J Comput Aided Mol Des (2008) 22:873–883
DOI 10.1007/s10822-008-9219-2

QSID Tool: a new three-dimensional QSAR environmental tool

Dong Sun Park · Jae Min Kim · Young Bok Lee · Chang Ho Ahn

Received: 4 December 2007 / Accepted: 24 April 2008 / Published online: 30 May 2008

Abstract QSID Tool (Quantitative structure–activity relationship tool for Innovative Discovery) was developed to provide an easy-to-use, robust and high quality environmental tool for 3D QSAR. Predictive models developed with QSID Tool can accelerate the discovery of lead compounds by enabling researchers to formulate and test hypotheses for optimizing efficacy and increasing drug safety and bioavailability early in the process of drug discovery. QSID Tool was evaluated by comparison with SYBYL® using two different datasets derived from the inhibitors of Trypsin (Böhm et al., J Med Chem 42:458, 1999) and p38-MAPK (Liverton et al., J Med Chem 42:2180, 1999; Romeiro et al., J Comput Aided Mol Des 19:385, 2005; Romeiro et al., J Mol Model 12:855, 2006). The results suggest that QSID Tool is a useful model for the prediction of new analogue activities.

Keywords 3D QSAR · Binding affinity · PLS · Cross-validation · Molecular alignment · Neural network

Introduction

Quantitative structure–activity relationship (QSAR) [1] is an area of computational chemistry that builds statistical models to predict quantities such as binding affinity, acute toxicity or pharmacokinetic parameters of existing or hypothetical molecules. In more detail, QSAR is a mathematical relationship between the biological activity of a molecule and its geometric and physicochemical characteristics and this relationship can be used to evaluate the activity of new molecules. QSAR also represents an attempt to correlate structural or property descriptors of molecules with activities. These physicochemical descriptors, which include accounts for hydrophobicity, topology, hydrogen bonding, electrostatic field force and steric effects, are determined empirically. Activities used in QSAR calculations include chemical measurements and biological or biochemical data. QSAR is currently being applied in many disciplines, with many applications pertaining to drug design and prediction of drug toxicity. Drug design includes not only ligand design, but also pharmacokinetics and toxicity that are mostly beyond the possibilities of structure/computer-aided design. Researchers have used QSAR for many years to develop a drug that contains better efficacy or better solubility or less toxicity, for example. The use of QSAR consisted primarily of statistical correlations of structural descriptors with biological activities. Although easy access to computational resources was not available when these efforts were begun, scientists in these days can use high-performance computing power to understand and evaluate structural features of molecules that are correlated with binding affinities or other properties of interest. Of particular interest for biomedical research are 3D QSAR models because they allow the simulation of directional forces, such as hydrogen bonds, metal-ligand interactions, polarization effects and electric dipoles.

QSID Tool is dedicated to QSAR, as an integrated molecular modeling tool for drug discovery in order to build statistical models of relationships between molecular activity and structure, and to use these models to make

Electronic supplementary material The online version of this article (doi:10.1007/s10822-008-9219-2) contains supplementary material, which is available to authorized users.

D. S. Park (✉) · J. M. Kim · Y. B. Lee · C. H. Ahn
Rexahn Pharmaceuticals, Inc., 9620 Medical Center Dr., Rockville, MD 20850, USA
e-mail: parkds@rexahn.com

accurate predictions of the activity of untested molecules. QSID Tool organizes structures and their associated data as molecular descriptors, and performs sophisticated statistical analyses that reveal patterns in structure–activity data. This tool provides visualization as well as analysis of structure–activity relationships. It features multiple statistical methods for generating predictive models, including Partial Least Squares (PLS), as linear analysis, and neural network, as non-linear analysis, and provides cross-validation of models for confidence in predictability. Automatic calculation of molecular force fields with some parameters and results of statistical analyses can be provided with scatter plots, and distributions; a structures viewer and the spreadsheets interact with each other to facilitate exploration of the data. This tool designed to build reliable and predictive models of property from molecular structures also includes molecular field generation, PLS and non-linear analysis tools. In addition to steric and electrostatic fields, QSID Tool can calculate several types of fields, such as hydrophobic and hydrogen bonding force fields. Hydrogen bonding fields are created by assigning energies equal to the steric cutoff energy to grid points that are close to hydrogen accepting or donating atoms.

Although several web-based QSAR tools are available in the public domain [2, 3], they do not provide 3D force field calculations for molecular binding affinity and visualization with molecular properties in the packages. Those tools also have limitations in computing ability and speed on the Internet. QSID Tool was developed in JAVA technology and is a stand-alone application for drug discovery that can run on multi-platform systems, which is similar to SYBYL® of Tripos and Cerius2® of Accelrys. However, QSID Tool has integrated molecular modeling functionality modules to build statistical QSAR models with open-source libraries.

Materials and methods

Development of QSID Tool

Most public or commercial applications have their own functionalities and unique user interfaces on UNIX-based platforms. Therefore they need to adopt other tools to perform the program because of different operational requirements. Mathematical statistics concepts, molecular dynamics (MD), and basic chemical properties are needed to design and develop new QSAR-based molecular modeling environments. QSID Tool was developed with open-source libraries that enabled an integrated and easy to use user interface, and allowed more reliable analytical methods [4–7]. QSID Tool, which was written in JAVA programming language, runs on a virtual machine, irrespective of any hardware and operating system. It enables users to design new ligand-based molecules without additional tools by providing integrated analysis and data-related functions in a single application. QSID Tool is also integrated with several other modules, consisting of visualization for user interface, input collection, calculation, interpretation and prediction. Therefore, the main purpose of QSID Tool is to give users an easy and useful tool for ligand-based modeling by providing several modules in one package. The key modules in QSID Tool are described in more detail below.

Molecular description

QSID Tool supports several file formats, including MDL mol, SYBYL® mol2, and PDB, with respect to processing molecule files. While molecules are being loaded into QSID Tool, basic molecular properties, including atom type, atom weight, partial charge, contribution of log*P*, hydrogen bonding, and solvent-accessible surface area, are automatically calculated.

Basic molecular properties Descriptors that depend on the partial charge of each atom in a chemical structure require calculation of those partial charges. Partial charge is necessary for electrostatic force fields, which either come from the Tripos Mol2 file containing the charge of atoms or will be calculated automatically in QSID Tool if the charges are not included in a loading file. The AtomPartialCharge function implemented in JOELib is adopted for Gasteiger–Marsili empirical atomic partial charge calculation among numerous methods of calculating partial charges [8, 9]. Log*P* is the logarithm of the ratio of the concentrations of the un-ionized solute in two solvents, which is calculated according to Eq. 1, where *o* is octanol and *w* is un-ionized water.

$$\log P_{o/w} = \log\left(\frac{[\text{solute}]_o}{[\text{solute}]_w}\right) \quad (1)$$

The hydrophobic effect is the major driving force for the binding of drugs to their receptor targets in pharmacodynamics, and is based on the log*P* contribution of each atom. The contribution of log*P* is calculated using the GroupContribution function in JOELib [10]. Source 1 below shows how to apply the JOELib library for the calculation of contribution of log*P*. Each atom in a molecule contributes to the log*P* by the amount of its atomic parameter multiplied by the degree of exposure to the surrounding solvent. This degree of exposure is calculated from the solvent-accessible surface area (SASA) [11].

```
GroupContributions gc = null;
gc = JOEGroupContribution.instance().getGroupContributions("logP");
atomValues[atoms[0] - 1] = ( (Double) gc.atomContributions.get(i)).doubleValue();
```

Source 1 Calculation of contribution of log*P* for each atom using JOELib library

SASA also can be calculated (or assigned) for each atom using the 'rolling ball' algorithm developed by Shrake and Rupley [12]. This algorithm uses a sphere of solvent of a particular radius to 'probe' the surface of the molecule. The choice of the 'probe radius' does have an effect on the observed surface area, as using a smaller probe radius detects more surface details and therefore reports a larger surface. A typical value is 1.4 Å, which approximates the radius of a water molecule. Another factor that affects the results is the definition of the van der Waals (VDW) radii of the atoms in the molecule under study. QSID Tool has adopted C-based SASA written by Le Grand [13].

Calculation of force fields for molecular binding affinity A force field refers to the function and parameter sets used to describe the potential energy on each lattice around molecules. Basically, five functions of force-field calculation are adopted in the tool derived from both empirical parameters and pharmacokinetic calculations. The force calculation encapsulates non-bonded interactions with each lattice on the grid describing the electrostatic and steric forces. Steric force calculation is usually computed with a Lennard-Jones potential. The Lennard-Jones potential (E_{lj}) [14] is an effective potential that describes the interaction between two uncharged atoms and is described by Eq. 2. It is mildly attractive as two uncharged atoms approach one another from a distance, but strongly repulsive when they approach too close. In Eq. 2, ε and σ are the specific Lennard-Jones parameters, where ε is the depth of the potential well and σ is the distance at which the potential is zero, and r is the distance between two uncharged atoms.

$$E_{lj} = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \tag{2}$$

The electrostatic potential (E_{el}) defined in Eq. 4 is computed with the forces exerted by static electric charge field upon each charged atom. Electrostatics involves the build-up of charge in objects due to contact between non-conductive surfaces. These charges are generally built up through the flow of electrons from one object to another. These charges then remain in the object until a force is exerted that causes the charges to balance. The fundamental equation of electrostatic interactions is Coulomb's law (F) in Eq. 3, which describes the force between two point charges Q_1 and Q_2 with distance r.

$$F = \frac{Q_1 Q_2}{4\pi\varepsilon r^2} \tag{3}$$

The electric field is defined as the force per unit charge. In QSID Tool, unit charge (Q_1) is defined for carbon with +1 charge. From this definition and Coulomb's law, the electrostatic potential of a point charge Q_2 is given by Eq. 4.

$$E_{el} = \frac{Q_2}{4\pi\varepsilon r} \tag{4}$$

The applied hydrophobic calculation is an atomistic approach to estimate molecular hydrophobicity. It takes into account the proximity effect of substituent groups as well as the importance of solute-solvent interaction in the partition phenomena. This method reassigns atomic parameters when the molecule is fully exposed to the surrounding solvent. Each atom in a molecule contributes to the log*P* by an amount of its atomic parameter multiplied by the degree of exposure to the surrounding solvent. This degree of exposure is calculated from SASA. Source 2 shows the calculation of hydrophobic force for each atom with SASA and the contribution of log*P*.

The contribution of hydrogen bonding is adopted from the empirical method in SCORE28 [15]. Hydrogen bonding is one of the key features for a specific binding process. Such interaction may happen when two atoms get close enough and form a donor–acceptor pair. A hydrogen bond donor is defined as a nitrogen or oxygen atom with hydrogen attached; while an acceptor is defined as a nitrogen, oxygen, or fluorine atom with at least one vacant valence to accept a hydrogen atom. The geometry of a hydrogen bond is characterized by two parameters: the bond length, i.e. the distance between donor and acceptor, and the bond angle, i.e. the angle among donor-hydrogen-acceptor. A hydrogen bond is possible only when the bond length is shorter than the sum of a van der Waals radii of donor and acceptor. Typical kinds of hydrogen bond angles are not likely to be less than 70°. An angle cutoff is set in the algorithm for defining a hydrogen bond. The atoms involved in hydrogen bonding are assigned by JOELib [10].

```
Atom[] targetAtoms = mol.getAllAtoms();
  for (int i = 0; i < targetAtoms.length; i++) {
    distance = targetAtoms[i].getPoint3d().distance(probe.getPoint3d());
        hydrophobicE += targetAtoms[i].getContriblogP ()
      * targetAtoms[i].getSasa()
      * Math.exp( -1 * distance);
  }
```

Source 2 Calculate hydrophobic force of each atom using SASA and contribution of Log*P*

In SCORE [15] empirical methods, the distance dependence of hydrogen bonding strength is gauged by using a step function, which is defined as Eq. 5,

$$\begin{aligned} \mathrm{SHB}(d) &= 1 \quad d < d_0 - 0.60 \\ &= 0 \quad \text{otherwise} \\ \mathrm{MHB}(d) &= 1 \quad d_0 - 0.60 \le d < d_0 - 0.30 \\ &= 0 \quad \text{otherwise} \\ \mathrm{WHB}(d) &= 1 \quad d_0 - 0.30 \le d < d_0 \\ &= 0 \quad \text{otherwise} \end{aligned} \tag{5}$$

where d represents the distance between donor and acceptor; d_0 represents the sum of a van der Waals radii of donor and acceptor. SHB, MHB, and WHB are indicators for strong, moderate, and weak hydrogen bonds respectively. In QSID Tool, the contribution of hydrogen bonding is the sum over all hydrogen bonds from the ligand and a probe as shown in Eq. 6.

$$\begin{aligned} K_{\mathrm{hbond}} &= K_{\mathrm{SHB}} + K_{\mathrm{MHB}} + K_{\mathrm{WHB}} \\ &= \sum_i \mathrm{SHB}(d_i) + \sum_i \mathrm{MHB}(d_i) + \sum_i \mathrm{WHB}(d_i) \end{aligned} \tag{6}$$

Building a QSAR model with cross-validation

Backpropagation neural network techniques with Java object oriented neural engine (Joone) as an open-source library [6] and PLS regression (programmed by us) have been implemented in QSID Tool to generate statistical models for molecular structure and biological activity relationships. Cross-validation methods are included in the analysis to obtain better predictive models.

PLS regression as linear analysis PLS regression [16–19] is a technique that generalizes and combines features from principal component analysis and multiple regressions. It is particularly useful when there is a need to predict a set of dependent variables (**Y**, bioactivity in QSAR) from a very large set of independent variables (**X**, molecular descriptors specifically in QSAR). It is used to build a linear model and to find the relations between two matrices by the following Eq. 7.

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{7}$$

where **Y** is an n by m variables response matrix, **X** is an n by p variables predictor matrix, **B** is a p by m regression coefficient matrix, n is the number of samples, m is the number of dependent variables, p is the number of independent variables, and **E** is an error for the model, which has the same dimensions as **Y**.

Neural network as non-linear analysis An artificial neural network (ANN) [20] is a computational model derived from a simplified concept of the brain, which can be used to model non-linear complex relationships between molecular descriptors and bioactivity or to find patterns in data. The most popular and widely used algorithm of ANN in QSAR studies is the backpropagation neural network (BNN) [21, 22] of the class of supervised learning techniques since the early 1990s [23]. A BNN is constructed from simple processing units called neurons which are arranged in a series of layers bounded by input and output layers encompassing a variable number of hidden layers. Each neuron is connected to other neurons in the network through connections of different strengths or weights as shown in Fig. 1. The most important feature of a neural network (NN) is its ability to learn from a training set and the learnt information is stored across the network weights. The adjustment of the weights is performed by delta-rule with momentum after presentation of each training pattern (online pattern training). The systematic search of the best values for the learning rate and the momentum are generally performed by means of a trial-and-error procedure.

The number of hidden layers in a NN depends on the complexity of the problem to solve, but in most cases one hidden layer consisting of an optional number of sigmoid neurons is sufficient to build a QSAR model. In QSID Tool, one hidden layer was initially selected and a trial-and-error procedure was employed to decide the number of neurons in the hidden layer. This approach uses a bottom-up strategy, which is starting with a small number of neurons and then increasing the number if needed.

A problem of overfitting occurs during the training with neural networks when an optimal time of training is exceeded (overtraining). It makes the predictive ability of the NN model worse. A crucial step in the training of a NN is the decision to stop the learning cycle at the right time to prevent overtraining. Cross-validation early stopping, which consists of a training set for adjusting the values of the weights and a validation set for deciding the time to cease the training phase, can be used to avoid overtraining [24]. The role of the validation set is to track the predictive abilities of the NN model by means of the root mean square error (RMSE) and the training can be stopped if there is no

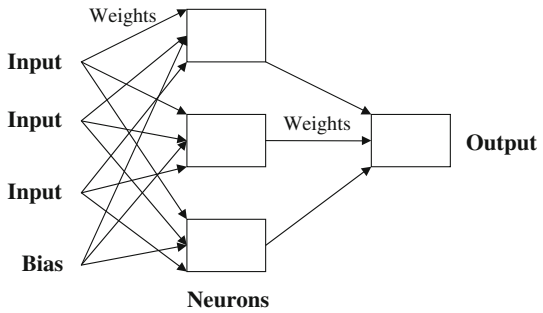

Fig. 1 Simple neural network showing connections for three inputs and three neurons

improvement. Another external testing set could be used to estimate the predictive ability of the NN model.

Cross-validation The predictive ability of the model generated is determined by using statistical cross-validation. Cross-validation is the statistical practice of partitioning a sample of data into N subsets to track the predictive abilities of the model. $(N - 1)$ subsets of these are initially used as training set in training a model, while the remaining subset is generally retained for subsequent use for the validation of the model. Holdout, K-fold and leave-one-out (LOO) cross-validation methods were implemented in QSID Tool.

Analysis of contour maps

QSID Tool can be represented as a 3D coefficient contour map. To visualize the steric, electrostatic, hydrogen bonding, hydrophobic fields from the analysis, contour maps of the product of the standard deviation associated with the molecular property column and the coefficient at each lattice point were generated. The contour maps were plotted as a percentage contribution to the QSAR equation and were associated with differences in biological activity.

Interpretation and validation of the model

While developing models, cross-validations with statistical coefficients (r^2, q^2, p^2) were selected to verify the predictive abilities of the generated models. The r^2 value was used to verify predictive performance for the training set, q^2 for cross-validation, and p^2 for the test set. The statistical coefficients were calculated by Eq. 8.

$$r^2 = q^2 = p^2 = \frac{\text{SSD} - \text{PRESS}}{\text{SSD}} \tag{8}$$

where SSD (sum of squared deviations) is $\sum [\text{target value} - \text{target mean}]^2$, and PRESS (predictive sum of squares) is $\sum [\text{target value} - \text{predicted value}]^2$.

The interpretation of the model, which means understanding of the rationale of a 3D QSAR model, depends on the modeling technique and the molecular descriptors/properties involved in the model. QSID Tool provides a descriptor contribution based on eigenvalues and eigenvectors in eigenspace from descriptor variables along with the PLS models developed. The descriptor contribution describes the more important descriptors involved. One major disadvantage in the NN models is the difficulty in interpreting the model, but it shows a reliable predictive power of a model. In order to gain any insight into a NN model, one should consider the weights to see the connections to or from a layer of a NN. One of the ways to represent and analyze the weights of a NN model is to use graphics. Among graphics, the Hinton diagram is widely used [25].

Prediction of the activity of new compounds

After generating the best model built with good statistical results, the biological activity of new compounds can be predicted. For accurate results, new molecules should be analogs of molecules that are used in training set.

Evaluation of QSID Tool

The evaluation presents both linear and nonlinear models to predict biological activity (specially, pIC_{50} or pK_i) for two sets of analogs. The purpose of the evaluation is to compare QSAR models developed using QSID Tool with those developed with SYBYL®, one of the leading tools on the market, in terms of interpretive and predictive functionalities.

Dataset

Seventy two analogs of trypsin inhibitors derived from 3-amidinophenylalanine, with pK_i values available at a SYBYL® reference (named dataset A, supplemental Table 1) [26] and 28 analogs of p38-MAPK inhibitors derived from pyridinyl-imidazole at the Laboratory of Molecular Modeling and Design, with pIC_{50} values (named dataset B, supplemental Tables 2 and 3) [27–29] were selected as a training set to evaluate the QSID Tool, as it was developed. Sixteen molecules from dataset A and five compounds from dataset B were not included in the training set, but were reserved as a test set to check the predictive power of the models obtained. The molecules of the test set were built and aligned by the same protocol as the inhibitors of the training set.

3D conformation and molecular alignment of the compounds

The development of a 3D QSAR model requires the alignment of the 3D structures of the molecules to a reasonable conformation, possibly the bioactive conformation. 3D structures of all molecules were generated using CORINA [30] and OBFit alignment of OpenBabel [7] or alignment of SYBYL® was used with root-mean square-deviation (RMSD) to obtain reasonable molecular alignments. This is considered to be a most crucial step in order to achieve reliable 3D QSAR models.

3D QSAR studies

A molecular force field consists of 3D grids, large enough to enclose all the aligned molecules, where in each grid

point interactions between a probe atom and each molecule are calculated. The interaction values in the grid points are thus utilized as variables in the following analysis methods.

Lowest energy conformations of all molecules were initially aligned and subsequently surrounded by a 3D grid with 2 Å in all directions. The surroundings of each molecule were mapped by calculating the interactions between probe atoms and each molecule at each grid point. The resulting grid, filled with interaction values, is called a molecular field. Molecular force fields were calculated using an sp3 carbon probe with a van der Waals radius of 1.52 Å and a charge of +1.0 to generate steric and electrostatic fields using a distance dependent dielectric at each grid point. An energy cutoff of 30 kcal/mol was applied at a point in the steric force field and electrostatic points were dropped with steric cutoff. The selected force fields were scaled by built-in transformation functions. The hydrogen donor and acceptor bonding were obtained from implementation of the empirical method in SCORE [15]. The PLS and NN methods were employed for the analysis of two datasets with a LOO cross-validation.

Biological data

The experimentally determined biological activities of the 88 trypsin inhibitors in dataset A are given as pK_i values that range from 3.00 to 7.70, as shown in supplemental Table 1. The potency of 33 imidazole p38-MAPK inhibitors in dataset B are given as IC_{50}, which were measured with the purified enzyme and range from 0.11 to 2,100 nM. The IC_{50} values were converted to molar units, and then expressed in negative logarithmic units, pIC_{50} that range from 5.68 to 9.96 as shown in supplemental Table 3.

Results and discussion

The new 3D QSAR environmental tool

QSID Tool was developed as a comprehensive QSAR molecular modeling system that enables one to establish structure–activity relationships, to predict quantitative biological activity, to analyze the contribution of structural properties, and to accelerate the discovery of lead compounds. QSID Tool provides basic molecular properties, including partial charges, log$P$ contribution, and solvent accessible surface area, of each atom in a molecule. Also this tool provides five force field calculations for molecular binding affinity and over 200 built-in traditional molecular descriptors (Table 1) for rapid processing of thousands of molecules. To build a QSAR model with cross-validation, PLS and neural network techniques were implemented and a genetic algorithm was employed to choose, automatically, the best set of descriptors to create 2D QSAR models from a given dataset.

Table 1 Traditional molecular descriptors in QSID Tool

Molecular descriptors	Count
Topological descriptors	116
Molecular properties	24
Geometrical descriptors	70

The following other useful functionalities are included in QSID Tool (see Fig. 2):

1. Multiple workspaces environment: QSID Tool supports multiple workspaces, resulting in an easy organization of the workflow and more working space. Instead of being limited to a single desktop, the user can quickly switch (under the file menu in the QSID Tool) between multiple workspaces to get some free room for more work.
2. Molecular viewer and properties: The molecular viewer built into QSID Tool allows zooming and rotating a molecular image and allows multiple molecules to be displayed within a single view panel, including the molecular property data, such as partial atom charge, ionization potential and molecular weights.
3. Alignment of the molecules: This tool has a function to superimpose molecules based on a SMARTS pattern [31]. The atoms used to fit the two molecules are defined by the SMARTS pattern. It is useful to align congeneric series of molecules on common structural scaffolds for 3D QSAR studies and to display the results of conformational generation.

Evaluation of QSID Tool

The aligned 3D molecular conformations of two datasets were obtained by superimposing common groups using CORINA and OBFit, before building QSAR models. To superimpose molecules, common groups that have a fixed position were selected from a common structural scaffold in the datasets. SMART pattern-based OBFit aligned each molecule with the core structure. The result for each dataset is shown in Fig. 3. Grid points were selected from a cubic size big enough to include all molecules but to exclude points out of the boundary margin of molecules. The setup parameters of both datasets included 2.0 Å intervals between grids and a 2.0 Å boundary margin as the default.

The major objective while working with QSID Tool is to build the best predictive model for design of new analogs and to figure out which force field around molecule affects its binding affinity. The 3D QSAR studies with inhibitors binding to Trypsin (dataset A) and p38-MAPK (dataset B)

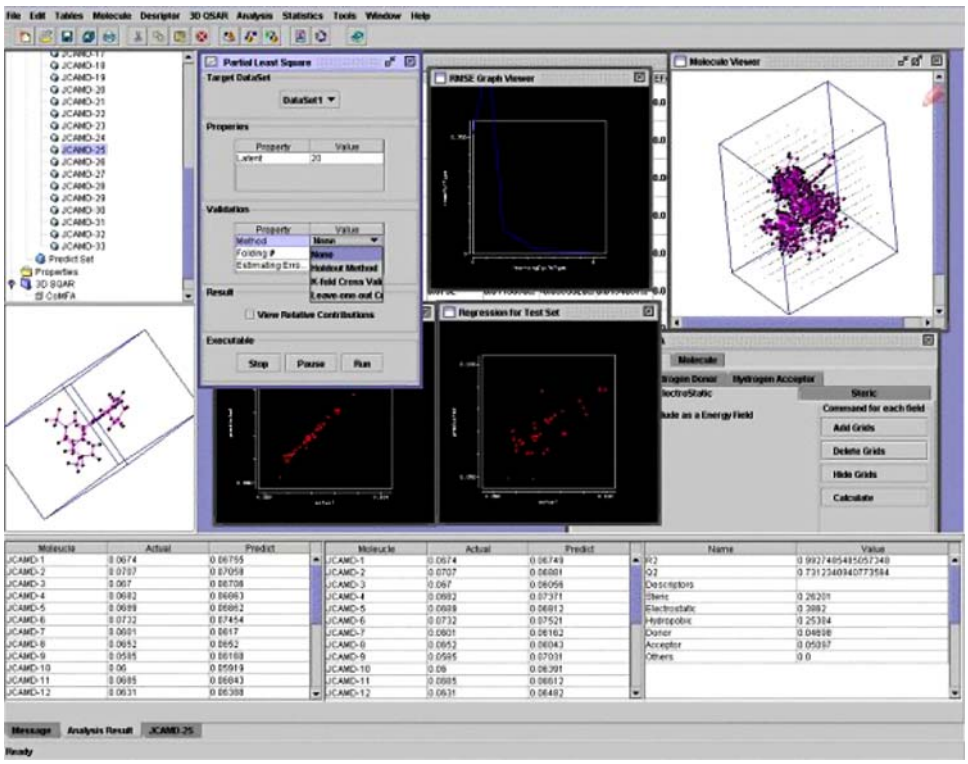

Fig. 2 The screenshot of QSID Tool working with dataset B

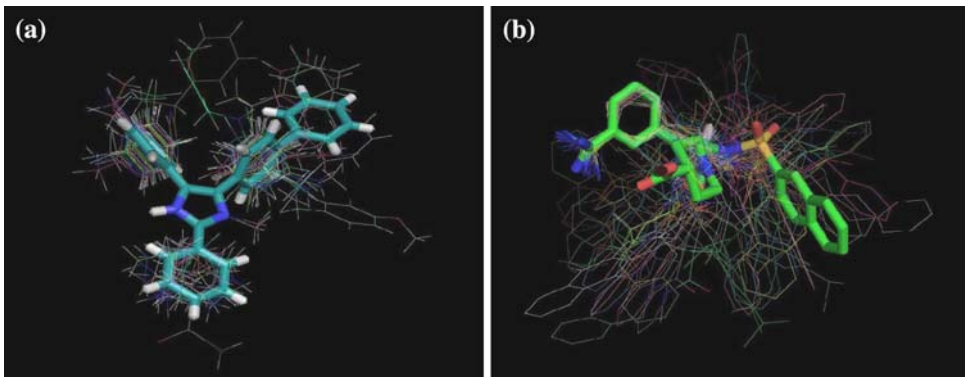

Fig. 3 Aligned confirmations generated and aligned by CORINA and OBFit (**a**) with dataset A, (**b**) with dataset B

were performed with PLS and neural network as methods of data analysis. Trypsin inhibitors in dataset A were chosen to compare predictive QSAR models with QSAR analysis tools in SYBYL®.

PLS linear model

The linear model was obtained by using PLS analysis. Several PLS models, with varying grid space, energy cut-off, or physicochemical parameters, were tried to improve the statistical results and were externally validated by LOO cross-validation. First, 88 compounds from trypsin inhibitors [26] were used for the evaluation and testing of QSID Tool. The test set (compounds 73–88, supplemental Table 1) was not included in the training set (compounds 1–72, supplemental Table 1) for model development, but was used to evaluate the predictability of the PLS model. The PLS models calculated from the combination of force fields, such as steric or electrostatic or hydrophobic or hydrogen bonding, for dataset A are shown in Table 2 with the statistical results and the contribution of each function. The best predictive PLS model, with $r^2 = 0.99$ and $q^2 = 0.65$ in those combination models, was obtained when all four fields were used. (1)-a of Fig. 4 shows a graphical representation of the predictability of the best predictive PLS model for the training set of 72 compounds in dataset A. For the 16 test compounds, a p^2 value of 0.74 was obtained for the best PLS model in dataset A, as shown in (1)-b of Fig. 4. Table 2 also provides statistical variances to understand the contribution of force fields affecting the molecular binding affinity. The variance for the best PLS model in dataset A is composed of 33.1% of

Table 2 Statistical results and contributions from the two models calculated with SYBYL®, and the six models calculated with the QSID Tool with dataset A

	SYBYL®[a]		QSID Tool					
	CoMFA	CoMSIA	Steric + electrostatic		Steric + electrostatic + hydrophobic		Steric + electrostatics + hydrophobic + H-bonding	
	PLS	PLS	PLS	NN	PLS	NN	PLS	NN
q^2	0.63	0.75	0.52	0.55	0.64	0.83	0.65	0.93
p^2	0.65	0.84	0.51	0.64	0.99	0.65	0.74	0.87
r^2	0.92	0.97	0.99	0.98	0.99	0.99	0.99	0.99
SEE	0.26	0.16	0.12	0.21	0.12	0.08	0.12	0.12
SSE	–	–	0.98	3.09	1.05	0.47	0.96	1.08
F	144.4	240.1						
N	5	9						
Fraction								
Steric	0.66	0.17	0.49		0.32		0.24	
Electrostatic	0.34	0.16	0.51		0.32		0.28	
Hydrophobic		0.30			0.36		0.33	
H-donor		0.10					0.09	
H-acceptor		0.28					0.06	

[a] Statistic values from Ref. [26]

PLS: partial least squares; NN: neural network; SEE: standard error of estimate; SSE: sum of squares of residuals; Correlation coefficients: r^2 for training set, q^2 for cross-validation, p^2 for test set

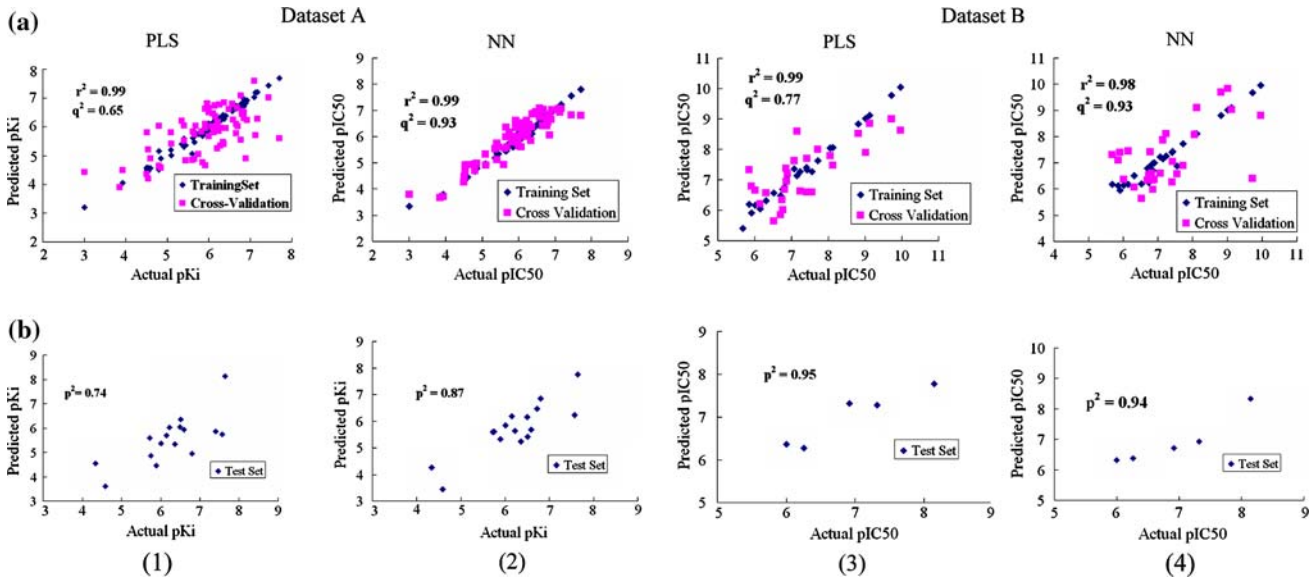

Fig 4 Graphical comparison of (**a**) predicted versus actual inhibition constant (pK_i or pIC_{50}) for the training set using the best predictive models with LOO cross-validation and (**b**) predicted versus actual inhibition constant for the test set using the best models. (1) PLS model with steric + electrostatic + hydrophobic + H-bonding force fields for dataset A. (2) NN model with steric + electrostatic + hydrophobic + H-bonding force fields for dataset A. (3) PLS model with steric + electrostatic force fields for dataset B. (4) NN model with steric + electrostatic force fields for dataset B

the hydrophobic field, 28.2% of the electrostatic field and 24.3% of the steric field; this suggests that these three fields are the major contributing factors in this best PLS model in dataset A. In the combination model of three fields, e.g. steric, electrostatic and hydrophobic, r^2 of 0.99 was obtained and q^2 was 0.64, as shown in Table 2 and (2)-A of supplemental Fig. 1 for the training set of 72 compounds. The predictability of this model is close to the combination model of four fields. The combination model of three fields showed a higher p^2 value (0.99) than the combination model of four fields (0.74). The combination model of steric and electrostatic fields in PLS analysis showed the lowest predictability among the three PLS models for dataset A. On the other hand, the CoMFA model gave values of 0.92 for r^2, 0.63 for q^2 and 0.65 for p^2 with more contribution from the steric function; the CoMSIA model yielded 0.97 for r^2, 0.75 for q^2 and 0.84 for p^2, which is close to the predictability of the PLS models from the combination of three or four fields. The results from dataset A suggest that QSID Tool is comparable to the PLS models (CoMFA and CoMSIA) of SYBYL® and as effective in prediction, as shown in Table 2.

Secondly, 33 inhibitors for p38-MAPK [27–29] were selected for further evaluation and testing of QSID Tool (supplemental Tables 2 and 3, dataset B). Among them, 28 compounds were designated as the training set and 5 compounds (compounds 7, 11, 14, 23 and 27) as the test set. Table 3 summarizes the statistical results of PLS models for dataset B in which the result was obtained by the combination of each field. The best PLS model in dataset B in Table 3 is the combination of steric and electrostatic fields with 0.99 for r^2 and 0.77 for q^2. (3)-A in Fig. 4 shows a graphical representation of the predictability of the best predictive PLS model for the training set of 28 compounds in dataset B. For the five test compounds, $p^2 = 0.95$ was obtained from the best PLS model for dataset B, as shown in Table 3 and (3)-b of Fig. 4. The variance for the best PLS model in dataset B is composed of 50.5% electrostatic field and 49.5% steric field and this suggests that these two fields are the major contributing factors in this best PLS model in dataset B. But the other two PLS models from the combination of three or four force fields in dataset B resulted in low predictability compared to the PLS model from the combination of steric and electrostatic fields, as shown in Table 3.

Neural network nonlinear model

The nonlinear model was obtained by using the neural network analysis in QSID Tool. The neural network models calculated from the combination of force fields, such as steric or electrostatic or hydrophobic, or hydrogen bonding for dataset A are shown in Table 2 for dataset A and in Table 3 for dataset B. Because the preprocessing of the input data plays a key role in the ability of NN models, normalization was executed by z-score (Eq. 9) before

Table 3 Statistical results and contributions from the six models generated with the QSID Tool with dataset B

	QSID Tool					
	Steric + electrostatic		Steric + electrostatic + hydrophobic		Steric + electrostatic + hydrophobic + H-bonding	
	PLS	NN	PLS	NN	PLS	NN
q^2	0.77	0.93	0.69	0.92	0.70	0.91
p^2	0.95	0.94	0.70	0.93	0.64	0.94
r^2	0.99	0.98	0.99	0.99	0.99	0.99
SEE	0.16	0.22	0.20	0.16	0.15	0.19
SSE	0.70	1.30	1.08	0.71	0.65	1.01
Fraction						
Steric	0.495		0.41		0.37	
Electrostatic	0.505		0.33		0.29	
Hydrophobic			0.26		0.23	
H-donor					0.05	
H-acceptor					0.06	

PLS: partial least squares; NN: neural network; Correlation coefficients: r^2 for training set, q^2 for cross-validation, p^2 for test set

building models. μ is the mean and σ is the standard deviation of the probability distribution of X in the following Eq. 9.

$$Z = \frac{X - \mu}{\sigma} \tag{9}$$

Several attempts were made to improve the statistical results by varying the learning rate, momentum, learning cycle, and the number of neurons in a hidden layer. The best neural network model for dataset A was obtained from the combination of steric, electrostatic, hydrophobic and hydrogen bonding fields with $r^2 = 0.99$, $q^2 = 0.93$ and $p^2 = 0.87$, as shown in Table 2 and (2)-a in Fig. 4. This combination model was generated with nine neurons in the hidden layer, learning rate of 0.3, momentum of 0.4, by a trial-and-error procedure, and the phase of the best LOO cross-validation result (q^2) among 500 of learning cycle. The second best model is the combination of steric, electrostatic and hydrophobic fields with $r^2 = 0.99$, $q^2 = 0.83$ and $p^2 = 0.65$, as shown in Table 2. Supplemental Fig. 3 shows the scatter plot of the predicted pK_i against experimental (actual) pK_i for the trypsin inhibitors of training set and test set.

The NN models from the combination of force field calculations for dataset B were also generated with nine neurons in the hidden layer, learning rate of 0.5, momentum of 0.6, by a trial-and-error procedure, and the phase of the best LOO cross-validation result (q^2) among 500 of learning cycle. Table 3 summarizes the statistical results of the obtained best NN models from the combination of each field. The best NN model among those combination models was chosen from the combination of steric and electrostatic fields, with $r^2 = 0.98$, $q^2 = 0.93$, and $p^2 = 0.94$, as shown in Table 3 and (4)-a in Fig. 4. The other two NN models also showed high predictability with the compounds in dataset B, as shown in Table 3 and supplemental Fig. 4. The combination of steric, electrostatic, hydrophobic, and hydrogen bonding fields yielded $r^2 = 0.99$, $q^2 = 0.91$ and $p^2 = 0.94$, while the combination of steric, electrostatic and hydrophobic fields resulted in $r^2 = 0.99$, $q^2 = 0.92$ and $p^2 = 0.93$.

The results from the neural network models in Tables 2 and 3 suggest that the application of NN models could produce a reliable value for prediction of the activity of unknown compounds.

Conclusion

The QSID Tool is a novel ligand-based molecular modeling system for QSAR modeling adopting non-linear and multilinear modeling methods. This system has been introduced successfully, with a high predictability, as a new tool for 3D QSAR. QSID Tool is also a rapid, highly predictive system compared to SYBYL®'s CoMFA and CoMSIA. Neural network analysis as non-linear modeling built into QSID Tool works effectively to examine the datasets that have complex relationships between bioactivity and molecular properties. QSID Tool makes it possible to discover new compounds by screening thousands of unknown compounds in a batch mode against a given predictive model. The sophisticated computational engine built into QSID Tool also accelerates research and discovery by providing reliable computational analysis tools with basic molecular properties and calculations of force field in a user-friendly environment. We are currently updating the application for more statistical functionalities and visualization.

Acknowledgments This work is part of an internal project supported by Rexahn Pharmaceutical, Inc. The authors extend their appreciation to Drs. John Orban, Edith C. Wolff, James Song and Yonil Park for discussion and critical reviews.

References

1. Hansch C, Klein T (1986) Acc Chem Res 19:392
2. Virtual Computational Chemistry Laboratory http://www.vcclab.org. Accessed Feb. 2008
3. Cheminformatics Modeling Laboratory http://eccr.stat.ncsu.edu/ChemModLab/. Accessed Feb. 2008
4. Hicklin J, Moler C, Webb P, Boisvert R, Miller B, Pozo R, Remington K Jama: a Java matrix package. http://math.nist.gov/javanumerics/jama. Accessed Nov. 2007
5. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E (2003) J Chem Inf Comput Sci (JCICS) 43:1077
6. Paolo Marrone Joone: Java object oriented neural engine. http://www.jooneworld.com/. Accessed Jan. 2008
7. The Open Source Chemistry Toolbox http://openbabel.sourceforge.net. Accessed Nov. 2007
8. Gasteiger J, Marsili M (1978) Tetrahedron Lett 34:3181
9. Gasteiger J, Marsili M (1980) Tetrahedron 36:3219
10. JOELib: A Java based cheminformatics library http://www-ra.informatik.uni-tuebingen.de/software/joelib/. Accessed Feb. 2008
11. Masuda T, Jikihara T, Nakamura K, Kimura A, Takagi T, Fujiwara H (2000) J Pharm Sci 86:57
12. Shrake A, Rupley JA (1973) J Mol Biol 79:351
13. Le Grand S, Merz K (1993) J Comp Chem 14:349
14. Lennard-Jones JE (1931) Cohesion. Proc Phys Soc 43:461
15. Wang R, Liu L, Lai L, Tang Y (1998) J Mol Model 4:379
16. Clark M, Cramer RD III (1993) Quant Struct Act Relat 12:137
17. Ian H, Eibe F (2005) Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco
18. Herve Abdi, LEAST Squares (PLS) Regression, The University of Texas at Dallas
19. Hoskuldsson A (1988) J Chemometr 2:211
20. http://en.wikipedia.org/wiki/Neural_network. Accessed Nov. 2007
21. Werbos P (1974) Beyond regression: new tools for prediction and analysis in the behavioural science. PhD dissertation, Committee on Application Mathematics, Harvard University, Cambridge, MA
22. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning internal representations by error propagation, Parallel distributed processing. MIT Press, Cambridge
23. Wikel JH, Dow ER, Heathman M Interpretative neural networks for QSAR. http://www.netsci.org/Science/Compchem/feature02.html. Accessed Nov. 2007
24. Tetko IV, Livingstone DJ, Luik AI (1995) J Chem Inf Comput Sci 35:826
25. Gorman RP, Sejnowski TJ (1988) Neural Nets 1:75
26. Böhm M, Stürzebecher J, Klebe G (1999) J Med Chem 42:458
27. Liverton NJ, Butcher JW, Claiborne CF, Claremon DA, Libby BE, Nguyen KT, Pitzenberger SM, Selnick HG, Smith GR, Tebben A, Vacca JP, Varga SL, Agarwal L, Dancheck K, Forsyth AJ, Fletcher DS, Frantz B, Hanlon WA, Harper CF, Hofsess SJ, Kostura M, Lin J, Luell S, O'Neill EA, O'Keefe SJ (1999) J Med Chem 42:2180
28. Romeiro NC, Albuquerque MG, de Alencastro RB, Ravi M, Hopfinger AJ (2005) J Comput Aided Mol Des 19:385
29. Romeiro NC, Albuquerque MG, de Alencastro RB, Ravi M, Hopfinger AJ (2006) J Mol Model 12:855
30. Molecular Networks http://www.molecular-networks.com/software/corina. Accessed Nov. 2007
31. A language for describing molecular Patterns http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html. Accessed Nov. 2007