QSAR Modeling Using Chirality Descriptors Derived from Molecular Topology

Alexander Golbraikh and Alexander Tropsha*

Laboratory for Molecular Modeling, Division of Medicinal Chemistry and Natural Products, School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-7360

Received March 22, 2002

Topological descriptors of chemical structures (such as molecular connectivity indices) are widely used in Quantitative Structure−Activity Relationships (QSAR) studies. Unfortunately, these descriptors lack the ability to discriminate between stereoisomers, which limits their application in QSAR. To circumvent this problem, we recently introduced chirality descriptors derived from molecular graphs and applied them in QSAR studies of ecdysteroids (Golbraikh A.; Bonchev, D.; Tropsha, A. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 147−158). In this paper, we extend our earlier work by applying chirality descriptors to four data sets containing chiral compounds. All models were derived with the *k*-nearest neighbors (kNN) QSAR method developed in our laboratory (Zheng, W.; Tropsha, A. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185−194). They were validated using the same training and test sets that were employed in various, mostly 3D-QSAR, investigations published by other authors. We show that for all data sets 2D-QSAR models that use a combination of chirality descriptors with conventional (chirality insensitive) topological descriptors afford better or similar predictive ability as compared to models generated with 3D-QSAR approaches. The results presented in this paper reassure that 2D-QSAR modeling provides a powerful alternative to 3D-QSAR.

1. INTRODUCTION

It is well-known that many biological molecules are asymmetric (or chiral), i.e., that they contain atoms in one of the two possible spatial configurations, which are mirror images of each other. Amino acids, carbohydrates, and lipids as well as many other natural and artificial receptor ligands are chiral.[2,3] Many biochemical processes and phenomena are stereospecific. For instance, L− and D−enantiomers of amino acids have different tastes,[3,4] enantiomers of some compounds have different odors,[5,6] many medicinal preparations have physiological properties different from those of their enantiomers,[7−9] many insecticides are stereospecific,[10] etc. In 1999, the worldwide annual sales of chiral drugs exceeded $100 billion that constituted almost one-third of all drug sales,[11] and in 2000, these numbers were $133 billion and 40%, respectively.[12] If a drug candidate is a racemate, the regulations of Food & Drug Administration (FDA) require a detailed study of both enantiomers.[11]

To establish the mathematical relations between biological activities of molecules and their physicochemical properties, quantitative structure−activity relationship (QSAR) methods are used. Due to stereospecificity of biological effects, QSAR methods must be capable of taking into account atomic chiralities. Indeed, one of the most popular 3D-QSAR methods, Comparative Molecular Field Analysis (CoMFA), developed in mid-1980s[13,14] and other CoMFA-like methods (some of them are discussed in refs 16−19) take into account chirality by default, since molecular fields of chiral isomers are different. However, as was outlined in our previous paper,[19] CoMFA and many other 3D-QSAR methods have several shortcomings. In many cases, it is impossible to precisely define a pharmacophore model. If all molecules in a data set are flexible, their unique structural alignment is impossible. If a spatial structure of a binding site is known, docking procedures such as DOCK,[20] AutoDock,[21,22] FlexiDock,[23,24] FlexX[26] and FlexE[27] can provide the alignment of molecules, which could be used in 3D-QSAR. However, docking algorithms are very time-consuming and not accurate enough to rank the binding energies of molecules.[27,28] This can lead to nonoptimal alignment of ligands and eventually can introduce errors in QSAR analysis. On the other hand, QSAR methods based on descriptors calculated using molecular graphs, i.e., structural formulas of compounds (2D-QSAR methods), provide an appealing alternative to 3D-QSAR since the former methods are alignment-free, faster and easier to implement in an automated fashion and are typically characterized by the same or better statistics.[29,30]

Despite their computational efficiency, until recently, due to the lack of adequate descriptors 2D-QSAR methods were unable to take chirality into account. The first attempt to add chirality descriptors to conventional 2D (Molconn-Z[31]) descriptors for QSAR studies was reported in 1998.[32] 2D-QSAR models that took into account chiral atoms were built for ligands of D_2 dopamine and σ receptors.[32] The 2D QSAR models appeared to have better statistics and predictive ability than those based on conventional nonchiral descriptors only.[32] Using chirality descriptors, a series of chiral barbiturates were correctly classified as sedatives or stimulants.[32]

In our previous paper,[19] several series of novel chirality molecular descriptors were introduced and their mathematical properties were considered. Most of the new descriptors are modified conventional (nonchiral) topological descriptors, which can be calculated using vertex degrees of the hydrogen-suppressed molecular graphs. In this report, we discuss the application of the chirality descriptors to QSAR

* Corresponding author phone: (919)966-2955; fax: (919)966-6919; e-mail: tropsha@email.unc.edu.

10.1021/ci025516b CCC: $25.00
Published on Web 12/05/2002

studies of several data sets of compounds. One of these data sets, a series of ecdysteroid analogues of 20-hydroxyecdysone (20E), the steroid hormone responsible for onset and regulation of molting in almost all arthropods,[33] was used in independent CoMFA studies[34] and was already considered in our previous publications.[19,35] Here we analyze this data set in more detail. Additional data sets include a "QSAR benchmark"[36] data set of 31 Cramer's steroids,[13,18,37−39] a set of 66 histamine H_1 receptor ligands,[40] and a set of 49 HIV-1 protease inhibitors.[15,23] In all of these studies chirality descriptors were used in combination with conventional (chirality insensitive) topological descriptors. We demonstrate that our models are characterized by similar or better statistics and predictive power as compared with CoMFA and/or other 3D-QSAR models reported in the literature for the same data sets. The success of our modeling studies expands the range of applicability of 2D-QSAR modeling and reassures that this approach provides a powerful alternative to more popular 3D-QSAR methods.

2. METHODS

2.1. Chirality Descriptors. We give a rather brief definition of chirality descriptors used in this work (cf. ref 19 for a complete discussion). These descriptors include modified overall Zagreb indices,[19,42,43] molecular connectivity indices,[43−45] extended connectivity indices,[46] and overall connectivity indices.[47,48] All of the indices make use of so-called chirality correction, which can be a real or imaginary number added to or subtracted from vertex degrees of a hydrogen-depleted molecular graph corresponding to atoms in R- and S-configurations, respectively. For example, conventional index ${}^1\chi$ is defined as ${}^1\chi = \sum_{\text{All edges ij}} (a_i a_j)^{-0.5}$, where a_i and a_j are vertex degrees of adjacent atoms i and j. Chirality index ${}^1\chi^{chir}$ is defined as ${}^1\chi^{chir} = \sum_{\text{All edges ij}} [(a_i \pm c_i)(a_j \pm c_j)]^{-0.5}$, where c_i is the chirality correction for atom i. The plus sign is used, if atom is in R-configuration, and the minus sign is used, if atom is in S-configuration. For achiral atoms, chirality correction is zero.

The chirality correction can be a real or imaginary number. Chirality descriptors based on the real chirality correction are real numbers. These descriptors are referred to as class I of chirality descriptors. Chirality descriptors based on the imaginary chirality correction are referred to as class II of descriptors. Four subclasses of class II of chirality descriptors, which are complex numbers, were defined earlier.[19] Subclass IIa descriptors are equal to these complex numbers. Subclass IIb descriptors are real and imaginary parts of these complex numbers. Subclass IIc descriptors are defined according to the following formula: $D_{IIc} = Re(d) + Im(d)$, where D is a subclass IIc descriptor, d is the corresponding complex number. Subclass IId descriptors are defined according to the formula: $D_{IId} = \arctan(Re(d), Im(d))$, where arctan is arctangent dependent on signs of both arguments, and thus it can be defined in segment $[-\pi, \pi)$. In this work, all subclasses of chirality descriptors have been used, except for subclass IIa, since QSAR software available is not adapted to complex descriptors.

2.2. Conventional Descriptors. Chirality descriptors do not take into account atom and bond types. To overcome this limitation, the chirality descriptors were combined with conventional topological descriptors such as overall Zagreb indices,[19,41,42] molecular connectivity indices,[43−45] extended connectivity indices[46] and overall connectivity indices,[47,48] and descriptors obtained using Molconn-Z[31] program. All descriptors were normalized by range-scaling, so that all normalized descriptors had values within the interval [0,1]. The total volume V occupied by the representative points in the normalized descriptor space is equal to one.

2.3. Training and Test Set Compounds Selection. Descriptors defined in the previous two sections have been used for the development of QSAR models for examples described in the Introduction (refs 15, 23, 34, 37−40). To compare predictive ability of the models developed herein with respective original QSAR models for the same data sets,[15,23,34,37−40] the same training and test sets of compounds as in the original reports were used.

For two examples, additional QSAR models were developed with other training and test sets. The following sphere-exclusion algorithm similar to that described in refs 50 and 51 for the selection of a representative subset of compounds from the whole data set was used to divide a set of N compounds into training and test sets.[51]

1. A total volume V occupied by the representative points of compounds in the descriptor space was estimated as in ref 52. (In this paper, V = 1; see the previous section). A volume corresponding to one representative point is equal to V/N = 1/N.

2. Select a compound with the highest activity.

3. Include this compound into the training set.

4. Construct a sphere with the center in the representative point of this compound with radius $R = c(V/N)^{1/K}$. Here, K denotes the number of descriptors (dimensionality of the descriptor space), and c is the dissimilarity level.[52] (Dissimilarity level was varied to construct more examples with different counts of compounds in training and test sets.)

5. Include compounds, corresponding to representative points within this sphere, except for the center of it, in the test set.

6. Exclude all points within this sphere from the initial set of compounds.

7. Let n be the number of remaining compounds. If n = 0, go to step 11; otherwise go to step 8.

8. Let m be the number of spheres already constructed. Calculate the distances d_{ij}, i = 1,...,n, j = 1,...,m of the representative points of the compounds left to the sphere centers.

9. Select a compound with the smallest d_{ij}.

10. Go to step 3.

11. Stop.

This algorithm allows constructing training sets covering all descriptor space areas occupied by representative points. The higher the dissimilarity level c is, the smaller the training set is and the larger the test set is. It is expected that the predictive ability of QSAR models generally decreases when the dissimilarity level increases.

2.4. k-Nearest Neighbors QSAR. K-nearest neighbors (*k*NN) QSAR method[30] uses leave-one-out (LOO) cross-validation procedure and an evolutionary simulated-annealing algorithm for descriptor selection. The procedure starts with the random selection of a predefined number of descriptors out of all descriptors. Activities of compounds excluded in LOO procedure are estimated using the following formula

$$\hat{y} = \frac{\sum_{\text{nearest neihbors}} y_i \exp(-d_i)}{\sum_{\text{nearest neighbors}} \exp(-d_i)} \quad (1)$$

where d_i are the distances between nearest neighbors and this compound. After each run, cross-validated R^2 (q^2) is calculated

$$q^2 = \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad (2)$$

where y_i, $\hat{y}_i$ and $\bar{y}$ are the actual, predicted and mean values of activity. The summation in (2) is performed over all compounds. After each run, a predefined number M of descriptors are randomly replaced by other descriptors from the original pool, and the new value of q^2 is obtained. If q^2(new) = q^2(old), the new set of descriptors is accepted with probability p = exp(q^2(new) − q^2(old))/T and rejected with probability (1-p), where T is a simulated "temperature" annealing parameter. During this process, T is gradually decreasing from T_{max} to T_{min}:$T_{next} = d \times T_{curr}$, where d is a parameter of the algorithm. Thus, q^2 is optimized (see ref 30 for additional details). In the prediction process, the final set of descriptors selected is used, and expression (1) is applied to predict activities of the test set compounds.

In all calculations in this work, the maximum number of nearest neighbors was equal to five, $T_{max} = 100$, $T_{min} = 10^{-9}$, $d = 0.9$, and M = 3. The number of selected descriptors was varied from 10 to 40 with step two.

2.5. Estimation of the Predictive Ability of a Model. We have shown earlier that a QSAR model with high predictive power must satisfy the following three conditions.[54]

1. High value of cross-validated R^2(q^2).

2. Correlation coefficient R between the predicted and observed activities of compounds from an external test set close to one. At least one (but better both) of the correlation coefficients for regressions through the origin (in the literature they are also referred to as coefficients of determination[54]) comparing predicted versus observed activities, (or observed versus predicted activities) R_0^2 (or $R_0'^2$) close to R^2.

3. At least one slope k or k' of regression lines through the origin close to one. (It will correspond to R_0^2 or $R_0'^2$ that is more similar to R^2).

To find the boundary significance level α, the following equation must be solved

$$F_{1,n-2,\alpha} = F \quad (3)$$

where F is the F-ratio,$F_{1,n-2,\alpha}$ is the F-distribution function with one and n-2 degrees of freedom. The higher α, the better is the model. α values were obtained using MATLAB[56] fcdf function. Frequently, p-values (defined as p-value = 1 − α-value) are used instead of α-values.

Additionally, our models were tested for robustness. This test implies comparison of models built with real data with models built with randomized activities (i.e., with shuffled target properties). The original model is considered acceptable if (i) it has significantly higher q^2 value than models developed with randomized activities or (ii) if random models have no predictive ability. Sometimes the first of the above conditions cannot be satisfied, particularly in the case of structural redundancy of the training set, or if the total number of descriptors is higher than or comparable with the number of compounds.[56]

2.6. Calculations. Descriptor Sets. For each example, the calculations were performed with the following values of chirality correction (see Section 2.1): 0.5, 1.0, 1.5, 2.0, and 2.5. The following subsets of chirality descriptors were used to develop QSAR models (see Section 2.1).

1. Only class I descriptors.
2. Only subclass IIb descriptors.
3. Only subclass IIc descriptors.
4. Only subclass IId descriptors.
5. Class I with subclass IIb descriptors.
6. Class I with subclass IIc descriptors.
7. Class I with subclass IId descriptors.
8. Class I and class II descriptors except for subclass IIa descriptors.

All descriptors with zero variance and duplicate descriptors were excluded.

Statistical Characteristics of Models. Using the *k*NN-QSAR procedure, for each of the eight subsets of descriptors and for each of the 16 sizes of descriptor sets (see Section 2.4), 10 QSAR models were built. Thus, for each example for each chirality correction value, the total number of models built was 16 × 10 × 8 = 1280. Models with $q^2 > 0.5$ were selected for further validation using external test sets of compounds. The models were considered statistically significant if they satisfied the following criteria (see section 2.5): $q^2 > 0.5$; $R^2 > 0.6$; R_0^2 or $R_0'^2$ close to R^2, i.e., $(R^2 - R_0^2)/R^2 < 0.1$ or $(R^2 - R_0'^2)/R^2 < 0.1$; k or k' lie within the interval [0.85, 1.15]. We sought models that satisfied these conditions for each chirality correction value. The models with the highest correlation coefficient R were considered as best. All satisfactory models were also characterized by the F-ratio and the α- or p-values.

For comparison, we also generated additional models using nonchiral descriptors only. In this case, for each data set and each set of descriptors 16 × 10 = 160 models were built.

Division into Training and Test Sets. The procedure for dividing a data set into training and test sets[51] (Section 2.3) was repeated with different dissimilarity levels. The starting dissimilarity level was equal to 0.5 and then increased with step 0.5. Consequently, the number of compounds in subsequent training sets was gradually decreasing, and the number of compounds in the corresponding test sets was gradually increasing. It is expected that as training sets become smaller the predictive ability of QSAR models must generally decrease. The minimum dissimilarity level was defined such that the test set would contain at least five compounds. The maximum dissimilarity level was defined such that the test set would contain at least half of all compounds yet no good QSAR model would be obtained.

RS configurations for chiral atoms were assigned using a SYBYL Programming Language (SPL) script. Following the input data preparation, the process of QSAR model development was completely automated. Calculation time depended on the number of compounds in a data set. For the largest data set (78 ecdysteroids) all calculations for the same

Table 1. Ecdysteroids: Training and Test Sets Contained 71 and 7 Compounds, Respectively[e]

model	chir. corr.	chir. descr	q^2	R	R^2	$R_0^{2\,a}$	k^a	RMSE[b]	F
1	0.5	I	0.62	0.93	0.86	0.85	0.89	0.28	31.7
2	0.5	I,IIb	0.57	0.94	0.89	0.82	0.89	0.16	38.8
3	0.5	I,IIb,IIc,IId	0.66	0.94	0.88	0.88	0.89	0.28	38.0
4	0.5	I,IIc	0.59	0.93	0.86	0.85	0.93	0.27	29.5
5	0.5	I,IId	0.66	0.91	0.82	0.82	0.87	0.43	23.5
6	0.5	IIb	0.71	0.97	0.94	0.88	0.93	0.09	72.8
7	0.5	IIc	0.69	0.88	0.78	0.71	0.92	0.24	17.2
8	0.5	IId	0.66	0.92	0.84	0.84	0.87	0.34	27.1
9	1.0	I	0.62	0.92	0.84	0.84	0.87	0.36	27.0
10	1.0	I,IIb	0.61	0.88	0.77	0.77	1.07	0.73	16.7
11	1.0	I,IIb,IIc,IId	0.74	0.85	0.73	0.72	1.13	0.90	13.5
12	1.0	I,IIc	0.65	0.90	0.80	0.80	0.90	0.48	20.1
13	1.0	I,IId	0.62	0.92	0.85	0.85	0.89	0.31	29.1
14	1.0	IIb	0.72	0.90	0.82	0.82	0.90	0.40	22.3
15	1.0	IIc	0.67	0.87	0.76	0.76	1.09	0.88	15.8
16	1.0	IId	0.60	0.90	0.81	0.81	0.92	0.34	21.5
17	1.5	I	0.62	0.91	0.83	0.82	0.88	0.35	23.6
18	1.5	I,IIb	0.63	0.89	0.79	0.76	0.91	0.30	19.3
19	1.5	I,IIb,IIc,IId	0.53	0.89	0.78	0.78	0.88	0.40	18.2
20	1.5	I,IIc	0.56	0.96	0.92	0.92	0.85	0.20	59.1
21	1.5	I,IId	0.62	0.87	0.75	0.74	0.87	0.42	15.3
22	1.5	IIb	0.58	0.89	0.79	0.78	0.88	0.37	19.0
23	1.5	IIc	0.64	0.89	0.79	0.79	0.87	0.53	18.4
24	1.5	IId	0.60	0.93	0.86	0.79	0.90	0.17	31.6
25	2.0	I	0.58	0.88	0.78	0.73	0.89	0.28	17.5
26	2.0	I,IIb	0.68	0.97	0.93	0.93	0.85	0.19	67.6
27	2.0	I,IIb,IIc,IId	0.71	0.98	0.96	0.95	0.85	0.11	110.5
28	2.0	I,IIc	0.75	0.90	0.81	0.81	1.14	0.56	21.8
29	2.0	I,IId	0.61	0.94	0.89	0.89	0.87	0.31	41.2
30	2.0	Iib	0.64	0.94	0.89	0.86	0.85	0.20	40.4
31	2.0	IIc	0.62	0.90	0.82	0.82	0.88	0.43	22.0
32	2.0	IId	0.59	0.87	0.76	0.76	0.89	0.44	16.0
33	2.5	I	0.60	0.86	0.74	0.73	1.13	0.82	14.2
34	2.5	I,IIb	0.59	0.92	0.84	0.84	0.89	0.44	26.9
35	2.5	I,IIb,IIc,IId	0.64	0.97	0.94	0.93	0.87	0.14	79.5
36	2.5	I,IIc	0.60	0.91	0.83	0.79	0.88	0.27	24.0
37	2.5	I,IId	0.62	0.88	0.77	0.76	0.86	0.42	17.1
38	2.5	Iib	0.66	0.97	0.95	0.94	0.86	0.14	88.6
39	2.5	IIc	0.59	0.94	0.88	0.84	0.86	0.22	34.9
40	2.5	IId	0.56	0.94	0.89	0.87	0.94	0.19	38.5
41	Molconn-Z descriptors[c]		0.74	0.98	0.96	0.92	0.93	0.06	133.6
42	Model A (ref 34)[d]		0.63	0.46(0.72)	0.21(0.52)	0.12(0.45)	0.99(1.06)	1.08(0.79)	1.36(4.39)
43	Model B (ref 34)[d]		0.69	0.34(0.62)	0.12(0.39)	−0.01(0.32)	0.96(1.03)	1.08(0.86)	0.67(2.55)

[a] The highest of R_0^2 vs $R_0'^2$ is given; the corresponding value of k or k' is given in the next column. [b] RMSE − residual mean square of error of prediction. [c] 13 compounds from σ-diastereomeric pairs have been excluded from the training set. [d] The values in parentheses are obtained after excluding one compound from the test set as an outlier.[34] [e] For each value of chirality correction and each descriptor subset kNN QSAR model with the highest predictive power is included. In the last two rows, statistics of CoMFA models from ref 34 is presented.

training and test sets as in ref 34 took two days on a Pentium III 500 MHz PC.

3. RESULTS AND DISCUSSION

3.1. 78 Ecdysteroids. Preliminary QSAR studies of this data set using chirality descriptors were reported elsewhere.[19,35] This data set was selected as a good first example to implement our chirality descriptors in QSAR studies[19,35] for the following reasons: (1) all molecules in this data set contain asymmetric atoms; (2) the data set contains 19 pairs of σ-diastereomers (most of them have only one atom in opposite spatial configurations); (3) results of bioassay studies of these compounds binding to ecdyson receptor are available;[34] (4) two CoMFA models based on 71 of these compounds were constructed to understand the binding of these compounds to ecdyson receptor;[34] and (5) these models were used to predict ED_{50} values for an external test set of seven compounds.[34]

Initially, QSAR models were developed using the same training and test sets as in ref 34. For each value of chirality correction and each subset of descriptors many predictive models were found, which satisfied all conditions considered in Sections 2.5 and 2.6. As many as 703 highly predictive models were generated; building one predictive model required on average slightly more than four minutes of computational time. Models with the highest predictive power for each value of chirality correction and each subset of descriptors are included in Table 1. All these models also have high α values: the lowest α value is 0.9856 (Model 11). It means that all our models (H_1 hypotheses) can be accepted with probabilities higher than 98.5%. The model with the highest predictive ability (Model 27) has $\alpha = 0.9999$. Figure 1a shows the plot of estimated vs observed ED_{50} values for the training set. The plot of predicted vs observed ED_{50} values for the test set (Figure 1b) demonstrates high predictive ability of this model. Note that the chirality

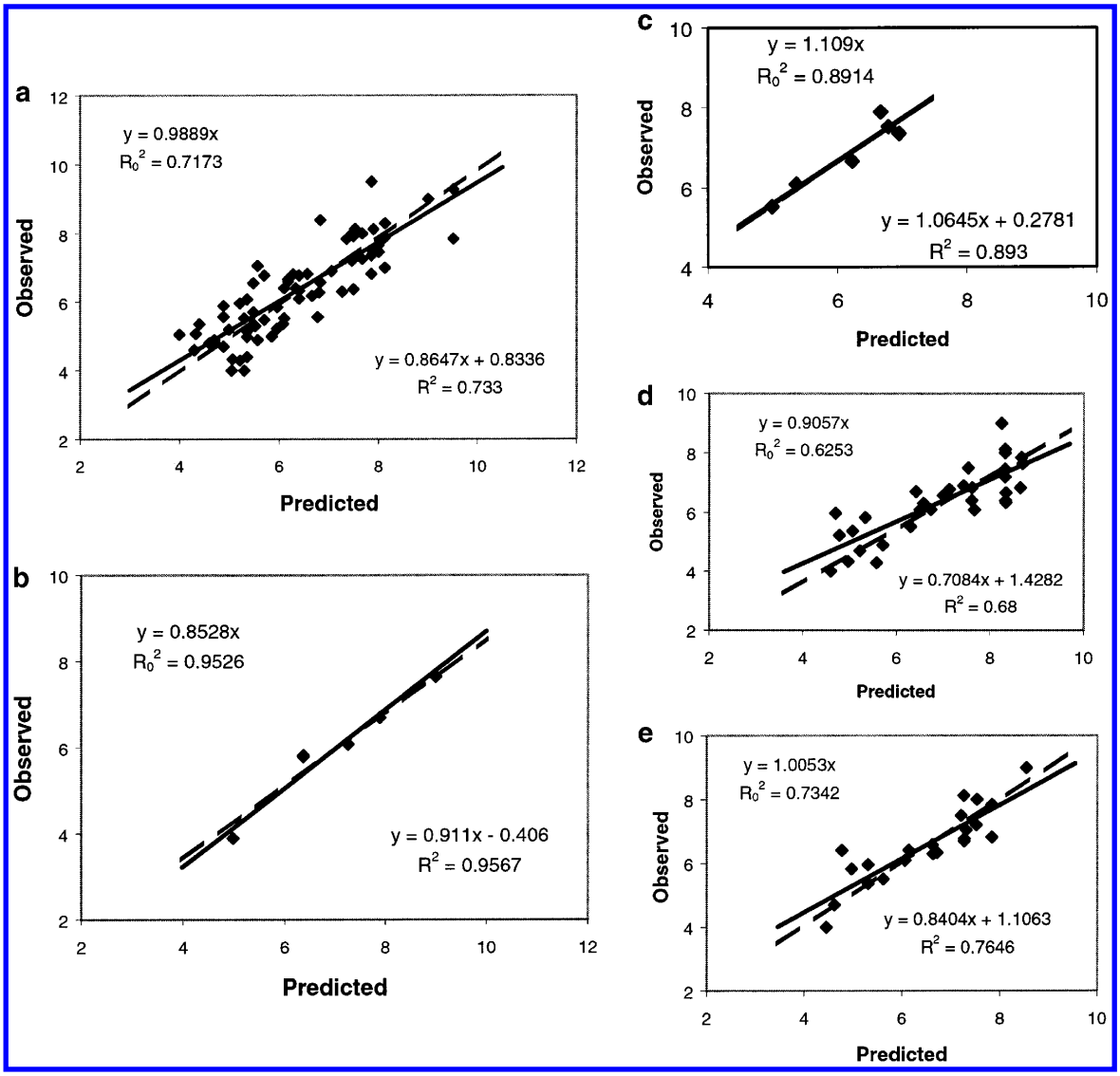

Figure 1. (a) Ecdysteroids. Model 27, Table 1. Estimated vs predicted -log(ED_{50}) values for the training set consisting of 71 compounds (training set was the same as in ref 34). (b) Ecdysteroids. Model 27, Table 1. Observed vs predicted -log(ED_{50}) values for the external test set consisting of 7 compounds (both training and test sets were the same as in ref 34). **(**c) Ecdysteroids. A model built for 72 compounds is able to distinguish between σ-diastereomers. Observed vs predicted -log(ED_{50}) values for the external test set consisting of 6 compounds (three pairs of σ-diastereomers). **(**d) Ecdysteroids. Observed vs predicted -log(ED_{50}) values for the external test set consisting of 32 compounds. (The model was built using the training set consisting of 46 compounds.) Training and test sets were built using the sphere-exclusion algorithm (Section 2.3). **(**e) Ecdysteroids. Observed vs predicted -log(ED_{50}) values for the external test set consisting of 22 compounds. (The model was built using the training set consisting of 56 compounds.) Training and test sets were built using the sphere-exclusion algorithm (Section 2.3).

correction of 2.0 was used to obtain this model. For comparison with our best models, the corresponding statistics for two CoMFA models A and B from ref 34 are also included in Table 1 (see Models 42 and 43). The latter models also have low α values of 0.7039 and 0.5497 for Models A and B, respectively. After deleting one outlier, the corresponding α values are 0.8958 and 0.8145, respectively. It means that even after deleting the outlier, these models cannot be accepted even with significance level 10%.

Additional calculations were performed using MolconnZ[31] descriptors only. The same *k*NN-QSAR procedure was applied. For these calculations, only more active σ-diastereomers from each pair have been retained in the training set. Thus, 13 compounds had to be excluded from the training set. Interestingly, one model was found to have a higher predictive power than the best models obtained with chirality descriptors (see Table 1). However, this model was apparently unable to distinguish between σ-diastereomeric compounds. For three pairs of σ-diastereomers with only one chiral atom in the opposite configurations, the differences between ED_{50} values were particularly large. In Table 2, both estimated (Model 27, Table 1) and observed ED_{50} values for these compounds are presented.

To demonstrate that QSAR models developed with chirality descriptors are capable of predicting the activities of σ-diastereomers in the test set, additional calculations were performed. For these calculations, the test set included six compounds (three pairs of σ-diastereomers, Table 2), and the training set consisted of the remaining 72 compounds. All five chirality correction values and eight sets of descriptors (see Section 2.6) were used in these calculations. For 20 out of these 40 cases, models satisfying conditions of Sections 2.5 and 2.6 were found. As in Table 1, the models generated with different subsets of descriptors even using

Table 2. Ecdysteroids: ED_{50} Values for Three Pairs of σ-Diastereomers Estimated in Cross-Validation Procedure (Model 27, Table 1) and Observed Experimentally

	no. of the compd	$-\log(ED_{50})$		no. of the compd	$-\log(ED_{50})$	
		estimated	obsd		estimated	obsd
1	2	7.85	7.35	38	5.30	5.52
2	18	6.18	6.66	19	8.12	7.89
3	66	7.85	7.52	71	5.36	6.08

single chirality correction value had different predictive ability. Models based on the most diverse set of descriptors, i.e., including all descriptor subsets, were not necessarily the best ones. Predictive power of the best model ($q^2 = 0.74$, $R^2 = 0.89$, $R_0^2 = 0.89$, $k = 1.11$, $\alpha = 0.9955$) is demonstrated in Figure 1c. This model was based on class I and subclass IIc descriptors calculated with chirality correction 2.0. Interestingly, the best model in Table 1 (Model 27) was also built with the chirality correction equal to two. However, this is only a coincidence. Indeed, we can find several other models in Table 1 almost as good as Model 27. Some of them are built using descriptors calculated with other values of chirality correction. Thus, for Model 6, chirality correction was 0.5, for Model 20 it was 1.5 and for Models 35 and 38 it was 2.5. The same is true for the models built for the training set containing 72 compounds. One model built with descriptors calculated with the chirality correction equal to one was almost as good as the best one obtained for this training set ($q^2 = 0.76$, $R^2 = 0.87$, $R_0^2 = 0.86$, $k = 1.00$, $F = 27.6$, and $\alpha = 0.9937$).

QSAR models built for training and test sets obtained using the sphere-exclusion algorithm described in Sections 2.3 and 2.6 also show that it is impossible to select a priori the best value of the chirality correction. Calculations were performed for all eight selections of descriptors considered above and for all five chirality correction values. The models for each value of the chirality correction, each descriptor set, and each training and test sets were tested as to whether they satisfied the conditions considered in Sections 2.5 and 2.6. We found, for instance, that for smaller training sets (and, therefore, larger test sets) models built using descriptors with the chirality correction 0.5 have better predictive ability than those built using descriptors with the chirality correction 2.0. For chirality correction 0.5, a predictive model was built using the training set containing 46 compounds (the test set included 32 compounds) ($q^2 = 0.61$, $R^2 = 0.68$, $R_0^2 = 0.67$, $k = 1.09$, $F = 63.8$ and α-value $= 1 - 6.4722 \times 10^{-9}$). Figure 1d demonstrates high predictive ability of this model. Two predictive models were built using training set containing 48 molecules ($q^2 = 0.69$, $R^2 = 0.65$, $R_0^2 = 0.63$, $k = 0.91$, $F = 51.5$ and $\alpha = 1 - 8.2466 \times 10^{-8}$ and $q^2 = 0.54$, $R^2 = 0.72$, $R_0^2 = 0.72$, $k = 0.92$, $F = 71.4$ and $\alpha = 1 - 3.4484 \times 10^{-9}$). At the same time, for chirality correction 2.0 no one model with the training set including less than 55 compounds appeared to have high predictive ability. Figure 1e demonstrates high predictive ability of another model built with chirality correction 0.5 but for a larger training set (56 compounds) and smaller test set (22 compounds) ($q^2 = 0.67$, $R^2 = 0.77$, $R_0^2 = 0.76$, $k = 0.99$, $F = 64.9$ and $\alpha = 1 - 1 - 1.0463 \times 10^{-7}$). Thus, chirality correction is an empirical parameter, and its best value depends on each particular case (descriptors used and compounds in the training and test sets). As we shall see, this conclusion is also true for other data sets. Therefore, to obtain a QSAR model with the highest predictive ability, extensive calculations with different selections of chirality correction and subsets of chirality descriptors must be performed.

In summary, for this example, using chirality descriptors we were able to build 2D-QSAR models that could distinguish between σ-diastereomers differing by the configuration of only one atom and rather accurately predict their biological activities. We showed also that our best models have much better predictive ability than those obtained with CoMFA.[34] Another advantage of 2D-QSAR vs 3D-QSAR is the time necessary to obtain a good model. As we mentioned above, on average one good model was generated in slightly more than four minutes on a PC. Preparation of input data for the development of 2D-QSAR models is also very fast. The limiting step is building chemical structural models, which can be done with a number of commercial software packages. Calculation of descriptors takes several minutes. Obtaining chiralities of all chiral atoms in molecules with Sybyl takes less than a minute. In contrast with 3D-QSAR no time-consuming conformational analysis, alignment or docking procedure is necessary. These results encouraged us to employ chirality descriptors in QSAR analysis of other series of compounds and compare our results with those obtained using other QSAR methods.

3.2. 31 Cramer's Steroids. This data set was introduced by Cramer et al. in 1988[13] and since then has become a benchmark for the assessment of novel QSAR methods.[36,39] All compounds in this data set contain chiral atoms, and binding affinities of these compounds to corticosteroid binding globulin are available.[13] Different methods were used to develop 3D-QSAR models for this data set, including CoMFA,[13,18] Comparative Molecular Similarity Indices Analysis (CoMSIA),[18] Quantitative Similarity-Activity Relationships (QSiAR) Analysis,[37] Pseudo-Atomic Receptor Model QSAR (PARM),[38] and Comparative Molecular Moment Analysis (CoMMA).[39]

We have employed the same training and test sets as in the above-mentioned publications:[13,18,37−39] the training set included compounds 1 to 21 and the test set contained compounds 22 to 31 (the *first* training and test set selection). Additionally, following ref 18 the models were constructed based on compounds 1 to 12 and 23 to 31, while compounds 13 to 22 were included into the test set (the *second* training and test set selection). QSAR models for each of the five chirality correction values and each of the eight subsets of descriptors (see Section 2.6) were generated. For the *first* training and test set selection, for 11 out of these 40 combinations of descriptor sets and chirality corrections satisfactory (Sections 2.5 and 2.6) were found, while for the *second* training and test set selection predictive models were found for 15 combinations. The total number of models with high predictive ability was 37 and 35, for the *first* and *second* training and test set selection, respectively. Best models for the *first* training and test set selection had the following statistics: $q^2 = 0.83$, $R^2 = 0.89$, $R_0^2 = 0.81$, $k = 0.93$, $F = 67.7$ (Model 1) and $\alpha = 1 - 3.5656 \times 10^{-5}$ and $q^2 = 0.74$, $R^2 = 0.86$, $R_0^2 = 0.82$, $k = 0.98$, $F = 49.3$ and $\alpha = 0.9999$ (Model 2). (The first model has higher R^2 than the second one, whereas k for the second model is closer to one.) The first model was built using nonchiral, and class I and subclass IId chirality descriptors were calculated with the chirality

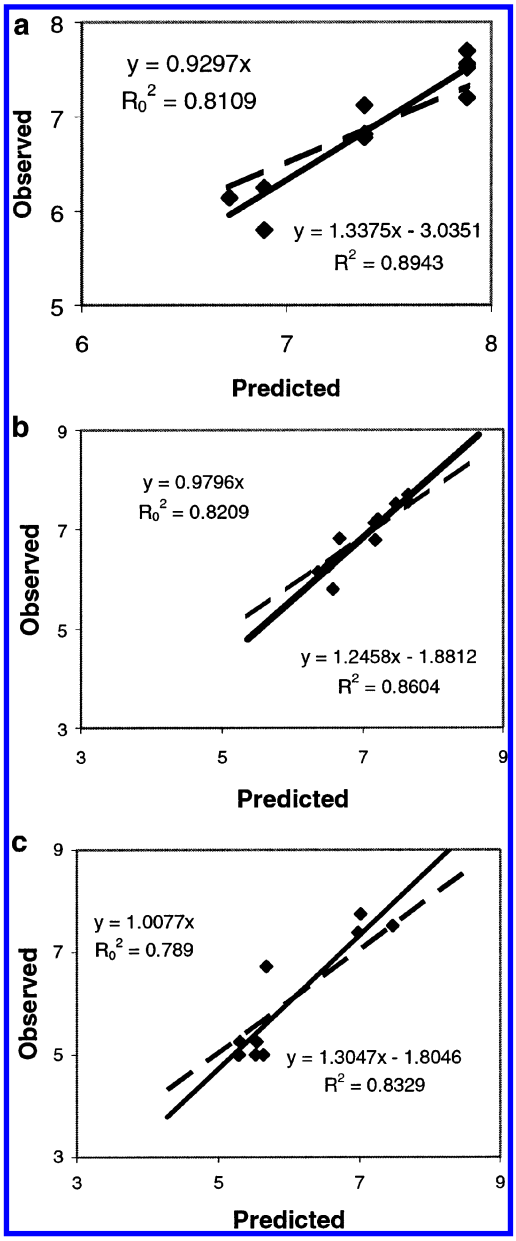

Figure 2. (a) Cramer's steroids. Observed vs predicted binding affinities for the external test set including compounds 22−31 (Model 1 was built for the training set including compounds 1−21). **(**b) Cramer's steroids. Observed vs predicted binding affinities for the external test set including compounds 22−31 (Model 2 was built for the training set including compounds 1−21). **(**c) Cramer's steroids. Observed vs predicted binding affinities for the external test set including compounds 13−22 (The model was built for the training set including compounds 1−12 and 23−31).

correction 1.0. The second model was built using all descriptors (nonchiral, I, IIb, IIc and IId) calculated with the chirality correction 1.5. In Model 1, all nearest neighbors of all 10 compounds of the test set included only five different compounds, two of them having the same binding affinity. Predictive abilities of these two models are demonstrated in Figure 2a,b.

The model with the highest predictive ability for the second training and test set selection had the following statistics: $q^2 = 0.83$, $R^2 = 0.83$, $R_0^2 = 0.79$, $k = 1.01$, $F = 39.9$ and $\alpha = 0.9998$. This model was built using nonchiral and subclass IId chirality descriptors calculated with chirality correction 2.0. Predictive ability of this model is demonstrated in Figure 2c.

Table 3. Cramer's Steroids—Training Set: Compounds 1−21 and Test Set: Compounds 22−31[c]

method	alignment	q^2	R^2 (pred)	ref
CoMFA	rigid	0.66	0.30	18
CoMFA	SEAL	0.60	0.36	18
CoMSIA	SEAL	0.67	0.40	18
QSiAR	rigid	0.59	0.78	37
QSiAR	SEAL	0.56	0.78	37
PARM		0.81	0.33	38
CoMMA		0.94	0.12[a] (0.88)[a,b]	39
Model 1		0.83	0.89	
Model 2		0.74	0.86	

[a] Estimated from Figure 2 in ref 39. [b] After exclusion of two outliers. [c] Comparison of our best models with best models obtained with other methods.

Table 4. Cramer's Steroids—Training Set: Compounds 1−12 and 23−31 and Test Set: Compounds 13−22[a]

method	alignment	q^2	R^2 (pred)	ref
QSiAR	rigid	0.56	0.84	18
QSiAR	SEAL	0.65	0.86	18
Our model		0.82	0.83	

[a] Comparison of our best model with the best models obtained with other methods.

For both training and test set selections, calculations were performed using Molconn-Z[31] descriptors only and Molconn-Z[31] descriptors along with chirality-insensitive overall Zagreb indices,[19,41,42] molecular connectivity indices,[43−45] extended connectivity indices[46] and overall connectivity indices.[47,48] No one model was found to satisfy conditions considered in Sections 2.5 and 2.6. We note that models based on different subsets of descriptors even for one chirality correction have different predictive ability, and models based on all subsets of descriptors are not necessarily the best ones. Thus, we come to the same conclusion as for the previous example. Indeed, it is relatively fast to obtain one QSAR model with high predictive ability. However, extensive calculations with different values of chirality correction and selections of subsets of chirality descriptors should be performed in order to find the QSAR model with the highest predictive power.

In Tables 3 and 4 our best 2D-QSAR models are compared with those obtained by other authors using various 3D-QSAR methods. Obviously, our best models built for the first training and test set selection are superior to almost all other models with respect to their predictive ability (Table 3). They appear to be comparable (but still better) to the models based on QSiAR approach (Table 3). In case of the second training and test selection, our best models are almost as good as ones obtained with the QSiAR approach (Table 4).

In summary, we have demonstrated that in the case of Cramer's steroids chirality descriptors implemented in 2D-QSAR appear to be useful. Our models compared favorably with most of other QSAR models obtained with different methods.

3.3. 66 Histamine H_1 Receptor Ligands. This data set includes 35 analogues of 1-phenyl-3-amino-1,2,3,4-tetrahydronaphthalenes [1-phenyl-3-aminotetralins (PATs)] and 31 non-PATs.[40] 52 of 66 molecules in this data set contain chiral

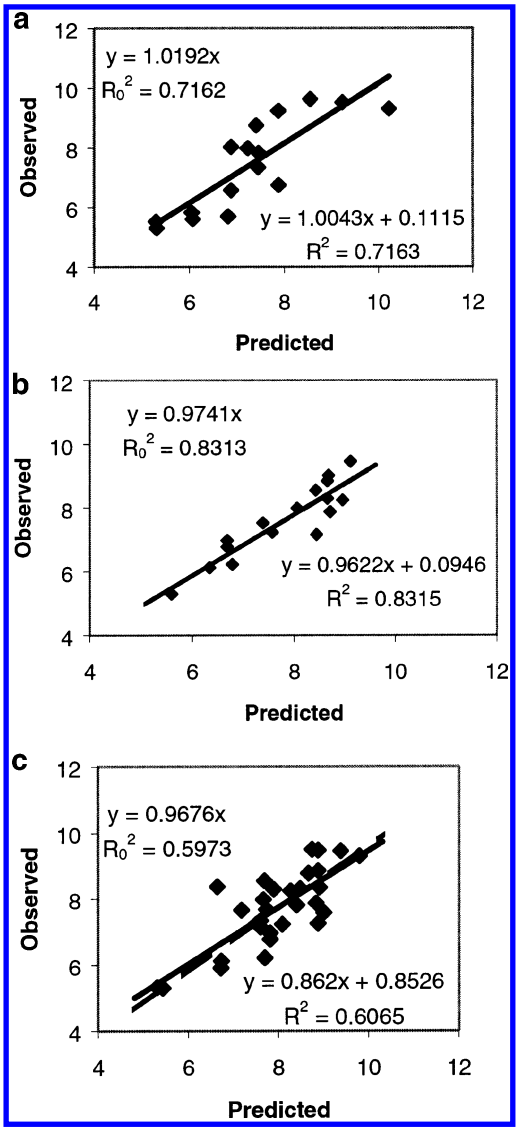

Figure 3. (a) Histamine H_1 receptor ligands. Observed vs predicted $K_{0.5}$ values for the test set including 16 compounds (both training and test sets were the same as in ref 40). **(**b) Histamine H_1 receptor ligands. Observed vs predicted $K_{0.5}$ values for the test set. The model was built using the training set (50 compounds) and validated using the test set (16 compounds) obtained using the sphere-exclusion algorithm (Section 2.3). **(**c) Histamine H_1 receptor ligands. Observed vs predicted $K_{0.5}$ values for the test set. The model was built using the training set (36 compounds) and validated using the test set (30 compounds) obtained using the sphere-exclusion algorithm (Section 2.3).

atoms. Among chiral molecules, 15 pairs of enantiomers and σ-diastereomers differ by spatial configuration of only one atom. Binding affinities $K_{0.5}$ to Histamine H_1 receptors of these compounds are available.[40] Standard CoMFA and CoMFA/q^2-GRS[57] models were built using the training set containing 50 of these compounds.[40] The models were validated using the predictions of $K_{0.5}$ values for the external test set consisted of the remaining 16 compounds.[40] The models for each chirality correction value and each subset of chirality descriptors were verified, whether they satisfied the conditions considered in Sections 2.5 and 2.6, and seven models satisfying these conditions were found. Model with the highest predictive ability had the following statistics: $q^2 = 0.69$, $R^2 = 0.72$, $R_0^2 = 0.72$, $k = 1.02$, $F = 35.4$ and $\alpha = 1-3.5489 \times 10^{-5}$. It was built using conventional and chirality descriptors of subclass IId calculated with the chirality correction 0.5. Its predictive ability is demonstrated in Figure 3a. Again, the results show that the best value of the chirality correction must be established experimentally for each particular data set.

Table 5. Histamine H_1 Receptor Ligands[a]

method	alignment	q^2	R^2 (pred)	ref
CoMFA	rigid	0.46	0.89	40
CoMFA/q^2-GRS	rigid	0.51	0.77	40
Our model		0.69	0.72	

[a] Comparison of our best model with models obtained with other methods.

QSAR models were also built using Molconn-Z[31] descriptors only. One of these models was found almost as good as the best models in which chirality descriptors were used ($q^2 = 0.65$, $R^2 = = 0.70$, $R_0^2 = = 0.68$, $k = 0.87$, $F = 33.0$ and $\alpha = 0.9996$). For the development of this model, the less active compound from each enantiomeric and σ-diastereomeric pair was excluded from the training set. Of course, this model was unable to distinguish between enantiomers and σ-diastereomers. Comparison with other 3D-QSAR models is given in Table 5. Apparently, the predictive power of 3D-QSAR models appears to be slightly better than that of our 2D-QSAR models. On the other hand, 3D-QSAR models have much lower q^2 than our models. The model based on standard CoMFA had $q^2 < 0.5$ which implies that this model is not robust (in this work, we did not even try to make predictions using models with q^2 less than 0.5).

Additional QSAR models were developed with the sphere-exclusion algorithm described in Sections 2.3 and 2.6. The models were obtained for each value of the chirality correction, each descriptor subset, and each training and test set. Several models built for the training set containing about 50 compounds have much better predictive ability than the models built using the training set from ref 40. Predictive power of one of these models is characterized by $q^2 = 0.55$, $R^2 = 0.83$, $R_0^2 = 0.83$, $k = 0.97$, $F = 69.1$, and $\alpha = 1-8.7402 \times 10^{-7}$ (see also Figure 3b). All descriptor subsets (nonchiral, I, IIb, IIc, and IId) were used to build this model; chirality correction was equal to 1.5. Better predictive ability of these models can be explained by better (rational) division of all compounds into the training and test sets used in these calculations. The rational methodology for the selection of training and test sets affords predictive QSAR models for relatively large test sets. Thus, one model built with the training set containing only 36 molecules and validated using the test set containing 30 compounds had the following statistics: $q^2 = 0.67$, $R^2 = 0.61$, $R_0^2 = 0.60$, $k = 0.97$, $F = 43.2$, and $\alpha = 1-3.9767 \times 10^{-7}$ (see also Figure 3c). Nonchiral and subclass IId chirality descriptors were used to develop this model; chirality correction was equal to 0.5. As in the previous examples, we emphasize the dependence of the best values of the chirality correction on the choice of compounds for the training and test sets and the descriptor subset used to build the model.

Thus, in the case of Histamine H_1 receptor ligands, we were able to build relatively good 2D-QSAR models for the prediction of $K_{0.5}$ values. We also emphasize that in this case, the data set included both chiral and achiral compounds. Chiral compounds from this data set contained one to four

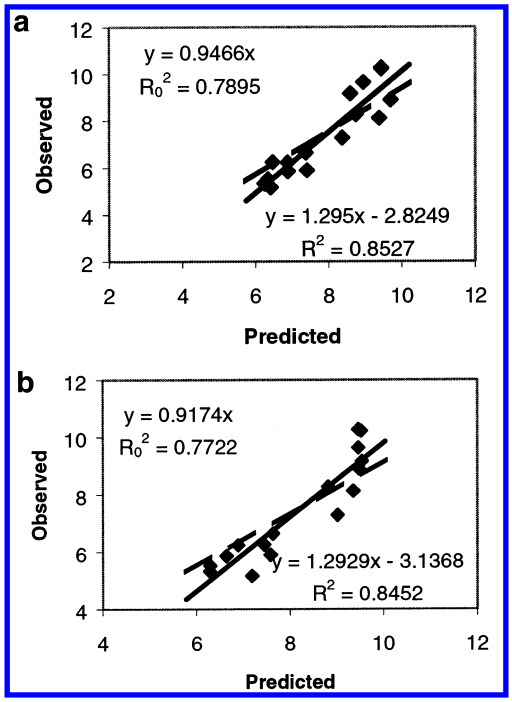

Figure 4. (a) Protease inhibitors. Model 1. Observed vs predicted IC_{50} values for the external test set, which contained 16 compounds. Training and test sets were the same as in refs 15 and 23. (b) Protease inhibitors. Model 2. Observed vs predicted IC_{50} values for the external test set, which contained 16 compounds. Training and test sets were the same as in refs 15 and 23.

chiral atoms, while Cramer's steroids contained four to eight, and ecdysteroids contained eight to fifteen chiral atoms. Our descriptors appeared to be applicable to QSAR studies of all these examples.

3.4. 49 HIV-1 Protease Inhibitors. All molecules in this data set contain chiral atoms. Activity data (IC_{50} values) for these compounds were published.[15,23] The correlation between the calculated binding energy and IC_{50} values has been established and regression analysis was used to predict activities of compounds from the test set.[23] Comparative Binding Energy Analysis (COMBINE) model was built and used to predict activities of compounds from the test set.[15] The same training set (33 compounds) and test set (16 compounds) as in refs 15 and 23 were used for building 2D-QSAR models. For each chirality correction value and each subset of descriptors the models were found that satisfied the conditions considered in Sections 2.5 and 2.6. The total number of models satisfying these conditions was as high as 606. Two models with the highest predictive ability had the following statistics: $q^2 = 0.77$, $R^2 = 0.85$, $R_0^2 = 0.79$, $k = 0.95$, $F = = 81.1$ and $\alpha = 1-3.3633 \times 10^{-7}$ (Model 1) and $q^2 = 0.79$, $R^2 = 0.85$, $R_0^2 = 0.77$, $k = 0.92$, $F = 76.4$ and $\alpha = 1-4.8142 \times 10^{-7}$ (Model 2). The first of these models was generated using nonchiral and Id chirality descriptors calculated with the chirality correction 2.5. For the second model nonchiral and all classes and subclasses of chirality descriptors were used; the chirality correction was equal to 0.5. Predictive power of these models is demonstrated in Figure 4a,b. Models using only Molconn-Z[31] descriptors were constructed as well. (There was only one stereoisomeric pair in this data set, and the compound with the lower activity from this pair was excluded from the test set.) In this case, one model based on Molconn-Z[31] descriptors was found, which was only slightly worse than the best models built using chirality descriptors ($q^2 = 0.79$, $R^2 = 0.7^{18}8$, $R_0^2 = 0.77$, $k = 0.96$, $F = 47.4$, and $\alpha = 1-7.5045 \times 10^{-6}$). Apparently, this model cannot discriminate between stereoisomers. In ref 23, the average absolute error for the prediction was 1.01 log units. After deleting one outlier, the absolute error of prediction was 0.79 log units. In contrast, the average absolute error for prediction using our best Models 1 and 2 were 0.79 log units (without outliers) and 0.85 log units (without outliers), respectively. The COMBINE model in ref 15 gave SDEP = 0.59, while our models gave SDEP = 0.61 and SDEP = 0.51, respectively.

Thus, in this case as in all previous cases we were able to obtain predictive 2D-QSAR models using chirality descriptors. Predictive ability of our best models appeared to be at least comparable with those of two different 3D-QSAR models.

4. CONCLUSIONS

We have employed chirality descriptors developed in our laboratory earlier[19] for 2D-QSAR studies of four data sets of ligands to different receptors. In all cases, the kNN QSAR procedure[29] with leave-one-out cross-validation was used. We compared our models with those reported in the literature for the same data sets using different QSAR approaches. The main criterion of comparison was the predictive power of the models estimated by their ability to predict accurately the activities of the external test set compounds. The training and test sets for all examples were the same as in the original publications. In all cases, our best models were either comparable to other (mostly, 3D) QSAR models or outperformed them. Thus, we have demonstrated that topological chirality descriptors can be successfully used to generate 2D-QSAR models for data sets containing stereoisomers.

We have established that for each particular data set, different chirality correction values should be used to obtain models with the highest predictive power. No preferable chirality correction value was found for all data sets. The best values of chirality correction were generally different even when models were built for different training and test sets derived from the same data set of compounds. This result suggests that this value has to be found experimentally for any given data set.

Our studies demonstrate that chirality descriptors calculated on the basis of chemical graph theory augmented by the knowledge of the absolute configurations of chiral atoms can be successfully applied in 2D-QSAR studies of data sets, which include chiral molecules. 2D-QSAR methods have several important advantages as compared to 3D-QSAR methods, i.e., they do not require extensive conformational analysis and spatial alignment of molecules; they can be easily automated and are much more computationally efficient. Therefore, we suggest that 2D-QSAR methods enhanced by chirality descriptors present a powerful alternative to popular 3D-QSAR approaches. A program to calculate chirality descriptors is available from authors upon request.

ACKNOWLEDGMENT

The support for this project was provided in part by the National Institutes of Health (grant number MH60328). The

authors are grateful to Prof. Danail Bonchev who has inspired and contributed to the development of chirality descriptors.

REFERENCES AND NOTES

(1) Moran, L. A.; Scrimgeour, K. G.; Horton, H. R.; Ochs, R. S.; Rawn, J. D. *Biochemistry*; Neil Patterson Publishers Prentice Hall: Englewood Cliffs, NJ, 1994.

(2) Potapov, V. M. *Stereochemistry*; Khimia: Moscow, 1988.

(3) Solms, J.; Vuataz, L.; Egli, R. H. The taste of L- and D-amino acids. *Experiencia* **1965**, *21*, 692−694.

(4) Schiffman, S. S.; Clark, T. B. 3d; Gagnon, J. Influence of chirality of amino acids on the growth of perceived taste intensity with concentration. *Physiol. Behav.* **1982**, *28*, 457−465.

(5) Laska, M.; Teubner, P. Olfactory discrimination ability of human subjects for 10 pairs of enantiomers. *Chem Senses*. **1999**, *24*, 161−70.

(6) Polak, E. H.; Fombon, A. M.; Tilquin, C.; Punter, P. H. Sensory evidence for olfactory receptors with opposite chiral selectivity. *Behav. Brain. Res*. **1989**, *31*, 199−206.

(7) DeCamp, W. H. The FDA Perspective on the development of stereoisomers. *Chirality* **1989**, *1*, 2−6.

(8) Hutt, A. J.; Tan, S. C. Drug chirality and its clinical significance. *Drugs* **1996**, *52*, 1−12.

(9) Wnendt, S.; Zwingenberger, K. Thalidomide's chirality. *Nature* **1997**, *385*, 303−304.

(10) Kurihara, N.; Miyamoto, J.; Paulson, G. D.; Zeeh, B.; Skidmore, M. W.; Hollingworth, R. M.; Kuiper, H. A. IUPAC reports on pesticides .37. Chirality in synthetic agrochemicals: Bioactivity and safety consideration. *Pure Appl. Chem.* **1997**, *69*, 1335−1348.

(11) Stinson, S. C. Chiral Drugs. *Chem. Eng. News* **2000**, *78*, 43.

(12) Stinson, S. C. Chiral Pahrmaceuticals. *Chem. Eng. News* **2001**, *79*, 79−97.

(13) Cramer, R. D., III.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Amer.Chem. Soc.* **1988**, *110*, 5959−5967.

(14) Marshall, G. R.; Cramer, R. D., III Three-Dimensional Structure−Activity Relationships. *Trends Pharmacol. Sci.* **1988**, *9*, 285−289.

(15) Pérez, C.; Pastor, M.; Ortiz, A. R.; Gago, F. Comparative Binding Energy Analysis of HIV-1 Protease Inhibitors: Incorporation of Solvent Effects and Validation as a Powerful Tool in Receptor-Based Drug Design. *J. Med. Chem.* **1998**, *41*, 836−852.

(16) Cho, S. J.; Tropsha, A. Cross-Validated R^2 Guided Region Selection for Comparative Molecular Field Analysis (CoMFA): A Simple Method to Achieve Consistent Results, *J. Med. Chem.* **1995**, *38*, 1060−1066.

(17) Klebe, G. Comparative Molecular Similarity Indices Analysis − CoMSIA. In *3D QSAR in Drug Design. Volume 3. Recent Advances*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Kluwer/ESCOM: Dordrecht, 1998; pp 87−104.

(18) Kubinyi, H.; Hamprecht, F. A., Mietzner, T. Three-Dimensional Quantitative Similarity-Activity Relationships (3D QSiAR) from SEAL Similarity Matrices. *J. Med. Chem*. **1998**, *41*, 2553−2564.

(19) Golbraikh, A.; Bonchev, D.; Tropsha, A. Novel chirality descriptors derived from molecular topology. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 147−158.

(20) Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 411−28.

(21) Osterberg, F.; Morris, G. M.; Sanner, M. F.; Olson, A. J.; Goodsell, D. S. Automated docking to multiple target structures: Incorporation of protein mobility and structural water heterogeneity in AutoDock. *Proteins* **2002**, *46*, 34−40.

(22) Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J. Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. *J. Comput.-Aided Mol. Des.* **1996**, *4*, 293−304.

(23) Holloway, M. K.; Wai, J. M.; Halgren, T. A.; Fitzgerald, P. M. D.; Vacca, J. P.; Dorsey, B. D.; Levin, R. B.; Thompson, W. J.; Chen, L. J.; deSolms, S. J.; Gaffin, N.; Ghosh, A. K.; Giuliani, E. A.; Graham, S. L.; Guare, J. P.; Hungate, R. W.; Lyle, T. A.; Sanders, W. M.; Tucker, T. J.; Wiggins, M.; Wiscount, C. M.; Woltersdorf, O. W.; Young, S. D.; Darke, P. L.; Zugay, J. A. A Priori Prediction of Activity for HIV-1 Protease Inhibitors Employing Energy Minimization in the Active Site. *J. Med. Chem.* **1995**, *38*, 305−317.

(24) Judson, R. Genetic Algorithms and Their Use in Chemistry. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: 1997; Vol. 10.

(25) Kramer, B.; Rarey, M.; Lengauer T. Evaluation of the FLEXX incremental construction algorithm for protein−ligand docking. *Proteins* **1999**, *37*, 228−41.

(26) Claussen, H.; Buning, C.; Rarey, M.; Lengauer T. FlexE: efficient molecular docking considering protein structure variations. *J. Mol. Biol.* **2001**, *27*, 377−95.

(27) Cho, S. J.; Serrano, M. G.; Bier, J.; Tropsha, A. structure based alignment and comparative molecular field analysis of acetylcholinesterase inhibitors. *J. Med. Chem.* **1996**, *39*, 5064−5071.

(28) Pilger, C.; Bartolucci, C.; Lamba, D.; Tropsha, A.; Fels, G. Accurate prediction of the bound conformation of galanthamine in the active site of Torpedo Californica acetylcholinesterase using molecular docking. *J. Mol. Graphics Modeling* **2001**, *19*, 288−296, 374−378.

(29) Hoffman, B.; Cho, S. J.; Zheng, W.; Wyrick, S. D.; Nichols, D. E.; Mailman, R. B.; Tropsha, A. Quantitative structure−activity relationship modeling of dopamine D-1 antagonists using comparative molecular field analysis, genetic algorithms-partial least-squares, and K nearest neighbor methods. *J. Med. Chem.* **1999**, *42*, 3217−3226.

(30) Zheng, W.; Tropsha, A. novel variable selection quantitative structure−property relationship approach based on the k-nearest-neighbor principle. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185−194.

(31) Molconnz. http://www.eslc.vabiotech.com/.

(32) Julián-Ortiz, J. V. de; Alapont, C. de G.; Ríos-Santamarina, I.; García-Doménech, R.; Gálvez, J. Prediction of properties of chiral compounds by molecular topology. *J. Mol. Graphics Mod.* **1998**, *16*, 14−18.

(33) *Ecdysone: From chemistry to mode of action*; Koolman J., Ed.; Thieme, 1989.

(34) Dinan, L.; Hormann, R. E.; Fujimoto, T. An extensive ecdysteroid CoMFA. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 185−207.

(35) Golbraikh, A.; Bonchev, D.; Xiao, Y.-D.; Tropsha, A. Novel chiral topological descriptors and their applications to QSAR. In Rational Approaches to Drug Design. Proceedings of the 13th European Symposium on quantitative Structure−Activity relationships, Prous Science, 2001, pp 219−223.

(36) Coats, E. A. The CoMFA steroids as a benchmark data set for development of 3D QSAR methods. *In 3D QSAR in Drug Design. V.3.*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Kluwer/ESCOM: Dordrecht, 1998; pp 199−213.

(37) Klebe, G.; Abraham, U.; Mietzner, T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their bioogical activity. *J. Med. Chem.* **1994**, *37*, 4130−4146.

(38) Chen, H.; Zhou, J.; Xie, G. PARM: A genetic evolved algorithm to predict bioactivity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 243−250.

(39) Silverman, B. D. The thirty-one benchmark steroids revisited: comparative molecular moment analysis (CoMMA) with principal component regression. *Quant. Struct.-Act. Relat.* **2000**, *19*, 237−246.

(40) Bucholz, E.; Brown, R. L.; Tropsha, A.; Booth, R. G.; Wyrick, S. D. Synthesis, evolution, and comparative molecular field analysis of 1-phenyl-3-amino-1,2,3,4-tetrahydronaphthalenes as ligands for Histamine H1 receptors. *J. Med. Chem.* **1999**, *42*, 3041−3054.

(41) Bonchev, D. Overall connectivity − a next generation molecular connectivity. *J. Mol. Graph. Model.* **2001**, 20 Sp. Iss. SI, 65−75.

(42) Gutman I.; Ruscić, B.; Trinajstić, N.; Wilcox, C. F., Jr. Graph theory and molecular orbitals. XII. Acyclic polyenes. *J. Chem. Phys*. **1975**, *62*, 3399.

(43) Randić, M. On characterization on molecular branching. *J. Am. Chem. Soc*. **1975**, *97*, 6609−6615.

(44) Kier, L. B.; Hall, L. H. *Molecular connectivity in chemistry and drug research*; Academic Press: New York, 1976.

(45) Kier, L. B.; Hall, L. H. *Molecular connectivity in structure−activity analysis*; Wiley: New York, 1986.

(46) Rücker, G.; Rücker, C. Counts of all walks as atomic and molecular descriptors. *J. Chem. Inf. Comput. Sci*. **1993**, *33*, 683−695.

(47) Bonchev, D. Overall connectivity and molecular complexity. In *Topological indices and related descriptors*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach: Reading, U.K., 1999; pp 361−401.

(48) Bonchev, D. Novel indices for the topological complexity of molecules. *SAR/QSAR Environ. Res*. **1997**, *7*, 23−43.

(49) Wootton, R.; Cranfield, G.; Sheppey, C.; Goodford, P. J. Physicochemical-activity relationships in practice. 2. Rational selection of benzenoid substituents. *J. Med. Chem.* **1975**, *18*, 607−612.

(50) Snarey, M.; Terrett, N. K.; Willett, P.; Wilton, D. J. Comparison of algorithms for dissimilarity-based compound selection. *J. Mol. Graphics. Model.* **1997**, *15*, 372−385.

(51) Golbraikh, A.; Tropsha, A. Predictive QSAR modeling based on rational division of experimental datasets into diverse training and test sets. *J. Comput.-Aided Mol. Des.*, in press.

(52) Golbraikh A. Molecular dataset diversity indices and their applications to comparison of chemical databases and QSAR analysis. *J. Chem. Inform. Comput. Sci.* **2000**, *40*, 414−425.

(53) Golbraikh, A.; Tropsha, A. Beware of q^2!. *J. Mol. Graphics Mod.* **2002**, *20*, 269−276.

(54) Sachs, L. *Applied statistics. A handbook of techniques*; Springer-Verlag: 1984.

(55) MATLAB. http://www.mathworks.com/products/matlab/.

(56) Clark, R. D.; Sprous, D. G.; Leonard, J. M. Validating models based on large datasets. In Rational approaches to drug design. Proceedings of the 13th European Symposium on Quantitative Structure−Activity Relationships. Aug 17 − Sept 1, Duesseldorf, Germany. Hoeltje, H.-D., Sippl, W., Eds.; Prous Science: 2001.

(57) Cho, S. J.; Tropsha, A. Cross-Validated R^2 Guided Region Selection for Comparative Molecular Field Analysis (CoMFA): A Simple Method to Achieve Consistent Results. *J. Med. Chem.* **1995**, *38*, 1060−1066.

CI025516B