Development of QSAR models using C-QSAR program: a regression program that has dual databases of over 21,000 QSAR models

Rajeshwar P. Verma (rverma@pomona.edu)
Department of Chemistry, Pomona College, California 91711, USA.
Corwin Hansch
Department of Chemistry, Pomona College, California 91711, USA.

Method Article

Keywords: C-QSAR program, quantitative structure-activity relationship (QSAR)

Posted Date: March 5th, 2007

DOI: https://doi.org/10.1038/nprot.2007.125

Abstract

Introduction

The interest in the application of quantitative structure-activity relationships has steadily increased in recent decades because it has repeatedly proven itself to be a low-cost, high-return investment. Potential use of QSAR models for screening of chemical databases or virtual libraries before their synthesis appears equally attractive to chemical manufacturers, pharmaceutical companies and government agencies. In the present protocol, we describe the use of C-QSAR program, a regression program used in the development of QSAR models for drug designers \(especially for those who do not have extensive experience in statistics), which handles linear, parabolic and bi-linear equations with various transformation of variables, and has 'jack-knifing' capability on all types of equations. The program has a very user-friendly method of data entry as well as of verification of structures and parameters. Auto-loading of preferred parameters is a time-saving feature of C-QSAR program. C-QSAR also produces a variety of 2-D graphs as output. C-QSAR[1] is a regression program used in the development of QSAR models for drug designers \(especially for those who do not have extensive experience in statistics), which handles linear, parabolic and bi-linear equations with various transformation of variables, and has 'jack-knifing' capability on all types of equations. The program has a very user-friendly method of data entry as well as of verification of structures and parameters. Auto-loading of preferred parameters is a time-saving feature of C-QSAR program. It also produces a variety of 2-D graphs as output. The C-QSAR program has dual databases of over 21,000 QSAR equations relating bio- and physico-chemical activities to structural parameters. Presently, BIO contains 12,950+ and PHYS 8,900+ equations. This is valuable for the users to validate new QSAR equations as they are being developed that is to see if the emerging structure-activity relationship bears a resemblance to others with known mechanisms. There are a number of commercial and free software programs available, which may be used to calculate descriptors and/or develop a QSAR model and are listed in **Table 1**. In C-QSAR program, the information associated with each data set, which is either in biological \(BIO) or physical \(PHYS) database, has been shown in **Table 2**.

Equipment

■ Modern personal computer \(Operating system should be Windows or Macintosh) ■ Reflection for ReGIS Graphics \(It is a terminal emulation program for Windows that allows any PC user to access a Unix or OpenVMS system, emulating terminals ranging from VT52 to VT400, WYSE to UNISYS). ■ Versa Term Pro \(It is for text and color graphics program for Macintosh, emulating DEC VT220 and Tektronix terminals, with automatic switching between them. ■ C-QSAR Program \(see EQUIPMENT SETUP) **Equipment Setup** ■ C-QSAR Program: The details about the C-QSAR package are available at "http://www.biobyte.com/bb/prod/cqsar.html":http://www.biobyte.com/bb/prod/cqsar.html and their installation instructions at "http://www.biobyte.com/bb/prod/qsarinstall.pdf":http://www.biobyte.com/bb/prod/qsarinstall.pdf. ■

For Table 1-4, the classification of C-QSAR program, and the details about the physicochemical parameters, please see the attached pdf \(pages 44-58).

Procedure

\(_By using Reflection for ReGIS Graphics from a personal computer with Microsoft Windows XP_) 1. Download the Reflection for ReGIS Graphics software and contact BioByte to provide the access of C-QSAR program and to create login and password. 2. To start C-QSAR, click Start → Programs → WRQ Reflection → Host-ReGIS Graphics → A window \(WRQ Reflection for ReGIS Graphiocs) will appear. Connection → Connect to Best Network → cqsar.com → click 'OK' BioByte AlphaServer 500/333 **Username**: \[write down the user name and then press ENTER] **Password**: \[write down the password and then press ENTER] **$** **$** **qsar** \[press ENTER] **qsar>** **qsar>** **data bio** \(or **data** **phys**) \[press ENTER] **QSAR Password**: \[press ENTER] **qsar: bio>** The main menu window of C-QSAR program for BIO when the program is started \(cqsar.com-WRQ Reflection for ReGIS Graphiocs) will appear \(**Fig. 1**). \[Similarly, one can obtain the main menu window of C-QSAR for PHYS by using \(**qsar> data phys**) instead of \(**qsar> data bio**)] 3. Summary of BIO-Database: From the above main menu window \(**Fig. 1**) **qsar: bio> 1** \[press ENTER] A summary of bio-database \(**Table 3**) will appear. Similarly, one can obtain the summary of physical-database from the main menu window of PHYS \(**Table 3**). **? TROUBLESHOOTING** Commands are shown in bold for only identification purposes and not for actual use. In practice it should be used as normal typing, users can use it in the capital letter too. h3. A. SEARCHING THE C-QSAR DATABASE In this protocol, we will restrict the discussion about the search mode of this program. For those who become interested are referred to the earlier publications.[2,3] Briefly, the search mode can be approached in three ways: \(i) String searching, based on words \(ii) Search on parameters \(iii) Chemical structure/molecule searching, using SMILES **1. Searching Bio-database** From the main menu window of C-QSAR program of Bio-database \(**Fig. 1**) **qsar: bio> 3** \[press ENTER] A search menu window for Bio-database will appear \(**Fig. 2**). \(i) **String searching:** String search is an important search mode, but one has to be careful. Some of the examples are as below: \(**a**) **Search using system \(1)**: users can search the database by using **system \(1)** from the search menu window \(**Fig. 2**) **qsar:bio> 1 goldfish** \[press ENTER] followed by **qsar:bio>sea** \[press ENTER] gave 17 hits. Similarly, one can search on the following terms: See figure in Figures section. *String search for the system shows the number of hits for the particular system as well as related to that system. Example: The search for the system 'Human' has 1989 hits; it means that the system must contained a word 'Human' that is, Human, Human Liver, Human Lymphocytes, Human Red Cell, Human Skin, Human Hemoglobin, Human Intestine, Human Platelets, Human Plasma, Human Thrombin etc. \(**b**) **Search using class \(2)**: **qsar:bio> 2 B4C** \[press ENTER] followed by **sea** command gave 2247 hits. Similarly, See figure in Figures section. \(**c**) **Search using compound \(3)**: **qsar:bio> 3 phenol** \[press ENTER] followed by **qsar:bio>sea** \[press ENTER] gave 281 hits, which will be narrow down as: **qsar:bio> 3 not miscellaneous** \[press ENTER] followed by **qsar:bio>sea** \[press ENTER] gave **257** hits. \(**d**) **Search using action \(4)**: See figure in Figures section. \(**e**) **Search using reference \(5)**: **Example-1**: If users want

to know about the number of QSARs in the bio-database from _Journal of Biochemistry_ \(stored as **J.BIOCHEM.**), the program will find all the sets from the journals where **BIOCHEM** occurs as leading or trailing or in between the names as shown as follows: **BIOCHEM** as in J. **BIOCHEM**. **BIOCHEM** as in CAN.J. **BIOCHEM**. **BIOCHEM** as in EUR.J. **BIOCHEM**. **BIOCHEM** as in IND.J. **BIOCHEM** .BIOPHYS. **BIOCHEM** as in J. **BIOCHEM** .MOL.BIOL.BIOPHYS., etc The best way for any character search can be negated by prefacing it with **NOT**. This causes the result to be the reverse \(logical complement) of what it would otherwise be. Thus, the search that picks only the word '**J.BIOCHEM.**' in the reference should be carried out from the search menu window of bio-database \(**Fig. 2**). **qsar:bio> 5 j.biochem.** \[press ENTER] followed by **qsar:bio> sea** \[press ENTER] gave **46** hits. **qsar:bio> 5 not can.j.biochem** \[press ENTER] followed by **qsar:bio> sea \[press ENTER]** gave **44** hits. **qsar:bio> 5 not eur.j.biochem** \[press ENTER] followed by **qsar:bio> sea** \[press ENTER] gave **22** hits. **qsar:bio> 5 not ind.j.biochem.biophys.** \[press ENTER] followed by **qsar:bio> sea** \[press ENTER] gave **21** hits. **qsar:bio> 5 not j.biochem.mol.biol.biophys.** \[press ENTER] followed by **qsar:bio> sea** \[press ENTER] gave **20** hits. **Example-2**: If user is interested to find out a particular paper \[Selassie, C.D., Kapur, S., Verma, R.P. & Rosario, M. _J. Med. Chem._ **48**, 7234-7242 \(2005)], which is available in the bio-database. It can be done from the search menu window of bio-database \(**Fig. 2**) as follows: **qsar:bio> 5 selassie** \[press ENTER] followed by **qsar:bio>sea** \[press ENTER] gave 57 hits. **qsar:bio> 5 2005** \[press ENTER] followed by **qsar:bio> sea** \[press ENTER] gave 7 hits. **qsar:bio> sh 5** \[press ENTER] lists the same reference of Selassie et al. with R7518 and seven sets with their set numbers 12865, 12866, 12867, 12868, 12869, 12870 & 12871. The desired summary of these sets can be obtained by show command \(**sh**) as follows: **qsar:bio> sh 1 2 3 4 16 18** \[press ENTER] will give the following details for all seven sets, but here we are giving the details for only one set: See figure in Figures section. Where, C represents the concentration of X-phenol that induces caspase-mediated apoptosis by 50%. $B5_2$ is Verloop's sterimol descriptor and is a measure of the maximum width of the substituents in the ortho position, while $B5_3$ represents the maximum width of the substituents in the meta position. The best way to open a data set from the main menu/regression mode that will be discussed later. \(**f**) **Search using source \(6):** Person who entered the data sets. **qsar:bio> 6 verma** \[press ENTER] followed by **qsar:bio> sea** \[press ENTER] gave **1679** hits. \(**g**) **Search using analysis \(7)**: Person who analyzed \(checked) the data sets. **qsar:bio> 7 verma** \[press ENTER] followed by **qsar:bio> sea** \[press ENTER] gave **63** hits. \(**h**) **Search using Prm max/min \(14)**: In this search, one can finds all the sets in which every compound that has a log 1/C of _n_ or greater and also the possibility to find out the sets in which at least one compound has a log 1/C of _n_ or greater. The search in the biological data base is as follows: **qsar:bio> 14 log1/C>9** \[press ENTER] followed by **sea** command gave **28** hits. This indicates that the bio-database has a total number of 28 sets in which every compound that has a log 1/C of **9** or greater. Similarly, the other command; **qsar:bio> 14 log1/C@max>9** \[press ENTER] followed by sea command gave **770** hits. This indicates that the bio-database has a total number of 770 sets in which at least one compound has a log 1/C of **9** or greater. \(**i**) **Search using Statistics \(18)**: **qsar:bio> 18 2<terms<4** \[press ENTER] followed by **sea** command gave **1909** hits – isolates all QSARs

having 3 terms in the bio-database. **qsar:bio> 18 n>75** \[press ENTER] followed by **sea** command gave **78** hits – isolates all QSARs based on more than 75 data-points. **qsar:bio> 18 r>.99** \[press ENTER] followed by **sea** command gave **1075** hits – selects all QSARs with _r_ greater than 0.99 \(**ii**) **Search on parameters**: Six parameters \(Clog _P_, Mlog _P_, CMR, NVE, MgVol, and MW) as well as forty-four parameters in **Table 4** can be automatically loaded for QSAR calculations. S stands for Hammett sigma σ; -P and -M stand for para and meta values, respectively. In the broader sense para values are used for aromatic substituents conjugated with the reaction center and meta values for non-conjugated aromatic systems. These Hammett-type parameters \[σ, σ+, σ-, σ** \(S-star), and σ_I \(S-inductive)] are obtained over half a century of study and testing on simple organic reaction mechanisms. Their use in the formulation of biological QSAR has already been discussed.[4] The resonance parameters \(R) and field/inductive \(F) have also been reviewed.[5] Molecular orbital parameters continue to be explored for the use in both biological and physical QSAR since there are many instances where Hammett constants cannot be used.[6-8] Searching the biological database from their search menu window \(**Fig. 2*): **qsar:bio> 10 HOMO LUMO** \[press ENTER] followed by **qsar:bio>sea** \[press ENTER] gave **166** hits. \(HOMO or LUMO was tested) **qsar:bio> 15 HOMO LUMO** \[press ENTER] followed by **qsar:bio>sea** \[press ENTER] gave **78** hits. \(HOMO or LUMO was used) This figure **78** shows that in **88** of the examples, the molecular orbital parameters \(HOMO or LUMO) were tested but found to be not as sound as Hammett constants. However, this statistic must be considered with caution since not all calculations were made with some of the more rigorous computational programs now available. The crucial parameter for the initial success of the biological QSAR was the numerical value of hydrophobic interactions.[9] Despite the great complexity of studies of all types of chemicals reacting with various kinds of biological systems \(from DNA to whole animals), the n-octanol/water partition coefficient used in log terms provides surprising insights. The hydrophobic parameter for the substituents \(π) can be of great importance in delineating local hydrophobic interactions at the receptor level.[4] Partition coefficients are rarely measured these days because it is costlier as well as time-consuming process. In the other hand, the use of data from the literature to formulate QSAR means that the compounds are not usually available for the measurement of their partition coefficients. C-QSAR program contains 12958 QSARs in bio-database, 6953 contain log _P_ terms and 930 have π terms; hence, it is very important to have the best possible means for their calculations. There are now a variety of methods for the calculation of log _P_ .[10] The most extensively used method is that of Leo.[10,11] The other important parameters are steric parameters i.e. MR-SUB, CMR, MgVol, E_S, _L_, B1, and B5. In the biological QSAR log 1/C is in molar terms except in a few cases marked by log 1/C'. The following approaches are used in the search of QSARs of particular type, which contain the desired parameter and are illustrated as: See figure in Figures section. The first step ensures that 1/C values are standard. The second step eliminates all QSARs with nonlinear terms, and the third step ensures that we have only n-octanol/water log _P_ values. Searches 4 and 5 eliminate parameters other than log _P_. Step 6 selects only those QSARs where the coefficient with log _P_ is between 0.6 and 1.0, and 7 eliminates QSARs whose intercept is outside of 0 and 0.5. Similarly, we can search the non-linear QSARs: \(**a**) **Optimal hydrophobicity**: See figure in Figures section. In this search, logP****2 represents log $_P_^2$. The third

step narrows the catch to log P_0 \(Optimal hydrophobicity) values between 1.5 and 2.5. Now, We can use the show command \(**sh**) that is; **qsar:bio> sh 5** \[press ENTER] list the results. For parabolic equations, log P_0 is displayed with its confidence limits, when it is possible to calculate them. One of the advantages of the parabolic model is that an estimate of log P_0 can be obtained without having data-points on the down side of the curve, which is necessary to derive the bilinear model. \(**b**) **Parabolic QSARs in terms of Clog _P_**: See figure in Figures section. Now, users can use the show command \(**sh**): **qsar:bio> sh 5** \[press ENTER], which list the results. \(**c**) **Inverted Parabolic QSARs in terms of Clog _P_**: See figure in Figures section. Inverted parabolic QSAR may correspond to an allosteric reaction. \(**d**) **Bilinear QSARs in terms of Clog _P_**: See figure in Figures section. \(**e**) **Inverted Bilinear QSARs in terms of Clog _P_**: See figure in Figures section. \(**iii**) **Chemical structure/molecule searching, using SMILES**: The SMILES search can be approached in two ways. One can identify every QSAR that contain a _specific_ molecule, or one can use a MERLIN search that finds all _derivatives_ of a given structures. For example: search for phenol \(Oc1ccccc1) in bio-database **qsar:bio> 12 Oc1ccccc1** \[press ENTER] followed by **sea** command gave **332** hits in the bio-database – finds **332** datasets that include un-substituted phenol. qsar:bio> 13 Oc1ccccc1 \[press ENTER] followed by sea command gave 7805 hits – finds 7805 datasets that include at least one derivative of phenol. **2. Searching Phys-database** From the main window of C-QSAR program of Phys-database **qsar: phys> 3** \[press ENTER] A search menu window for Phys-database will appear \(**Fig. 3**). Now, users can search the physical database as similar to that of bio-database using the search menu window \(**Fig. 3**). **? TROUBLESHOOTING** If one used the search command for one kind followed by **sea** command to check the number of the hits and not used the show command \(**sh**) then only blank command \(**bl**) will be needed to return into the normal search window and other type of search should be followed. On the other hand, if one used the show command then quit command \(**q**) will give the main menu, 3 \[ENTER] will give the previous search window \(not the normal search window), now the use of blank command will return into the normal search window. h3. B. LOADING DATA SET FROM 'DATABASE SEARCH' INTO 'WORK SPACE' After a data set of interest has been found by means of the **'sea'** and **'sh'** commands, it can be further examined in details if transferred into regression mode via its set number. This is done by entering regression command \(**reg**) from either the Bio or Phys database search menu window \(**Fig. 2** or **3**): **qsar:bio> reg** \[press ENTER] **qsar>** \(Note: This mode can also be obtained from either the main menu of Bio or Phys database using regression command \(**reg**)) **If users want to open a set no. 12675 from bio-database, which is for the binding of thioacridone derivatives \(I) to DNA, then we need to follow the following steps in the regression mode:** See figure in Figures section. **qsar> load /d 12675** \[press ENTER] will transfer the set to the workspace **qsar> sum** \[press ENTER] will list the key items of the data set as follows: See figure in Figures section. qsar> seeeq \[press ENTER] will list the following equation: See figure in Figures section. **qsar> pred** \[press ENTER] will list the data in tabular form: See figure in Figures section. In this set, 'Parameter' shows all parameters considered in this study. 'Ypred' is the predicted value from the stored equation. 'Dev' is the difference between this figure and the log of the observed value. In the QSAR equation, _n_ is the number of data points, _r_ is the correlation coefficient between observed values of the dependent and the values calculated from the equation, $_r_^2$ is

the square of the correlation coefficient represents the goodness of fit, _q_^{2} is the cross-validated _r_^{2} \(a measure of the quality of the QSAR model and obtained by using leave-one-out procedure[12]), and _s_ is the standard deviation. DF is the number of degree of freedom. SS1 is the sum of squares about the mean of the dependent variable and SS2 is the sum of squares from the deviations from the regression line. DEV+ and DEV- are the number of positive and negative deviations respectively from the QSAR. **To see the SMILES generated structures for the compounds of a data set:** **qsar> depict** **,** \[press ENTER] will depict sequentially all of the compound structures as one presses **ENTER** after each panel of 4 structures \(**Fig. 4**). The process can be stopped at any point by entering **q**. This can be very important in dealing with a large data set, say > 100 compounds. To view any particular structure enter **depict #** \(compound number in the set). To check all structures following compound 5, enter **depict 5**, to view all structures up to 7, enter **depict ,7**. To see those between 4 and 7, enter **depict 4,7**. h3. C. DERIVATION OF A QSAR MODEL For several reasons it is advisable to begin the regression program in the proper area, either database bio or database phys. Searching for the similar equations can then be carried out directly and if the developed equation is useful, it is easier to save in that area. **qsar:bio> reg** \[press ENTER] **qsar> clear** \[press ENTER] To be sure the workspace is empty \(**i**) **Title Information \(Set No. B12870)** **qsar> name Selassie-1** \[press ENTER] **qsar> t/sys CCRF \(sensitive cell)** \[press ENTER] **qsar> t/comp 4-X-phenols** \[press ENTER] **qsar> t/act log1/C: Cytotoxicity** \[press ENTER] **qsar> t/ref Selassie,C.D., Kapur,S., Verma,R.P., Rosario,M., J. Med. Chem. 48,7234-7242 \(2005)** \[press ENTER] **qsar> t/source Rajeshwar Prasad Verma** \[press ENTER] **qsar> t/class b4c** \[press ENTER] **qsar> sum** \[press ENTER] To check the correctness of the above information \(**ii**) **Naming Parameters:** Automatic loading will be demonstrated in this example, and so only the dependent variable need to be entered. **qsar> getp** \[press ENTER] **Label for parameter 3: log1/C** \[press ENTER] \(Since parameter 1 is reserved for predicted values and 2 is used for the deviation.) Since, M.O. parameters, BDE, or pKa are not auto-loaded, they will be entered manually. Thus, in the present example the BDE parameter will be entered as parameter 4. This can also be entered in later by using **newp** command, Label for parameter #: '**BDE**' and finally the values of this parameter for each compound. **Label for parameter 4: BDE** \[press ENTER] **Label for parameter 5: end** \[press ENTER] **? TROUBLESHOOTING**: In general the biological activity data is not published in logarithmic form with negative sign and in molar concentration. They can be entered as such and then converted into this form by using gettran \(**gett**) command. For example, if the biological activity is given in micro mole \(μM), then the data first entered as such under 'C' for the Label of parameter **3** and then converted into log1/C and in molar concentration by using gettran command as follows: **Label for parameter 3: C** \[press ENTER] **Label for parameter 4: end** \[press ENTER] **qsar> newsub** \[press ENTER] **Label for substituent 1: Substituent 1 \(or compound)** \[press ENTER] **C – parameter value 3: activity value** \[press ENTER] **Label for substituent 2: Substituent 2 \(or compound)** \[press ENTER] **C – parameter value 3: activity value** \[press ENTER] Similarly, all the labels and parameter 3 values should be enter and then **Label for substituent #: end** \[press ENTER] **qsar> seed** \[press ENTER] To see the correctness of the data **qsar> gett** \[press ENTER] A box \(Enter new label) will appeared. Type **log1/C** in the box and press ENTER. Now a second box \(Enter transformation) will obtain. In this box, type the following: **6 – log C** \[press ENTER] The transformed

activity will be in log1/C \(molar concentration) and present in the next **Label for parameter** \(i.e. 4). Now, one wants to delete the parameter 'C' \(Label for parameter 3). It can be done by the following command: **qsar> del /para 3** \[press ENTER] **3 C** **All these parameters will be deleted. OK? \(yes/no):** press **y** **The number of items deleted: 1** **qsar> save** \[press ENTER] To save the data Similarly, the other transformation should be carried out by using gettran command \(**gett**). See figure in Figures section. If the biological activity concentration is in microgram/ml \(mcg/ml), then the data first entered as such under 'X' for the Label of parameter 3 and then converted into 'C' in micro molar concentration by using gettran command as follows: **qsar> gett** \[press ENTER] A box \(Enter new label) will appear. Type C in the box and press ENTER. Now a second box \(Enter transformation) will be obtained. In this box, type the following: **X / MW** \[press ENTER] The transformed activity will be in C \(micro molar concentration) and present in the next Label for parameter \(i.e. 4). Now this will be converted into log1/C \(molar concentration) as shown above. \(**iii**) **Naming and Entering Substituents:** **qsar> newsub** \[press ENTER] **Label for substituent 1: NH2** \[press ENTER] **Log1/C – parameter value 3: 4.61** \[press ENTER] **BDE – parameter value 4: -9.25** \[press ENTER] **Label for substituent 2: OC6H13** \[press ENTER] **Log1/C – parameter value 3: 5.34** \[press ENTER] **BDE – parameter value 4: -6.30** \[press ENTER] Similarly, all the labels and parameters 3 and 4 values should be entered and then **Label for substituent 11: end** \[press ENTER] **? TROUBLESHOOTING:** It is important to note that there must be no spaces within the label, i.e., 2,4-di-CL not 2 4 di-CL. **qsar> seed** \[press ENTER] Yields the following table: See figure in Figures section. **? TROUBLESHOOTING:** If entry errors need to be corrected, users can enter **editsub** for editing substituents or **editdata** for editing data. **Editing Substituent:** Example- If 2nd substituent OC_6H_{13} should be OC_6H_5, then **qsar> editsub 2** \[press ENTER] A substituent label box will be obtained, Now one can simply correct the substituent and then press ENTER. **qsar> save** \[press ENTER] Editing Data: Example- If log1/C value for cpd #1 should be 3.61 and not 4.61, then **qsar> editdata 1** \[press ENTER] A substituent box \(having typed 1) will be obtained. Again pressing ENTER gave a blank parameter box, type 3 in the box \(because log1/C is the parameter 3) and presses ENTER. This will give a blank box for the parameter value. Now user can simply type the correct value of log1/C and press ENTER. **qsar> save** \[press ENTER] It is advisable to save the data entry frequently by entering save command. \(**iv**) **Entering Structures via SMILES:** \(**a**) If the data set is not based on a parent structure, then the SMILES for each compound should be entered one at a time and auto-loading of parameters is not possible. \(_This step should be avoided if the data set is based on parent structure_) **qsar> getsmi** \[press ENTER] will provide a panel with a prompt for structure 1. When the SMILES of structure #1 is entered, 'pressing ENTER' will display the 2-D structure. If the structure needs editing, enter 'y' if not, enter n and the prompt for the second SMILES will appear. Since there are a large number of SMILES stored in the database together with name\(s), the name can be entered at this point and the SMILES will be picked up from the database. When all the SMILES have been added, then enter end at the panel for next compound and \[press ENTER], the prompt will return to qsar \(qsar>). **qsar> save** \[press ENTER] **? TROUBLESHOOTING:** Common name of the compound/drug should be used to enter their structure, i.e. acetic acid not ethanoic acid; p-chlorophenol not 4-chlorophenol. This can be a very time saving procedure for complex structures such as strychnine. \(**b**) If the data set is based on a

parent structure, then the SMILES for the parent compound should be entered and auto-loading of parameters is possible. The present example is a set of 4-X-phenols, automatic loading is to be used, and the parent structure is entered via **getsmi /parent** . A panel is displayed into which one enters the SMILES with an * for each substituent position. For the present example: **qsar> getsmi /p** \[press ENTER] A panel is displayed into which one enters the SMILES with an ** for the substituent at 4-position, and a proper SMILES for the parent is: Oc1ccc\(**)cc1. Pressing ENTER should then display: See figure in Figures section. If the parent structure is not correct, enter y for editing. When the parent structure is correct, enter n and the prompt return to qsar \(qsar>). **qsar> getsmi** \[press ENTER] A panel is displayed for the entry of first compound. *N \[press ENTER] to see 4-aminophenol. If editing is not needed, enter n and then the panel will ask to enter the second structure and on so on: See figure in Figures section. Now enter **end** at the panel for next compound and \[press ENTER], the prompt will return to qsar \(**qsar>**). **qsar> depict** , \[press ENTER] To check that all SMILES correspond to the substituent name. If a particular SMILES is incorrect, then **qsar> editsmi #** \(the compound number) \[press ENTER] and make the correction **qsar> save** \[press ENTER] **? TROUBLESHOOTING**: The SMILES \(Simplified Molecular Input Line Entry System) is a language for linear entry of complex structures of organic compounds into computer, which was invented by David Weininger.13 \(**v**) **Auto-Loading of Parameters:** \(**a**) **Physicochemical parameters for the whole molecule \(Clog P, Mlog P, CMR, NVE, MgVol, and MW):** Auto-loading of these six parameters is for the both types of data sets, which is either based on the parent structure or not. **qsar> addcal** \[press ENTER] This will auto-load four parameters \(Clog P, CMR, NVE, and MgVol) **qsar> add mlogp** \[press ENTER] Mlog P will be auto-loaded **qsar> add mw** \[press ENTER] MW will be auto-loaded **qsar> seed** \[press ENTER] To see all the parameters in tabular form **qsar> save** \[press ENTER] To save the data. \(**b**) Physicochemical parameters for the substituents \(forty-four parameters): These physicochemical parameters for the substituents \(**Table 4**) can be auto-loaded by using fetch command \(**f**) if the structures of the data set are entered by the use of PARENT SMILES. For the present example it can be demonstrated as: **qsar> f** \[press ENTER] This will show: See figure in Figures section. **Pick one or more positions to parameterize: 1** \[press ENTER] \(For this example, only one choice is possible) A list of parameters obtained \(**Table 4**). **Enter rangelist of parameters: 16** \[press ENTER] \(users can enter here more parameter numbers as required, but one space must be between two parameter numbers. Pressing ENTER will auto-load all these parameters) **10 values added for S-P+-1** \(The values of σ+ for the substituents at 4-positions are auto-loaded) Now we need to edit **S-P+-1** to **S+**, which will be possible as follows: **qsar> editp 11 \(parameter level 11)** \[press ENTER] A box of parameter label will obtained, one should edit **S-P+-1** to **S+** by simply the use of BACKSPACE. **qsar> save** \[press ENTER] To save the data. \(**vi**) **Plotting Data**: Any two parameters can be plotted against each other by the command: **qsar> x gr y , i.e. 3 gr 4** \[press ENTER] gives some idea about the nature of fit to the most important variables that is linear, parabolic, or bilinear. In the present instance this is of little help. \(**vii**) **Permuting**: **qsar> 3 perm 4 5 6 7 8 10 11** \[press ENTER] derives all possible equations for 1, 2, and 3 variables. S.D. is the standard deviation. As this value decreases, the quality of the fit increases. CONST is the value of the constant in each QSAR equation. Clearly, BDE is the most important parameter and then σ+ and Clog P. **1 TERM REGRESSIONS** See figure in Figures section.

2 TERM REGRESSIONS See figure in Figures section. **3 TERM REGRESSIONS** See figure in Figures section. \(**viii**) **Checking for Parameter Collinearity and derivation of QSAR:** We can check for overall collinearity problems as follows: **qsar> corr 4 5 6 7 8 10 11** \[press ENTER] **CORRELATION MATRIX: R^2 and N** See figure in Figures section. There is a high collinearity between BDE and σ+ and also the high collinearity among Clog _P_ , CMR, NVE, MgVol, and MW. The upper part of the matrix gives correlation among the 10 different substituents. Next exploring the two best terms, the following QSAR can be derive: **qsar> 3 reg 5 4** \[press ENTER] will give the following equation: See figure in Figures section. Since, there is a high collinearity between BDE and σ+, BDE can be replaced by σ+, thus **qsar> 3 reg 5 11** \[press ENTER] will give the following equation: See figure in Figures section. **qsar> pred** \[press ENTER] will list the data in tabular form: \! http://www.nature.com/protocolexchange/system/uploads/328/original/328.jpg\! **? TROUBLESHOOTING:** The following commands are used for the derivation of different types of QSAR in the MLR analysis: \(i) Linear Model **qsar> x reg y** \[press ENTER] \(x and y are the parameter numbers) i.e. **qsar> 3 reg 4** \[press ENTER] The QSAR may be obtained by the use of more than one parameter i.e. **qsar> 3 reg 4 5 6 7 8** \[press ENTER], etc \(ii) Parabolic / Inverted Parabolic Model **qsar> x reg Py** \[press ENTER] \(x and y are the parameter numbers, and P is the command for parabolic derivation of the QSAR), i.e., **qsar> 3 reg P4** \[press ENTER] The QSAR for more than one parameter can be derive as: **qsar> x reg Py z** \[press ENTER] \(iii) Bilinear / Inverted Bilinear Model **qsar> x reg By** \[press ENTER] \(x and y are the parameter numbers, and B is the command for bilinear derivation of the QSAR), i.e., **qsar> 3 reg B4** \[press ENTER] The QSAR for more than one parameter can be derive as: **qsar> x reg By z** \[press ENTER] \(**ix**) **Jackknifing:** This process is used for the detection and removal of the outliers. Since the present example \(QSAR 3 or 4) is the good model and not have any outlier, jackknifing will not be used. To understand this process, we consider the data set B11019. **qsar> load /d 11019** \[press ENTER] will transfer the set to the workspace **qsar> star /d** \[press ENTER] Asterisks will be removed to restore all the data points for study. **qsar> sum** \[press ENTER] will list the key items of the data set as follows: See figure in Figures section. **qsar> 3 reg B4** \[press ENTER] will give the following equation: See figure in Figures section. This is obviously not a good equation. We can now use jackknifing by the following command: **qsar> 3 j B4** \[press ENTER] will derive all possible regression equations by dropping a different data point in each instance. See figure in Figures section. Dropping compound #19 yields $_r_^2$ = 0.782. To delete this compound, the following command will be use: **qsar> star /a 19** \[press ENTER] This will place an asterisk on the data point #19 and it will not be used in deriving future equations. **qsar> 3 reg B4** \[press ENTER] will give the following equation: See figure in Figures section. **qsar> 3 j B4** \[press ENTER] will derive all possible regression equations by dropping a different data point in each instance. See figure in Figures section. **qsar> star /a 22** \[press ENTER] This will place an asterisk on the data point #22 and it will not be used in deriving future equations. **qsar> 3 reg B4** \[press ENTER] will give the following equation: See figure in Figures section. **qsar> 3 j B4** \[press ENTER] will derive all possible regression equations by dropping a different data point in each instance. See figure in Figures section. Dropping compound #23 yields $_r_^2$ = 0.865. **qsar> star /a 23** \[press ENTER] This will place an asterisk on the data point #23 and it will not be used in deriving future equations. **qsar> 3 reg B4** \[press ENTER] will

give the following equation: See figure in Figures section. Note that _q_^{2} is now approaching _r_^{2} in value. This implies that dropping another point would have little effect on the correlation. Thus, this equation \(8) can be considered as the final QSAR model for this data set. **qsar> pred** \[press ENTER] will list the data in tabular form. **? TROUBLESHOOTING:** Any number of data points can be omitted from this fashion: **qsar> star /a 1 4 8 10,15** \[press ENTER] would place asterisks on the data points # 1, 4, 8, and 10-15. An asterisk can be removed by the following command: **qsar> star /d 15** \[press ENTER] will remove asterisk and restore the data point #15 **qsar> star /d 10,15** \[press ENTER] will remove asterisks and restore the data point #10-15 **qsar> star /d** \[press ENTER] will remove all asterisks from the data set and restore all the data points. \(**x**) **Editing:** This is the important tool to correct the errors in the data set. **qsar> editset** \[press ENTER] This displays all the data and puts one into a general edit mode allowing: \(1) movement of the cursor by the arrow key, \(2) the use of the delete key to erase the character to the left of the cursor, and \(3) the insertion of new characters at that spot. A new variable can also be added by first assigning it a symbol and then entering the values. When finished, one must exit from this mode, using control Z. For the minor change a quick mode is also available utilizing the following command: **qsar> editdata** \[press ENTER] will provide the prompt for substituent number. Entering that number and pressing ENTER will give the prompt for parameter number and the current value will display. Entering the parameter number and pressing ENTER will give a box for parameter value. The correct value of that parameter should be entered and then presses ENTER. After editing is complete, the command seedata \(**seed**) enables one to check the results. \(**xi**) **Deleting**: It is convenient to use the delete command to clean up the set. However, parameters cannot be deleted if an equation has been saved. To check for saved equations: **qsar> eq run** \[press ENTER] will list all the saved equations for the present data set. **qsar> del /eq** \[press ENTER] will delete all the saved equations from the present data set. **qsar> del /eq 2** \[press ENTER] will delete only the 2nd equation from the present data set. **qsar> seep** \[press ENTER] will display the parameter numbers **qsar> del /para 4** \[press ENTER] will delete the parameter 4. **qsar> del /para 4 8 10,16** \[press ENTER] will delete the parameters 4, 8, and 10-16. To delete a data point and the information associated with it including the SMILES, the following command should be used. **qsar> del /sub 4** \[press ENTER] will delete the data point #4 \(compound #4). Similarly, more data points can be deleted. h3. D. VALIDATION OF QSAR MODEL \(EXAMPLE: QSAR 8; SET NO. B11019) 1. Statistical Diagnostics: \(i) Number of descriptors: The ideal ratio = data points \(compounds)/descriptor ≥ 4. **Number of data points/number of descriptors = 20/3 = 6.67** \(ii) Squared correlation coefficient \(_r_^{2}): Closer the value of _r_^{2} to unity, the better the QSAR model. According to the literature, the predictive QSAR model must have _r_^{2} > 0.60.[14] **_r_^{2} = 0.865** \(iii) Standard Deviation \(s): It is believed that the smaller the value of s \(s ≤ 0.3), the better the QSAR model.[15] **s = 0.096** **2. Internal Validation:** \(i) Cross-validated _r_^{2} \(_q_^{2}): It has been suggested that the value of q2 must be greater than 0.50 for a predictive QSAR model.[14] **_q_^{2} = 0.819** \(ii) Quality factor \(Q): The Q is the quality factor \(quality ratio), where Q = r/s. **_Q = r/s_ = 0.930/0.096 = 9.688** \(iii) Fischer statistics \(F): The F-value \(Fischer ratio) is the ratio between explained and unexplained variance for a given number of degree of freedom. It indicates a true relationship, or the significance level for the MLR models. See figure in Figures section. \(iv) Y-

randomization Test: At present, we not have the facility to do this test automatically from this program, but it can be done manually. In this test, the dependent-variable vector \(Y-vector) is randomly shuffled and a new QSAR model is developed using the original independent variable matrix. This process is repeated several times. It is expected that the resulting QSAR models should have low r^2 and low q^2 values. Alternatively, this test can be done automatically from our Bio-Loom program \("http://www.bio-loom.com":http://www.bio-loom.com or "http://www.biobyte.com":http://www.biobyte.com) for any data set saved in C-QSAR program. For the present example the Y-randomization test results obtained from the Bio-Loom program are as follows: See figure in Figures section. **3. External Validation:** It is a better method that removing a percentage of the training set into a test set. The QSAR model is derived using the reduced training set, and the properties of the test set predicted using this model. Following are the methods for the selection of training and test sets: \(i) Random selection \(ii) Selection based on biological activity data \(iii) Various systematic clustering techniques \(iv) Self-organizing map \(SOM) \(v) Kennard Stone method \(vi) Formal statistical experimental design \(factorial and D-Optimal) \(vii) Sphere-exclusion algorithms

References

1. C-QSAR Program, BioByte Corp., 201W. 4th st., Suit 204, Claremont, CA 91711, USA. www.biobyte.com 2. Hansch, C., Hoekman, D., Leo, A., Weininger, D. & Selassie, C.D. Chem-bioinformatics: Comparative QSAR at the interface between chemistry and biology. _Chem. Rev._ **102**, 783-812 \(2002). 3. Hansch, C. Hoekman, D. & Gao, H. Comparative QSAR: Toward a deeper understanding of chemicobiological interactions. _Chem. Rev._ **96**, 1045-1075 \(1996). 4. Hansch, C. & Leo, A. Exploring QSAR: Fundamentals and Applications in Chemistry and Biology, American Chemical Society, Washington, D.C. \(1995). 5. Hansch, C. Leo, A. & Taft, R.W. A survey of Hammett substituent constants and resonance and field parameters. _Chem. Rev._ **91**, 165-195 \(1991). 6. Schultz, T.W. & Cronin, M.T.D. Response-Surface Analyses for Toxicity to Tetrahymena pyriformis: Reactive Carbonyl-Containing Aliphatic Chemicals. _J. Chem. Inf. Comput. Sci._ **39**, 304-309 \(1999). 7. Zhang, L., Gao, H., Hansch, C. & Selassie, C.D. Molecular orbital parameters and comparative QSAR in the analysis of phenol toxicity to leukemia cells. _J. Chem. Soc. Perkin Trans_ **2**, 2553-2556 \(1998). 8. Hu, J. Eriksson, L., Bergman, A., Jakobsson, E., Kolehmainen, E., Knuutinen, J., Suontamo, R. & Wei, X. Molecular orbital studies on brominated diphenyl ethers. Part II-reactivity and quantitative structure-activity \(property) relationships. _Chemosphere_ **59**,1043-1057\(2005). 9. Hansch, C., Maloney, P.P., Fujita, T. & Muir, R.M. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. _Nature_ **194**, 178-180 \(1962). 10. Leo, A.J. Calculating log Poct from structures. _Chem. Rev._ **93**, 1281-306 \(1993). 11. Leo, A.J. & Hansch, C. Role of hydrophobic effects in mechanistic QSAR. _Perspect. Drug Discovery Des._ **17**, 1-25 \(1999) 12. Cramer III, R.D., Bunce, J.D., Patterson, D.E. & Frank, I.E. Cross validation, Bootstrapping and partial least squares compared with multiple regression in conventional QSAR studies. _Quant. Struct.-Act. Relat._ **7**, 18-25 \(1988). 13. Weininger, D. SMILES, a chemical language and information system.1. Introduction to methodology and encoding rules. _J. Chem. Inf. Comput. Sci._ **28**, 31-36 \(1988). 14. Golbraikh, A. & Tropsha, A. Beware

of q2\! _J. Mol. Graph. Modl._ **20**, 269-276 \(2002). 15. Wold, S. Validation of QSAR's. _Quant. Struct.-Act. Relat._ **10**, 191-193 \(1991).

Figures

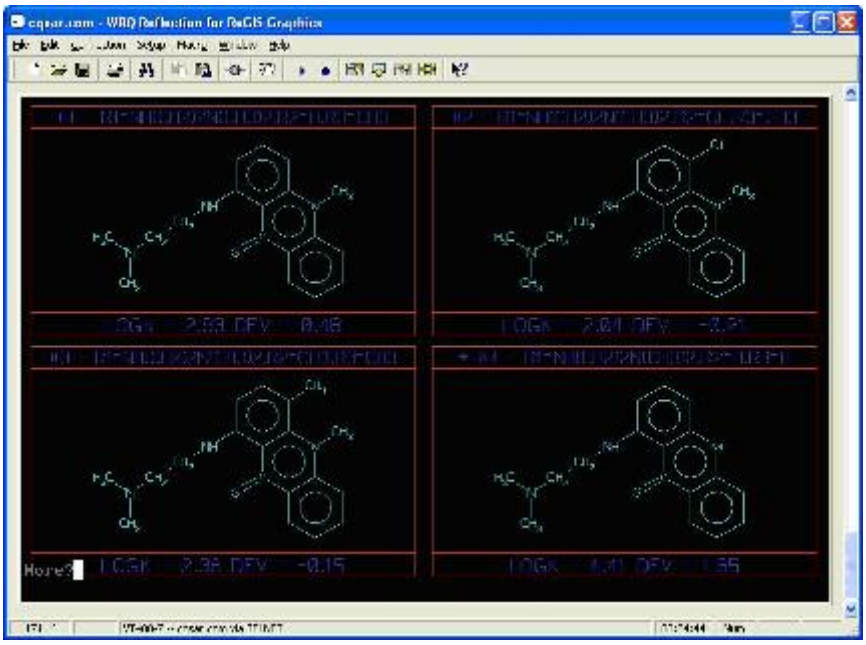

Figure 1

Figure 4 Sequentially depiction of four structures of a data set (#12675)

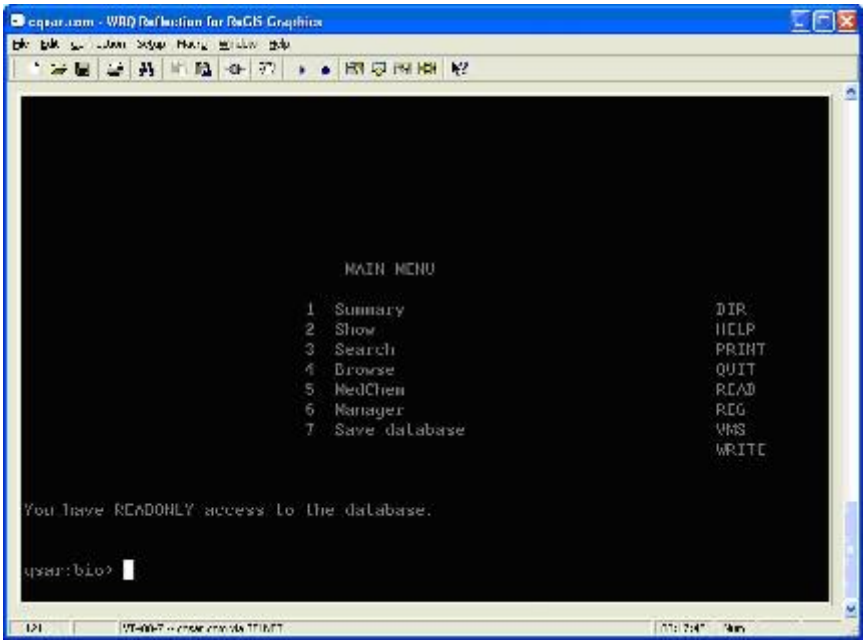

Figure 2

Figure 1 The main window of C-QSAR program when the program is started

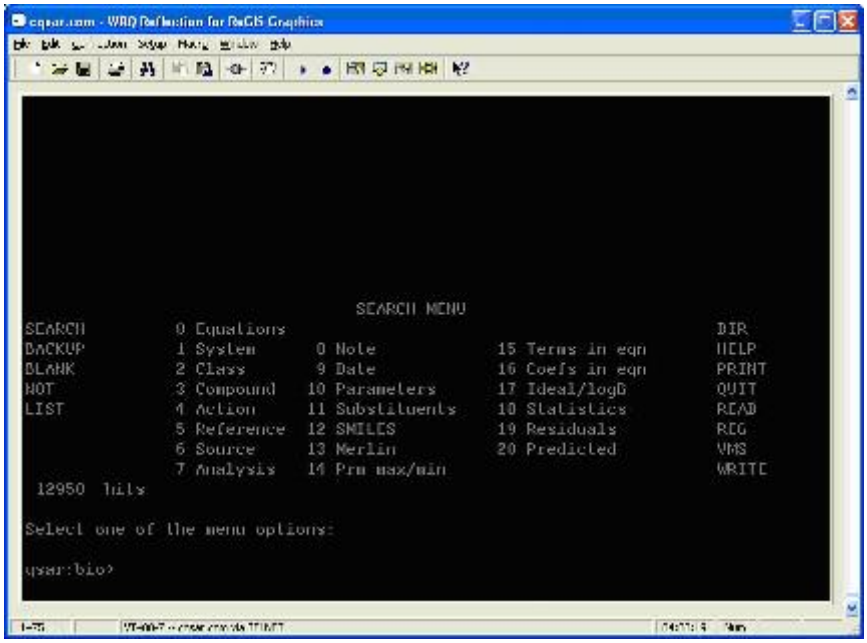

Figure 3

Figure 2 The search window of C-QSAR program for Bio-database

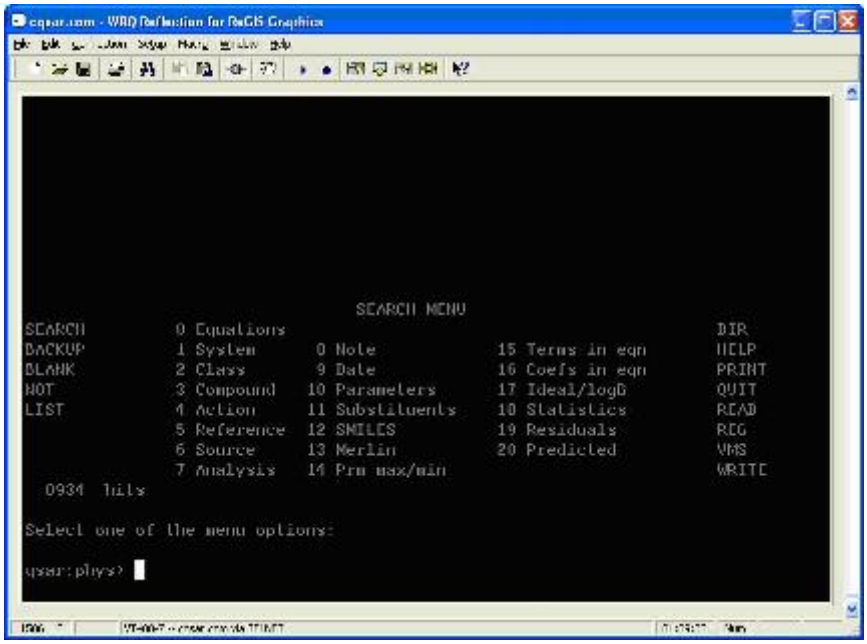

Figure 4

Figure 3 The search window of C-QSAR program for Phys-database