Bioorganic & Medicinal Chemistry 18 (2010) 7078–7084

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

QSAR-based solubility model for drug-like compounds

Rafael Gozalbes *, Antonio Pineda-Lucena *

Structural Biochemistry Laboratory, Department of Medicinal Chemistry, Centro de Investigación Príncipe Felipe (CIPF), Avda. Autopista del Saler 16, 46012 Valencia, Spain

ARTICLE INFO

Article history:
Received 13 April 2010
Revised 19 July 2010
Accepted 2 August 2010
Available online 6 August 2010

Keywords:
Chemical space
Computational chemistry
Drug discovery
Quantitative structure–activity relationships (QSAR)
Solubility

ABSTRACT

Solubility plays a very important role in the selection of compounds for drug screening. In this context, a QSAR model was developed for predicting water solubility of drug-like compounds. First, a set of relevant parameters for establishing a drug-like chemical space was defined. The comparison of chemical structures from the FDAMDD and PHYSPROP databases allowed the selection of properties that were more efficient in discriminating drug-like compounds from other chemicals. These filters were later on applied to the PHYSPROP database and 1174 chemicals fulfilling these criteria and with experimental solubility information available at 25 °C were retained. Several QSAR solubility models were developed from this set of compounds, and the best one was selected based on the accuracy of correct classifications obtained for randomly chosen training and validation subsets. Further validation of the model was performed with a set of 102 drugs for which experimental solubility data have been recently reported. A good agreement between the predictions and the experimental values confirmed the reliability of the QSAR model.

1. Introduction

Solubility in water is a key physicochemical parameter when selecting compounds in drug discovery. It plays a very important role on different ADME properties, such as oral absorption, distribution or bioavailability of drugs. Moreover, compound solubility is particularly relevant in screening assays, since insolubility and aggregation of molecules are behind many of the failures in primary screens.[1] Solubility issues are normally evaluated at the early stages of drug discovery, especially when screening campaigns based on biophysical techniques with low sensitivity, such as NMR spectroscopy, will be pursued.[2]

Experimental determination of compound solubility is not easily manageable, or even possible, when working with large chemical libraries. In these cases, computational approaches for the solubility prediction of chemicals in water would be very useful. A large number of in silico models and approaches have been reported covering this issue, most of them based on QSAR methods relying on very different chemical descriptors and statistical approaches.[1,3,4] Also, different chemical software packages and internet-based resources are devoted to this task.[3,5,6] Nevertheless, from our point of view, many of these approaches present significant limitations that restrain their reliability and use. For example, very often, the models have been developed based on databases composed by a reduced number of structures, thus resulting in a low confidence on their predictability.[7] In other cases, the problem is the inadequate applicability domain of the models, a very well known problem for QSAR practitioners.[7,8] Many QSAR solubility models have been developed from databases such as AQUASOL[9] or the Physical Properties Database—PHYSPROP.[10] These databases contain information mostly related to organic compounds, not necessarily drugs or drug-like chemicals. This means that, very often, these models cover a different chemical space to the one they were intended for and explain why they often fail to accurately predict the solubility of drug-like compounds.[11] Only in few cases, the models have been based on information from large and homogeneous drug-like databases. These models have been developed by pharmaceutical companies or software developers,[12,13] and usually there is not an easy access to the structures used to build the models, the experimental solubility data used, or the technical details of the models.

Our group is heavily involved in the application of NMR screening to fragments and drug-like compounds in the context of discovery projects. Our experience is that compounds with low molecular weights tend to have good aqueous solubilities, and that solubility is not often a problem in fragment-based screening. On the contrary, it is a critical issue when dealing with drug-like compounds considering that NMR screening methods usually require much higher aqueous compound solubility than that required by other conventional screening techniques.[2] Trying to circumvent the limitations of the solubility models previously described, a new solubility QSAR predictor focused on the identification of highly soluble drug-like chemicals for NMR studies was developed.

The first step of the process was the selection of drugs, and drug-like compounds, with available experimental solubility data to build the QSAR model. The definition of the drug-like space

* Tel.: +34 96 328 96 80; fax: +34 96 328 97 01.
E-mail addresses: rgozalbes@cipf.es (R. Gozalbes), apineda@cipf.es (A. Pineda-Lucena).

0968-0896/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2010.08.003

has traditionally been based on the Lipinski's 'rule of five',[14] but this classification has some limitations. For example, Frimurer et al. showed that these filters accept 74% of the ACD compounds, but only 66% of the MDL Drug Data Report (MDDR) drugs.[15] In another study, Oprea demonstrated that the 'rule of five' does not clearly discriminate drugs from non-drugs.[16] Trying to identify a set of common drug-like relevant properties, a comparison was performed, using a set of physical parameters and cut-off values (Table 1A), between drugs from the Food and Drug Administration Maximum Recommended Daily Dose database (FDAMDD)[17] and generic organic compounds from the PHYSPROP database. Drug-like filters were defined as those properties able to maximize the difference between the retention and the rejection of compounds from the FDAMDD and PHYSPROP databases, respectively. The QSAR predictive model was based on those compounds from the PHYSPROP database fulfilling the drug-like criteria and with experimental solubility data available at 25 °C.

Most of the solubility models developed so far, with the exception perhaps of the Recursive Partitioning Model from Lamanna et al.,[13] have been based on the prediction of precise solubility values. However, the model presented here is intended to be used as a decision tool. For that purpose, the PHYSPROP compounds retained for the development of the QSAR model were distributed into three solubility categories: compounds with solubility $\leqslant 10$ mg/L were considered low soluble chemicals (LS), compounds with solubility values $\geqslant 1000$ mg/L as high soluble structures (HS), and those with solubility values falling between these two limits were included in the medium solubility (MS) category (Figure 1). All the structures from this database were characterized using a panel of more than 1200 different chemical descriptors, and a training and validation set were randomly chosen (Tset and Vset, respectively). Different QSAR methods were applied to these compounds and the best model was selected based on the statistical parameters and percentages of accuracy of predictions on both, the Tset and Vset.

Table 1A
List of properties and counts used to characterize the structures from the FDAMDD and PHYSPROP databases

Category	Property
Physical properties	Molecular weight
	Total charge (sum of formal charges)
	Number of reactive groups
Atom counts	Number of atoms (including implicit hydrogens)
	Number of carbon atoms
	Number of hydrogen atoms (including implicit hydrogens)
	Number of heteroatoms
	Ratio between the number of carbon atoms and the number of heteroatoms
	Number of heavy atoms
	Number of aromatic atoms
	Number of nitrogen atoms
	Number of oxygen atoms
	Number of hydrogen bond acceptors (number of nitrogen plus oxygen atoms)
	Number of hydrogen bond donors (number of OH and NH atoms)
	Number of fluorine atoms
	Number of chlorine atoms
	Number of bromine atoms
	Number of iodine atoms
	Number of halide atoms
	Number of phosphorous atoms
	Number of sulfur atoms
	Number of P and S atoms
	Number of chiral centers
	Absence of atoms different to C, O, N, S, P, F, Cl, Br, I, Li, Na, K, Mg, Ca
Bond counts	Number of bonds (including implicit hydrogens)
	Number of single bonds (including implicit hydrogens)
	Number of double bonds
	Number of triple bonds
	Number of bonds between heavy atoms
	Number of rotatable single bonds
	Number of rigid bonds
	Number of aromatic bonds
	Number of rings
	Absence of $-(CH_2)_6CH_3$ chains
Adjacency and distance matrix descriptors[30]	Diameter
	Radius
	Petitjean descriptor

Finally, further validation of the model was pursued by its application to an external set (Eset) composed by 102 drugs with experimentally available solubility information.[18,19] The quality of this validation dataset was very high as intrinsic solubility values were determined by the same research group.[20]

2. Results and discussion

2.1. Drug-like chemical space

Based on previous studies[16,21], different cut-off values for the physicochemical parameters listed in Table 1A were applied to the FDAMDD and PHYSPROP datasets trying to identify a set of common drug properties. Table 1B lists the properties and cut-off values selected as being more efficient for discriminating drug-like from not drug-like compounds when comparing both datasets. The application of these filters retained 91.0% of the FDAMDD compounds, almost the double than those retained from the PHYSPROP database (46.2%). Thus, it can be concluded that the parameters and limits selected for this classification can be considered very effective filters to select drug-like structures. A PCA analysis of the drug-like chemical space was also performed using the filters from Table 1B. Figure 2A shows that the PHYSPROP compounds classified as not drug-like structures, based on our approach, occupy a different region of the chemical space than the drugs retained from the FDAMDD database. Also, Figure 2B displays the comparison between drug-like and not drug-like compounds from the PHYSPROP dataset. It can be observed that the partitioning of compounds is roughly similar in both figures, thus demonstrating the ability of these filters to identify drug-like features and to define a more general drug-like chemical space.

Interestingly, the comparison of these findings with other studies[16,21,22] indicates that some of the parameters found in our study to be critical for the definition of a drug-like chemical space had not been previously reported. For example, as shown in Table 1B, the most effective filters were the number of hydrogen atoms and rigid bonds, both properties discarding around 25% more structures from the PHYSPROP subset than from the FDAMDD database. However, none of these filters had been reported before to be so important when describing drug characteristics. On the contrary, we found that other properties previously reported to be relevant, such as the number of hydrogen bond donors and rotatable bonds,[21] were not so significant. Thus, only 121 structures (1.8%) from the PHYSPROP chemicals and 14 compounds (1.3%) from the FDAMMD database contained more than five hydrogen bond donors, indicating that this parameter did not effectively discriminate drug-like from not drug-like structures. Similarly, 79 compounds (1.2%) from the PHYSPROP database and 18 (1.6%) from the FDAMDD compounds have more than 15 rotatable bonds.

Finally, 3070 compounds out of the 6647 PHYSPROP chemicals with available solubility data were considered 'drug-like' based on the filters defined in Table 1B, that is, 46.2% of the total ensemble. Out of these 3070 chemicals, a subset of 1174 compounds with

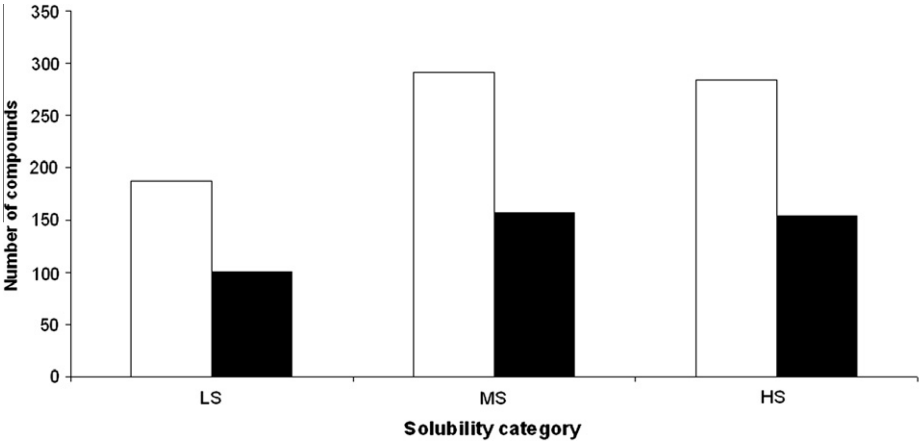

Figure 1. Distribution of the Tset (white bars) and Vset (black bars) compounds selected from the PHYSPROP database into solubility categories (LS: low soluble; MS: medium solubility; HS: high soluble).

experimental solubility values determined at 25 °C were selected to build the QSAR models. Compounds were distributed in a training set (Tset, 762 chemicals, 64.9%) and a validation set (Vset, 412 chemicals, 35.1%) and classified in different solubility categories as LS (288 chemicals, 24.5%), MS (448, 38.2%) and HS (438, 37.3%).

2.2. QSAR model

Several QSAR models were developed based on different combinations of descriptors and statistical approaches. The reliability of the different models was evaluated by comparing their ability to properly classify the solubility of compounds from the Tset and Vset. The best model turned out to be an Axis- Parallel Decision Tree model developed with 36 descriptors (Table 2) belonging to different categories: properties-based (two descriptors), atom/bond counts (three descriptors), and 2D-based (6 BCUT descriptors, 13 2D Autocorrelation indices, 10 Topological indices, and 2 Information Content descriptors). No 3D descriptors were included in this model. The choice of descriptors in the best model was further validated using the Y-scrambling procedure (see Section 4 for details).

The statistics, based on this model, for the Tset and Vset are summarized in Table 3, and they show a very good agreement between modeled and experimentally-obtained solubility classifications, with overall accuracy percentages of 86.4% and 67.0% for both sets, respectively. As it could be expected, it was difficult to separate the MS compounds from the other two groups. However, the percentage of compounds with experimental LS compounds classified as HS, or vice versa, was really low: only 16 compounds in the Tset (2.1%) and in the Vset (3.9%).

Table 1B
List of selected parameters and cut-off values

Drug-like filter	FDAMDD chemicals[a]	PHYSPROP chemicals[a]
MW $\geqslant$ 100 Da	5 (0.4%)	454 (6.8%)
Number of carbons $\geqslant$ 5	23 (2.1%)	918 (13.8%)
Number of heteroatoms $\geqslant$ 1	0 (0.0%)	263 (4.0%)
Carbon-heteroatoms ratio $\geqslant$ 0.7	18 (1.6%)	451 (6.8%)
Halide count $\leqslant$ 5	2 (0.2%)	165 (2.5%)
Number of hydrogen bond acceptors $\geqslant$ 1	1 (0.1%)	769 (11.6%)
Diameter $\geqslant$ 4	8 (0.7%)	490 (7.4%)
Number of hydrogens $\geqslant$ 8	36 (3.2%)	1791 (26.9%)
Number of rigid bonds $\geqslant$ 5	61 (5.5%)	2038 (30.6%)
Absence of $-(CH_2)_6CH_3$ chains	7 (0.6%)	156 (2.3%)
Absence of atoms different to C, O, N, S, P, F, Cl, Br, I, Na, K, Mg, Ca, Li	0 (0.0%)	150 (2.3%)
Total number:[b]	100 (9.0%)	3577 (53.8%)

[a] Number and percentage of compounds discarded by using a particular filter.

[b] This value is not necessarily the sum of all the compounds discarded by the application of all the filters, as one particular compound could be discarded by more than one filter.

Further analysis of the data was done by simulating a 'real' case in which a decision about selecting or not compounds for screening would have to be made. In this context, the analysis of the Tset revealed that if the 181 structures predicted to have low solubility were discarded, 551 compounds out of the 581 remaining chemicals would be experimentally soluble or highly soluble (94.8%). For the Vset, it was found that 286 out of the 323 compounds predicted to be soluble would be experimentally soluble (88.5%).

An external validation of the QSAR model was sought using experimentally available information for 102 drugs (Eset) obtained from the so-called 'solubility challenge'.[18,19] The results of this evaluation (Tables 4 and 5) show a clear trend of correctly classified compounds. Thus, 20 out of the 21 structures predicted as being LS have experimental solubilities (Sexp) <100 mg/L (10 of them under 10 mg/L). On the other hand, 23 out of the 30 compounds predicted as being HS have Sexp >100.0 mg/L (12 of them over 1000 mg/L). A histogram displaying the analysis of the results in terms of solubility categories and experimental values is shown in Figure 3.

Finally, the analysis of the 'solubility challenge'[19] results revealed that most of the models used in that study were not very reliable when the predictions involve compounds with very low solubilities (Sexp <4 mg/L). Interestingly, the QSAR model presented here was able to correctly classify, as belonging to the LS category, the three compounds (meclofenamic acid, diflunisal and dipyridamole) with very low intrinsic solubilities present in the Eset.

3. Conclusions

This paper describes a QSAR solubility model based on freely available information and robust enough to correctly identify soluble drug-like compounds in large collections of chemicals. This model could be very useful in the early phases of the drug discovery process when drug-like compounds are considered for carrying out biophysical screening campaigns.

The QSAR predictor relies on information obtained from drug-like compounds, and thus it is expected that this model will be generally applicable to other drug-like structures. In this study, a specific set of filters, physicochemical parameters and cut-off values, for defining the drug-likeness of chemicals was obtained by comparing the FDAMDD and the PHYSPROP databases. The

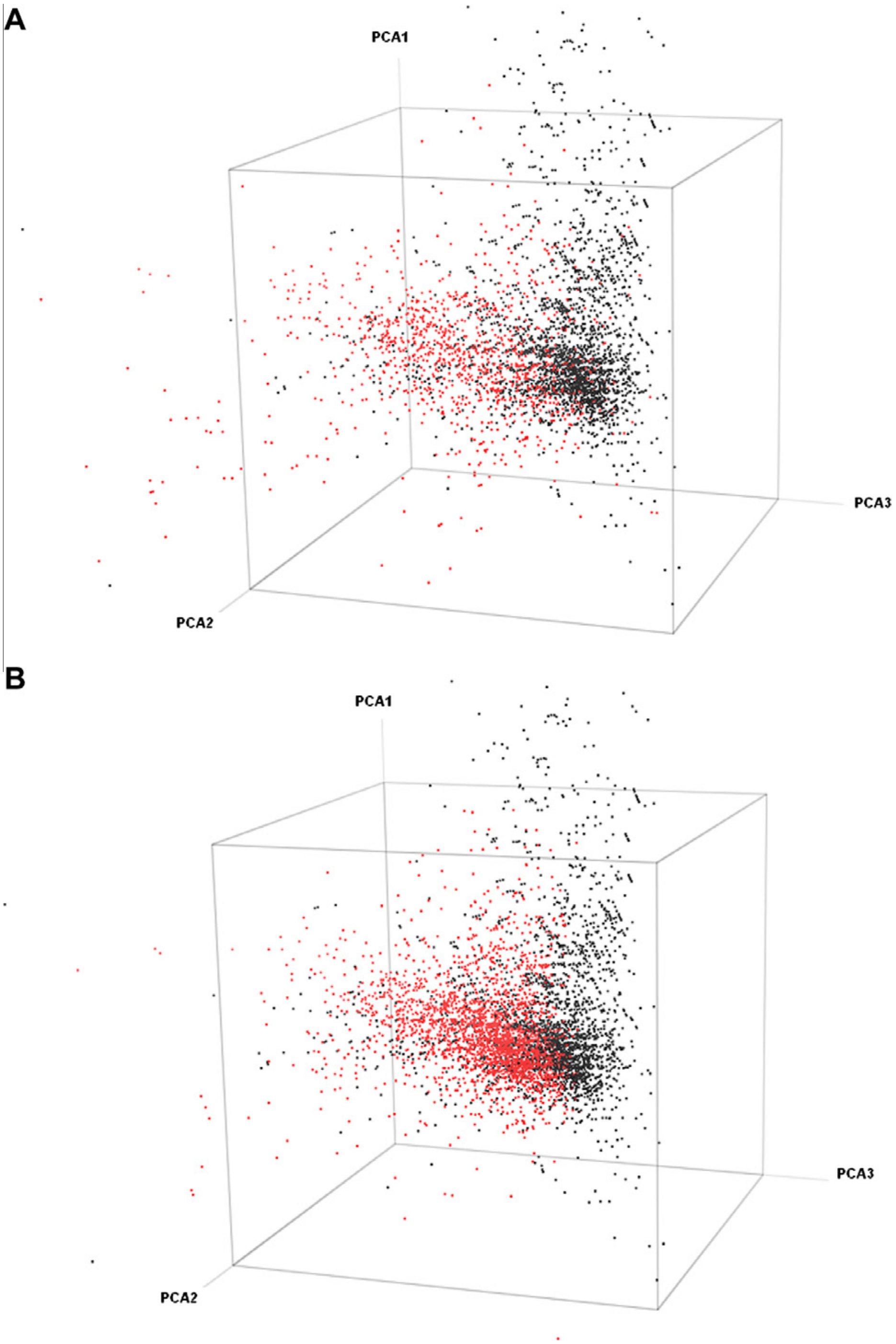

Figure 2. PCA analysis of the drug-like chemical space: (A) FDAMDD drugs (red dots) and PHYSPROP compounds (black dots) retained and rejected, respectively, after the application of the filters; (B) Drug-like (red dots) and not drug-like (black dots) compounds from the PHYSPROP database based on the selected filters.

comparison confirmed that some of the properties usually employed for the selection of drug-like compounds are not as discriminant as expected. On the contrary, it was found there were some unexpected parameters (e.g., the total number of hydrogen atoms and rigid bonds) that showed to be more relevant. We believe that the combination of parameters and cut-off values described in this work could be very helpful in future studies concerning the definition of a chemical space restricted to drugs.

As it would be expected,[1,3] lipophilicity (e.g., MLOGP and mlogpCX) and molecular weight and size descriptors (e.g., nC, petitjean, piPC02) were selected by the model, acknowledging the role these descriptors play in the water solubility of drug-like compounds. In addition to that, topological indices and E-state parameters, as those included in the Axis-Parallel Decision Tree, have also been described previously[1,3] as efficient descriptors for solubility prediction.

The model stands out by its simplicity and rapidity of calculation, since it relies on 1D and 2D descriptors easy to calculate. Furthermore, it does not require any experimentally determined value, thus offering a significant advantage over other standardized methods that usually demand the experimental measurement of melting points.[23] Another peculiarity of the model is the prediction of solubility ranges (i.e., LS, MS, and HS), as opposed to the calculation of precise solubility values. Although unusual, this approach

Table 2
Descriptors included in the Axis-Parallel Decision Tree model

Abbreviation	Descriptor definition
MLOGP	Moriguchi octanol–water partition coefficient (Log *P*)
mlogpCX	Moriguchi based lipophilicity descriptor (carbon and halogen atoms)
b-single	Number of single bonds (including implicit hydrogens)
nC	Number of carbon atoms
nHAcc	Number of hydrogen bond acceptors
BEHm4	BCUT descriptor (mass weighted)
BEHm5	BCUT descriptor (mass weighted)
BEHp1	BCUT descriptor (polarizability weighted)
BEHv4	BCUT descriptor (van der Waals volume weighted)
BEHv5	BCUT descriptor (van der Waals volume weighted)
BELp1	BCUT descriptor (van der Waals polarizability weighted)
ATS3e	Broto-Moreau autocorrelation descriptor—lag 3 (weighted by atomic Sanderson electronegativities)
ATS3v	Broto-Moreau autocorrelation descriptor—lag 3 (weighted by atomic van de Waals volumes)
ATS5e	Broto-Moreau autocorrelation descriptor—lag 5 (weighted by atomic Sanderson electronegativities)
ATS5v	Broto-Moreau autocorrelation descriptor—lag 5 (weighted by atomic van de Waals volumes)
ATS6m	Broto-Moreau autocorrelation descriptor—lag 6 (weighted by atomic masses)
ATS7p	Broto-Moreau autocorrelation descriptor—lag 6 (weighted by atomic polarizabilities)
MATS3e	Moran autocorrelation descriptor lag 3/weighted by atomic Sanderson electronegativities
MATS3p	Moran autocorrelation descriptor lag 3/weighted by atomic polarizabilities
MATS4e	Moran autocorrelation descriptor lag 4/weighted by atomic Sanderson electronegativities
MATS5e	Moran autocorrelation descriptor lag 5/weighted by atomic Sanderson electronegativities
MATS5p	Moran autocorrelation descriptor lag 5/weighted by atomic polarizabilities
MATS6m	Moran autocorrelation descriptor lag 5/weighted by atomic masses
MATS8p	Moran autocorrelation descriptor lag 8/weighted by atomic van der Waals volumes
GGI5	Topological charge index of order 5
GGI10	Topological charge index of order 10
Petitjean	Value of (diameter − radius)/diameter
piPC02	Molecular path count of order 10
R–CH–R	Atom centered fragment descriptors
R–CR–R	Atom centered fragment descriptors
TIE	E-state topological parameter
Vindex	Balaban V index
X1A	Average connectivity index chi-0
X3Av	Average valence connectivity index chi-3
BIC1	Bond information content descriptor (neighborhood symmetry of order 1)
BIC2	Bond information content descriptor (neighborhood symmetry of order 2)

Table 3
Confusion matrix summarizing the number and percentages of compounds correctly classified by the QSAR solubility model

		QSAR classification			
		LS_{pred}	MS_{pred}	HS_{pred}	Recall (%)
Tset	LS_{exp}	157	24	6	84.0
	MS_{exp}	14	251	26	86.3
	HS_{exp}	10	24	250	88.0
	Precision	86.7%	83.9%	88.7%	86.4
Vset	LS_{exp}	64	29	8	63.4
	MS_{exp}	17	106	34	67.5
	HS_{exp}	8	40	106	68.8
	Precision	71.9%	60.6%	71.6%	67.0

is very well suited for the selection of soluble compounds for drug discovery campaigns.

Despite the fact that the QSAR solubility model was developed based on an inhomogeneous database, it exhibited an excellent performance when applied to internal and external validation sets. The results of this validation confirmed the reliability of this QSAR solubility predictor as an efficient tool for the selection of soluble compounds to be used in drug screening campaigns based on low sensitivity techniques.

4. Computational details

4.1. Data set

The solubility model was based on the PHYSPROP database[10] (version March 2009) that includes 43,386 different chemical structures, 6647 out of them with information regarding experimentally determined solubility values. Given the fact that the model was intended to be applied to drug-like compounds and considering the importance of working in an homogeneous chemical space,[7,8] those PHYSPROP compounds that could not be considered drug-like structures were discarded. The distinction between drug-like and not drug-like compounds has traditionally been based on the definition of limits for some physicochemical properties, and thus different parameters and cut-off values have been proposed.[16,21,22] Based on previous studies, and trying to circumvent the limitations of other already proposed classifications, a comparison was performed between the 6647 structures with experimental solubility data from the PHYSPROP database and the 1216 pharmaceuticals included in the FDAMDD database.[17] A detailed inspection of the FDAMDD database revealed that there were a number of compounds (100 structures) that could not be clearly classified as drugs (e.g., acidulants, alkalizers, coenzymes, dyes, food additives, pigmentation agents, rat poisons, rodenticides, etc.) and they were not included in this comparison. All the remaining structures were identified using different sets of properties (i.e., physicochemical values, atom and bond counts, adjacency, and distance matrix descriptors) calculated with the VIDA module (version 3.0.0) from OpenEye Scientific Software[24] and the QuaSAR-descriptor tool from Molecular Operating Environment (MOE)[25] (Table 1A). Different cut-off values for these parameters were assayed to determine the combination leading to an optimal discrimination between drug-like and not drug-like compounds from the PHYSPROP database.

4.2. QSAR development

In addition to the descriptors shown in Table 1A, and in order to obtain an in-depth description of the structures for QSAR analysis, the selected drug-like compounds from the PHYSPROP database were characterized using several complementary subsets of numerical identifiers implemented in MOE and in the SArchitect™ Designer program.[26] These identifiers belonged to three main groups: constitutional descriptors (i.e., counts and property-based parameters), 2D descriptors reflecting the connection of atoms in the molecule, and 3D descriptors containing three-dimensional information about each molecule. The standard list of 166 MACCS keys[27] reporting the presence/absence of specific atoms or chemical groups was also calculated. At the end of this process, each compound was characterized by more than 1200 descriptors. Different combinations of descriptors were used to develop the mathematical models, either by individually dealing with specific classes (1D, 2D, MACCS, 3D) or by grouping descriptors from different classes. These lists were refined by discarding those descriptors with a low variance (standard deviation $\leqslant 0.1$), as well as those highly correlated among them (correlation coefficient $\geqslant 0.9$). Further reduction of the number of descriptors was achieved by applying sequential selection (forward/backward) and/or genetic algorithms that discarded descriptors not relevant for water solubility classification.

Table 4
Solubility challenge[18,19] compounds included in the Eset

Rank	Drug name	Solubility$_{exp}$ (mg/L)	Solubility$_{pred}$	Rank	Drug name	Solubility$_{exp}$ (mg/L)	Solubility$_{pred}$
1	Amiodarone	<0.01	LS	52	Trazodone	127.00	MS
2	Loperamide	0.04	LS	53	Nitrofurantoin	137.00	HS
3	Mefenamic acid	0.04	MS	54	Chlorpropamide	156.00	MS
4	Meclizine	0.13	MS	55	Ketoprofen	157.00	MS
5	Meclofenamic acid	0.16	LS	56	Sparfloxacin	167.00	MS
6	Diflunisal	0.29	LS	57	Azathioprine	172.00	MS
7	Sulfasalazine	0.29	MS	58	2-Amino-5-bromobenzoic acid	182.00	HS
8	Chlorprothixene	0.43	MS	59	Clozapine	188.90	LS
9	Amodiaquine	0.57	LS	60	Sulfamerazine	200.00	HS
10	Glipizide	1.45	HS	61	Enrofloxacin	237.00	MS
11	Chlorpromazine	2.70	MS	62	Tetracaine	258.00	MS
12	Dipyridamole	3.46	LS	63	Sarafloxacin	280.00	MS
13	Miconazole	3.50	LS	64	Diphenhydramine	289.00	MS
14	Probenecid	3.90	MS	65	Hydroflumethiazide	360.00	HS
15	Sertraline	4.50	MS	66	Naloxone	414.00	MS
16	Trimipramine	4.80	LS	67	Sulfamethizole	450.00	MS
17	Warfarin	5.10	MS	68	Danofloxacin	450.40	MS
18	Piroxicam	5.20	MS	69	Cephalothin	457.00	MS
19	Carprofen	5.50	MS	70	Sulfamethazine	508.00	MS
20	Maprotiline	5.60	LS	71	Norfloxacin	559.00	MS
21	Guanine	5.60	HS	72	5-Bromo-2,4-dihydroxybenzoic acid	559.00	HS
22	Benzthiazide	6.40	MS	73	Deprenyl	575.00	MS
23	Amitriptyline	7.80	LS	74	Alprenolol	580.00	MS
24	Sulindac	11.00	MS	75	Chlorpheniramine	590.00	MS
25	Phenylbutazone	12.54	LS	76	Acebutolol	711.00	HS
26	Phenazopyridine	13.70	MS	77	Famotidine	760.00	HS
27	Dibucaine	14.00	LS	78	Benzocaine	780.00	MS
28	Flurbiprofen	17.20	MS	79	Acetazolamide	816.00	HS
29	Pyrimethamine	19.40	LS	80	Benzylimidazole	874.00	HS
30	Furosemide	19.60	HS	81	Thymol	979.00	MS
31	Tolmetin	21.00	MS	82	Ranitidine	990.00	HS
32	Imipramine	22.00	LS	83	1-Naphthol	1500.00	MS
33	Bromogramine	22.30	MS	84	Salicylic acid	1620.00	HS
34	Carvedilol	22.60	LS	85	Lomefloxacin	1631.00	HS
35	Nortriptyline	25.00	LS	86	Amantadine	2120.0	MS
36	Naphthoic acid	28.96	MS	87	Lidocaine	3130.00	MS
37	5,5-Diphenylhydantoin	35.00	LS	88	Amoxicillin	3900.00	HS
38	Pindolol	40.00	HS	89	Phthalic acid	4130.00	HS
39	Papaverine	46.00	LS	90	4-Iodophenol	4250.00	HS
40	Flumequine	48.00	MS	91	Procaine	4500.00	MS
41	Fenoprofen	48.00	HS	92	4-Hydroxybenzoic acid	4740.00	HS
42	Verapamil	48.10	LS	94	Sulfacetamide	6470.00	HS
43	Nalidixic Acid	57.00	LS	94	Ofloxacin	19,600.00	MS
44	Desipramine	63.00	MS	95	Levofloxacin	TStM*	MS
45	Thiabendazole	66.00	MS	96	Marbofloxacin	TStM*	MS
46	Tryptamine	80.00	HS	97	Orbifloxacin	TStM*	MS
47	Ciprofloxacin	84.00	MS	98	2-Chloromandelic acid	TStM*	HS
48	Tolbutamide	93.00	MS	99	5-Fluorouracil	TStM*	HS
49	Difloxacin	100.00	MS	100	Ephedrine	TStM*	HS
50	Diazoxide	100.00	HS	101	Procainamide	TStM*	HS
51	Trichlormethiazide	113.00	HS	102	Pseudophedrine	TStM*	HS

* Compound too soluble to measure.[18]

Table 5
Distribution of the Eset compounds by solubility prediction categories and experimental values[18,19]

		Experimental solubilities			
		<10 mg/L	10–100 mg/L	100–1000 mg/L	>1000 mg/L
Eset	LS$_{pred}$	10	10	1	0
	MS$_{pred}$	11	12	20	8
	HS$_{pred}$	2	5	11	12

The development of the QSAR models was preceded by the definition, based on a random distribution, of a training set (Tset) containing 64.9% of the compounds retained for this analysis and a validation set (Vset, 35.1% of the compounds). Both groups maintained a similar distribution of LS, MS and HS compounds (Figure 1). The different statistical methods implemented in SArchitect (Naïve Bayes, Axis-Parallel Decision Trees, Neural Networks, Support Vector Machines and Decision Forest) were applied to the Tset in order to obtain several QSAR models. A detailed description of the different QSAR methods employed and their implementation in the SArchitect software is available from the SArchitect tutorial.[26]

All the models were validated by N-fold cross-validation.[28] In this particular case, the number of folds was three, one of them being retained as validation data. This process was iteratively repeated 10 times. Quality assessment of the models was obtained using the Y-scrambling or Response Permutation Testing method.[29] This procedure was used to ensure that the descriptors selected by each model were meaningful and that the models were not obtained by chance. The best model was selected based on the accuracy of correct classifications obtained for both set of compounds, the Tset and the Vset.

4.3. External validation set (Eset)

A good practice for further validating the QSAR models is their application to external sets of compounds.[7,8] In this case, the best

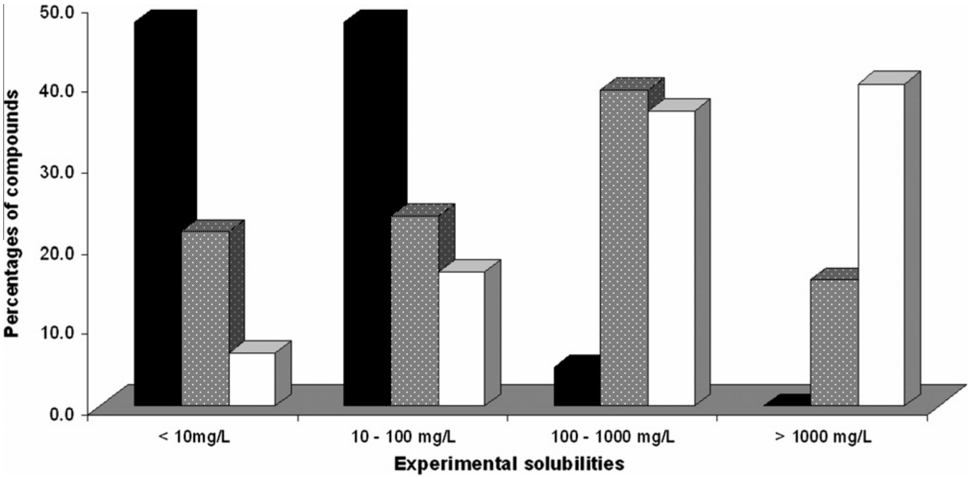

Figure 3. Percentage distribution of the 102 drugs (Eset) from the "solubility challenge" [18,19] into solubility categories (LS: black bars; MS: grey bars; HS: white bars) based on experimental values.

QSAR model was applied to a set of drugs (Eset) with experimentally available solubility information used in the so-called 'solubility challenge'.[18,19] The quality of this validation dataset was very high as intrinsic solubility values (i.e., solubility of compounds in its free acid or base forms, which does not depends on pH) were determined by the same research group, using the same experimental approach[20] and conditions (25 °C, ionic strength of 0.15 M). Thirty out of the 132 drugs from this database were not included in the external validation exercise as they were already present in the Tset and Vset.

Acknowledgments

The authors wish to thank the Spanish Ministerio de Ciencia e Innovación (MCINN, SAF2008-01845) and the Generalitat Valenciana (ACOMP/2009/048) for their economic support.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmc.2010.08.003.

References and notes

1. Delaney, J. S. *Drug Discovery Today* **2005**, *10*, 289.
2. Lepre, C. A. *Drug Discovery Today* **2001**, *6*, 133.
3. Balakin, K. V.; Savchuk, N. P.; Tetko, I. V. *Curr. Med. Chem.* **2006**, *13*, 223.
4. Faller, B.; Ertl, P. *Adv. Drug Delivery Rev.* **2007**, *59*, 533.
5. Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 246.
6. Geronikaki, A.; Druzhilovsky, D.; Zakharov, A.; Poroikov, V. S. A. R. *QSAR Environ. Res.* **2008**, *19*, 27.
7. Dearden, J. C.; Cronin, M. T.; Kaiser, K. L. *SAR QSAR Environ. Res.* **2009**, *20*, 241.
8. Tropsha, A.; Golbraikh, A. *Curr. Pharm. Des.* **2007**, *13*, 3494.
9. AQUASOL database (University of Arizona, http://www.pharmacy.arizona.edu/outreach/aquasol/).
10. Physical Properties Database—PHYSPROP (Syracuse Research Corporation, http://www.syrres.com/what-we-do/product.aspx?id=133).
11. Taskinen, J.; Yliruusi, J. *Adv. Drug Delivery Rev.* **2003**, *55*, 1163.
12. Du-Cuny, L.; Huwyler, J.; Wiese, M.; Kansy, M. *Eur. J. Med. Chem.* **2008**, *43*, 501.
13. Lamanna, C.; Bellini, M.; Padova, A.; Westerberg, G.; Maccari, L. *J. Med. Chem.* **2008**, *51*, 2891.
14. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Delivery Rev.* **1997**, *23*, 3.
15. Frimurer, T. M.; Bywater, R.; Naerum, L.; Lauritsen, L. N.; Brunak, S. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1315.
16. Oprea, T. I. *J. Comput. Aided Mol. Des.* **2000**, *14*, 251.
17. Matthews, E. J.; Kruhlak, N. L.; Benz, R. D.; Contrera, J. F. *Curr. Drug Discov. Technol.* **2004**, *1*, 61 (http://www.epa.gov/NCCT/dsstox/sdf_fdamdd.html; version 3b).
18. Llinas, A.; Glen, R. C.; Goodman, J. M. *J. Chem. Inf. Model.* **2008**, *48*, 1289.
19. Hopfinger, A. J.; Esposito, E. X.; Llinas, A.; Glen, R. C.; Goodman, J. M. *J. Chem. Inf. Model.* **2009**, *49*, 1.
20. Stuart, M.; Box, K. *Anal. Chem.* **2005**, *77*, 983.
21. Monge, A.; Arrault, A.; Marot, C.; Morin-Allory, L. *Mol. Divers.* **2006**, *10*, 389.
22. Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. *J. Med. Chem.* **2002**, *45*, 2615.
23. Ran, Y.; Yalkowsky, S. H. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 354.
24. VIDA, OpenEye Scientific Software, Inc. (http://www.eyesopen.com), 2009.
25. Molecular Operating Environment (MOE), available from Chemical Computing Group Inc. (http://www.chemcomp.com; version 2008.10).
26. SArchitect Designer software, available from Strand Life Sciences (http://www.strandls.com/sarchitect; version 2.5.0).
27. Symyx Technologies (http://www.symyx.com).
28. Arlot, S.; Celisse, A. *Stat. Surv.* **2010**, *4*, 40.
29. Gramatica, P. *QSAR Comb. Sci.* **2007**, *26*, 694.
30. Petitjean, M. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 331.