Journal of Computer-Aided Molecular Design **18:** 175–187, 2004.

Ligand intramolecular motions in ligand-protein interaction: ALPHA, a novel dynamic descriptor and a QSAR study with extended steroid benchmark dataset

Kari Tuppurainen, Marja Viisas, Mikael Peräkylä & Reino Laatikainen*
University of Kuopio, Department of Chemistry, P.O. Box 1627, FIN-70211 Kuopio, Finland

Received 24 November 2003; accepted in revised form 16 April 2004

Key words: $1/f^{\alpha}$ noise, descriptor, molecular dynamics, molecular motions, QSAR, SOMFA, spectral exponents

Summary

The role of intramolecular motions in ligand–macromolecule interactions has been explored by developing and validating ALPHA, a novel QSAR (quantitative structure–activity relationship) descriptor. It is based on the spectral exponents (α), which measure the degree of $1/f^{\alpha}$ noise of coordinate fluctuations in molecular dynamics (MD) simulations. ALPHA is the first truly 'dynamic' QSAR descriptor, i.e., it can be derived directly from an MD trajectory. The performance of ALPHA was tested in detail employing the CBG (corticosteroid binding globulin) affinity of 31 benchmark steroids, supplemented with 11 steroids as an external test set. The only fair (42–50%) correlations of ALPHA with static 3D and electronic descriptors mean that ALPHA forms an independent molecular property. Furthermore, inclusion of ALPHA in the SOMFA/ESP model improves the correlation coefficient from 0.86 to 0.91, and $|\Delta|_{ave}$ from 0.46 to 0.36 for the benchmark dataset. The predictive ability of ALPHA can be interpreted as indirect evidence of the dynamic contribution to ligand–macromolecule interactions. The physical background of ALPHA is discussed and the importance of molecular motions for biological activity is anticipated.

Introduction

The importance of intramolecular motions of proteins and other macromolecules is well approved for their biological activity [1–4]. For example, it has been found that substrate turnover rates correlated strongly with fluctuations that occur on a time scale of hundreds of microseconds [5]. Often the term intramolecular motion is used to mean these or slower conformational change type motions in biomolecules [2, 3], ignoring that most of the thermal energy of a protein is stored in high-frequency vibration-like motions. These vibrations are invisible in standard NMR and X-ray measurements, but they have been proposed to have a significant effect on the properties of a protein [6]. In general, events like formation of the high energy transition state of a chemical reaction are very fast and it is obvious that these high-frequency motions could play a role, for example, in enzyme catalysis: if the intramolecular motion takes place in the direction of the reaction coordinate, the thermal energy can be transferred into activation energy. These motions have recently been called protein promoting vibrations [7] and they have been connected to hydride transfer [7, 8]. The very fast motions in proteins are localized to 'hot spots' [9], which are primarily related to protein folding and stability [9, 10]. However, these hot residues seem to be strongly coupled to functionally important motions [10]. Thus it is not surprising that even a small mutation in structure, far from the active site, can have a large effect on enzyme function [8]. This kind of considerations led us to formulate the hypothesis that 'molecular motions of native proteins are more correlated than those of mutants' [11, 12], in analogy with the thermodynamic hypothesis [13]. Principal component analysis of molecular dynamics data supported our hypothesis for two proteins of different type, lysozyme and BPTI.

*To whom correspondence should be addressed. Fax: +358-17-163259, E-mail: Reino.Laatikainen@uku.fi

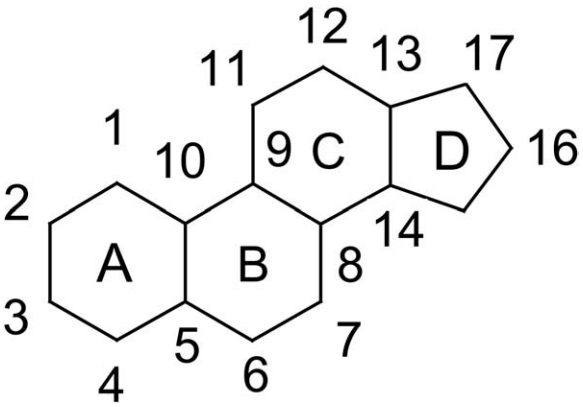

Figure 1. The numbering of the steroid backbone.

It is obvious that also intramolecular motions of a ligand can play a role in molecular regulation and recognition, as supported by our computations with BPTI [12]. Recently inhibitor binding has been shown to alter the directions of domain motions in HIV-1 reverse transcriptase [14]. In principle, if the ligand enhances the motions of enzyme in the direction of the reaction coordinate, it can act as an activator (or agonist); in the other case it can act as an inhibitor (or antagonist). Also ligand binding affinity is expected to depend on motions: if the ligand binding reduces the motions of the protein, this decreases the protein vibrational entropy and leads to lower binding affinity (by the way, it is easy to imagine systems where the stronger the binding, the larger the opposing entropy term: this could be an interpretation of the somewhat controversial 'entropy–enthalpy compensation principle' [15]). Alternatively, a ligand may tune the motions of a protein, although at the cost of binding affinity, so that its motions correspond better to the motions of another protein and thus enhance protein–protein recognition. In this case a poor binding affinity may mean an agonistic function.
Previously, the intramolecular motions of ligands have been almost completely neglected in (quantitative) structure–activity (SAR/QSAR) studies. Of course, there have been numerous attempts to account for conformational flexibility, i.e., the ability of molecules to change conformation, employing molecular shape analysis [16–18], conformational entropy and energy values [19–21], a substituent entropy constant [22] and flexibility indices [23]. The culmination of this approach, 4D-QSAR [24], considers both conformational and alignment freedom in the derivation of common grid-type 3D QSAR models.

Table 1. Benchmark data set of steroids with observed and LOO-CV predicted CBG affinities.

No.	Compound	Activity	
		Observed	Predicted
1	Aldosterone	6.28	6.58
2	Androstanediol	5.00	4.74
3	Androstenediol	5.00	6.59
4	Androstenedione	5.76	5.48
5	Androsterone	5.61	5.36
6	Corticosterone	7.88	7.37
7	Cortisol	7.88	7.45
8	Cortisone	6.89	6.60
9	Dehydroepiandrosterone	5.00	6.03
10	Deoxycorticosterone	7.65	7.61
11	Deoxycortisol	7.88	7.06
12	Dihydrotestosterone	5.92	6.10
13	Estradiol	5.00	6.37
14	Estriol	5.00	5.78
15	Estrone	5.00	4.81
16	Etiocholanolone	5.26	5.16
17	Pregnenolone	5.26	5.99
18	17-Hydroxypregnenolone	5.00	6.07
19	Progesterone	7.38	7.02
20	17-Hydroxyprogesterone	7.74	7.24
21	Testosterone	6.72	6.69
22	Prednisolone	7.51	6.27
23	Cortisol 21-acetate	7.55	7.31
24	4-Pregnene-3,11,20-trione	6.78	6.18
25	Epicorticosterone	7.20	6.31
26	19-Nortestosterone	6.14	5.04
27	16α,17-Dihydroxy-4-pregnene-3,20-dione	6.25	7.15
28	16-Methyl-4-pregnene-3,20-dione	7.12	7.20
29	19-Norprogesterone	6.82	6.72
30	11β,17,21-Trihydroxy-2α-methyl-4-pregnene-3, 20-dione,	7.69	7.20
31	11β,17,21-Trihydroxy-2α-methyl-9α-fluoro-4-pregnene-3,20-dione	5.80	6.30

PLS statistics for the LOO CV: $S_{press} = 0.73$ $q^2 = 0.57$, $c = 2$; for the fitted model: $r^2 = 0.79$, $SE = 0.51$, $F = 52.8$.

In this study, we worked in order to develop 'dynamic' SAR/QSAR descriptors based on molecular dynamics (MD) simulations. The large amount of information of the MD trajectories can be compressed and analyzed by statistical methods such as principal components analysis or cluster analysis. It has proven difficult to distinguish biologically relevant motions from thermal fluctuations that are common for all molecular matter and, although MD simulations have been used to derive 'time-averaged' QSAR descriptors

Table 2. The observed and predicted CBG affinities for the ALPHA Model 1 (training set **1–21**, test set **22–31**) and the corresponding PLS statistics.

Molecule	CBG affinity		
	Observed	Predicted	Deviation
22	7.51	6.20	1.32
23	7.55	7.08	0.47
24	6.78	6.28	0.50
25	7.20	6.38	0.82
26	6.14	4.91	1.23
27	6.25	6.96	−0.71
28	7.12	7.02	−0.10
29	6.82	6.80	0.02
30	7.69	7.26	0.43
31	5.80	6.06	−0.26

PLS statistics for the LOO CV: $S_{press} = 0.76$, $q^2 = 0.63$, $c = 2$; for the fitted model: $r^2 = 0.85$, $SE = 0.48$, $F = 50.3$; for the external prediction: $r^2_{ex} = 0.32$ (0.30), $|\Delta|_{ave} = 0.58$ (0.62), $SDEP = 0.71$ (0.75), $pr{-}r^2 = 0.45$ (0.45). The numbers in parentheses correspond to predictions without **31**.

for a set of pyrethroid analogs [25], previous studies have not provided truly dynamic descriptors that could be derived directly from an MD trajectory. The objective of this work was, first, to examine whether it is possible to derive such descriptors and, second, to explore the significance of motions in steroid–protein interaction by employing the corticosteroid binding globulin (CBG) activity dataset [26, 27], which has repeatedly served as a benchmark in evaluating the performance of QSAR methods [28].

Materials and methods

The dataset of 31 steroids with corticosteroid binding globulin (CBG) affinity is given in Table 1. The numbering of the steroid backbone is given in Figure 1. Molecular dynamics simulations were done with the SYBYL molecular modeling package [29] using the Tripos force field. For the MD simulations the geometries of the steroids were first energy-minimized for 2000 steps with the conjugate gradient method. The MD simulations were then started by equilibrating the ligands for 30 ps at 300 K using a time step of 1 fs. During the 30 ps production simulations structures were saved every 10 fs (3000 structures) for analyses. The PLS analyses were done using MATLAB (The MathWorks, Natick, MA) scripts written by the authors. The scripts are based on an efficient modification of the PLS algorithm, SVDPLS (singular value decomposition PLS), which facilitates very rapid cross-validation runs [30].

Results and discussion

Spectral exponents as dynamical QSAR descriptors

Any task to derive useful QSAR descriptors from an MD trajectory is challenging. First, a substantial compression of data is a prerequisite. Second, the comparison between the molecules would require an alignment of molecules, which would be difficult due to the 'dynamic' nature of data. Of course, a restriction to the common atoms such as to the molecular backbone would make the comparison between the molecules possible, but it does not solve the alignment problem and the potentially important side chain information would be missing. In order to avoid these difficulties, we sought methods to derive effective 'unassigned' descriptors from the MD trajectory. By the term 'unassigned' we mean in this context that there is no need to assign the descriptor to a specific atom or coordinate.

As the first attempt, based on our success with the proteins [12,13] we tried principal component analysis of the MD trajectories. No regularity between the eigenvalues and activity was discovered, but a notable point is that the motions of all the steroids could be well described by only three principal components. In the second attempt, we applied more successfully a technique of Gaussian smoothing, previously used in deriving 'spectroscopic' QSAR descriptors [31–34]. In the following, we first describe the derivation of the ALPHA descriptor.

Flickering noise in MD simulations and alpha-exponents

If the power spectrum of a time series depends on the frequency (n.b. sampling frequency), the fluctuation is said to be $1/f^{\alpha}$-like ('colored' or flickering) noise. The spectral exponents α approximately range from 0 to 3, and the terms 'white' ($\alpha = 0$), 'pink' ($\alpha = 1$) and 'brown' ($\alpha = 2$) noise are in common use in the field of signal analysis, time series analysis and non-linear dynamics [35]. It has been previously shown that the potential energy fluctuation during an MD simulation of a protein (plastocyanin) is 1/f noise [36, 37]. This work was originally based on the postulate that the same holds for small molecules and, in particular, for

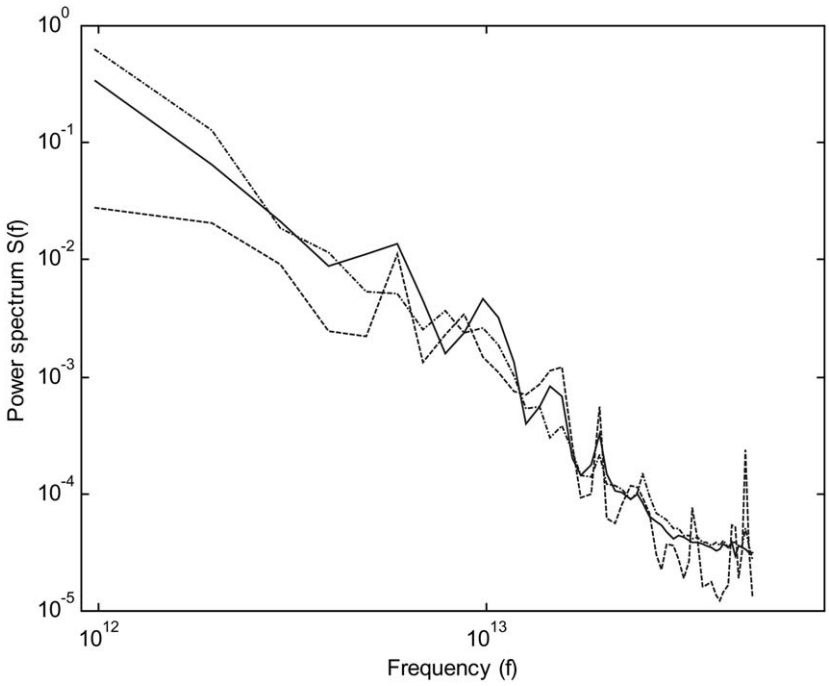

Figure 2. Calculation of α-exponents by plotting the power spectra as a function of frequency (atom O19, attached to C3, of corticosterone as an example). x-Coordinate (—), slope 2.60±0.07; y-coordinate (- - -) slope 2.25±0.13; z-coordinate (- · - · - ·), slope 2.58±0.06.

Table 3. Comparison of ALPHA with other 3D QSAR models for test set **22**–**31**[a].

Method [ref.]	r^2_{ex}	$\|\Delta\|_{ave}$	*SDEP*	$pr{-}r^2$
EEVA [41]	0.36 (0.58)	0.41 (0.30)	0.58 (0.40)	0.64 (0.85)
COMPASS [42]	0.16 (0.69)	0.46 (0.29)	0.70 (0.34)	0.46 (0.89)
MS-WHIM [43]	0.28 (0.63)	0.44 (0.30)	0.66 (0.41)	0.52 (0.83)
PARM [44]	0.33 (0.30)	0.52 (0.56)	0.71 (0.74)	0.45 (0.45)
TQSAR [45]	0.16 (0.36)	0.59 (0.46)	0.76 (0.56)	0.37 (0.69)
SOMFA [46]	0.20 (0.62)	0.43 (0.32)	0.58 (0.36)	0.63 (0.87)
EVA [47]	0.36 (0.34)	0.42 (0.39)	0.53 (0.51)	0.69 (0.74)
CoMFA [47]	0.25 (0.75)	0.46 (0.30)	0.71 (0.40)	0.45 (0.84)
GRIND [48]	– (0.88)	– (0.23)	– (0.26)	– (0.93)
MFTA [49]	0.87 (0.82)	0.21 (0.23)	0.30 (0.31)	0.90 (0.90)
COMSA [50]	0.09 (0.41)	0.52 (0.38)	0.70 (0.44)	0.47 (0.81)
MEDV [52]	0.45 (0.57)	0.54 (0.48)	0.65 (0.59)	0.54 (0.66)
QS-SM [52]	0.36 (0.22)	0.47 (0.42)	0.54 (0.49)	0.68 (0.76)
ALPHA[c]	0.32 (0.30)	0.58 (0.62)	0.71 (0.75)	0.44 (0.45)

Abbreviations: EEVA, Electronic Eigenvalue; COMPASS, Comparative Surface Similarity; MS-WHIM, Molecular Surface – Weighted Holistic Invariants of Molecules; PARM, Pseudo Atomic Receptor Model; TQSAR, Tuned QSAR; SOMFA, Self–Organizing Molecular Field Analysis; EVA, Eigenvalue, i.e. vibrational normal mode in this context; CoMFA, Comparative Molecular Field Analysis; GRIND, Grid Independent Descriptors; MFTA, Molecular Field Topology Analysis; COMSA, Comparative Surface Analysis; MEDV, Molecular Electronegativity Distance Vector; QS-SM, Quantum Self–Similarity Measures.
[a] The values in parentheses are predictions without **31**.
[b]The predicted value of **31** was not reported (stated as an outlier).
[c]This work, Model 1.

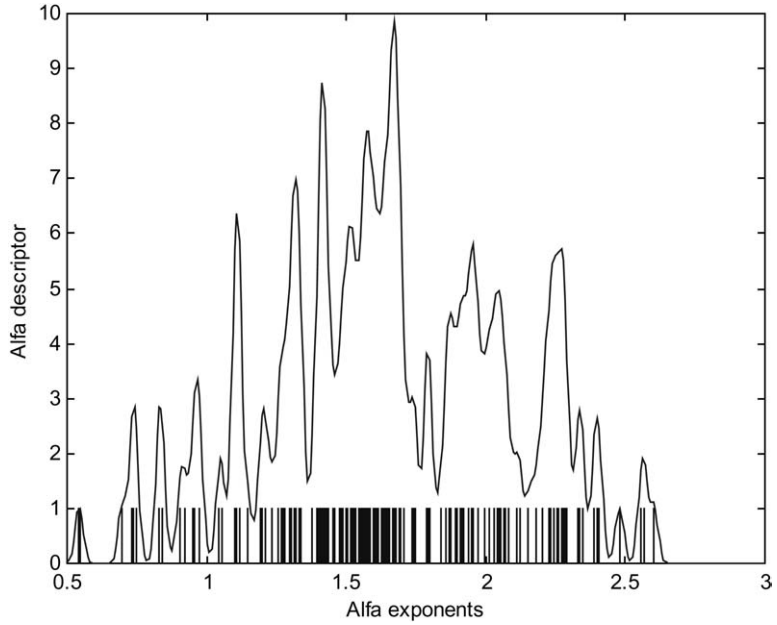

Figure 3. Distribution of the α-exponents and the corresponding ALPHA descriptor generated by Gaussian smoothing.

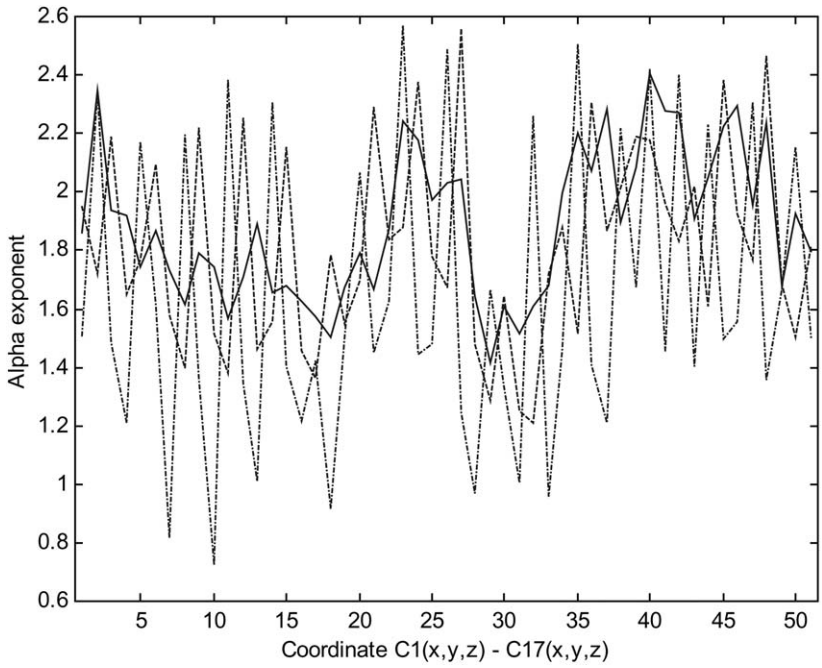

Figure 4. Variation of the α-exponents along the molecular backbone of aldosterone **1** (- - -), androstanediol **2** (- · - · - ·) and corticosterone **6** (—), the most active molecule.

each and every coordinate. In practice, the slope of the best-fit line of the power spectrum as a function of frequency, both expressed on a logarithmic scale, gives a least-squares estimate for the spectral exponent α of each coordinate, as illustrated in Figure 2.

In the Gaussian smoothing [31–34], the α-exponents of a molecule are first transformed to a bounded range (0.5–3 is feasible in this case). Then a Gaussian kernel of fixed standard deviation σ (a parameter to be optimized, see below) is placed over each α-exponent. Finally, the summation of overlaid kernels at intervals of L (usually L is set at σ/2) gives a (pseudo)-spectrum (Figure 3), which can be used as a QSAR descriptor (Equation 1):

$$ALPHA(x) = \sum_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\alpha_i)^2/2\sigma^2} \quad (1)$$

The dimensionality of the ALPHA descriptor is high (depending strongly on the value of σ) and, thus, the PLS method was used to compress the data.

In order to rule out the possibility that the following results depend on the force field, we calculated the α-exponents also using the MMFF94 force field for several molecules. These results showed that the ALPHAs are practically independent of the force field. The correlations between the exponents computed with the different methods were 95–99%.

Data analysis

PLS (partial least-squares or projections to latent structures) is a bilinear projection method to model complex relationships in a set of data (for mathematical details, see for example Höskuldsson [38]). PLS decomposes two data matrices, X (independent variables) and Y (dependent variables), into new variables (PLS components), and creates simultaneously a predictive relationship between them. The optimum number of PLS components is determined by cross-validation (CV). LOO (leave-one-out) CV proceeds by omitting one sample of input data, re-deriving the PLS model, and predicting the Y value(s) of the omitted sample; this cycle continues until all Y values have been predicted exactly once. The S_{press} value and CV correlation coefficient (q^2) are calculated from:

$$S_{press} = \sqrt{\frac{PRESS}{n-c-1}} \quad (2)$$

$$q^2 = 1 - \frac{\sum (y_{obs} - y_{pred})^2}{\sum (y_{obs} - y_{mean})^2} = 1 - \frac{PRESS}{\sum (y_{obs} - y_{mean})^2} \quad (3)$$

where n is the number of compounds and c is the number of principal components extracted. The S_{press} value is weighted so that it penalizes models with a large number of principal components. In this study, the internal predictability was assessed with LOO CV, and the optimum number of principal components was selected on the basis of the first S_{press} minimum, with the constraint [39] that the maximum number of principal components does not exceed n/4. The final, non-cross-validated (fitted) models were derived using the optimum number of PLS components. Their statistical significance was evaluated using conventional statistical indicators: r^2 = squared correlation coefficient, SE = standard error, and F = Fischer test for significance. For external test sets, conventional squared correlation coefficients (r^2_{ex}), mean absolute deviations ($|\Delta|_{ave}$), predictive r^2-scores ($pr-r^2$, Equation 4) and $SDEP$ values (Equation 5) were calculated:

$$pr - r^2 = \frac{SD - PRESS}{SD} = 1 - \frac{PRESS}{SD} \quad (4)$$

$$SDEP = \sqrt{\frac{PRESS}{n}} \quad (5)$$

where SD is the sum of the squared deviations between the activities of molecules in the test set and the mean affinity of the training set molecules. In general, PLS is more resistant against chance correlations than the conventional multiple linear regression.

Colored noise in MD simulations of steroids

The results indicate that the fluctuations of the coordinates of the steroids are $1/f^{\alpha}$ noise. As can be seen from Figure 4, α-exponents differ always from zero, indicating that the atomic fluctuations are not Gaussian ('white') but 'colored' noise, i.e. the power spectra of the fluctuations depend very clearly on the frequency. As expected, the values of the α-exponents vary considerably along the molecule and between the molecules (Figure 4). In fact, the overall correlations are high enough so that PLS modeling, i.e., the desired 'dynamic' QSAR, is feasible. However, it is necessary for a predictive QSAR model that the side chain information is taken into account, as discussed above.

Qualitatively, molecules with light substituents such as oxygen and hydroxyl at position 17 of a steroid skeleton lead to low CBG activity, whereas the presence of a bulky chain such as $COCH_2OH$ enhances the activity. It is of interest that the α-exponents of C17 do

Table 4. The observed and predicted CBG affinities for the ALPHA Model 2 (training set **1–12**, **23–31**, test set **13–22**) and the corresponding PLS statistics.

Molecule	CBG affinity		
	Observed	Predicted	Deviation
13	5.00	5.98	−0.98
14	5.00	5.83	−0.83
15	5.00	5.17	−0.17
16	5.26	4.87	0.38
17	5.26	6.27	−1.01
18	5.00	6.35	−1.35
19	7.38	6.94	0.44
20	7.74	7.00	0.74
21	6.72	6.41	0.32
22	7.51	6.19	1.33

PLS statistics for the LOO CV: $S_{press} = 0.76$, $q^2 = 0.48$, $c = 2$; for the fitted model: $r^2 = 0.85$, $SE = 0.42$, $F = 49.2$; for the external prediction: $r^2_{ex} = 0.46$, $|\Delta|_{ave} = 0.75$, $SDEP = 0.85$, $pr{-}r^2 = 0.56$.

not correlate with the CBG activity at all, suggesting that steric and dynamic factors may be separable.

QSAR with α-exponents

The Gaussian width σ in Equation 1, the main parameter of ALPHA, must be optimized first. Previously, it has been found that related 'spectroscopic' QSAR methods, EVA [40] and EEVA [41], are very sensitive to the value of σ, and the optimum value is a dataset dependent feature. In this work, a total of 31 σ values were scanned (0.0025, 0.0050, 0.0075, 0.010, 0.015, 0.02, 0.025, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.10, 0.11, 0.12, 0.13, 0.14, 0.15, 0.16, 0.17, 0.18, 0.19, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45 and 0.50). A fairly large number (100) of PLS models were derived for each σ value by randomly choosing 21 steroids for the training set and placing the remaining 10 in the test set, resulting in 3100 different PLS models. The best performing σ values can be selected by examining the PLS statistics as a function of σ (Figure 5). Obviously, the range of 0.010–0.020 is optimal, although the overall dependence on σ is not very strong in this case.

Second, the internal predictability of ALPHA for the whole dataset was assessed with the LOO CV and the results are given in Table 1. Judging from this test, the overall performance of ALPHA can be considered as satisfactory.

Third, a large number of PLS models (500, Figures 6a–d) were derived as described above with the optimum value of σ (0.015). These runs were done in order to ensure that chance correlations are absent and to explore how dataset-dependent the PLS results are. The results indicate that both internal and external predictability of the ALPHA models varies from fair to good, depending on how the compounds **1–31** were divided into the training and test sets. The scattering of PLS statistics is considerable (Figures 6a–d), a reminder of how important well-balanced datasets are in QSAR studies (see below).

Finally, the reliability of ALPHA models was verified by scrambling the y variable several times (500), i.e. the activities of the training set compounds were mixed so that each y value is no longer assigned to the right ALPHA descriptor and repeating the LOO CV run. It appeared that no random combination yielded PLS statistics even close to the correct one: usually $q^2 < 0$ and only in very few cases q^2 values were above zero (mean $= -0.40$).

Standard benchmark tests for steroids

In line with most previous publications, the binding affinities for CBG of steroids **1–21** were used to derive the standard model, and steroids **22–31** formed the corresponding test set (Model 1). The predicted affinities and PLS statistics of Model 1 are given in Table 2. A comparison of ALPHA with some established QSAR methods [41–52] is presented in Table 3. Overall, the performance of ALPHA is comparable with most 3D QSAR methods. In particular, the notorious outlier in most previous QSAR studies, compound **31**, is almost correctly predicted with ALPHA (Table 2). On the other hand, the difference between the observed and predicted activity is quite large for compounds **22** and **26**, although the structural features of these compounds are not exceptional.

It has been emphasized by Kubinyi [53] that the molecules repeatedly used in the training set (compounds **1–21**) do not cover all structural features within the series. Instead, the recommended training set comprises molecules **1–12**, **23–31** and test set molecules **13–22**. This model will be referred to as ALPHA Model 2 (Table 4). The performance of ALPHA is slightly better with this new training set composition, in particular if the external predictability is considered.

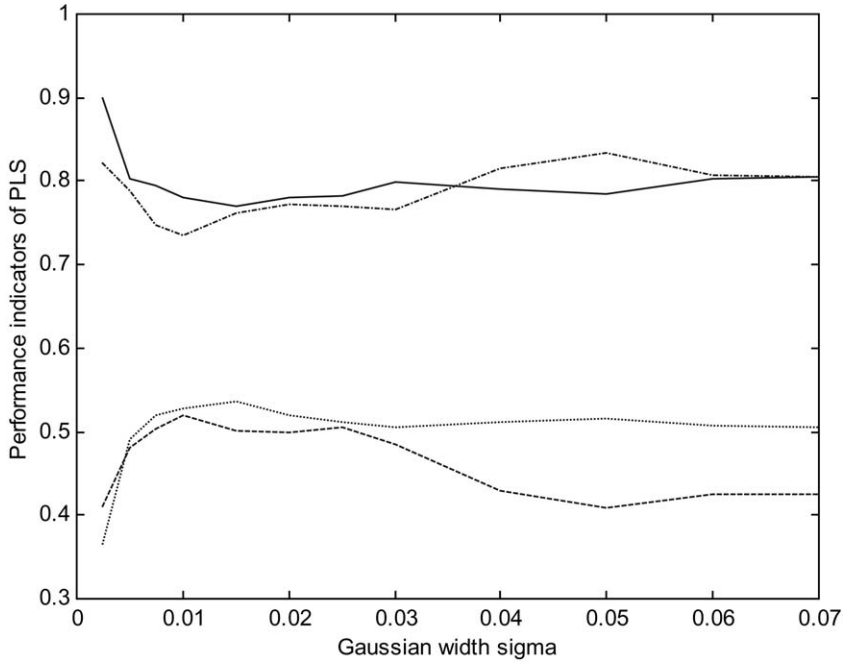

Figure 5. Performance indicators of PLS statistics as a function of σ: S_{press} (—), q^2 (· · · · ··), *SDEP* (-· - - - ·) and $pr{-}r^2$ (- - -).

Table 5. Correlations between the descriptor vectors.

	EVA	EEVA	SHAPE	ESP	POLAR	Free-Wilson	ALPHA
EVA	1.00	0.90	0.95	0.94	0.96	0.81	**0.52**
EEVA		1.00	0.87	0.91	0.90	0.92	**0.52**
SHAPE			1.00	0.98	0.99	0.74	**0.43**
ESP				1.00	0.98	0.83	**0.47**
POLAR					1.00	0.88	**0.46**
Free-Wilson						1.00	**0.47**
ALPHA							**1.00**

ALPHA vs. EVA, EEVA and SOMFA descriptors: α-SOMFA

Overall, the above results show inevitably that there exists a correlation between the activity and the ALPHA descriptor. However, this does not prove that there exists a sound physical relationship between the activity and molecular motions. On the basis of the above results we cannot say whether the ALPHA descriptor just reflects the motional effects on the activity or just structural and physical differences that correlate strongly with the ALPHA descriptor. To examine the problem more closely we computed the correlations between the ALPHA and EVA [40], EEVA [41], and 3D SOMFA (Self-Organizing Molecular Field Analysis) [46] descriptors (Table 5). In order to compare molecular descriptors of different length, all descriptor matrices were first compressed into one-dimensional vectors employing 'self-organizing regression'; for details, the reader is referred to the original SOMFA publication [46]. However, it should be emphasized that the predictive ability of EVA, EEVA and ALPHA is reduced considerably in the compression process (Table 6; the corresponding PLS models are presented for comparison). The results show, firstly, that the co-linearity between EVA, EEVA, and SOMFA descriptors is very large. This result is somewhat surprising, as their origins are completely different. Secondly, the modest (0.40–0.52) correlations of ALPHA with the other descriptors propose

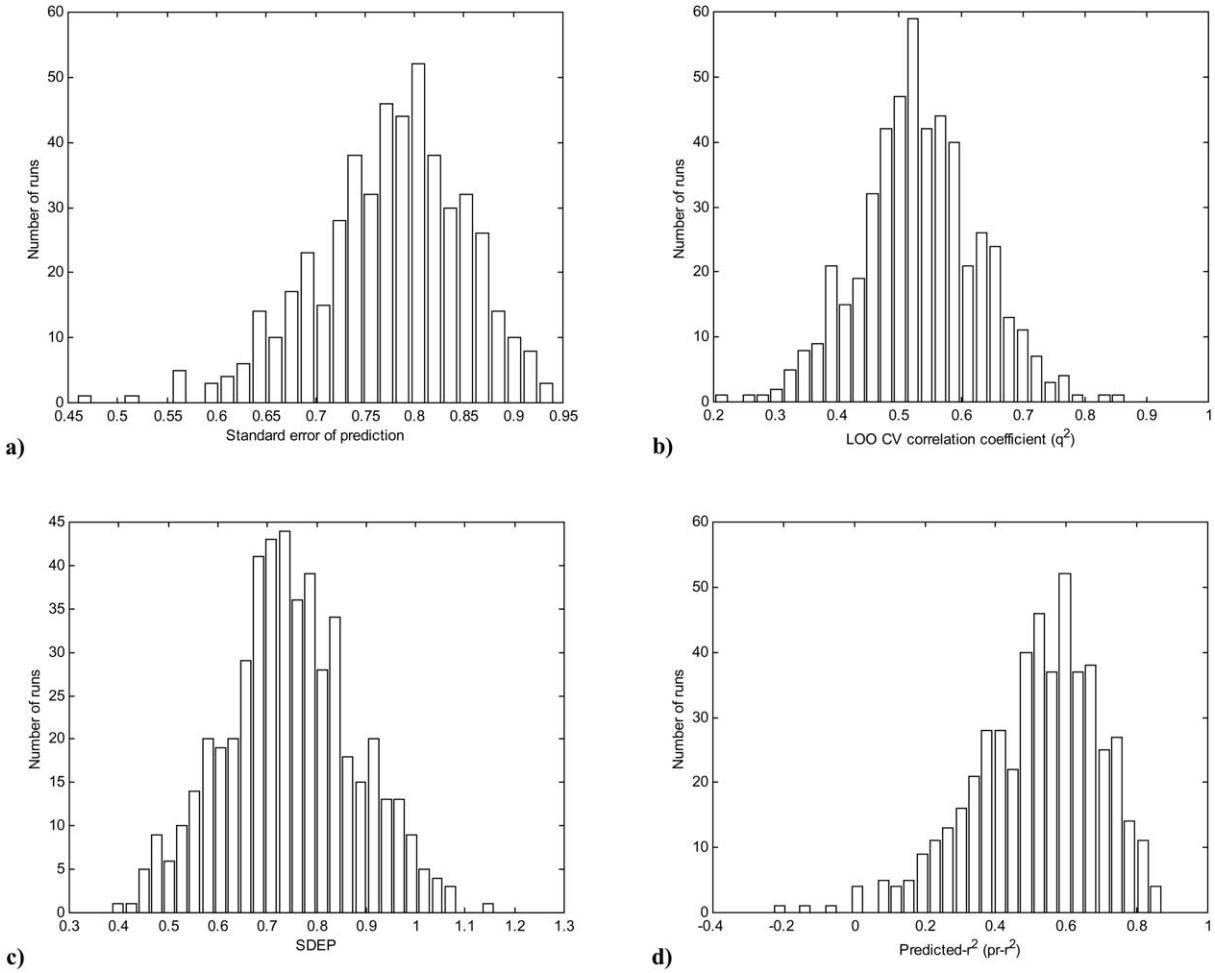

Figure 6. Histograms of the variation of S_{press} (a), q^2 (b), *SDEP* (c) and $pr{-}r^2$ (d) of 500 PLS models with $\sigma = 0.015$.

Table 6. EVA, EEVA, SOMFA/shape/esp/polar, and ALPHA models with PLS and SOR (self-organizing regression) statistics.

Mode	PLS			SOR	
	S_{press}	q^2	PC	S_{press}	q^2
EVA	0.57	0.77	5	0.69	0.60
EEVA	0.51	0.81	4	0.70	0.59
SOMFA/shape	0.62	0.70	3	0.63	0.67
SOMFA/esp	0.61	0.74	6	0.57	0.73
SOMFA/polar	0.61	0.71	2	0.64	0.65
ALPHA	0.73	0.57	2	0.86	0.39
EVA + ALPHA	–	–	–	0.55	0.75
EEVA + ALPHA	–	–	–	0.64	0.66
SOMFA/shape + ALPHA	–	–	–	0.50	0.80
SOMFA/esp + ALPHA	–	–	–	0.46	0.82
SOMFA/polar + ALPHA	–	–	–	0.54	0.76

Table 7. Comparison between ALPHA, SOMFA/ESP and α-SOMFA/ESP.

Experimental	ALPHA	|Δ|	SOMFA/ESP	|Δ|	α-SOMFA/ESP	|Δ|
6.28	6.97	0.69	6.48	0.20	6.76	0.48
5.00	5.09	0.09	5.82	0.82	5.27	0.27
5.00	5.31	0.31	5.57	0.57	5.14	0.14
5.76	5.45	0.32	6.43	0.67	5.97	0.20
5.61	6.42	0.81	5.49	0.12	5.67	0.06
7.88	6.86	1.02	7.49	0.39	7.55	0.33
7.88	6.96	0.92	7.26	0.62	7.42	0.46
6.89	7.19	0.30	7.34	0.44	7.59	0.70
5.00	5.29	0.29	5.26	0.26	4.86	0.14
7.65	7.15	0.51	7.09	0.56	7.37	0.28
7.88	7.25	0.64	7.30	0.58	7.60	0.28
5.92	5.70	0.22	5.73	0.19	5.49	0.43
5.00	6.46	1.46	4.45	0.55	4.86	0.14
5.00	6.97	1.97	4.52	0.48	5.23	0.23
5.00	7.28	2.28	4.54	0.46	5.44	0.44
5.25	5.49	0.23	5.53	0.28	5.21	0.04
5.25	5.90	0.64	5.46	0.21	5.37	0.11
5.00	5.34	0.34	5.71	0.71	5.29	0.29
7.38	6.95	0.43	7.00	0.38	7.18	0.20
7.74	6.88	0.86	7.23	0.51	7.35	0.39
6.72	6.68	0.04	6.58	0.15	6.69	0.03
7.51	7.05	0.46	7.14	0.37	7.36	0.15
7.55	6.82	0.74	7.30	0.25	7.37	0.18
6.78	7.21	0.43	7.03	0.25	7.34	0.56
7.20	6.08	1.12	6.78	0.42	6.56	0.64
6.14	5.17	0.97	5.91	0.23	5.36	0.78
6.25	6.30	0.05	7.44	1.20	7.22	0.97
7.12	6.13	0.99	6.86	0.26	6.65	0.47
6.82	6.62	0.20	6.20	0.62	6.35	0.46
7.69	7.16	0.53	7.59	0.09	7.80	0.11
5.80	5.85	0.06	7.29	1.49	6.91	1.11
$\|\Delta\|_{ave}$		0.64		0.46		0.36
$\|\Delta\|_{\max}$		2.28		1.49		1.11
Corr. coeff.		0.63		0.86		0.91

that ALPHA really represents a new molecular dimension and the information provided by ALPHA is only partially redundant with other descriptors for the present dataset. On the other hand, even the modest correlations mean that molecular motions are to a good extent accounted implicitly in the models based on the other descriptors. An additional piece of evidence that the ALPHA descriptor really represents a new independent molecular property is obtained from Tables 6 and 7: inclusion of the ALPHA descriptor in the EVA, EEVA and SOMFA models very clearly improves the predictive ability. The best combined model, α-SOMFA/ESP, is presented in Table 7.

Cramer's steroids as a benchmark – further validation with extended steroid data set

Cramer's steroids have become a *de facto* benchmark test set for evaluating the performance of different QSAR methods. We have recognized the criticism on the use of this set as a benchmark [28, 50, 54]. As emphasized by a reviewer, 'anything goes' with Cramer's steroids, i.e. it is difficult even to find a

Table 8. Comparison between ALPHA, SOMFA/ESP and α-SOMFA/ESP for ALPHA Model 3 (training set **1**–**31**, test set **32**–**42**).

No.	Compound	Observed	SOMFA/ESP	ALPHA	α-SOMFA/ESP
32	Danazol	6.81	5.92	6.38	6.02
33	Estradiol benzoate	5.00	4.72	6.53	5.61
34	Ethisterone	5.32	6.85	6.31	6.62
35	Fluoxymesterone	5.00	6.45	6.75	6.47
36	Medroxypregesterone	6.91	7.29	7.09	7.22
37	Methyltestosterone	5.00	6.75	6.35	6.48
38	Methyltrienolone	5.36	5.93	5.66	5.92
39	Norethindrone	5.26	6.24	5.72	6.19
40	Norethynodrel	5.00	5.43	6.50	5.68
41	Norgestrel	5.00	6.47	5.88	6.27
42	Predinsone	6.51	7.26	6.69	7.23
	Corr coeff.		0.44	0.44	0.56
	SDEP		1.07	1.03	0.99

QSAR method that would not work with this particular data set. For example, Kubinyi has shown [28, 53] that a simple one-parameter Free–Wilson analysis, with the presence or absence of the cycloaliphatic 4,5-double bond being used as an independent variable, leads to a highly predictive model. However, the correlation between the 4,5-double bond Free–Wilson and ALPHA parameters is only 0.47, while the correlations with the other parameters given in Table 5 are much higher. This indicates that ALPHA represents a new dimension also in this respect.

Another object of the critics is the use of the arbitrary values (log K = 5.00) that have been assigned for the seven inactive molecules in the training set, for which no binding could be actually observed. The problem is that all the methods involved in this study (EVA, EEVA, SOMFA, and ALPHA, most data not shown) predict some inactive compounds too active, especially for the extended set. In this context, we examined the influence of the activity assignments for inactive compounds (in the training set) by replacing the value of 5.00 with 4.00 and 4.50, respectively. This procedure led only to a slight improvement in q^2, r^2_{ex} and $pr{-}r^2$ values, being, however, clearly detrimental for S_{press}, Δ_{ave} and *SEDP* values, and so it was abandoned.

For the above reasons, and also in line with the original SOMFA study [46], we also used an additional external test set of 11 steroids (for some unknown reason, compound No. 22 and 42, prednisolone, was present in duplicate with slightly different activities, and so it was dropped). The results are given in Table 8. Again, a clear improvement was achieved by combining the SOMFA/ESP and ALPHA descriptors.

In general, the applicability of ALPHA (or any other QSAR method) requires that the correlations between descriptors and biological activities are large enough to facilitate the derivation of predictive QSAR models. This can be examined preliminarily, for example, by employing a simple correlation analysis and a subsequent graphic representation of the correlations between descriptor vectors to see if the training set compounds are classed together correctly with their biological activities. This indicates that the validity of ALPHA is very good for the GBG steroids. However, extension of the present method to other data sets than steroids is not straightforward because it obviously, in its present formulation, demands comparison of motions of well-defined centers like steroid backbone carbon atoms.

Physical nature of ALPHA descriptor

In general, physical systems in which many particles are relaxing from high-energy or excited states at different rates are well known generators of 1/f-like noise [55]. The relevance of this general scheme to the atomic fluctuations during an MD simulation is almost self-evident. The dominant factors in MD simulations resulting in the $1/f^{\alpha}$ noise are (i) the nonlinear interactions inherent in the force field, (ii) the restraints on temperature, bond lengths, bond angles and dihedral angles, and (iii) the stochastic forces generated by the

solvent (not appropriate in this case, as the simulations were done in vacuum). Recently, it has been shown that the potential energy fluctuations during an MD simulation of a protein (plastocyanin) are 1/f noise [36, 37], and the same phenomenon has been observed in a helix–coil transition of polypeptides [56]. Our results indicate that the same holds for small organic molecules (ligands), both for energy (data not shown) and, more importantly, for individual coordinates. The presence of $1/f^{\alpha}$ noise in an MD time series is a hallmark of self-organization and indicates that the fluctuations actually reduce to a few collective degrees of freedom. In fact, principal components analysis of the coordinate fluctuations has independently corroborated this on several occasions [1].

From a physical point of view, the presence of $1/f^{\alpha}$ noise is closely related to other scale-invariance properties of dynamical systems, i.e., the memory effects (or complexity or self-similarity) described by the Hurst exponent (H), and the fractal-like behavior described by the fractal dimension (D) [57]. In fact, all scaling exponents α, H and D of a time series can be estimated employing the same mathematical procedure described above. The relations between the exponents are straightforward, i.e., $H = (\alpha - 1)/2$ [58] and $D = (5 - \alpha)/2$ [59], remembering that H makes no sense if α is smaller than 1 and D makes no sense unless α is in the range of (1, 3). Moreover, which kind of scaling is actually present must be determined separately by using other information than the original time series data. Anyhow, the nature of scaling does not have an influence on the use of the scaling exponents as QSAR descriptor. Thus, the physical meaning of α-exponents seems relatively straightforward.

The uniqueness of the ALPHA descriptor probably lies in that it is not directly related to the substituent effect, which forms the basis of almost all other QSAR methods, i.e. α-exponent is a holistic molecular property. This also means that it is not possible to predict the influence of a substituent on the α-exponents *a priori*, but they must be calculated using the procedure described above. The dimension represented by the ALPHA descriptor could also be called the 5^{th} dimension, or the color, of the molecule, because it is not simply related to the 3D structure or the motions of the molecule. To go on this analogy, if some data set there has no correlation with ALPHA, it just means that the biological system behind the data is color-blind. Finally, it should be emphasized that the calculations of this study were done in vacuo, but they could be easily extended to solvent–ligand–receptor systems. In the latter case, α-exponents may be employed to explore whether there are e.g. resonance-type interactions between ligands and receptors – an intriguing possibility bearing in mind that the most specific recognition mechanisms in nature are of the resonance type.

Conclusions

Our results propose that there really exists a correlation between the ligand motions (or the property represented by ALPHA) and binding affinity: the dynamic ALPHA descriptor based on α-exponents clearly improves the models obtained with standard static 3D QSAR descriptors. The only modest (42–50%) correlations of ALPHA with static 3D and electronic descriptors mean that ALPHA forms an independent molecular property, which offers a mechanism for molecular regulation and recognition. This new dimension increases the otherwise apparently poor structural diversity of steroids, bringing a new dimension also to their design. On the other hand, although the correlations of the ALPHA and the other QSAR descriptors were modest, they are large enough so that the ligand motional properties are actually implicitly included in the static 3D QSAR models and vice versa. In general, ALPHA is computationally simple, easy to use and invariant as to the alignment of the structures concerned. It is our hope that the results, conjectures and interpretative comments of this study would stimulate further work in this area – including molecular design.

Acknowledgements

This work was supported by the Academy of Finland (grant 74097 to M.P.).

References

1. Kitao, A. and Go, N., Curr. Opin. Struct. Biol., 9 (1999) 164 and references therein.
2. Teaque, S.J., Nat. Rev., 2 (2003) 527.
3. Editorial, Nat. Struct. Biol., 7 (2000) 701.
4. Falke, J.J., Science, 295 (2002) 480.
5. Eisenmesser, E.Z., Bosco, D.A., Akke, M. and Kern, D., Science, 22 (2002) 1520.
6. Miller, D.W. and Agard, D.A., J. Mol. Biol., 286 (1999) 267.
7. Mincer, J.S. and Schwartz, S.D., J. Proteome Res., 2 (2003) 437.

8. Watney, J.B., Agarwal, P.K. and Hammes-Schiffer, S., J. Am. Chem. Soc., 125 (2003) 3745.
9. Demirel, M.C., Atilgan, A.R., Jernigan, R.L., Erman, B. and Bahar, I., Protein Sci., 7 (1998) 2522.
10. Bahar, I., Atilgan, A.R., Demirel, M.C. and Erman, B., Phys. Rev. Lett., 80 (1998) 2733.
11. Laatikainen, R., Saarela, J., Tuppurainen, K. and Hassinen, T., Biophys. Chem., 73 (1998) 1.
12. Saarela, J.A., Tuppurainen, K., Peräkylä, M., Santa, H. and Laatikainen, R., Biophys. Chem., 95 (2002) 49.
13. Anfinsen, C.B., Science, 181 (1973) 223.
14. Temiz, N.A. and Bahar, I., Proteins Struct. Funct. Genet., 49 (2002) 61.
15. For a critical approach, see Sharp, K., Protein Sci., 10 (2001) 661; Cornish-Bowden, A., J. Biosci., 27 (2002) 121.
16. Hopfinger, A.J., J. Am. Chem. Soc., 102 (1980) 7196.
17. Hopfinger, A.J. and Potenzone, R.J. Jr., Mol. Pharmacol., 21 (1982) 187.
18. Mabilia, M., Pearlstein, R.A. and Hopfinger, A.J., Eur. J. Med. Chem., 20 (1985) 163.
19. Avbelj, F. and Hadzi, D., Mol. Pharmacol., 27 (1985) 466.
20. Lopez de Compadre, R.L., Pearlstein, R.A., Hopfinger, A.J. and Seydel, J.K., J. Med. Chem., 30 (1987) 900.
21. Hopfinger, A.J., Lopez de Compadre, R.L., Koehler, M.G., Emery, S. and Seydel, J.K., Quant. Struct.-Act. Relat., 6 (1987) 111.
22. Sasaki, Y., Takagi, T., Yamasoto, Y., Iwata, A. and Kawaki, H., Chem. Pharm. Bull., 29 (1981) 3073.
23. Kier, L.B., Quant. Struct.-Act. Relat., 8 (1989) 221.
24. Hopfinger, A.J., Wang, S., Tokarski, J.S., Jin, B., Albuquerque, M., Madhav, J. and Duraiswan, C., J. Am. Chem. Soc., 119 (1997) 10509.
25. Hudson, B.D., George, A., Ford, M.G. and Livingstone, D.J., J. Comput.-Aided Mol. Design, 6 (1992) 191.
26. Dunn, J.F., Nisula, B.C. and Rodbard, D., J. Clin. Endocrin. Metab., 53 (1981) 58.
27. Mickelson, K.E., Forsthoefel, J. and Westphal, U., Biochemistry, 20 (1981) 6211.
28. Coats, E., Perspect. Drug Discov. Design, 12/13/14 (1998) 199.
29. SYBYL, Tripos Inc., St. Louis, MO.
30. Wang, T.W., Khettry, A., Berry, M. and Batra, J., The First International Chemometrics InterNet Conference, InCINC'94. (The MATLAB code of SVDPLS can be found from the WWW site http://www.emsl.pnl.gov:2080/docs/incinc/papers/wang_pls/sectionstar3_9.html)
31. Tuppurainen, K., SAR QSAR Environ. Res., 10 (1999) 39.
32. Tuppurainen, K. and Ruuskanen, J., Chemosphere, 41 (2000) 843.
33. Turner, D.B. and Willett, P., Eur. J. Med. Chem., 35 (2000) 367.
34. Bursi, R., Dao, T., van Wijk, T., de Gooyer, M., Kellenbach, E. and Verwer, P., J. Chem. Inf. Comput. Sci., 39 (1999) 861.
35. Kantz, H. and Schreiber, T., Nonlinear Time Series Analysis, Cambridge University Press, New York, 1997.
36. Bizzarri, A.R. and Cannistraro, S., Phys. Lett. A, 236 (1997) 596.
37. Bizzarri, A.R. and Cannistraro, S., Physica A, 267 (1999) 257.
38. Höskuldsson, A., J. Chemometrics, 2 (1988) 211.
39. Wold, S., Johansson, E. and Cocchi, M., In Kubinyi, H. (Ed.), 3D QSAR in Drug Design: Theory, Method and Applications, ESCOM, Leiden, The Netherlands, 1993, p. 523.
40. Turner, D.B, Willett, P., Ferguson, A.M. and Heritage, T., J. Comput.-Aided Mol. Design, 11 (1997) 409.
41. Tuppurainen, K., Viisas, M., Peräkylä, M. and Laatikainen, R., J. Chem. Inf. Comput. Sci., 42 (2002) 607.
42. Jain, A.N., Koile, K. and Chapman, D., J. Med. Chem., 37 (1994) 2315.
43. Bravi, G., Gancia, E., Mascagni, P., Pegna, M., Todeschini, R. and Zaliani, A., J. Comput.-Aided Mol. Design, 11 (1997) 79.
44. Chen, H., Zhou, J. and Xie, G., J. Chem. Inf. Comput. Sci., 38 (1998) 243.
45. Robert, D., Amat, L. and Carbo-Dorca, R., J. Chem. Inf. Comput. Sci., 39 (1999) 333.
46. Robinson, D.D., Winn, P.J., Lyne, P.D. and Richards, W.G., J. Med. Chem., 42 (1999) 573.
47. Turner, D.B., Willett, P., Ferguson, A.M. and Heritage, T.W., J. Comput.-Aided Mol. Design, 13 (1999) 271.
48. Pastor, M., Cruciani, G., McLay, I., Pickett, S. and Clementi, S., J. Med. Chem., 43 (2000) 3233.
49. Palyulin, V.A., Radchenko, E.V. and Zefirov, N.S., J. Chem. Inf. Comput. Sci., 40 (2000) 659.
50. Polanski, J. and Walczak, B., Comput. Chem., 24 (2000) 615.
51. Liu, S.-S., Yin, C.-S., Li, Z. and Cai, S.-X., J. Chem. Inf. Comput. Sci., 41 (2001) 321.
52. Amat, L., Besalu, E., Carbo-Dorca, R. and Ponec, R., J. Chem. Inf. Comput. Sci., 41 (2001) 978.
53. Kubinyi, H., In van de Waterbeemd, H., Testa, B. and Folkers, G. (Eds.), Computer-Assisted Lead Finding and Optimization: Current Tools of Medicinal Chemistry, VHCA, Basel, Switzerland, and Wiley-VCH, Weinheim, Germany, 1997, p. 7.
54. Wagener, M., Sadowski, J. and Gasteiger, J., J. Am. Chem. Soc., 117 (1995) 7769.
55. Schroeder, M., Fractals, Chaos, Power Laws, W. Freeman, New York, 1991.
56. Takano, M., Takahashi, T. and Nagayama, K., Phys. Rev. Lett., 80 (1998) 5691.
57. Feder, J., Fractals, Plenum Press, New York, 1988.
58. Voss, R.F., In Peitgen, H.-O. and Saupe, D. (Eds.), The Science of Fractal Images, Springer-Verlag, New York, 1988, p. 47.
59. Saupe, D., In Peitgen, H.-O. and Saupe, D. (Eds.), The Science of Fractal Images, Springer-Verlag, New York, 1988, p. 91.