Journal of Computer-Aided Molecular Design **18:** 437–449, 2004.
DOI 10.1007/s10822-004-4060-8

An automated PLS search for biologically relevant QSAR descriptors

Marius Olah[a], Cristian Bologa[a] & Tudor I. Oprea[a,b,*]
[a]*Division of Biocomputing, University of New Mexico School of Medicine, 1 University of New Mexico, MSC08 4560, Albuquerque, NM 87131, USA*
[b]*Sunset Molecular Discovery LLC, 1704 B Llano Street, S-te 140, Santa Fe, NM 87505, USA*

Received 19 April 2004; accepted in revised form 8 September 2004

Key words: data-mining, fingerprints, PLS, QSAR, SMARTS, SMILES, topological indices, WOMBAT

Summary

An automated PLS engine, WB-PLS, was applied to 1632 QSAR series with at least 25 compounds per series extracted from WOMBAT (WOrld of Molecular BioAcTivity). WB-PLS extracts a single Y variable per series, as well as pre-computed X variables from a table. The table contained 2D descriptors, the drug-like MDL 320 keys as implemented in the Mesa A&C Fingerprint module, and in-house generated topological-pharmacophore SMARTS counts and fingerprints. Each descriptor type was treated as a block, with or without scaling. Cross-validation, variable importance on projections (VIP) above 0.8 and $q^2 \geq 0.3$ were applied for model significance. Among cross-validation methods, leave-one-in-seven-out (CV7) is a better measure of model significance, compared to leave-one-out (measuring redundancy) and leave-half-out (too restrictive). SMARTS counts overlap with 2D descriptors (having a more quantitative nature), whereas MDL keys overlap with in-house fingerprints (both are more qualitative). The SMARTS counts is the most effective descriptor system, when compared to the other three. At the individual level, size-related descriptors and topological indices (in the 2D property space), and branched SMARTS, aromatic and ring atom types and halogens are found to be most relevant according to the VIP criterion.

Introduction

Started by Corwin Hansch in 1948, the area of biological QSARs has evolved from the Hansch–Fujita [1] and Free–Wilson [2] paradigm – where models were based on a few descriptors with clear physicochemical meaning – to models based on large numbers of abstract descriptors, in parallel with the increase in CPU power and availability. There are currently over 3000 molecular descriptors [3] that have been used in QSAR (Quantitative Structure Activity Relationship) studies [4, 5]. If one takes into account the number of descriptors, their possible combinations and the number of statistical techniques available, the number of theoretical models one could derive can quickly reach 10^{10} [6]. While the situation may appear to benefit the researchers in the field, it is more likely to create confusion, yielding computer-driven explanatory (not predictive) models where no hypotheses were tested, and where human input is reduced to a mere table with structures and activities. In this paper, we limit ourselves to the discussion of one statistical technique and a relatively small number of descriptor classes, and explore the value of automated QSAR in the attempt to identify effective descriptor systems for coarse-grained (as opposed to fine-tuned) QSAR modeling.

Two-dimensional (2D) descriptors, i.e., descriptors that do not use information related to the three-dimensional (3D) characteristics of model compounds, can be classified as:

*To whom correspondence should be addressed. Fax: +1-505-272-8738; E-mail: toprea@salud.unm.edu

1. *Size-related:* molecular weight – MW; calculated molecular refractivity – CMR [7]; molecular volume and molecular surface area, pre-computed from tabulated values (e.g., using Van der Waals radii), etc.;
2. *Hydrophobicity-related:* the logarithm of the octanol–water partition coefficient, Log *P* [8] – besides CLOGP [9], several other Log *P* estimating programs are available [10]; the π fragmental constant [1]; the logarithm of the (molar) aqueous solubility [11, 12] (Log S_w);
3. Descriptors related to *electronic effects:* CMR; the (tabulated) estimated polarizability [13]; Hückel-level estimates of the highest-occupied, and lowest-unoccupied, molecular orbitals; partial atomic charges based on electronegativity equilibration schemes [14, 15], e.g., Gasteiger–Marsili [16]; counts of positive or negative ionic centers; total number of valence electrons [17], etc.
4. *Hydrogen bonding descriptors* that estimate the basicity or acidity factors, e.g., the HYBOT [18, 19] or Abraham descriptors [20], or electrotopological (E-state) descriptors [21], or counts [22] of hydrogen bond acceptors or donors;
5. *Topological indices* [23] derived from connectivity [24] matrices [25, 26].

The above types of descriptors have been successfully used to derive QSAR models for the past four decades. The third dimension is typically investigated by analyzing conformers via 3D-QSAR methods: CoMFA [27] (comparative molecular field analysis), the GRID [28] – PLS (partial least squares) [29] combination and other 3D-QSAR methods [30–32] try to explain the variance in biological activity by monitoring variations in the 3D structures of chemical compounds. Graphical analysis [33] remains one of the main advantages of 3D-QSAR methods.

Even though variable selection has been used with some success in QSAR technologies [34–38], the question: ''Which descriptors should I use first, when seeking a structure–activity relationship?'' has not been properly addressed. This question is relevant if one considers that there are hundreds of combinations of descriptors and methodologies that have been developed to map the chemistry:biology interface. Variable selection operates under the assumption that, if out of K descriptors, L (where $L \leq K$) yield significant QSARs, then these L descriptors are appropriate to use in that series, or class of series, or perhaps class of biological targets.

However, this type of reasoning does not address the question: How to tackle an unknown series/target, or which *initial* set of descriptors to use before running comparative analyses for all possible combinations of *available* descriptors. There is no comprehensive study that evaluates the value of various descriptor systems for *initial* relevance. In this paper, we begin to address this question by studying two major categories of descriptors: the 2D descriptors, and the '1D' descriptors, based on chemical substructures. The term '1D' relates to the fact that many such chemical substructures can be derived from a linear notation, e.g., SMILES [39]. We performed an automated PLS analysis using WB-PLS [40] on 1632 QSAR series from the WOMBAT 2004.1 database [41], having $N \geq 25$ compounds per series. We used 2D descriptors, the 'drug-like' MDL 320 keys [42], and in-house generated SMARTS [43] counts (termed Q504) and binary (F504) descriptors. We compare the success of each method across all series that yield significant QSARs according to three cross-validation methods, and attempt to answer the question of *initial* relevance in biological QSARs by showing which descriptors, or classes of descriptors, are more successful.

Materials and methods

The WOMBAT database [41], version 2004.1, was used as input basis for all further calculations. Each root record in this database consists of one *chemical structure* and one or more associated *biological activity data* sub-records (Figure 1). At the root level, information about the original bibliographic reference from where the data was indexed is recorded, together with several calculated and experimental properties for each structure, e.g., counts of miscellaneous atom types, Lipinski's rule-of-five (RO5) parameters [44], log *P*s [8], water solubilities [11], polar/non-polar surface descriptors (PSA, NPSA), etc. One field, series ID (SID), has the same value for, and links all the root records indexed from one reference (article). In WOMBAT.2004.1, SID takes values between 1 and 3039. Each activity sub-record has the following fields: the activity identifier (AID,

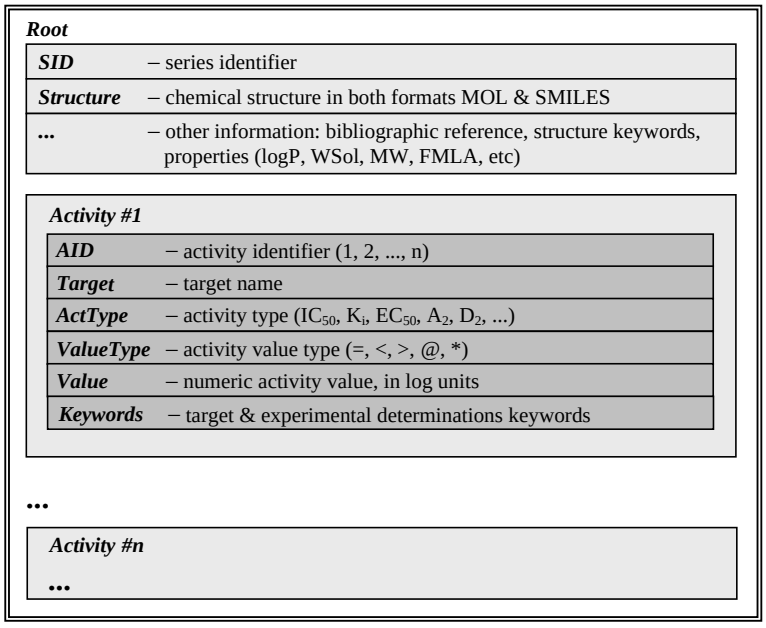

Figure 1. Simplified WOMBAT database tree.

with values from 1 to n, where n is the number of biological activity determination sets), *Target* (the target name on which the activity was measured), *ActType* (the activity type), *ValueType* (which can be one of: exact (=), lower than (<), greater than (>), % inhibition at a given concentration (@), or inactive (*)), *Value* (the numeric value of activity, in negative logarithmic units) and *Keywords* (referring to the target and experimental determination like tissue or cell type, used radioligand, etc.). For one series (with one SID value for all structures), each activity block $(1, \ldots, n)$ has the same *Target*, *ActType*, *ValueType* and *Keywords* associated data. An integrated methodology (see Figure 2) was applied for automatic extraction of the individual QSAR datasets from the WOMBAT database and the descriptors matrix, and for PLS processing of these datasets.

A preliminary filtering operation was applied on the database in order to extract only records with exact activity values and the following activity types: IC_{50}, K_i, EC_{50}, A_2, and D_2. The number of root records with at least one biological activity sub-record was thus reduced from 76,176 to 70,058. From the reduced dataset, the descriptors matrix was calculated in *step 1* (Figures 3 and 4). Using WB-PLS, QSAR series were automatically extracted from the reduced WOMBAT dataset. Two constraints were applied in this step: (a) at least 25 structures with biological activity records in one QSAR series; (b) one activity column present for all molecules in each QSAR series. For series with multiple activities, one QSAR series was generated for each activity block with respecting the first condition. The 1632 QSAR series satisfying these criteria were used as input to the PLS engine.

Descriptors

The 2DProp descriptors

Eighty descriptors of different types were computed starting from 2D structure: size-related descriptors included MW, the number of heavy atoms, the number of carbons, and CMR [7]. Polarizability is estimated by CMR and by an atom-based scheme [13]. Flexibility and rigidity are estimated [22] by counting the total number of bonds, the number of rings and the number of rotatable bonds and the number of rigid bonds, and by several topological indices that estimate other properties [26] as well. The Wiener, Balaban, Randič and Motoc indices, as well as the Kier and Hall suite of connectivity descriptors [24] are also computed. Furthermore, the descriptors set contains simple counts for oxygen, nitrogen, H-bond

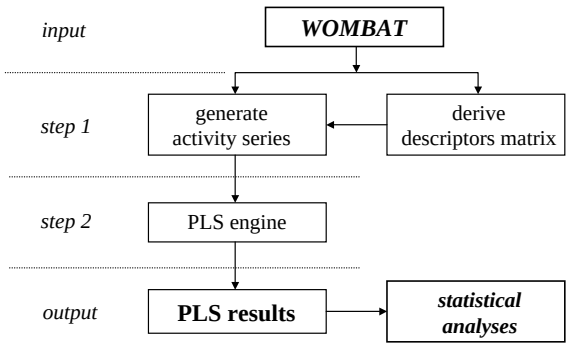

Figure 2. The WB-PLS flowchart.

donors and H-bond acceptors, positive and negative ionization centers, as well as the maximum positive and negative charge, as calculated using the Gasteiger–Marsili method [16]. Hydrophobicity was estimated using CLOGP [9]. Hydrogen-bond donor and acceptor counts are derived from a look-up table of fragments known to be involved in hydrogen bonding, limited to nitrogen and oxygen [22]. Other donors, such as thiols, or acceptors, such as halides, are ignored. Donor counts include all N–H and O–H fragments, excepting all acids (considered deprotonated). Amide and amide-like (e.g., urea, sulfonamide) nitrogens, as well as tertiary amines, are not counted as H-bond acceptors. Since no pK_a estimator is included in this scheme, protonation states are not considered (e.g., amines are not protonated). However, donors and acceptors were counted separately, meaning that an O–H group can be both a donor and an acceptor [22].

MDL 320 fingerprints keys (F320)

Binary representations for chemical structures characterization are called *molecular fingerprints*. The MDL fingerprints (also termed keys or keybits) use a pre-defined set of definitions and derive fingerprints based on pattern matching of the structure to the defined 'keyset' [42]. The keysets used as descriptors were generated with the Mesa Analytics and Computing *Fingerprint* Module, which reads SMILES strings as input and creates 320 bit structural MDL-MACCS key representations [45].

Frequency and fingerprint descriptors (Q504 and F504)

Inspired by the work from MDL [42] and by the CATS (chemically advanced template search) concept [46, 47], we explored various SMARTS [43] subsets, in an effort to systematically capture topologically relevant pharmacophore-based substructures. F504 is a fingerprint-like descriptor system (binary), whereas Q504 includes counts for each substructure – for a total of 504 different SMARTS. The number of matches of these SMARTS patterns in the WOMBAT SMILES were computed using an in-house routine based on functions available in the Daylight Toolkit [48] and OEChem [49] libraries – see Figure 5.

The PLS engine

We implemented the PLS-method [50–52] (NIPALS algorithm) in WB-PLS, in order to process each QSAR series automatically. Cross-validation was used for model validation and predictivity [53, 54]. In WB-PLS, no human intervention was required – a process that has some shortcomings, as discussed below.

Data preprocessing

While PLS tolerates moderate numbers of missing values both in the X and the Y blocks [29], not all descriptors could be calculated for all the

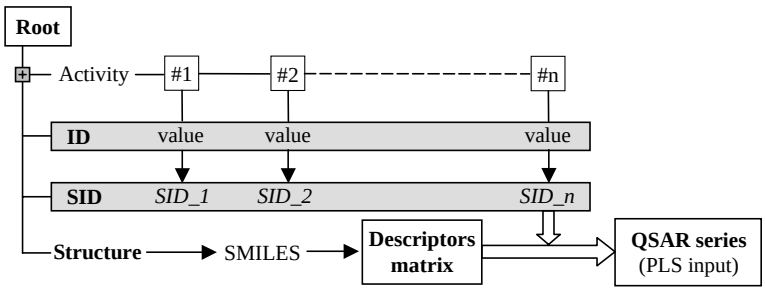

Figure 3. The QSAR series generation process.

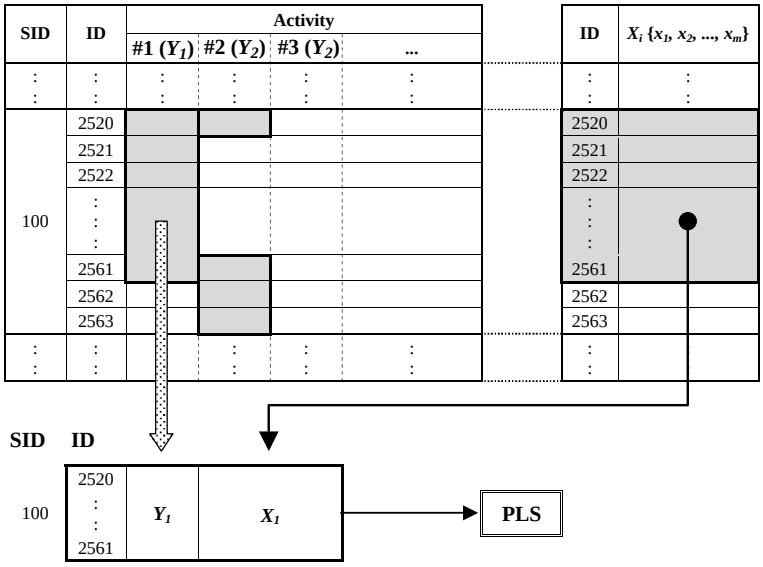

Figure 4. Example of QSAR series generation. SID, series identifier; ID, chemical structure identifier, used as link field between the two tables; Y_i, numeric values for activities; X_i, descriptor values. In this example the database series SID = 100 has two activity blocks. Activity #1 has 42 values and activity #2 has only 4. Only the QSAR series corresponding to activity #1 will be generated because it satisfies the minimum series length condition.

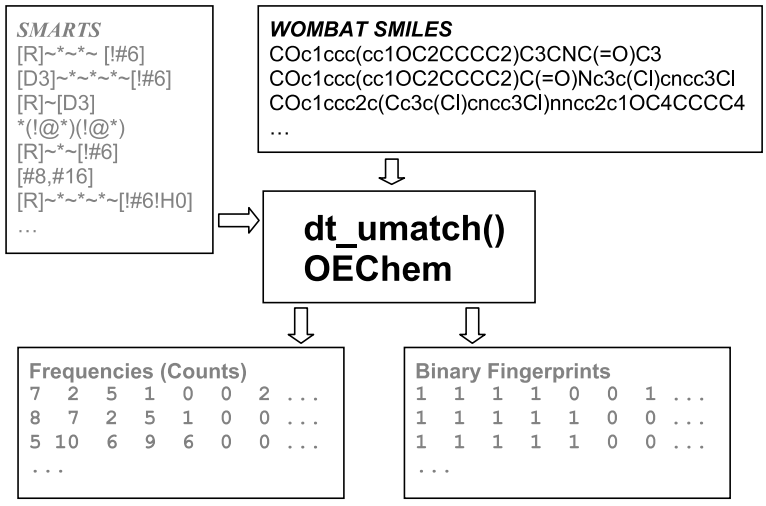

Figure 5. Fingerprints and frequencies generating process.

structures (~0.2% for 2DProp descriptors). The missing elements were substituted with the average value calculated by columns from the existing elements. The unit variance scaling was used for 2DProp and Q504 descriptors, while no scaling was applied for binary descriptors (F320 and F504). No other weighting or scaling was applied, nor any form of block scaling. We autoscaled Q504 columns because some SMARTS patterns occur more frequently than others, hence some columns have values between 0 and 20 (e.g., $-CH_3$), whereas others will have a much smaller range (e.g., –SH is likely to be between 0 and 3). Without scaling, the entire analysis would be biased towards more frequently occurring SMARTS.

Autofit, cross-validation and statistical parameters

The optimal dimensionality (*PC*#) of each model was determined from the evolution of the cross-validation statistic parameter (q^2) as a function

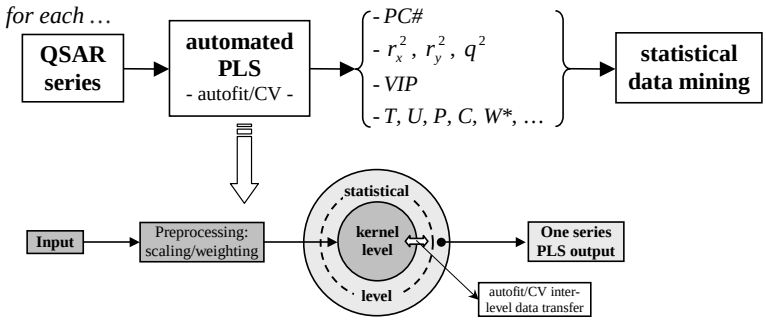

Figure 6. The automated PLS scheme.

of the number of latent variables extracted. The iteration process stops when for the extracting component, q^2 is negative, or has an increase less than 5% relative to the previous component. Cross-validation is a model evaluation method that is better than goodness of fit (r^2). The problem with residual evaluations is that they do not give an indication of how well the QSAR performs when it is asked to make predictions (i.e., for data it has not already seen). One way to overcome this problem is not using the entire data set in the training set. By removing some of the data before deriving the QSAR, this data can be used to test the performance of the model ('cross-validation'). The algorithm can be described as follows:

(a) split randomly by rows the dataset, with N observables in a given number of groups of approximately the same size; based on group size, three situations are possible: LOO – leave-one-out, LHO – leave-half-out, named here CV2, and LnO – leave-n-out. In this study, $n \cong N/7$, hence this method is termed CV7.
(b) derive a model from the *training set* obtained by eliminating one *test* group;
(c) compute the predicted values for the *test set*;
(d) repeat steps (b) and (c) until every element from the whole dataset is excluded once and only once;
(e) repeat steps (a)–(d) 100 times for CV2 and 30 times for CV7;
(f) get the final model using the whole dataset for the dimensionality found to be optimal through the entire previous process.

The performance statistics [55] for the model's characterization were computed as follows:

r_x^2, r_y^2 – the cumulative sum of squares of all X's and Y's explained by all extracted components;
q^2 – the cumulative fraction of the variance in the set-aside Y's that can be predicted by all components;
VIP – the variable importance in the projection. VIP values reflect the importance of terms in the model; according to Eriksson et al., X-variables could be classified according to their relevance in explaining Y: VIP > 1.0 (highly influential), 0.8 < VIP < 1.0 (moderately influential) and VIP < 0.8 (less influential) [56].

The PLS module

The module is structured in two levels. Level 1 – *kernel* – is a very efficient implementation of the standard NIPALS algorithm [57] used for successive principal components extraction. The following matrices are calculated and stored: T (X-scores matrix), U (Y-scores matrix), P (X-loadings matrix), C (Y-weights matrix), W (X-weights matrix), W^* (transformed PLS weights matrix). Level 2 – *statistics* – is used for statistical parameter calculations based on the data deduced in level 1 (Figure 6). The WB-PLS program was written in C++ and compiled for execution on a Pentium 4/2 GHz system under the Windows XP operating system. About one QSAR per second was generated including autofit with cross-validation.

Y-distribution analysis

The distribution of Y values across all series was analyzed in order to understand its influence on QSARs. Instead of applying Y-distribution cut-offs to eliminate series from the study, we wanted to monitor how Y values influence the outcome of

PLS analyses. Therefore, the discussion about Y-distribution and its influence follows the main discussion part. To check the Y-distribution, we applied two successive tests to each series: (a) the Y range, ΔY, as difference between the maximum and the minimum activity values must be at least two log units; (b) the bin test verifies the minimum number of Y activity values on each logarithmic unit as follows:

$$n_{\min} = \begin{cases} 4, & 2.0 \leq \Delta Y \leq 2.5, \\ 3, & 2.5 < \Delta Y \leq 3.5, \\ 2, & 3.5 < \Delta Y. \end{cases}$$

If for each logarithmic unit in the ΔY range the counted number of values was above or equal to $n_{\min}$, the series passed the test. A second binning for visual inspection of possibly unsuitable series was produced in parallel. In this case the ΔY range was divided into 10 equal intervals and the number of Y values per interval was counted.

Results and discussion

The selection of the relevant QSAR equations was done based on q^2 cross-validation values. Analyzing the distribution of q^2 values (Figure 7) we have chosen a threshold value of 0.3, and all further analysis takes into account the equations having a q^2 bigger than this value (Table 1). The choice of $q^2 \geq 0.3$ as a critical threshold value is based on the empirical observation that by eliminating between 1 and 5 outliers from an initial set of 25 or more objects, one can boost the q^2 value to 0.5 or higher. In this paper, we did not attempt to locate or eliminate outliers, as emphasis was placed on automated QSAR generation and processing.

The (relative) value of cross-validation

From the analysis shown in Figure 7 and Table 1, one can note that CV2 is too restrictive, since it dramatically reduces the number of QSARs with $q^2 \geq 0.3$. On the other hand, except for 2DProp descriptors, LOO does not increase considerably the number of equations, compared to CV7. A more detailed comparison between cross-validation methods reveals that out of 375 unique series with q^2(LOO) ≥ 0.5, 45 (or 12%) have q^2(CV7) < 0.3; and that out of 300 unique series with q^2(CV7) ≥ 0.5, 160 (or 53.3%) have q^2(CV2) < 0.3. While this second result indicates that CV7 is less reliable compared to CV2, we note that CV2 is too restrictive (as confirmed by Table 1 and Figure 7). Since more than 10% of the QSARs with q^2(LOO) ≥ 0.5 are not confirmed by other cross-validation methods, we suggest that, on occasion, LOO measures the redundancies in the training set, rather than model significance. Unless noted, the remainder of this section refers to CV7 as the measure to evaluate descriptor system performance in QSAR.

Descriptor system performance in biological QSAR

The best QSAR performance is obtained using the counts of the 504 SMARTS patterns (Q504), as

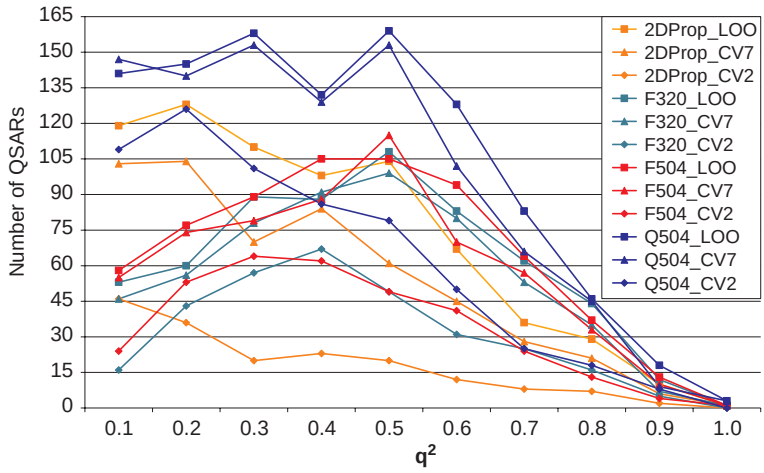

Figure 7. Distribution of cross-validated q^2 (using 0.1 intervals).

Table 1. The number of QSAR equations having $q^2 \geq 0.3$.

	2DProp	F320	F504	Q504
LOO	345	398	419	569
CV7	246	366	374	507
CV2	72	193	194	266

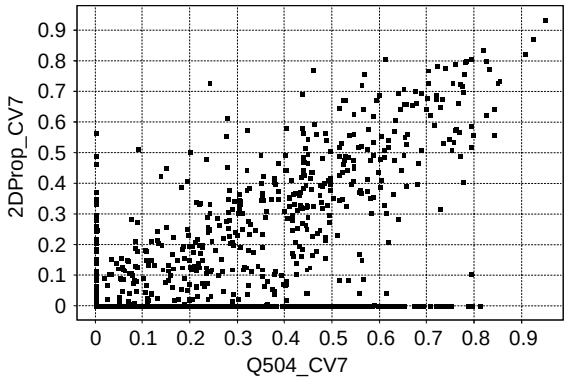

Figure 8. Model q^2 overlap for CV7 cross-validation, comparing 2Dprop and Q504.

illustrated in Figure 7 and Table 1. 2DProp (physico-chemical properties and topological indices) are the least effective descriptor combination. Binary methods (F320 and F504) have rather similar results across the three cross-validation methods – with F504 showing somewhat better performance. This may be related to noise- rather than signal-fitting, since 36% increase in the number of descriptors does not yield a significant improvement – in contrast to the difference between F320 and Q504.

A more quantitative measure of descriptor system performance can be obtained by correlating q^2 values across all models (1632 series) at the CV7 cross-validation level (see Figures 8–10). Comparing 2Dprop to Q504 ($R^2 = 0.378$, Figure 8), we found 28 QSARs with q^2(2Dprop) ≥ 0.3 and q^2(Q504) < 0.3, and 287 models with q^2(Q504) ≥ 0.3 and q^2(2Dprop) < 0.3, hence the odds of finding significant QSARs using Q504 are better, compared to 2Dprop. Evaluating F320 and Q504 ($R^2 = 0.545$, Figure 9), there are 54 models for which q^2(F320) ≥ 0.3 and q^2(Q504) < 0.3, and 187 models for which q^2(Q504) ≥ 0.3 and q^2(F320) < 0.3. The odds of finding 'significant' QSARs are better using Q504, compared to F320. Finally, between F504 and Q504 QSARs ($R^2 = 0.573$,

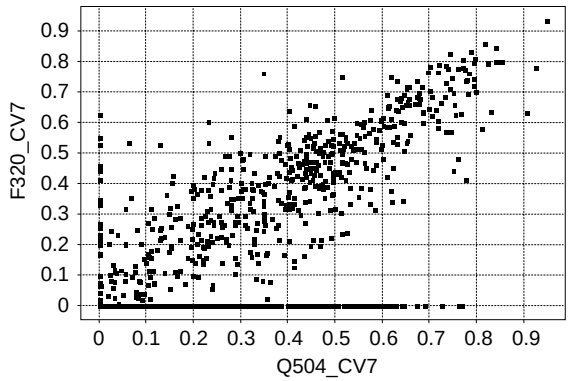

Figure 9. Model q^2 overlap for CV7 cross-validation, comparing F320 and Q504.

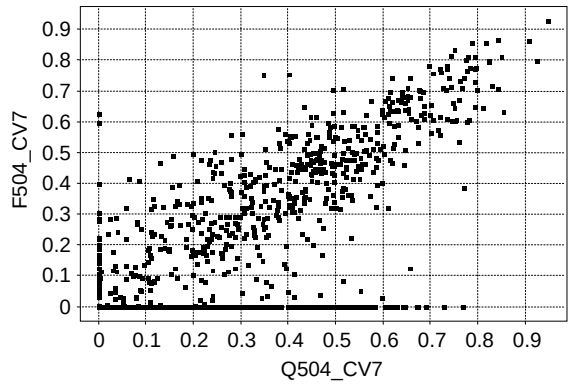

Figure 10. Model q^2 overlap for CV7 cross-validation, comparing F504 and Q504.

Figure 10), there are 46 models for which q^2(F504) ≥ 0.3 and q^2(Q504) < 0.3 for Q504, and 181 models for which q^2(Q504) ≥ 0.3 and q^2(F504) < 0.3 – so the odds of finding significant QSARs using Q504 are somewhat better than when using F504. Overall, 92 models satisfy the condition q^2(*any* method except Q504) ≥ 0.3 and q^2(Q504) < 0.3, while 110 models satisfy the condition q^2(Q504) ≥ 0.3 and q^2(*all* methods except Q504) < 0.3. Thus, when compared to all three other descriptor systems, Q504 alone shows better performance. We conclude that the probability to find significant QSARs is higher when using Q504, compared to all other three descriptor systems. However, 1033 out of 1632 QSARs satisfy the condition q^2(all methods) < 0.3. In other words, none of the above descriptor methods has a chance above 50% to find relevant (CV7) QSARs for an arbitrary series, at least in the absence of additional data (see also Conclusions).

Table 2. Ranking of 2DProp descriptors ($q^2 \geq 0.3$).

Rank	Sum	VIP ≥ 0.8	VIP ≥ 1.0
1	CMR (261.3)	Kier Chi5p (206)	CMR (155)
2	Non-polar atoms (258.2)	Molecular volume (201)	Atom count (152)
3	Polarizability (257.6)	Kier Chi3p (201)	Randic index (151)
4	Bond count (257.3)	Kier Chi6p (201)	Bond count (151)
5	Molecular volume (257.2)	CMR (200)	Polarizability (149)
6	MW (256.6)	Randic index (199)	Carbon count (149)
7	Atom count (256.0)	Kier Chi4p (199)	Kier Chi0 (148)
8	Kier Chi0 (255.8)	Polarizability (198)	Molecular volume (147)
9	Molecular surface (255.7)	Molecular surface (198)	K & H Kappa1 (147)
10	Carbon count (255.6)	Bond count (197)	Molecular surface (145)
11	Kier Chi3p (255.2)	Kier Chi2 (197)	Wiener index (143)
12	Randic index (254.6)	Non-polar surface (196)	Kier Chi5p (141)
13	Non-polar surface (254.1)	Carbon count (194)	Kier Chi3p (140)
14	Kier Chi5p (253.8)	Atom count (193)	Kier Chi6p (139)
15	Kier Chi2 (253.7)	Wiener index (193)	Kier Chi2 (139)
16	K & H Kappa1 (253.0)	MW (192)	Non-polar atoms (139)
17	Kier Chi6p (252.1)	Kier Chi0 (192)	MW (138)
18	Wiener index (250.8)	Motoc index (192)	Motoc index (137)
19	HMO pi-energy (250.7)	Non-polar atoms (191)	Non-polar surface (133)
20	Motoc index (248.1)	K & H Kappa1 (190)	Kier Chi4p (128)

Table 3. Ranking of SMARTS based descriptors.

Rank	Fingerprints		Counts	
	Descriptor	SMARTS [a]	Descriptor	SMARTS [a]
1	f441 (333.4)	[R]~*~*-[CH3]	q458 (597.2)	[R]~[D3]
2	f440 (328.3)	[a]~*~*-[CH3]	q155 (569.3)	[!$(*#*)&!D1]-!@[!$(*#*)&!D1]
3	f147 (316.1)	[!#6]-[CH3]	q453 (562.7)	[!#6]~*~*~*~[D3]
4	f444 (310.5)	[R]~*-[CH3]	q137 (542.7)	[!#6]~*~*~[R]
5	f389 (305.7)	[#9,#17,#35,#53]	q139 (536.4)	[!#6]~*~[R]
6	f393 (302.9)	[#9,#17,#35,#53]~*(~*)~*	q476 (521.9)	[R]
7	f407 (300.6)	[$([#6]-!@[A!C!H]-!@[#6])]	q017 (519.0)	*(!@*)(!@*)
8	f005 (300.3)	*-!:[a]:*:*:[a]-!:*	q277 (515.0)	[#6X3v3-1,#6X3v4 + 0,#6X3v3 + 1]
9	f443 (298.8)	[a]~*-[CH3]	q358 (505.5)	[#8,#16]
10	f004 (289.4)	*!:[a]:*:*:[a]!:*	q483 (505.4)	[R](-*(-*))~*~*~[a]
11	f007 (285.6)	*-!:[a]:*:[a]-!:*	q481 (505.2)	[R](-*(-*))~*~*~*~[a]
12	f390 (285.5)	[#9,#17,#35,#53]!@*@*	q041 (504.1)	*-*-[R]~*~*:[a]
13	f392 (285.5)	[#9,#17,#35,#53]-[R]	q452 (503.9)	[CX4v4]
14	f456 (285.1)	[#9,#17,#35,#53]~[D3]	q039 (503.3)	*-*-[R]:[a]
15	f391 (282.2)	[#9,#17,#35,#53]-[a]	q040 (500.3)	*-*-[R]~*:[a]
16	f015 (279.7)	*!@[#8]!@*	q485 (499.2)	[R](-*(-*))~*~[a]
17	f413 (271.6)	[$([#8X2v2]([#6])[#6])]	q300 (492.7)	[#7]~*(~*)~*
18	f245 (269.2)	[#6!H0]~@[#6!H0]~@[#6!H0] ~@[#6!H0]~@[#6!H0]	q192 (487.3)	[!a](=*)~*~*~*~[R]
19	f244 (264.5)	[#6!H0]~@[#6!H0]~@[#6!H0] ~@[#6!H0]	q457 (483.7)	[#7]~[D3]
20	f301 (261.3)	[#7]~*~*~*~*~*~*~*~*~[#8]	q257 (482.8)	[#6]~!@[#8]

[a]The online Daylight SMARTS tutorial can be found at the web link from Ref. [43].

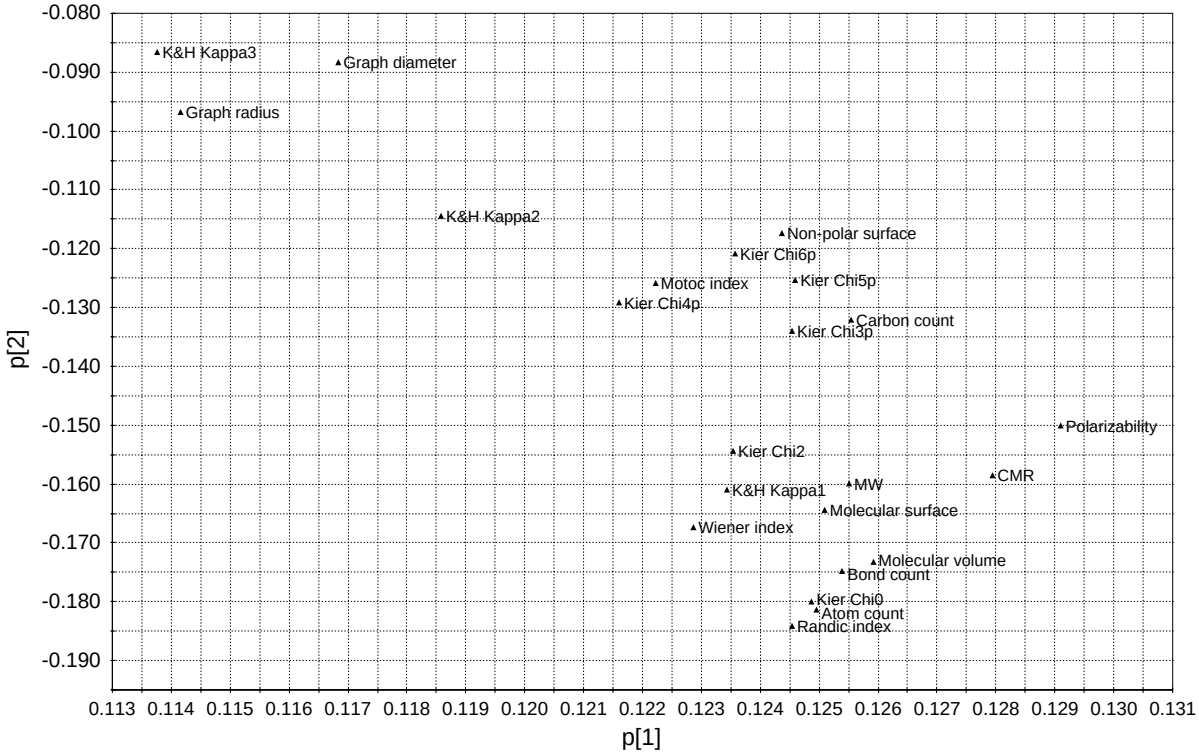

Figure 11. Loadings plot $p1/p2$ of 2DProp descriptors (partial view).

Y-distribution and its influence on QSARs

The *Y*-distribution test module was applied to each series as described above. We found that 1307 of 1632 QSAR series passed all the criteria, while for 168 series ΔY was less than 2.0. For the remaining 1307 series, we applied additional filters as follows: We looked at all QSARs with q^2(LOO) > 0.3 in any descriptor set, and found 682 QSARs meeting this condition; of these, 93 series did not pass the *Y*-distribution test. After filtering series with $\Delta Y \geq 2$, we found 41 series that were potentially unsuitable for QSAR modeling; further reducing this set to series for which q^2(CV7) ≥ 0.3, 34 series were left. These series produced acceptable QSARs according to q^2(CV7), yet they did not pass the *Y*-distribution test. These 'bad QSAR' series warrant individual analyses that are beyond the scope of this paper. None of them satisfied the q^2(CV2) ≥ 0.3 criterion. The 'bad QSARs' represent ~10% of the QSARs (see also Table 1), and thus will have little influence on our stated aim, i.e., to discover a suitable set of initial descriptors for QSAR modeling.

Individual descriptor performance in biological QSAR

To evaluate the *initial* usefulness of descriptors across all X-blocks, we plotted the distribution of the number of significant QSARs (with $q^2 \geq 0.3$) as a function of each descriptor's VIP, binned at 0.1 intervals. This data is not shown because the plot complexity is too high [58]. Except for a few insignificant descriptors (e.g., phosphorus and silicon atom counts), the VIP distribution is close for many significant descriptors, therefore a 'best' individual descriptor cannot be identified in this manner. This is also true if one ranks 2Dprop descriptors based on a specific cumulative threshold (e.g., VIP ≥ 0.8 or VIP ≥ 1.0 – see also Table 2). For these reasons, descriptors were ranked using the sum of the VIP scores, since the value is not threshold-dependent. The top 20 2Dprop descriptors, ranked according to the VIP sum and at two cumulative thresholds, is given in Table 2. Between 8 and 9 of these top 20 descriptors are size-related; this is not unexpected, since size

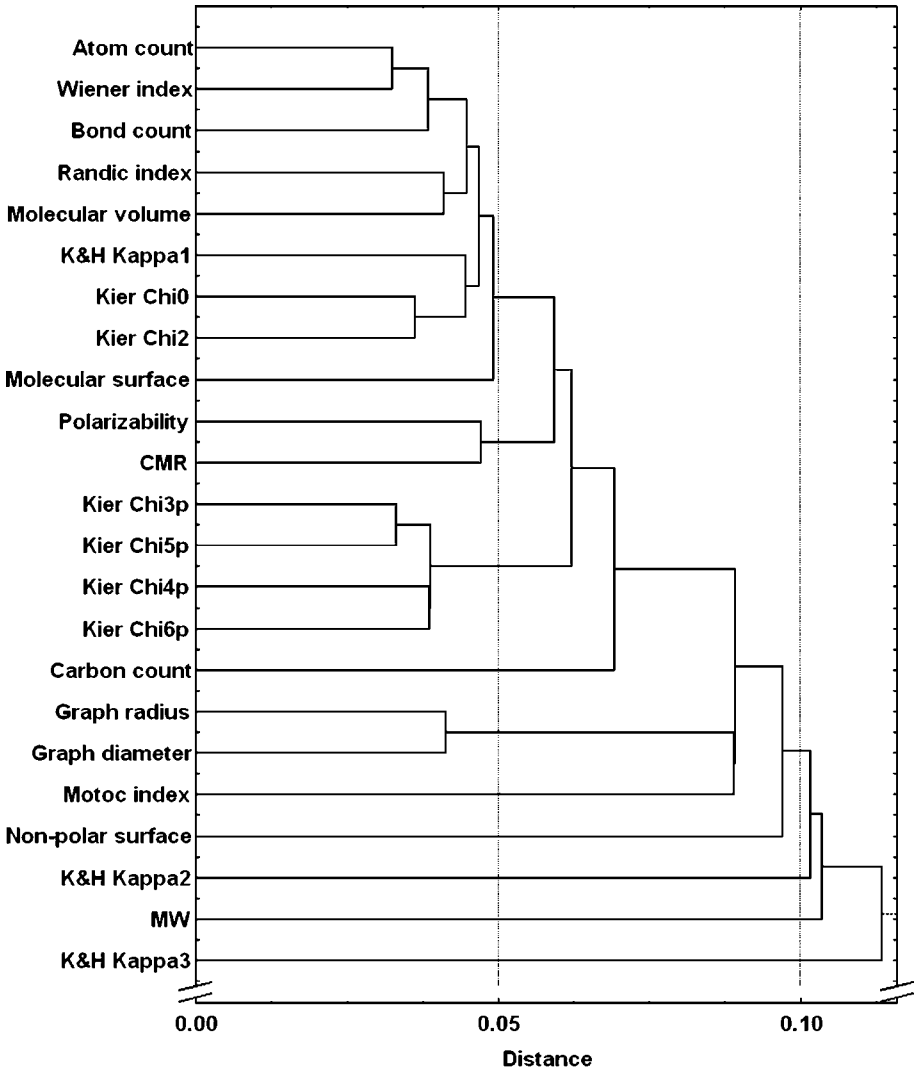

Figure 12. Clustering of the 2DProp descriptors (partial view).

covers ~60% of the first latent variable in chemistry-and-property independent spaces [59, 60]. Between 8 and 10 topological indices are also present in the top 20 list, whereas hydrophobicity is related to a maximum of three descriptors. The role of hydrophobicity and size in biological QSARs has been thoroughly investigated [61], and can be explained by the importance of such chemical descriptors in capturing a molecule's ability to permeate bio-membranes. On the other hand, the importance of topological indices can be related to the efficient way in which these descriptors capture structural patterns, and their relevance to bio-activity.

The F504 and Q504 descriptor sets were ranked using the sum of the VIP scores, as shown in Table 3. The relative importance of the SMARTS differs from fingerprints to counts, since none of the F504 top 20 substructures is found in the Q504 top 20. Six of the top 20 F504 descriptors are halogen-containing SMARTS, five of these contain a methyl (—CH_3) moiety, while several other SMARTS relate to aromatic atoms, or to hydrophobic moieties. In the Q504 list, ring atoms are frequent – three times branched, and seven times in linear connectivity. Flexible bond counts (second SMARTS from top) and non-carbon heavy atoms ([!#6]) are the remaining SMARTS atom

types in the top 5. Hydrophobicity (halogens, methyl groups, 4–5 CH-groups in rings) is better captured by the F504 top 20 list, whereas topological information (branched versus linear, in ring or outside, aromatic or polar) is better captured by the Q504 top 20 list.

In addition to evaluating the VIP parameter rankings, we also monitored the hidden relationship between descriptors and between the different QSAR series. These were obtained by applying PCA (principal component analysis) to the matrix of VIP scores resulting from the raw PLS results. This way, one can study the information overlap between different parameters in the descriptor space, as well as the similarity between QSAR series in the series (targets) space.

The PCA results for molecular descriptors show clusters of related descriptors (partial views are given in Figures 11 and 12 for 2Dprop), indicating a similarity in the modeling ability of these parameters on the studied biological QSARs. The clusters observed in the descriptor space indicate that by using a similarity metric, one can eliminate highly correlated (overlapping) descriptors, yielding a smaller, non-redundant, diverse set of descriptors that is likely to have similar predictive power to the original set, which could be more easily computed for QSARs. For example, CMR and polarizability are in the high-loading area for $p1/p2$ (Figure 11), and can be found in the vicinity of molecular weight, molecular volume and surface area, atom and bond count and the Randič, Kier Chi0 and Wiener indices. By clustering the PCA scores from the eight significant principal components (data not shown) using hierarchical clustering, we obtain similar information, e.g., CMR is related to polarizability, whereas atom count, the Wiener, Randič, Kier Chi0 and Kier Chi2 indices, molecular volume and surface are also related to molecular weight (Figure 12).

Conclusions

Of the 1632 QSARs, only a fraction show q^2 above 0.3, regardless of the cross-validation technique and descriptor set. The CV7 method provides most consistent results with respect to the number of successful QSARs [62]. CV2 appears to be too restrictive, while LOO can sometimes measure the redundancy in the training set. The SMARTS counts of the ~500 topological-pharmacophore patterns (Q504) outperform all other descriptor sets, whereas 2DProp yields the smallest number of successful QSARs. This may be related to the reduced number of descriptors in this particular system. Further studies, using a larger set of 2D/3D properties descriptors, could help us reevaluate these findings. We note that, in this study, Y-space distribution was evaluated independently from the actual QSAR modeling process. A good majority of the series were, in fact, not influenced by inadequate ΔY. Out of 1632 series, only 168 had $\Delta Y < 2.0$, and only 55 had $\Delta Y < 1.5$ – implying that there is a certain trend in medicinal chemistry literature to report a reasonable spread of Y values. In future studies, we will incorporate automatic Y-distribution detection before the actual QSAR modeling is performed – in order to avoid the ~10% 'bad QSARs'. We will also introduce Y-value scrambling as another validation method.

By using PCA on the VIP scores, we found interesting relationships between descriptors, and between different QSAR series. We are able to observe descriptors with high information overlap, both in 2D property space (shown above), and related to SMARTS patterns (data not shown). Clustering-based selections starting from larger sets of descriptors could thus prove useful in leading us to the most descriptive set, to be used as initial choice in biological QSARs. At least in the context of automated PLS analysis settings and the four particular descriptor sets compared, the most informative descriptor system identified is Q504 (topological-pharmacophore SMARTS counts). This seems likely to be a good choice for initial attempts to uncover relevant biological QSARs.

Acknowledgements

The authors wish to thank Mr. Tharun Kumar Allu (University of New Mexico) for providing technical support. This work was supported in part by New Mexico Tobacco Settlement funds.

References

1. Hansch, C. and Fujita, T., J. Am. Chem. Soc., 86 (1964) 1616.
2. Free, Jr. S.M. and Wilson, J.W., J. Med. Chem., 7 (1964) 395.

3. Todeschini, R. and Consonni, V., Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, 2000.
4. Hansch, C. and Leo, A., Exploring QSAR. Fundamentals and Applications in Chemistry and Biology, ACS Publishers, Washington, DC, 1995.
5. Livingstone, D.J., J. Chem. Inf. Comput. Sci., 40 (2000) 195.
6. Kubinyi, H., unpublished results.
7. Leo, A. and Weininger, D., CMR3. Daylight Chemical Information Systems, Santa Fe, New Mexico, htttp://www.daylight.com/, 1995.
8. Leo, A., Chem. Rev., 5 (1993) 1281.
9. Leo, A. and Weininger, D., CLOGP 4.0. Daylight Chemical Information Systems, Santa Fe, New Mexico, http://www.daylight.com/, 2001.
10. http://www.qsar.org/resource/software/htm, accessed in June 2002.
11. Ran, Y., Jain, N. and Yalkowsky, S.H., J. Chem. Inf. Comput. Sci., 41 (2001) 1208.
12. Livingstone, D.J., Ford, M.G., Huuskonen, J.J. and Salt, D.W., J. Comput.-Aided Mol. Design, 15 (2001) 741.
13. Glen, R.C., J. Comput.-Aided Mol. Design, 8 (1994) 457.
14. Hinze, J. and Jaffe, H.H., J. Am. Chem. Soc., 84 (1962) 540.
15. Hinze, J., Whitehead, M.A. and Jaffe, H.H., J. Am. Chem. Soc., 85 (1963) 148.
16. Gasteiger, J. and Marsili, M., Tetrahedron, 36 (1980) 3219.
17. Hansch, et al., J. Chem. Inf. Comput. Sci., 43 (2003) 120.
18. Raevsky, O.A., Grigor'ev, V.Yu., Kireev, D. and Zefirov, N.S., Quant. Struct.-Act. Relat., 11 (1992) 49.
19. HYBOT. TimTec Inc., Moscow, Russia, http://www.timtec.net/software/hybotplus.htm, 1998.
20. Zissimos, A.M., Abraham, M.H., Barker, M.C., Box, K.J. and Tam, K.Y., J. Chem. Soc., Perkin 2, 3 (2002) 470.
21. Kier, L.B. and Hall, L.H., Molecular Structure Description: The Electrotopological State, Academic Press, New York, 1999.
22. Oprea, T.I., J. Comput.-Aided Mol. Design, 14 (2000) 251.
23. Balaban, A.T., SAR QSAR Environ. Res., 8 (1998) 1.
24. Kier, L.B. and Hall, L.H., Molecular Connectivity in Structure-Activity Analysis, John Wiley, New York, 1986.
25. An analysis [26] using over 200 topological indices on over 1000 diverse structures revealed that these descriptors are grouped in 18 clusters that can be related to size, bond information, and molecular complexity (among other properties).
26. Basak, S.C., Balaban, A.T., Grunwald, G.D. and Gute B.D., J. Chem. Inf. Comput. Sci., 40 (2000) 891.
27. Cramer III, R.D., Patterson, D.E. and Bunce, J.D., J. Am. Chem. Soc., 110 (1988) 5959.
28. Goodford, P.J., J. Med. Chem., 28 (1985) 849.
29. Wold, S., Johansson, E. and Cocchi, M., In Kubinyi, H. (Ed), 3D QSAR in Drug Design: Theory, Methods and Applications, ESCOM, Leiden, 1993, pp. 523–550.
30. Kubinyi, H. (Ed), 3D QSAR in Drug Design: Theory, Methods and Applications, ESCOM, Leiden, 1993.
31. Kubinyi, H., Folkers G. and Martin Y.C., 3D QSAR in Drug Design, Vol. 2. Ligand Protein Interactions and Molecular Similarity, Kluwer/ESCOM, Dordrecht, 1998.
32. Kubinyi, H., Folkers, G. and Martin, Y.C., 3D QSAR in Drug Design, Vol. 3. Recent Advances, Kluwer/ESCOM, Dordrecht, 1998.
33. Cramer III, R.D. and Wold, S.B., US pat. 5025388 (1991). (CAN 115:135113).
34. Unger, S.H. and Hansch, C., J. Med. Chem., 16 (1973) 745.
35. Whitley, D.C., Ford, M.G. and Livingstone, D.J., J. Chem. Inf. Comput. Sci., 40 (2000) 1160.
36. Ferreira, M.M.C., Montanari, C.A. and Gaudio, A.C., Quimica Nova, 25 (2002) 439.
37. Nicolotti, O., Gillet, V.J., Fleming, P.J. and Green, D.V.S., J. Med. Chem., 45 (2002) 5069.
38. Golbraikh, A., Shen, M., Xiao, Z., Xiao, Y.-D., Lee, K.-H. and Tropsha, A., J. Comput.-Aided Mol. Design, 17 (2003) 241.
39. Weininger, D., J. Chem. Inf. Comput. Sci., 28 (1988) 31.
40. WB-PLS 1.0, developed at Sunset Molecular Discovery LLC, Santa Fe, New Mexico, http://www.sunsetmolecular.com/, 2004.
41. WOMBAT database, Sunset Molecular Discovery LLC, Santa Fe, New Mexico, http://www.sunsetmolecular.com/, 2004.
42. Durant, J.L., Leland, B.A., Henry, D.R. and Nourse, J.G., J. Chem. Inf. Comput. Sci., 42 (2002) 1273.
43. SMARTS, Daylight Chemical Information Systems, Santa Fe, New Mexico, http://www.daylight.com/dayhtml/doc/theory.smarts.html; online SMARTS tutorial: http://www.daylight.com/dayhtml/doc/theory.smarts.html, 2004.
44. Lipinski, C.A., Lombardo, F., Dominy, B.W. and Feeney, P.J., Adv. Drug Delivery Rev., 23 (1997) 3.
45. MacCuish, J. and MacCuish, N., Measures Software, Mesa Analytics and Computing LLC, Santa Fe, New Mexico.
46. Schneider, G., Neidhart, W., Giller, T. and Schmidt, G., Angew. Chem. Int. Ed. Engl., 38 (1999) 2894.
47. Byvatov, E., Fechner, U., Sadowski, J. and Schneider, G., J. Chem. Inf. Comput. Sci., 43 (2003) 1882.
48. Daylight Toolkit v4.81, Daylight Chemical Information Systems, Santa Fe, New Mexico, http://www.daylight.com/, 2003.
49. OEChem v1.2, Openeye Scientific Software, Santa Fe, New Mexico, http://www.eyesopen.com/, 2004.
50. Wold, S., Ruhe, A., Wold, H. and Dunn III, W.J. SIAM J. Sci. Stat. Comput., 5 (1984) 735.
51. Trygg, J., Parsimonious Multivariate Models, Umetrics Academy, Umeå, 2001.
52. Höskuldsson, A., J. Chemometr., 2 (1998) 211.
53. Cramer, R.D., Bunce, J.D., Patterson, D.E. and Frank, I.E., Quant. Struct. – Act. Relat., 7 (1988) 18.
54. Wold, S., Technometrics, 20 (1978) 397.
55. Statistical parameters are described in the SIMCA user manual; the software is available from Umetrics, Umeå, Sweden, web site: http://www.umetrics.com/.
56. Eriksson, L., Johansson, E., Kettaneh-Wold, N. and Wold, S., Multi- and Megavariate Data Analysis. Principles and Applications, Umetrics Academy, Umeå, 2001.
57. Zhu, E. and Barnes, R.M., J. Chemometr., 9 (1995) 363.
58. These figures are available from the authors upon request.
59. Oprea, T.I. and Gottfries, J., J. Comb. Chem., 3 (2001) 157.
60. Oprea, T.I., J. Braz. Chem. Soc., 13 (2002) 811.
61. Hansch, C., Hoekman, D., Leo, A., Weininger, D. and Selassie, C.D., Chem. Rev., 102 (2002) 783.
62. By default, for cross-validation the SIMCA-P software divides the original data into 7 groups; see the user manual or the document http://www.umetrics.com/download/KB/Multivariate%20FAQ.pdf, 2004.