Journal of Computer-Aided Molecular Design **17:** 187–196, 2003.

C-QSAR: a database of 18,000 QSARs and associated biological and physical data

Alka Kurup
Chemistry Department, Pomona College, Claremont, CA 91711, USA

MS received 29 October 2002; accepted for publication 15 November 2002

Key words: C-QSAR, QSAR, quantitative structure–activity analysis, parameters, biological database, physical–organic database, database search, SMILES, sub-structure searching

Summary

The C-QSAR program is used to develop and search a database of over 18,000 equations that relate biological or physico-chemical properties of molecules to various molecular descriptors. The data used to derive the quantitative structure activity relationships (QSAR) are taken from various high quality journals. C-QSAR comprises two databases, one for structure-activity information biological systems ($n = 9200$) and the other for physical organic systems. Users can search the data in 20 different fields; for example by structure or substructure of the compounds involved, by the type of property correlated, by molecular properties, or by properties of the QSAR equation. Various ways in which information can be obtained is briefly discussed. Initially the database is often used for data mining, to search lead molecules, for substituent selection and "model mining" for lateral validation. The regression analysis is useful when the user wants to derive a new QSAR using his structures and activity data.

Introduction

Scientists in the field of drug design perceive it necessary to optimize a 'lead' molecule. An essential feature of drug discovery is to synthesize analogs of lead molecules and test their biological activity in order to obtain progressively better analogs, better not only in terms of potency and efficacy but also in terms of pharmacokinetics and side effects (ADME). The principal hypothesis employed is that any change in the chemical structure produces a positive or negative change in the bioactivity. A systemic study of such cause and effect relation is called structure–activity relationship (SAR) study. The objective of SAR is to define the chemical consequences of changing the drug structure and subsequently to establish which changes in chemical structure and properties would produce better biological activities. The whole process, though apparently simple, is quite involved, mainly because of the complexity of the biological system that is the target of the drug action. However with the introduction of computers, SAR has been made more methodical, and with the use of the proper parameters (most often measured but sometimes calculated) can be made quantitative i.e. quantitative structure–activity relationship (QSAR).

QSAR approaches involve statistical analysis of various molecular descriptors for a series of biologically active molecules. It is an attempt to correlate variations in the biological activity of a series of congeners to these molecular descriptors, which depict the physicochemical properties of the molecules. The result of a QSAR study provides useful clues regarding type of the substituents that should be tried to improve the activity further. Hence QSAR plays a vital role in lead exploitation. One of the best-publicized examples is the transformation of nalidixic acid with the help of QSAR into an important family of drugs – quinolone carboxylates [1].

There are many approaches to QSAR, and there are various computer programs available in the market. One such program is C-QSAR [2], which has been used to generate a database of QSAR since its advent at Pomona College by Corwin Hansch and his

*Correspondence. E-mail: akurup@pomona.edu

group. It has been now 40 years since the first QSAR were developed at Pomona College. Although it is not possible to gather all the enormous amount of scientific information published in various journals, still an attempt has been made to develop a dynamic, integrated computerized system covering most (possible) chemical-biological and chemical-chemical interactions. We enter any information into the system that can be described in mathematical terms with reasonable statistics. It already contains more than 18,000 QSAR out of which more than 9300 are biological QSAR. It should be noted that every day new QSAR are added. Realizing the amount of valuable information it contains, we felt that chemists should become aware of it, and its application to their research efforts. It is not possible to cover all the details in this report, but I will try to give the salient features in brief.

Source of data

Before going into the salient features, relating to the organization and utility of the database (C-QSAR database), I would like to comment on the source of data. The 9300 biological QSAR in C-QSAR have been derived by using the C-QSAR program. The data was taken from well-known journals, such as the *Journal of Medicinal Chemistry, European Journal of Medicinal Chemistry, Journal of Biochemistry, Chemical Pharmaceutical Bulletin, Journal of American Chemical Society, Nature, Journal of Molecular Pharmacology, Quantitative Structure-Activity Relationship, Journal of Pharmaceutical Sciences etc.*. All the data published in the *Journal of Medicinal Chemistry*, starting from volume 1 (year 1959) and for which a valid QSAR could be derived, is present in this database. To date there are 3355 QSAR from *Journal of Medicinal Chemistry alone.* Similarly from *Bioorganic Medicinal Chemistry* there are 480 QSAR; from *Bioorganic Medicinal Chemistry Letters*, 392 QSAR; *European Journal of Medicinal Chemistry*, 347 QSAR; and so on.

The database contains more than 8870 physical–organic QSAR. derived from data from the *Journal of American Chemical Society* (894 QSAR), *Journal of Organic Chemistry* (1181 QSAR), *Journal of Chemical Society* (1712 QSAR), *Journal of Chemical Society Perkin Transaction 2* (754 QSAR), etc. It is interesting to note that good data are never obsolete. There are 2 QSAR for the data published in 1899 by Reid [3].

Having QSAR for both physical and biological data enables comparisons between biological QSAR, and QSAR for physical organic chemistry [4,5].

Structure of database

Table 1 contains the information that is associated with every QSAR. The systems are different for biological and physical data. For the former, a name (enzyme, cell, animal etc.) of the biological system is recorded, whereas for physical data, the name of the reaction solvent is recorded.

For each dataset one can conveniently examine the summary (**SUM**, which gives the information at a glance about the data, e.g., what is the type of compound (parent molecule) for which the specific activity (mentioned under action) is studied, the reference for the data, the number of compounds in the data set, the number of parameters used for regression analysis, how many compounds are actually used in the QSAR etc. The output data (fields 15–20) describe the regression equation. For example for QSAR 1 obtained for the activity of phenols on Gold Fish, the **SUM** is as shown in Chart 1.

$$\begin{aligned} \mathrm{Log1/C} &= 1.02(\pm 0.17)\,\mathrm{ClogP} + 1.21(\pm 0.53) \quad (1)\\ \mathrm{n} &= 12 \quad \mathrm{r} = 0.972 \quad \mathrm{r}^2 = 0.945\\ \mathrm{q}^2 &= 0.915 \quad \mathrm{s} = 0.225 \end{aligned}$$

Chart 1:
Dataset name: BIO_4139 (each set stored in the database is given a number)
Substituents: 13 Parameters: 6 SMILES: 13 Equation: 1 Active: 12 Starred: 1 Inactive: 0

System	Gold Fish
Class	B6F; Fish
Compound	Chlorophenols
Action	LC_{50}: Concentration of the chlorophenols in the media causing 50% mortality of goldfish at 5 Hr Exposure
Reference	Kishino, T., Kobayshi, K. Wat.Res. 30,393-399(1996) R3838
Source	Rajni Garg
Check	Unknown
Date	2002 August 31
Parameters	YPred. Dev Log1/C ClogP S S-

Table 1. Organization of sets

Field		Description
Input data		
1.	SYSTEM	biological or physical system
2.	CLASS	Pomona classification of system (tables 2 and 3)
3.	COMPOUND	parent compound (if any)
4.	ACTION	measured action or activity
5.	REFERENCE	journal reference or other source of data set
6.	SOURCE	person who entered data set
7.	CHECK	person who checked data set
8.	NOTE	additional information about data set
9.	DATE	date on which set was saved into database
10.	PARAMETERS	list of parameters*
11.	SUBSTITUENTS	labels of substituents
12.	SMILES	topological description of compounds
13.	DATA	table of data used to generate the QSAR
14.	PRM MAX/MIN	maximum and minimum of each parameter
Output data (equation)		
15.	TERMS IN EQN	parameters in regression equation
16.	EQUATION	regression coefficients for each parameter
17.	IDEAL	ideal (or optimal) logP, and confidence limits
18.	STATISTICS	n, df, r, s, etc.
19.	RESIDUALS	deviations between y-predicted and observed
20.	PREDICTED	predicted values of dependent parameter

*examined, even if not used in final equation.

The information we get from the **SUM** is that there are 13 compounds in total, one is starred (not included in deriving QSAR 1), hence total number of data points (*active*) =12 for which the QSAR is derived. The *system* is Gold fish, the *compounds* are chorophenols, the *action* is LC_{50} i.e., lethal concentration causing 50% mortality of Gold fish at 5 hour exposure.

Reference: The reference from which the data was taken

Source: The person who entered the data and the date on which it was entered

Parameters are, calculated Log1/C (YPred), deviation (Dev) between calculated and observed Log1/C, observed Log 1/C, and the parameters used in the regression analysis ClogP, S, and S-. ClogP is calculated log P (measure of hydrophobicity), S and S- are Hammett electronic constants designating σ and σ^- respectively. In the equation n is the number of data points, r is the correlation coefficient, q^2 is the quality of fit and s is the standard deviation.

In order to organize the data in such a way that one can request all the information related to a particular problem for either a biological system or physical system, the database has been divided into two sections (Table 2 and Table 3)

Table 2 shows how the biological data are categorized. There are 6 major classes and these are further divided into subclasses. Considering the large number of sub classifications in the major areas of such as biochemistry, medicinal chemistry, pharmacology, toxicology etc, these may not appear to be an adequate classification scheme but in actual practice they fulfill the purpose.

It is easier to organize the physical (phys) database in comparison to biological (bio) database, although here also there is a miscellaneous class.

Physicochemical parameters

The various molecular descriptors used, named as physiochemical parameters, are given in Table 4. Although there are 43 different parameters, which can

Table 2. Class Codes–Biological Database (Number of sets in parentheses)*

B0	**Unknown**(10)	**B4**	**Single-Celled Organisms**
		B4A	Algae (42)
B1	Nonenzymatic Macromolecules (DNA, Fibrin, Hemoglobin, soil, Albumin, etc) (308)	B4B	Bacteria (800)
		B4C	Cell in culture (1019)
		B44	Erythrocytes (79)
B2	**Enzymes**	B4F	Fungi, Molds (308)
B2A	Oxidoreductases (830)	B4P	Protozoa (122)
B2B	Transferases (287)	B4V	Viruses (207)
B2C	Hydrolases (1018)	B4Y	Yeasts (54)
B2D	Lyases (67)		
B2E	Isomerases (23)	**B5**	**Organs/Tissues**
B2F	Ligases (2)	B5C	Cancer (290)
B2G	Receptors (1785)	B5G	Gastro-intestinal tract (79)
		B5H	Heart (91)
B3	**Organelles**	B5I	Internal/soft organs (65)
B3A	Mitochondria (91)	B5N	Nerves, Brain, Muscles (369)
B3B	Microsomes (99)	B5S	Skin (55)
B3C	Chloroplasts (84)	B5L	Liver (31)
B3M	Membranes (118)		
B3R	Ribosomes (0)	**B6**	**Multi-Cellular Organisms**
B3S	Synaptosomes (23)	B6A	Animal (vertebrates) (700)
		B6B	Insects (234)
		B6F	Fish (202)
		B6H	Human (44)
		B6I	Invertebrates (non-insect) (114)
		B6P	Plants (125)

*These numbers are constantly changing as new data are added daily.

be auto-loaded in the system for deriving QSAR [2], Table 4 shows only those that are commonly used.

A brief definition and references follows.

1. PI (π). Hydrophobic parameter for substituents defined by partitioning of X-C_6H_5 between octanol and water. (P) [6].

$$\pi_x = \log \mathrm{P\,X\text{-}C_6H_5} - \log \mathrm{P\,C_6H_6}$$

2. CPI (calculated π) for the substituents.
3. MR- SUB. Molar refractivity of a substituent defined analogously to π.

$$\mathrm{MR} = (n^2 - 1/n^2 + 2)(\mathrm{MW/d})$$

where n = refractive index, MW = molecular weight and d = density.

MR is highly collinear with substituent volume [6].

4. F. Swain-Lupton inductive/field effect parameter for aromatic systems [7, 19].
5. E_S. Classic steric parameter for substituents defined by Taft [8].
6. L- STM. Verloop sterimol parameter for substituent length [6, 9].
7. B1- STM. Sterimol parameter for the width of the first atom of the substituent [6, 9].
8. B5-STM. An estimate of the overall width of the substituent [6, 9].
9. S-P (σ). Normal Hammett constant for para substituents. It is based on the ionization constants of benzoic acids [6].
10. S-P + (σ^+). Brown parameter where substituents delocalize a +charge or radical via resonance [6, 10].
11. S- P- (σ^-). Hammett constant where substituents delocalize a negative charge via resonance. It is derived from the ionization constants of phenols [6]
12. S-M (σ_m). Hammett constant for meta substituents (non conjugated substituents) [6].
13. S-INDUC (σ_I). for the field/inductive effect. Originally defined from 4-X-bicyclo[2.2.2]octane-1-carboxylic acids [6].

Table 3. Class Codes–Physical Database (Number of sets in parentheses)*

PT	**Theoretical** (47)	**P7**	**Addition**(155)
		P7D	Dimerization (12)
PO	**Unknown** (18)	P7E	Electrophilic Addition (151)
		P7N	Nucleophilic Addition (253)
P1	**Ionization** (1742)	P7P	Polymerization (10)
P1P	Ionization Potential (39)		
P1X	Proton Exchange (74)	**P8**	**Elimination** (173)
		P9	**Rearrangement** (220)
P2	**Hydrolysis** (1035)	**P10**	**Oxidation** (569)
		P12	**Radical Reactions** (613)
P3	**Solvolysis** (649)	**P13**	**Complex Formation** (105)
		P14	**Partitioning** (109)
P4	**Spectra**	P14C	Chromatography (21)
P4I	Ionization Spectra (60)		
P4E	ESR Spectra (6)		
P4M	Mass Spectra (12)	**P15**	**Pyrolysis** (90)
P4N	NMR Spectra (194)	**P16**	**H-Bonding** (35)
P4R	IR Spectra (9)	**P17**	**Electrochemical** (301)
P4U	UV Spectra (23)	**P18**	**BrØnsted** (121)
		P19	**Esterification** (238)
P5	**Miscellaneous Reactions** (525)	**P20**	**Photochemical** (48)
		P21	**Hydrogenation** (16)
P6	**Substitution**	**P22**	**Isokinetic**(2)
P6E	E Electrophilic Substitution (264)	**P23**	**Reduction**(94)
P6N	Nucleophilic Substitution (1209)		

*These numbers are constantly changing as new data are added daily.

Table 4. Physicochemical parameters available in C-QSAR

PI	Pi	S- P	sigma para
CPI	Calculated Pi Value	S- P+	sigma para plus
MR- SUB	substituent refractivity	S- P-	sigma para minus
F	field effect (from S-L)	S- M	sigma meta
ES	E(s) from Taft	S- INDUC	sigma inductive
L-STM	length sterimol	S- STAR	sigma star from Taft's
B1- STM	width sterimol	CMR-SUB	calculated MR for Sub
B5- STM	width sterimol		

14. S- STAR. Classic σ^* defined by Taft [6].
15. Calculated MR for substituents.

In a broad sense the parameters can be divided into three categories elucidating three important features of a chemical entity. These are hydrophobic, electronic and steric effects.

Hydrophobic – Clog P, π, MlogP
Electronic – σ, σ^-, σ^+, σ_I, σ^*, F
Steric – CMR, Bl, L, B5, E_S, MgVol.

These parameters along with their application have been discussed in previous publications [6, 11]. Very briefly ClogP is the calculated partition coefficient in octanol/water and is a measure of hydrophobicity of the whole molecule while MlogP is the measured partition coefficient. There are 30518 values for measured log P for the octanol/water solvent system in the database and 60472 measured logP values for different solvent pairs. There are various methods for calculating logP [12]. The most extensively supported method

is that of Leo [12, 13]. The quality of his method is illustrated by eq. 2 [14].

$$\text{MlogP} = 0.96, (\pm 0.003)\text{ClogP} + 0.07(\pm 0.008) \quad (2)$$
$$n = 12510, \quad r^2 = 0.973, \quad q^2 = 0.973,$$
$$s = 0.300$$

This equation shows the relationship between 12510 experimental and calculated ClogP values.

π is the hydrophobic parameter for substituents attached to benzene [11].

The electronic parameter σ, σ^+ and σ^- are Hammett constants, which apply to substituent effects on aromatic systems [15–18].

σ^* is Taft's electronic effects that apply to aliphatic systems [8]. σ_I is the measure of inductive effects of aliphatic substituents. F is the field (inductive) effect of an ortho substituent [19].

CMR is the calculated molar refractivity for the whole molecule. MR is calculated from Lorentz-Lorenz equation and is described as $= (n^2 - 1)/(n^2 + 2)(MW/d)$. n^2 accounts for polarizability. We have scaled our MR values by 0.1. MR can be used for substituents or for the whole molecule.

B1, B5 and L are Verloop's sterimol parameters for the substituents [9]. E_S is the Tafts's steric constant [8]. It is based on the acid catalyzed hydrolysis of α-substituted acetates i.e. X-$CH_2COOCH_3(C_2H_5)$, and represents the steric effect of intramolecular and intermolecular bulk, which hinders the reaction or binding.

MgVol is the molar volume calculated by the method of McGown [20].

Searching the C-QSAR database

There are two uses of the database

1. To search for information that more clearly defines the drug design problem.
2. To derive a QSAR that best 'explains' ones preliminary activity data and directs the synthesis of more active analogs.

In this report, I will restrict the discussion to the former use. For those who become interested in construction of their own regression equations, they may either contact the author or BioByte Corp [2]. Since it is not possible to cover all the possible ways in which the database can be searched, an attempt will now be made to give an overall view. Broadly the search modes can be divided into three styles.

1. String searching, based on words
2. Chemical structure/molecule searching, using SMILES notation
3. Numeric searching on the values of parameters

String searching is a significant search mode, but one has to be careful in selecting words. For example, if one wants to find all the QSAR for data from *Journal of Biochemistry* that is stored in the database, then, by entering the word **BIOCHEM** (the journal names are entered in abbreviated form without spaces between the words) at the search command for reference, the program will find all the sets where BIOCHEM occurs as leading or trailing or in between the names as shown below:

BIOCHEM as in J.**BIOCHEM**
BIOCHEM as in **BIOCHEM**ISTRY
BIOCHEM as in ARCH.**BIOCHEM**.BIOPHYS.
BIOCHEM as in **BIOCHEM**.J.
BIOCHEM as in **BIOCHEM**.PHARMACOL.
etc.

To properly narrow the search, starting/ending query with a quote and blank should be used. A quote and a blank before the set of searching letters restricts matches to the beginning of the letter string, while blank and quote after the string of letters restricts matches to the end of word. To match the exact specific words, leading and trailing blanks and quotes must be present. Hence making a search as BIOCHEM.J. would only pick the exact match as hits.

An alternative way is to negate by prefacing with **NOT**.

For instance **5 BIOCHEM** (number 5 is corresponding to the number for *reference* in table 5)

5 NOT BIOCHEMISTRY
5 NOT BIOCHEM.PHARMACOL
5 NOT ARCH.BIOCHEM.BIOPHYS.
and so on.

This would remove them from the search hits made by BIOCHEM. A string search is of more value for searching the database for specific system, action, compounds etc., and the **NOT** command to negate the hits is equally significant.

Searching molecular structure employs **SMILES**, a language invented by David Wieninger [6, 21, 22]. It was incorporated into C-QSAR while he was a member of the Pomona College Medchem Project. There are two ways that one can search for compounds in the database. The first uses **SMILES** to identify all the sets that contain that specific molecule. The second uses **MERLIN**, to find all the compounds that contain

Table 5. Fields that can be searched

0	Equations				
1	System	8	Note	15	Terms in eq.
2	Class	9	Date	16	Coeffs in eq.
3	Compound	10	Parameters	17	Ideal/logB
4	Action	11	Substituents	18	Statistics
5	Reference	12	SMILES	19	Residuals
6	Source	13	Merlin	20	Predicted,
7	Check	14	Prm max/min		

the parent on which substituents have replaced one or more hydrogen atoms.

For example entering a **SMILES** search for phenol (Oc1ccccc1) finds 311 datasets that include unsubstituted phenol, where as **MERLIN** search for phenol (Oc1ccccc1) gives 5329 hits in the bio database for any of its analogs.

The biodatabase contains over 52,500 common chemical names (structures?), as well as official names of drugs either currently on the market, discontinued, or interesting but not on the market. This means that one can do a MERLIN search on any one of these compounds to uncover QSAR on similar chemicals.

Numeric searching on a parameter value is used to compare QSAR and to select substituents for the next round of synthesis that are expected to have better activity based on the derived QSAR. This use is discussed below under Substituent Selection in Molecular Design.

The **SEARCH MENU** given in Table 5 lists the different possible searches that can be made using the three broadly discussed modes.

Searching on system (**1 SYSTEM**; whatever search one wants to make that field number has to be entered before writing the words for **string** searching) of interest will scan the entire database for the specific string.

For instance

Searching for a system:

(1) **1 KINASE** gives 144 hits. A subsequent search on these hits as **1 TYROSINE** gives 63 hits (So there are 63 QSAR for Tyrosine Kinase)

(2) **1 ANGIOTENSIN** gives 41 hits (question – is this a new search or does it follow on from tyrosine kinase?)

(3) **1 CYTOCHROME** gives 65 hits. (again – is this a new search?)

A second search made on these hits for **1 P450** gives 31 hits.

Searching for a class and system:

2 B5C (for cancer as a class in table 2) gives 290 hits

A second search made as **1 LEUKEMIA**, gives 56 hits. Hence out of 290 QSAR for cancer (anti-cancer activity) 56 are for datasets tested for their actions against Leukemia. This is another way of narrowing down the search. The following example illustrates this further.

1 SERUM 106 hits
1 ALBUMIN 62 hits
1 BOVINE 33 hits

To look for the sets which have been studied for binding to bovine serum albumin a subsequent search yields

4 BINDING 23 hits.

If the interest is to find type of compounds studied, one can do so by giving the command to show (**SH),** and the compounds will be listed along with the set numbers. Then each set can be loaded to look into to the details like the type of substituents, the equation stored, and the reference, etc.

Search on an action can be made in a similar fashion:

4 ABSORPTION 33 hits
4 METABOLISM 11 hits
4 EXCRETION 13 hits

Similarly one searches the "phys" database to abstract the information required. For instance to search the database for all the QSAR that depend on the electronic parameter, S (σ), one can proceed in the following way:

15 " S 8265 hits These sets might contain other terms that could be eliminated by **NOT** command as follows

15 NOT S- S+ S' SI 4345 hits
15 NOT MR 4333 hits
15 NOT B1 B5 4171 hits
15 NOT ES 4087 hits
15 NOT **2 BILIN 4047 hits
15 NOT LOGP PI 4033 hits

Thus 4033 QSAR are with σ alone.

One of the most important classes of data in the phys database is that of pKa values and ionization constants. These have been entered as reported in the original articles without modification to put them on a common scale. A search made as shown below illustrates the potential where one wants to find solutes whose *aqueous* pKa's fall between 10 and 12.

15 pK	1736 hits
15 NOT LOGK	1646 hits
1 AQUEOUS	1172 hits
1 NOT %	586 hits (to eliminate system with a percentage of aqueous with some other solvent)
14 10<pKa<12	9 hits (this command is to sequester sets with pKa in the specified range)

This indicates that to date there are 1646 sets with pKa values for the analogs in each set and that only 9 sets contain molecules with pKa values falling in the range of 10–12.

Hence the database can be searched for each of the 20 fields given in Table 5. However as mentioned earlier, it depends purely on the interest and requirements. Whether the search is for data, for system, for action or for QSAR models, it is all there.

One is not limited to separate searches of the Phys and Bio databases. They can be searched as a combined database by using command 'datadouble'. This is useful because QSAR from physical organic chemistry can provide an excellent basis for understanding and supporting the enormously more complex QSAR from bio system [23–27].

The following example illustrates the value of searching the double database:

1. **12 c1ccccc1N** 397 hits (SMILES for aniline)
2. **15 S+** 44 hits
3. **15 NOT **2 BILIN** 43 hits (to remove QSAR having bilinear term)
4. **3 NOT MISC** 42 hits (removes miscellaneous series)
5. **SH**
6. **/sort = 16 1 3 4 6 15 16 18** (these numbers corresponds to the sections in Table 5)
7. **Sort on S+**

The following hits are illustrative:

Oxidation of X-$C_6H_4NH_2$ by MnO_2 in aqueous solution

$$\text{Logk} = -3.80\,(\pm 1.28)\,\sigma^+ + 1.49\,(\pm 0.65)$$

$$n = 6,\ \ r^2 = 0.944,\ \ s = 0.567,\ \ q^2 = 0.876$$

Oxidation of X-$C_6H_4NH_2$ by vanadium V in 50% aqueous acetic acid

$$\text{Logk}_2 = -3.31\,(\pm 0.79)\,\sigma^+ + 0.58\,(\pm 0.41)$$

$$n = 7,\ \ r^2 = 0.958,\ \ s = 0.263,\ \ q^2 = 0.916$$

Hydrogen abstraction of X-$C_6H_4NH_2$ by α,α-diphenyl-β-picrylhydrazyl radical in CCl_4

$$\text{Logk}_3 = -2.83\,(\pm 0.63)\,\sigma^+ + 4.69\,(\pm 0.17)$$

$$n = 6,\ \ r^2 = 0.949,\ \ s = 0.138,\ \ q^2 = 0.932$$

Hydrogen abstraction of X-$C_6H_4NH_2$ by α,α-diphenyl-β-picrylhydrazyl radical in CCl_4 The above QSAR establish a radical mechanism for the oxidation of anilines with which the following biological QSAR can be compared.

Oxidation of X-$C_6H_4NH_2$ by horseradish peroxidase I

$$\text{Logk}_2 = -3.17\,(\pm 0.46)\,\sigma^+ + 0.34\,(\pm 0.08)\,\text{ClogP} + 5.14\,(\pm 0.21)$$

$$n = 9,\ \ r^2 = 0.990,\ \ s = 0.227,\ \ q^2 = 0.973$$

Oxidation of X-$C_6H_4NH_2$ by horseradish peroxidase II

$$\text{Logk}_2 = -3.00\,(\pm 0.53)\,\sigma^+ + 0.25\,(\pm 0.10)\,\text{ClogP} + 4.52\,(\pm 0.24)$$

$$n = 8,\ \ r^2 = 0.988,\ \ s = 0.240,\ \ q^2 = 0.972$$

Oxidation of X-$C_6H_4NH_2$ by cytochrome peroxidase

$$\text{Logk}_2 = -2.86(\pm 0.43)\sigma^+ + 1.09(\pm 0.21)$$

$$n = 7,\ \ r^2 = 0.984,\ \ s = 0.193,\ \ q^2 = 0.962$$

The toxic action of radicals is an important factor to understand in drug research and environmental toxicology. To help in this process, there are over 600 phys QSAR on radicals.

Oxidation of X-$C_6H_4NH_2$ by cytochrome peroxidase

Substituent selection in molecular design

This is a significant and important feature of C-QSAR. Once a QSAR is derived the next question that arises is how to select additional substituents, to design the next series of congeners, which will maximize the information obtained on enhanced performances. C-QSAR has a unique feature whereby entering the command for parameter selection one gets a list of substituents with known parameter values. Let's say one wants to know all the substituents with σ values in the range of −1 to 1. When in the regression mode, if one enters the

command **PARA** (parameter) and selecting 15 (for S-P, corresponding to number for σ_p, in the system) and then entering the minimum and the maximum limit (here -1.0 to 1.0), then within a minute a list of all the substituents with σ values falling in this range comes up. We can see the whole list by entering **SEED** (see data). For instance in the current database, this list consists of 1957 substituents in total. In fact there are 1997 σ values in the system. There are 398 σ^+ and 375 σ^- values, 1078 B5, 1080 B1 and 1083 L values and so on, in the database. That implies that not only the access for the choice of the substituents is convenient but it is time saving too.

Entering **PARA**METER/**NOLIMIT** does not ask for limit. This saves the time in specifying limits when many parameters are requested simultaneously. Entering **PARA/NOLIMIT** by selecting **/ 2 3 13 16 19**, (corresponds to MR-SUB, F, B5-STEM, S-P+, S-M) the **SEED** command produces a set of 256 substituents ordered on increasing value of first parameter selected, i.e. π (2). All these 256 substituents have π, MR, B1, $\sigma_p$and σ_m values. By entering **SEEP**, we find the parameter labels. We can get a correlation matrix for these 5 set of parameters by entering **CORR 3 4 5 6 7**.

It is very important to select the substituents wisely to design a series, which has not only enough variation to study the effect of the individual parameters but also to have enough spread in the values. We have noticed that sets reported in the literature are not usually designed keeping this factor in the mind. The QSAR that results from such data can be confusing or misleading. Corwin Hansch has tried to illustrate this aspect of QSAR in drug design in reference 28.

Search for new lead compounds

A. Based on QSAR searches on highly active compounds

Yet another important search that can be made using the C-QSAR database is basing the search on only the most active compounds exhibiting a specific activity. The large number of bio QSAR in the system were developed by our group from the data published by others where no attempt was made to formulate a QSAR. Hence it is unlikely that the most active congeners were discovered. A selective approach is to find QSAR that cover highly active compounds. The dependent variable, log1/C has been entered in molar terms, except where it is not possible to do so. Hence log1/C=9 implies activity at 10^{-9} molar (nanomolar) concentration.

To consider QSAR that cover most highly active compounds, two search modes are possible.

1. **14 " log1/C>9 "** makes 6 hits. All compounds in these 6 datasets have log1/C values > 9.0
2. **14 " log1/C "@max>9** makes 458 hits. In this case, sets having at least one congener with log1/C greater than 9 are selected. Lowering the standard to 8, we find 1235 hits.

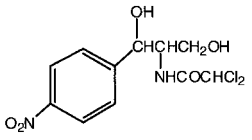
Another example to illustrate more a precise search is the following:

2 B4 B3 (cells and organelles)	2990 hits
14 " LOG1/C " @max>7	840 hits
5 (1990) (1991) (1992) (1993) (1994) (1995)	219 hits
13 IMIDAZOLE	11 hits

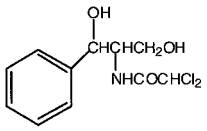
This searches all QSAR for cells and organelles developed for the data published in 1990–1995, which contain compounds containing an imidazole moiety, and have high activity.

Based on MERLIN search

A more general method for searching for compounds of interest is that of substructure searching. MERLIN is entered from the Bio (or Phys) regression mode. A panel follows this where the SMILES or the name of the compound is entered. The program identifies all the derivatives of the searched structure. The 2-D structures of these selected compounds can be seen by entering **DEPICT**. For example, a MERLIN search for derivatives of chloramphenicol in the bio database is illustrative. One way to make the search is by entering SMILES for chloramphenicol without nitro group, an easier way is to enter chloramphenicol by name, followed by choosing the option to modify the SMILES and then deleting the **N(=O)=O** at the end of the SMILES string to get structure as I:



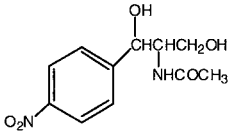
Chloramphenicol



I

On entering DEPICT, 24 examples with substitution on phenyl ring are found. If we wish instead to study the variations on the methyl of the acetamido

group, we can enter the SMILES for II



II OCC(NC(=O)C)C(O)c1ccc(cc1)N(=O)=O

OCC(NC(=O)C)C(O)c1ccc(cc1)N(=O)=O

Now 20 examples are found and are viewed by DEPICT mode. All are 4-NO_2-phenyl analogs with variations in the side chain.

Substructure searching often finds too many examples for consideration. For example, in the phys regression mode, on entering MERLIN and c1ccccc1C#N to find derivatives of benzonitrile 692 derivatives are identified. The same search in bio database yields 1015 derivatives. However, on limiting the search for derivatives of 4-chlorobenzonitriles (SMILES N#Cc1ccc(Cl)cc1), only 12 examples are found.

Conclusion

In this article I have given only a glimpse of the important features of the vast C-QSAR database. Only through actual use can one realize its potential to explore and use the information stored in the database. This information represents much time and effort by a number of researchers over a periods of 40 years.

Acknowledgements

I thank Corwin Hansch and Al Leo for their support and guidance in preparing this article, and Yvonne C. Martin for inspiring me to write this article.

References

1. Fujita, T. In Jolles, G. and Wooldridge, K.P.H. (Eds.), Drug Design: Fact or Fantasy?, Academic Press, 1984, p. 17.
2. C-QSAR program, BioByte Corp., 201 W, 4th St. Suite 204, Claremont, CA 91711. www.biobyte.com
3. (a) Reid, E.E., Am. Chem. J., 21 (1899) 340. (b) Reid, E.E., Am. Chem. J., 24 (1899) 397.
4. Selassie, C.D., Garg, R., Kapur, S., Kurup, A., Verma, R.P., Mekapati, S.B. and Hansch, C., Chem. Rev., 102 (2002) 2585.
5. Hansch, C., Hoekman, D., Leo, A., Weininger, D. and Selassie, C.D., Chem. Rev. 102 (2002) 783.
6. Hansch, C. and Leo, A., Exploring QSAR. Fundamentals and Applications in Chemistry and Biology; American Chemical Society, Washington, DC, 1995.
7. Hansch, C., Leo, A. and Taft, R.W., Chem. Rev., 99 (1991) 165.
8. Taft, R.W., Steric effects in organic Chemistry, Newman, M.S. (Ed.), Wiley, New York, 1956, p556.
9. (a) Verloop, A. The Sterimol approach to Drug Design, Marcel Dekker, New York, 1987. (b) Verloop, A., Hoogenstraaten, W. and Tipker, J., In Ariens, E.J. (Ed.), Drug Design, Vol VII, Academic Press, 1976, pp. 165-207.
10. Okamoto, Y. and Brown, H.C., J. Org. Chem., 22 (1957) 485.
11. Hansch, C., Leo, A. and Hoekman, D., Exploring QSAR: Hydrophobic, Electronic, and Steric Constants, American Chemical Society, Washington, DC, 1995.
12. Leo, A., Chem. Rev., 93 (1993) 1281.
13. Leo, A. and Hansch, C., Perspect. Drug Discov. Des., 17 (1999) 1.
14. Leo, A., unpublished results.
15. Jaffe, H.H., Chem. Rev., 53 (1953) 191.
16. Hammett, L.P., Physical Organic Chemistry, McGraw-Hill, New York, 1940.
17. Ritche, C.D. and Sagar, W.F., Prog. Phys. Org. Chem., 2 (1964) 323.
18. Brown, H.C. and Okamto, Y., J. Am. Chem. Soc., 80 (1958) 4979.
19. Swain, C.G. and Lupton, E.C. Jr., J. Am. Chem. Soc., 90 (1968) 4328.
20. Abraham, M. and McGown, J.A., J. Chrmotatographica, 23 (1987) 243.
21. Weininger, D., Weininger, A. and Weininger, J.L., J. Chem. Inf. Comp. Sci., 29 (1989) 97.
22. Weininger, D. and Weininger, J.L, In Hansch, C., Sammes, P.G. and Taylor, J.B. (Eds.) Comprehensive Medicinal Chemistry, Pergamon Press, Oxford, New York, Vol. 4., 1990, pp. 59.
23. Hansch, C., Hoekman, D. and Gao, H., Chem. Rev., 96 (1996) 1045.
24. Hansch, C., Acc. Chem. Res., 26 (1993) 147.
25. Garg, R., Gupta, S.P., Gao, H., Babu, M.S., Debnath, A.K. and Hansch, C., Chem. Rev., 99 (1999) 3525.
26. Hansch, C., Kurup, A., Garg, R. and Gao, H., Chem. Rev., 101 (2001) 619.
27. Hansch, C. In Hansch, C. and Fujita, T. (Eds.), Classical and Three dimensional QSAR in Agrochemistry, ACS Symposium Series, 606, American Chemical Society, Washington, DC, 1995, p. 254.
28. Mekapti, S.B. and Hansch, C., J. Chem. Inf. Comp. Sci., 42 (2002) 956.