighborhood of the positive maximum of the correlograms may carry exploitable information concerning the biological activities of the studied molecules, and the usage of the respective descriptors in a QSAR model is justified. For the endomorphin analogues, the descriptor prefiltering described in the Methodology resulted in 4, 26, 78, 147, and 233 parameters meeting the criteria of the multiple peak rejection factors of 0.5, 0.6, 0.7, 0.8, and 0.9, respectively.

The PLS model exhibited optimum prediction accuracy at the multiple peak rejection factor of 0.7 (Figure S1), where the values of r^2, LOO q^2, and L50%O q^2 were 0.67, 0.62, and 0.58, respectively. The input descriptors and the main results of the best PLS model are given in Table 2 and Figure 3. To test the possibility of improving the prediction accuracy, a Monte Carlo Annealing (MCA) optimization was performed in the descriptor space. The best three QSAR models extracted from the model population generated by the MCA-PLS protocol resulted in L50%O q^2 values in the range 0.62−0.65 (details in Table S1), indicating an improved prediction accuracy. The L50%O q^2 values generated on 3000 random prediction sets furnished a very rigorous assessment of the goodness of the models; in our case, they indicated that the generated linear relationships are stable in terms of the training set and prediction set selection. This can also be demonstrated by splitting the data set into a training set and an external test set. Here, we moved one-third of the data set into the external test set by utilizing a random split, because a test set balanced in the descriptor space may lead to underestimation of the true prediction error.[42] The QSAR models were generated with the components selected by the LMO optimization. The resulting predictive q^2 value was 0.73, indicating the reliability of the proposed QSAR approach in a realistic situation. The QSAR models were constructed for the uncorrected $\ln(1/K_i)$ values too, and the cross-validation results proved that the applied corrections do not affect the overall quality.

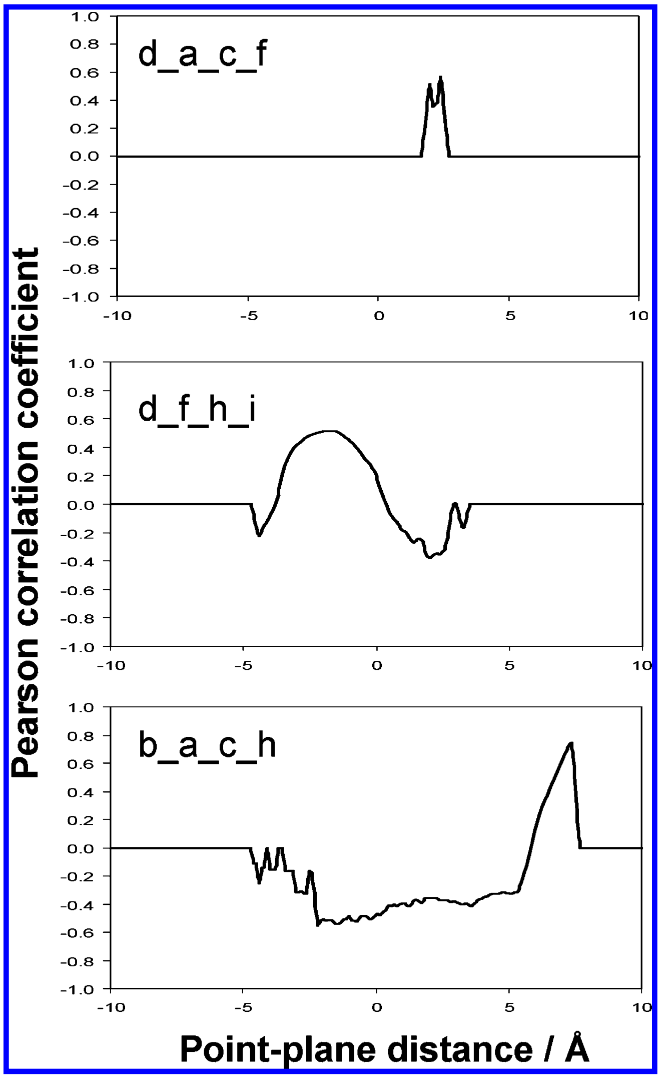


Figure 2. Selected plots of the Pearson correlations of the biological activity with the pharmacophore point-plane distances obtained for endomorphin analogues.

It must be emphasized that the descriptors afford information only on the molecular flexibility (i.e. the conformational free energy loss); it is not to be expected, therefore, that this 3D-QSAR model will explain the total variance in the biological activity. Because of the simplistic nature of the model, we must estimate the confidence at which we can state that the explained variance is statistically relevant and that a chance correlation can be ruled out. This was achieved with the random scrambling protocol described in the Methodology. The resulting L50%O q^2 distribution generated for the scrambled PLS models (Figure S2) revealed that the maximum value and 99% percentile were −2.04 and −4.32,

Table 2. Input Descriptors (Multiple Peak Rejection Factor = 0.7, Lower Limit for Pearson Correlation = 0.3) and the Results of the Best PLS Model (Rank = 1, r^2 = 0.67, LOO q^2 = 0.62, and L50%O q^2 = 0.58) for Endomorphin Analogues

PPQ	lower bound, Å	upper bound, Å	PLS coeffs	PPQ	lower bound, Å	upper bound, Å	PLS coeffs
a_b_c_d	−2.5	−2.3	0.0550	d_c_e_h	0.0	0.5	0.0883
a_b_c_h	−2.6	−2.1	0.0556	d_c_f_i	−0.5	0.8	0.0301
a_b_c_i	−1.7	−1.2	0.0466	d_e_f_g	0.6	1.0	0.0615
a_b_e_h	−1.9	−0.8	0.0521	d_e_f_h	−0.7	−0.5	0.0877
a_b_f_g	−2.7	−2.5	0.0274	d_e_f_i	−1.1	−0.9	0.0836
a_b_f_h	2.8	3.2	0.0239	d_e_g_h	−1.0	−0.6	0.0800
a_c_e_h	1.7	2.0	0.0720	d_e_g_i	−1.1	−1.0	0.0726
a_c_e_i	1.0	1.7	0.0538	d_f_h_i	−2.6	−1.1	0.0638
a_c_h_i	2.2	2.6	0.0631	d_g_h_i	−2.6	−1.0	0.0586
b_a_c_e	7.3	7.4	0.0662	e_a_b_f	1.3	1.6	0.0387
b_a_c_f	7.0	7.2	0.0754	e_a_b_h	−1.1	−0.2	0.0437
b_a_c_g	7.4	7.7	0.0694	e_a_c_h	1.8	2.3	0.0689
b_a_c_h	7.2	7.4	0.1003	e_a_c_i	1.4	2.3	0.0574
b_a_c_i	6.9	7.4	0.0971	e_b_c_f	1.7	2.3	0.0842
b_a_d_g	5.5	6.0	0.0289	e_b_g_h	−0.7	−0.3	0.0401
b_a_d_h	5.2	6.0	0.0522	e_d_f_h	0.3	1.1	0.0782
b_c_d_g	5.3	6.0	0.0492	e_d_g_h	0.9	1.1	0.0732
b_c_e_f	6.8	6.9	0.0949	f_a_c_d	−2.2	−2.2	0.0640
b_c_e_h	0.5	1.9	0.0357	f_b_c_e	−2.8	−2.3	0.0543
b_c_f_h	7.1	7.6	0.0615	f_b_d_h	3.6	3.8	0.0330
b_c_g_h	7.1	7.6	0.0611	f_b_e_g	−1.3	−1.0	0.0226
b_f_h_i	−9.3	−7.7	0.0630	f_b_h_i	3.9	4.5	0.0622
c_a_b_f	−2.1	−2.0	0.0207	g_b_d_h	5.4	6.0	0.0326
c_a_b_h	−2.7	−2.6	0.0587	g_b_d_i	5.2	5.5	0.0261
c_a_d_f	−1.0	−0.8	0.0815	g_b_h_i	6.0	6.8	0.0616
c_a_d_i	−0.8	−0.5	0.0893	h_a_b_e	0.9	2.1	0.0479
c_a_e_i	−1.8	−0.8	0.0557	h_a_d_g	0.3	1.4	0.0280
c_a_h_i	−2.3	−1.9	0.0401	h_a_e_i	−0.7	0.0	0.0331
c_b_e_f	−3.8	−3.0	0.0806	h_b_c_e	−1.1	−0.3	0.0386
c_b_e_g	−3.8	−2.6	0.0568	h_b_d_e	−3.4	−2.9	0.0513
c_b_f_h	−3.5	−2.6	0.0485	h_b_d_i	−0.8	−0.2	0.0338
c_b_f_i	−3.0	−1.9	0.0592	h_b_e_i	−1.2	−0.8	0.0291
c_b_g_h	−3.1	−2.4	0.0490	h_b_f_i	−2.5	−2.0	0.0382
c_b_g_i	−3.1	−2.1	0.0504	h_b_g_i	−2.7	−2.2	0.0327
d_a_b_g	−3.8	−3.6	0.0521	h_c_d_f	1.0	2.1	0.0282
d_a_c_f	2.0	2.4	0.1000	h_c_d_g	0.5	1.7	0.0291
d_a_e_h	−0.4	0.7	0.0793	i_a_b_c	3.1	4.9	0.0358
d_b_e_g	−0.1	0.5	0.0439	i_a_c_e	−4.8	−2.7	0.0723
d_b_f_g	2.2	3.0	0.0502				
d_b_h_i	0.2	1.3	0.0426	const			16.6544

respectively, indicating a negligible probability of a chance correlation for our final QSAR models, and the explained variation in biological activity is statistically significant. It is interesting that the probability of the chance correlation decreased strongly as compared with the value obtained for the original PPP distance-based flexibility descriptors. These results prove that the proposed CSF descriptor type furnishes a firm basis for the construction of an acceptably predictive flexibility 3D-QSAR model. In comparison wih our earlier results on the PPP distance-based descriptors, the prediction accuracy of the present models is slightly increased, suggesting that the PPQ representation gives a more comprehensive description even when direct enantiomeric pairs are not present in the active series.

The estimation of the relative importance of the selected descriptors facilitates the back-projection of the results and thereby the prediction of the active conformation. As is usual in grid-based 3D-QSAR field representations, the products of the PLS coefficients and the descriptor standard deviation over the active series were used to measure the relative importance of the descriptors (Figure S3). The three most important variables were selected as d_a_c_f(2.0−2.4 Å), c_a_d_i(−0.8 to −0.5 Å), and b_c_e_f(6.8−6.9 Å). Through use of the 3D attributes (the distance range labels) of the PLS-selected descriptors, the conformational databases were filtered for each molecule in the training set in order to identify the active conformation predicted by the 3D-QSAR model. All the molecules were able to meet the filtering conditions, demonstrating that the conformational distance restraints selected by the model can be satisfied simultaneously. The lowest energy structure of endomorphin-2 satisfying the PPQ distance requirements for the active conformation is depicted in Figure 4. The comparison with our earlier results obtained from the chirally nonsensitive PPP-based descriptors reveals that the global fold of the conformations and the position of the side chains are very similar, but the CSF descriptors predict a significantly different orientation for the second peptide bond, thereby providing a presumably refined picture of the active conformation. The active conformation predicted here is in good accordance with the results of 3D modeling of the opioid ligand−receptor complexes.[40,41]

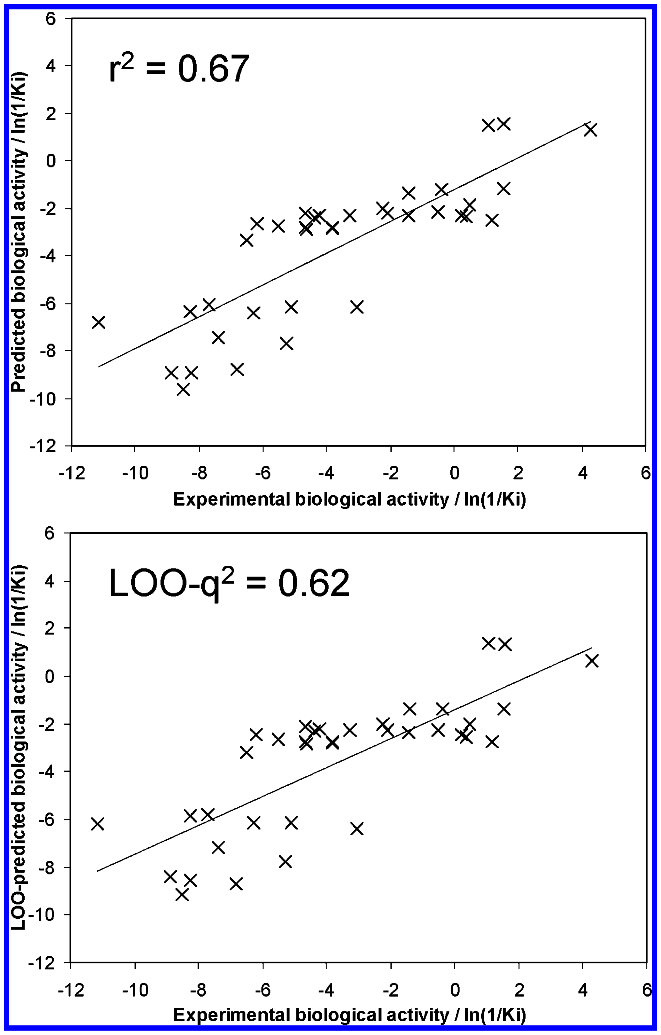


Figure 3. Predicted biological activities vs experimental biological activities calculated with the best PLS model without MCA optimization (Table 2) for endomorphin analogues.

The performance of the proposed CSF descriptors was tested on an active series comprising 38 PGF2α prostaglandins with antinidatory effects, which were also used to test the 4D-QSAR methodology in the work of Hopfinger et al.[19] and to test the 3+3D-QSAR approach[25] (Table 3 and Chart 3; see Figure S4 for the complete series). The trial set of

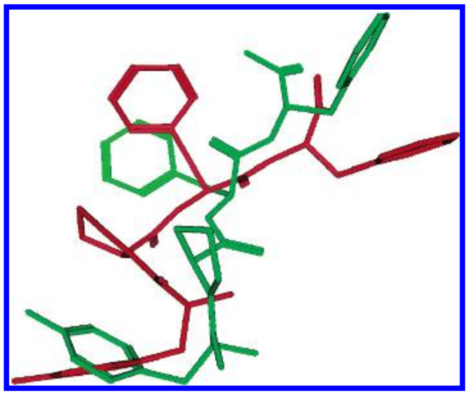


Figure 4. Overlay of the active conformations for endomorphin-2 predicted by the newly devised chirality-sensitive flexibility descriptors (green) and predicted by the pharmacophore point pair distance-based flexibility descricptors (red).

Table 3. Biological Activities[a] for the PGF2α Analogues[b]

compd	log(REL. ED_{50})	compd	log(REL. ED_{50})
1	1.699	**20**	2.481
2	0.081	**21**[c]	2.000
3[c]	2.301	**22**	2.000
4	0.279	**23**	2.000
5	0.664	**24**[c]	2.886
6[c]	2.301	**25**	2.301
7	0.669	**26**	0.000
8	1.000	**27**[c]	0.301
9[c]	0.398	**28**	1.699
10	0.398	**29**	1.699
11	1.000	**30**[c]	0.699
12[c]	0.699	**31**	0.699
13	−0.301	**32**	1.699
14	−0.602	**33**[c]	1.398
15[c]	0.699	**34**	−0.602
16	0.602	**35**	2.000
17	0.482	**36**[c]	0.000
18[c]	0.777	**37**	0.602
19	2.301	**38**	1.959

[a] Based on the relative ED_{50} values as described in ref 19. [b] For the chemical structures, see Figure S4. [c] The marked compounds were included in the external test set.

PPs was selected as displayed in Chart 3. The pharmacophore point **g** was assigned uniformly to the remote carbon on the *ω*-chain. After flexibility descriptor generation, the Pearson correlograms clearly demonstrated the same deterministic behavior as seen for the endomorphins (Figure 5). The prefiltering protocol reduced the number of descriptors to 2, 6, 18, and 37 for the rejection factor of 0.6, 0.7, 0.8, and 0.9, respectively. This reflects the fact that the studied PGF2α analogues are considerably simpler molecules, and the number of relevant geometrical factors determining the active conformation is much less than for the endomorphin analogues.

The PLS model exhibited optimum prediction accuracy at the multiple peak rejection factor of 0.8 (Figure S5), where the values of r^2, LOO q^2, and L50%O q^2 were 0.42, 0.36, and 0.33, respectively. The input descriptors and the main results of the best PLS model are given in Table 4 and Figure 6. The MCA-PLS method was applied to the input descriptors listed in Table 4, and the best model displayed a considerably higher prediction accuracy: $r^2 = 0.68$, LOO $q^2 = 0.63$, and L50%O $q^2 = 0.59$. This finding suggests that the MCA protocol is very effective in selecting the descriptors with the highest linear correlation. These results must be handled with caution, however, because the random stepwise regression approaches practically neglect the descriptor variance as a selection criterion, which

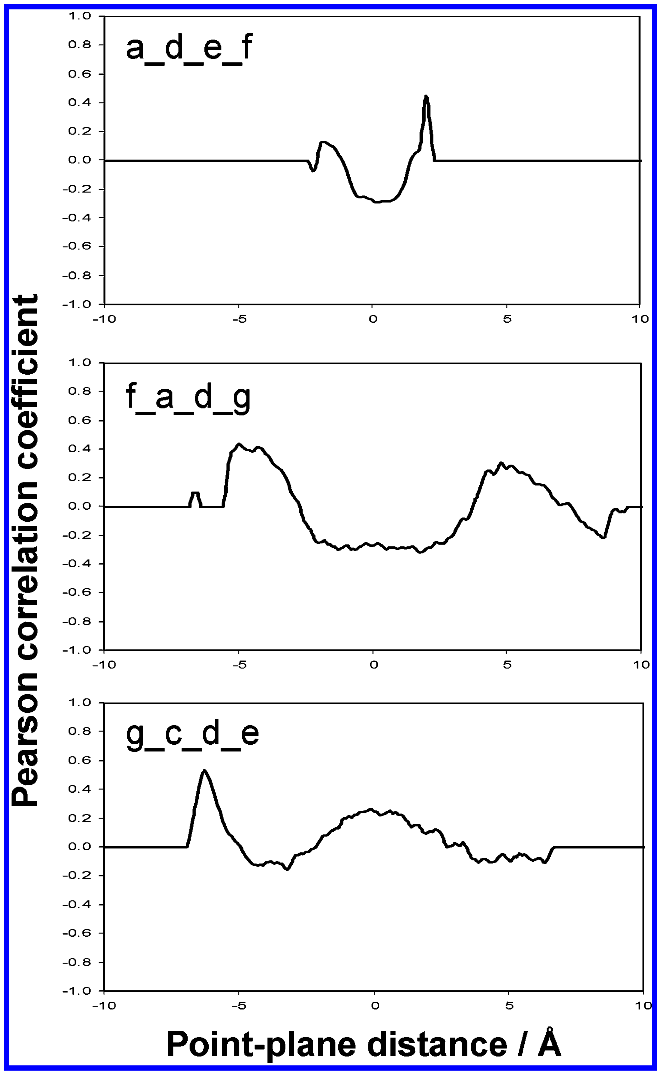
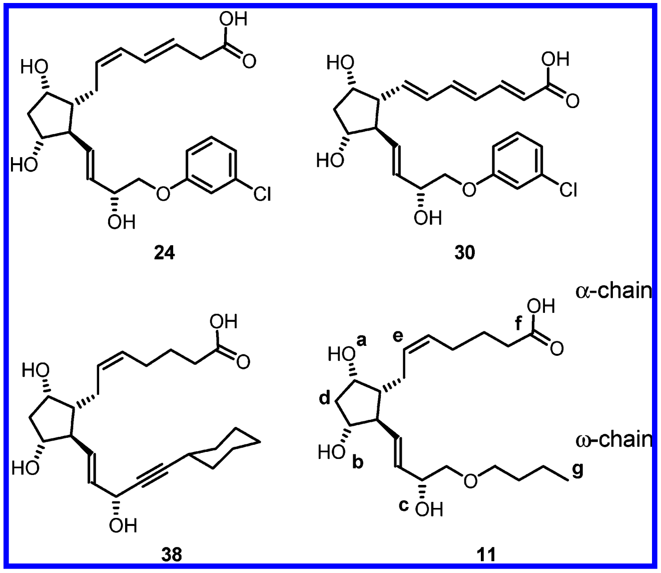


Figure 5. Selected plots of the Pearson correlations of the biological activity with the pharmacophore point-plane distances obtained for prostaglandin analogues.

Chart 3. Chemical Structures of Representative PGF2α Analogues and the Applied Pharmacophore Point Assignment



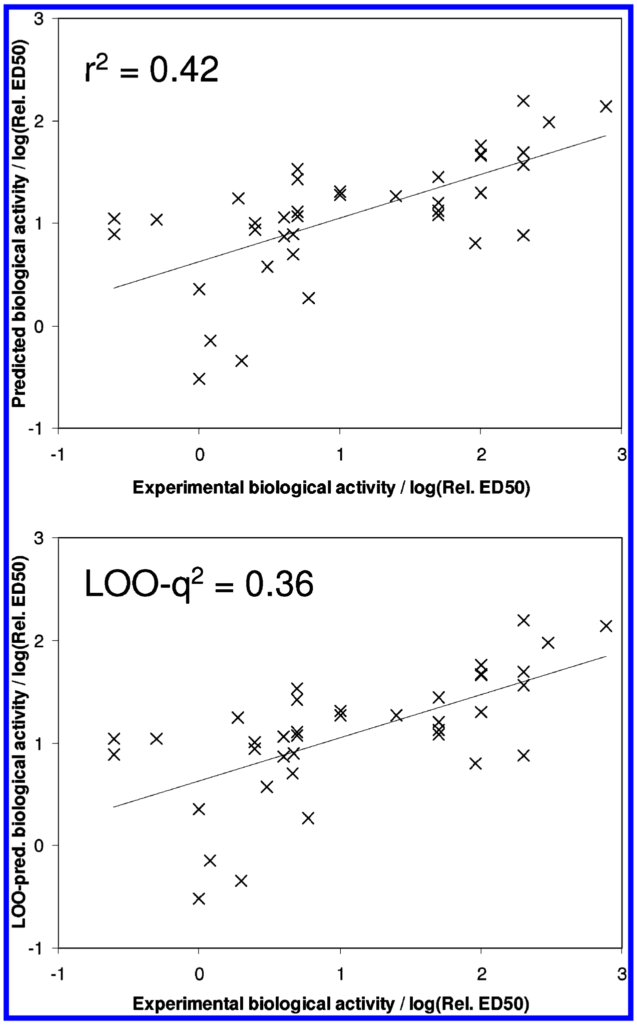


Figure 6. Predicted biological activities vs experimental biological activities calculated with the best PLS model without MCA optimization (Table 4) for prostaglandin analogues.

Table 4. Input Descriptors (Multiple Peak Rejection Factor = 0.8, Lower Limit for Pearson Correlation = 0.3) and the Results of the Best PLS Model (Rank = 1, r^2 = 0.42, LOO q^2 = 0.36, and L50%O q^2 = 0.33) for Prostaglandin Analogues

PPQ	lower bound, Å	upper bound, Å	PLS coeffs
a_b_f_g	−3.6	−3.2	0.1483
a_c_d_e	−0.1	0.0	0.0857
a_c_f_g	−7.5	−7.5	0.1469
a_d_e_f	2.0	2.0	0.1587
b_a_c_g	3.1	3.3	0.0588
b_c_e_g	−5.3	−5.1	0.0814
c_d_e_g	3.8	4.0	0.0331
c_d_f_g	3.7	4.0	0.0451
d_c_e_g	−4.9	−4.9	0.1084
e_a_d_g	1.7	2.1	0.0448
f_a_d_g	−5.1	−4.1	0.1296
f_b_c_d	−5.1	−4.9	0.0319
g_a_b_d	−10.4	−10	0.0780
g_a_d_e	−5.9	−4.9	0.0562
g_b_c_d	6.3	6.5	0.1031
g_b_c_e	6.0	6.4	0.0742
g_c_d_e	−6.4	−6.2	0.1015
g_c_e_f	−6.1	−5.5	0.0959
const			9.6273

in turn can lead to QSAR models independent of the physical background.[43]

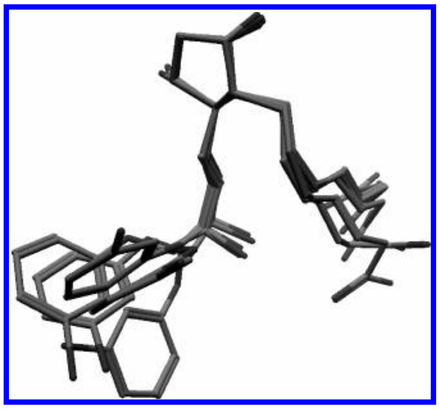


Figure 7. Active conformation for PGF2α analogue **24** predicted by the newly devised chirality-sensitive flexibility descriptors. Overlay of the five lowest energy conformations meeting the distance range criteria.

The full PLS model exhibited moderate prediction accuracy figures; the scrambling test was therefore of increased importance. The LMO q^2 distribution curve (Figure S6) indicated maximum and 99% percentile values of −0.14 and −0.98, proving the statistical significance of the QSAR model. The validation was complemented by constructing the model for the split training and external test sets (compounds marked with asterisks in Table 3) by using the same components as obtained from LMO optimization. The resulting predictive q^2 value was 0.40, which is in line with the LOO-estimated predicting power.

The relative importance of the flexibility descriptors from the PLS model was estimated as described (Figure S7), and the most relevant variables were selected as a_c_f_g(−7.5 Å), a_d_e_f(2.0 Å), and a_b_f_g(−3.6 to −3.2 Å). On use of the 3D attributes of the selected descriptors, the conformational databases for all the molecules in the active series set were filtered, and conformational hits were obtained in each case. The five lowest energy conformation meeting the distance range criteria was selected for the most active compound (**24**) and depicted in Figure 7. In contrast with the original PPP distance-based flexibility descriptors, the new CSF descriptors unambiguously determine the orientation of both the α- and ω-chains (allowing for the statistical uncertainty indicated by the validation figures).

CONCLUSIONS

The multidimensional QSAR techniques are devised to handle flexible molecules by incorporating the conformational sampling into the analysis. In the 3+3D-QSAR approach, separate flexibility descriptors are utilized, based on the internal coordinates of the ligands. In this work, a new flexibility descriptor has been proposed, which is generated by calculating the population distribution along a directed point-plane distance defined by four ordered pharmacophore points. This type of descriptor not only is sensitive to the conformational preferences in terms of internal distances but also is able to account for the chiral arrangement of the corresponding PPQ. Hence, it eliminates the drawbacks of the inherently achiral PPP distance-based descriptors. The tests performed on 37 endomorphin and 38 prostaglandin analogues revealed that the CSF descriptors provide a sufficiently predictive QSAR model to allow conclusions about the active conformation.

ACKNOWLEDGMENT

This work was supported by the Hungarian Research Foundation (TS04888) and the National Research and Development Office, Hungary (GVOP-311-2004-05-0255/3.0). T.A.M. acknowledges the award of a János Bolyai scholarship from the Hungarian Academy of Sciences.

Supporting Information Available: For the endomorphin analogues: the results of the MCA optimized PLS models; the dependence of L50%O-q^2 on the multiple peak rejection factor; the results of the scrambling test; and the relative importance of the CSF descriptors in the PLS model. For the prostaglandin analogues: the chemical structures in the active series; the dependence of L50%O-q^2 on the multiple peak rejection factor; the results of the scrambling test; and the relative importance of the CSF descriptors in the PLS model. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Goodford, P. J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28*, 849−857.

(2) Cramer, R. D., III.; Patterson, D. E.; Bunce, J. D. Comparative Molecular-Field Analysis (COMFA): Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(3) Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indexes in a Comparative-Analysis (COMSIA) of Drug Molecules to Correlate and Predict their Biological Activity. *J. Med. Chem.* **1994**, *37*, 4130.

(4) Doweyko, A. M. The Hypothetical Active-Site Lattice - an Approach to Modeling Active Sites from Data on Inhibitor Molecules. *J. Med. Chem.* **1988**, *31*, 1396−1406.

(5) Robinson, D. D.; Winn, P. J.; Lyne, P. D.; Richards W. G. Self-Organizing Molecular Field Analysis: A Tool for Structure-Activity Studies. *J. Med. Chem.* **1999**, *42*, 573−583.

(6) Cramer, R. D., III.; DePriest, S. A.; Patterson, D. E.; Hecht, P. The Developing Practice of CoMFA. In *3D-QSAR in Drug Design: Theory, Methods and Applications;* Kubinyi, H., Ed.; ESCOM: Leiden, 1993; pp 443−485.

(7) Doweyko, A. M. 3D-QSAR Illusions. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 587−596.

(8) Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. GRid-INdependent Descriptors (GRIND): A Novel Class of Alignment-Independent Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **2000**, *43*, 3233−3243.

(9) Ferguson, A. M.; Heritage, T.; Jonathon, P.; Pack, S. E.; Philips, L.; Rogan, J.; Snaith, P. J. EVA: A New Theoretically Based Molecular Descriptor for Use in QSAR/QSPR Analysis. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 143−152.

(10) Stiefl, N.; Baumann, K. Mapping Property Distributions of Molecular Surfaces: Algorithm and Evaluation of a Novel 3D Quantitative Structure−Activity Relationship Technique. *J. Med. Chem.* **2003**, *46*, 1390−1407.

(11) Schuur, J. H.; Selzer, P.; Gasteiger, J. The Coding of the Three-Dimensional Structure of Molecules by Molecular Transforms and its Application to Structure-Spectra Correlations and Studies of Biological Activity. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 334−344.

(12) Gancia, E.; Bravi, G.; Mascagni, P.; Zaliani, A. Global 3D-QSAR Methods: MS−WHIM and Autocorrelation. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 293−306.

(13) Baumann, K. Distance Profiles (DiP): A Translationally and Rotationally Invariant 3D Structure Descriptor Capturing Steric Properties of Molecules. *QSAR Comb. Sci.* **2002**, *21*, 507−519.

(14) Silverman, B. D.; Platt, D. E. Comparative Molecular Moment Analysis (CoMMA): 3D-QSAR Without Molecular Superposition. *J. Med. Chem.* **1996**, *39*, 2129−2140.

(15) Klein, C. T.; Kaiblinger, N.; Wolschann, P. Internally Defined Distances in 3D-Quantitative Structure−Activity Relationships. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 79−93.

(16) Kovatcheva, A.; Golbraikh, A.; Oloff, S.; Feng, J.; Zheng, W.; Tropsha, A. QSAR Modeling of Datasets with Enantioselective Compounds Using Chirality Sensitive Molecular Descriptors. *SAR QSAR Environ. Res.* **2005**, *16*, 93−102.

(17) Payne, J. W.; Grail, B. M.; Marshall, N. J. Molecular Recognition Templates of Peptides: Driving Force for Molecular Evolution of Peptide Transporters. *Biochem. Biophys. Res. Commun.* **2000**, *267*, 283−289.

(18) Kalászi, A.; Farkas, Ö. Lead Conformer Prediction Based on a Library of Flexible Molecules. *J.* Mol. Struct. (THEOCHEM) **2003**, *666*−667, 645−649.

(19) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B., Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509−10524.

(20) Duca, J. S.; Hopfinger, A. J. Estimation of Molecular Similarity Based on 4D-QSAR Analysis: Formalism and Validation. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 1367−1387.

(21) Santos-Filho, O. A.; Hopfinger, A. J. The 4D-QSAR Paradigm: Application to a Novel Set of Nonpeptidic HIV Protease Inhibitors. *Quant. Struct.−Act. Relat.* **2002**, *21*, 369−381.

(22) Vedani, A.; Briem, K.; Dobler, M.; Dollinger, H.; McMasters, D. R. Multiple-Conformation and Protonation-State Representation in 4D-QSAR: The Neurokinin-1 Receptor System. *J. Med. Chem.* **2000**, *43*, 4416−4427.

(23) Vedani A.; Dobler, M. 5D-QSAR: The Key for Simulating Induced Fit? *J. Med. Chem.* **2002**, *45*, 2139−2149.

(24) Vedani A.; Dobler, M.; Lill, M. A. Combining Protein Modeling and 6D-QSAR. Simulating the Binding of Structurally Diverse Ligands to the Estrogen Receptor. *J. Med. Chem.* **2005**, *48*, 3700−3703.

(25) Martinek, T. A.; Ötvös, F.; Dervarics, M.; Tóth, G.; Fülöp, F. Ligand-based Prediction of Active Conformation by 3D-QSAR Flexibility Descriptors and Their Application in 3+3D-QSAR Models. *J. Med. Chem.* **2005**, *48*, 3239−3250.

(26) Okada, Y.; Fukumizu, A.; Takahashi, M.; Shimizu, Y.; Tsuda, Y.; Yokoi, T.; Bryant, S. D.; Lazarus, L. H. Synthesis of Stereoisomeric Analogues of Endomorphin-2, H-Tyr-Pro-Phe-Phe-NH_2, and Examination of their Opioid Receptor Binding Activities and Solution Conformation. *Biochem. Biophys. Res. Commun.* **2000**, *276*, 7−11.

(27) Halgren, T. A. Merck Molecular Force Field: Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519, 520−522, 553−586, 587−615, 616−641.

(28) Halgren, T. A. MMFF VI. MMFF94s Option for Energy Minimization Studies. *J. Comput. Chem.* **1999**, *20*, 720−729.

(29) Leitgeb, B.; Ötvös, F.; Tóth, G. Conformational Analysis of Endomorphin-2 by Molecular Dynamics. *Biopolymers* **2003**, *68*, 497−511.

(30) Rannar, S.; Lindgren, F.; Geladi, P.; Wold, S. A PLS Kernel Algorithm for Data Sets with Many Variables and Fewer Objects: Theory and Algorithm *J. Chemom.* **1994**, *8*, 111−125.

(31) Helland, I. S. On the Structure of Partial Least-Squares Regression. *Commun. Stat.- Simul. Comput.* **1988**, *17*, 581−607.

(32) Fedders, M.; Ponder, J. W. *Program QSAR* **1996**. http://dasher.wustl.edu.

(33) Sutter, J. M.; Jurs, P. C. Automated Descriptor Selection for Quantitative Structure−Activity Relationships Using Generalized Simulated Annealing, *J. Chem.* Inf. Comput. Sci. **1995**, *35*, 77−84.

(34) Golbraikh, A.; Trophsa, A. Beware of q^2. *J. Mol. Graphics Modell.* **2002**, *20*, 269−276.

(35) Zadina, J. E.; Hackler, L.; Ge, L.-J.; Kastin, A. J. A Potent and Selective Endogenous Agonist for the Mu-Opiate Receptor. *Nature* **1997**, *386*, 499−502.

(36) Tóth, G.; Fülöp, F.; Péter, A.; Fábián, G.; Murányi, M.; Horváth, Gy.; Szûcs, M. New Endomorphin Analogues Using Beta-Amino Acids as Proline Mimetics in Position 2. In *Peptides 2002*; Benedetti, E., Pedone, C., Eds.; Edizione Ziino: Naples, 2002; pp 630−631.

(37) Tóth, G.; Keresztes, A.; Tömböly, Cs.; Péter, A.; Fülöp, F.; Tourwé, D.; Navratilova, D.; Varga, É.; Roeske, W. R.; Yamamura, H. J.; Szûcs, M.; Borsodi, A. New Endomorphin Analogs with Mu-Agonist and Delta-Antagonist Properties. *Pure Appl. Chem.* **2004**, *76*, 951−957.

(38) Pogozheva, I. D.; Lomize, A. L.; Mosberg, H. I. Opioid Receptor Three-Dimensional Structures from Distance Geometry Calculations with Hydrogen Bonding Constraints. *Biophys J.* **1998**, *75*, 612−634.

(39) Law, P. Y.; Loh, H. H. Regulation of Opioid Receptor Activities. *J. Pharmacol. Exp. Ther.* **1999**, *289*, 607−624.

(40) Mosberg, H. I.; Fowler, C. B. Development and Validation of Opioid Ligand−Receptor Interaction Models: The Structural Basis of Mu vs. Delta Selectivity. *J. Pept. Res.* **2002**, *60*, 329−335.

(41) Zhang, Y.; Sham, Y. Y.; Rajamani, R.; Gao, J.; Portoghese, P. S. Homology Modeling and Molecular Dynamics Simulations of the Mu Opioid Receptor in a Membrane-Aqueous System. *ChemBioChem* **2005**, *6*, 853−859.

(42) Roecker, E. B. Prediction Error And Its Estimation For Subset Selected Models. *Technometrics* **1991**, *33*, 459−468.

(43) Baumann, K. Chance Correlation in Variable Subset Regression: Influence of the Objective Function, the Selection Mechanism, the Ensemble Averaging. *QSAR Comb. Sci.* **2005**, *24*, 1033−1046.

CI0505574