J. Chemtracks 23 (1&2)1-12, 2021

Abstracted / Indexed in
A division of the American Chemical Society
Chemical Abstract U.S.A.

ISSN 0973-239X
CODEN:JCOHF7
IMPACT FACTOR : 4.950

QSPR nobel study of water solubility of diverse functional acyclic compounds

Leelawati Kumari *and Jitendra Mahato
*Department of Chemistry, P.K. Roy Memorial College, Dhanbad (Jharkhand)
Department of Chemistry, Rajganj Degree College, Rajganj, Dhanbad (Jharkhand)
Email : jitendramahato1974@gmail.com

Manuscript received online 02 November 2020, accepted on 15 February 2021

Abstract: This work required to assign the training set of compounds to one or more activity classes' Specific descriptor centers also can be defined if desired. The automated pharmacophore identification portion of the program then builds the knowledge base using the following steps: Identify all possible binding interaction centers for each compound in the data set; Generate topological (2D) or topographical (3D) distance matrices based on the set of descriptors; Identify possible pharmacophores from all pairs of molecules using clique selection algorithms;Classify these pharmacophores based upon their occurance in compounds in each activity class using Bayesian statistics and their nonchanceoccurance; Set thresholds for probability and reliability statistics associated with a pharmacophore so that all training set molecules are properly classified by the pharmacophore rules; Align compounds containing high probability pharmacophores on the pharmacophore.Basic macroscopic properties: Molar Volume (MV), Molar Refractivity (MR) and Parachor (Pr);Derived macroscopic properties: density (d), refractive index (n) and surface tension (M^1T^{-2}). Once the knowledge base has been constructed, the scientist can use it to predict biological activity for compounds not included in the training set.

(Keywords : Water Solubility; DiverseFunctional; Acyclic Compounds; QSPR; macroscopic properties)

Introduction

The combinatorial possibilities of severalstrategies for even simple systems can be explosive. As an example, the number of compounds required for synthesis in order to place 10 substituents on the four open positions of an asymmetrically disubstituted benzene ring system is approximately 10,000. The alternative to this labor intensive approach to compound optimization is to develop a theory that quantitatively relates variations in biological activity to changes in molecular descriptors which can easily be obtained for each compound. A Quantitative Structure Activity Relationship (QSAR) can then be utilized to help guide chemical synthesis. This chapter develops the concepts used to derive a QSAR and reviews the application of these techniques to medicinal research.A QSAR attempts to find consistent relationships between the variations in the values of molecular properties and the biological activity for a series of compounds so that these “rules” can be used to evaluate new chemical entities.A QSAR generally takes the form of a linear equation Biological Activity = Const + $(C_{1\boxtimes}P_1)$ + (C_2P_2) + (C_3P_3) + ...

Where the parameters P_1 through P_n are computed for each molecule in the series and the coefficients C_1 through C_n are calculated by fitting variations in the parameters and the biological activity. Since these relationships are generally discovered through the application of statistical techniques, a brief introduction to the principles behind the derivation of a QSAR follows.

The quality of any QSAR will only be as good as the quality of the data which is used to derive the model. Dose-response curves need to be smooth, contain enough points to assure accuracy and should span two or more orders of magnitude. Multiple readings for a given observation should be reproducible and have

relatively smaller errors. The issue being addressed is the signal-to-noise ratio. The variation of the readings obtained by repeatedly testing the same compound should be much smaller than the variation over the series. In cases where the data collected from biological experiments do not follow these guidelines, other methods of data analysis should be utilized since the QSAR models derived from the data will be questionable.

Once biological data has been collected, it is often found that the data is expressed in terms which cannot be used in a QSAR analysis. Since QSAR is based on the relationship of free energy to equilibrium constants, the data for a QSAR study must be expressed in terms of the free energy changes that occur during the biological response. When examining the potency of a drug (the dosage required producing a biological effect), the change in free energy can be calculated to be proportional to the inverse logarithm of the concentration of the compound.

$$G_0 = -2.3RT\log K = \log 1/[S]$$

Further, since biological data are generally found to be skewed, the log transformation moves the data to a nearly normal distribution. Thus, when measuring responses under equilibrium conditions, the most frequent transformation used is to express concentration values (such as IC_{50}, EC_{50}, etc.) as log[C] or log 1/[C].

There are several potential classes of parameters used in QSAR studies. Substituent constants and other physico-chemical parameters (such as Hammett sigma constants) measure the electronic effects of a group on the molecule. Fragment counts are used to enumerate the presence of specific substructures. Other parameters can include topological descriptors and values derived from quantum chemical calculations.

How much confidence should we place in this model? The first step to answering this question is to determine how well the equation predicts activities for known compounds in the series. The equation above estimates the average value for the EC_{50} based on the value for ; because assays vary, it is not surprising that individual values will differ from the regression estimate. The difference between the calculated values and the actual (or measured) values for each compound is termed the residual from the model.

In order to be an improved model, the standard deviation of the residuals calculated from the model should be smaller than the standard deviation of the original data.

There are several assumptions inherent in deriving a QSAR model for a series of compounds. First, it is assumed that parameters can be calculated (or measured in some cases) more accurately and cheaply than activity can be measured. Second, it is assumed that deviations from the best fit line follow a normal (Gaussian) distribution. Finally, it is assumed that any variation in the line described by the QSAR equation is independent of the magnitude of both the activity and the parameters. Given these assumptions, the quality of the model can be gauged using a variety of techniques.

Variation in the data is quantified by the correlation coefficient, r, which measures how closely the observed data tracks the fitted regression line. Errors in either the model or in the data will lead to a bad fit. This indicator of fit to the regression line is calculated as:

$$r^2 = \frac{\text{Sum-of-Squares of the deviations from the regression line}}{\text{Sum-of-Squares of the deviations from the mean}}$$

$$r^2 = \frac{\text{Regression Variance}}{\text{Original Variance}}$$

Where the Regression Variance is defined as the Original Variance minus the Variance around the regression line.

Multiple linear regression attempts to maximize the fit of the data to a regression equation

(minimize the squared deviations from the regression equation) for the biological activity (maximize the r^2 value) by adjusting each of the available parameters up or down. Regression programs often approach this task in a stepwise fashion. That is, successive regression equations will be derived in which parameters will be either added or removed until the r^2 and s values are optimized. The magnitude of the coefficients derived in this manner indicates the relative contribution of the associated parameter to biological activity.

There are two important caveats in applying multiple regression analysis. The first is based on the fact that, given enough parameters any data set can be fitted to a regression line. The consequence of this is that regression analysis generally requires significantly more compounds than parameters; a useful rule of thumb is three to six times the number of parameters under consideration. The difficulty is that regression analysis is most effective for interpolation and it is extrapolation that is most useful in a synthesis campaign (i.e., the region of experimental space described by the regression analysis has been explained, but projecting to a new, unanalyzed region can be problematic).

To judge the importance of a regression term, three items need to be considered.

1. Statistical significance of the regression coefficient.
2. The magnitude of the typical effect $b_i x_i$ (in this case, 0.011 25.36).
3. Any cross-correlation with other terms.

1. Computational Methodology & Approaches to Developing a QSAR

Drugs exert their biological effects by participating in a series of events which include transport, binding with the receptor and metabolism to an inactive species. Since the interaction mechanisms between the molecule and the putative receptor are unknown in most cases (i.e., no bound crystal structures), one is reduced to making inferences from properties which can easily be obtained (molecular properties and descriptors) to explain these interactions for known molecules. Once the relationship is defined, it can be used to aid in the prediction of new or unknown molecules.

The first approach to developing quantitative relationships which described activity as a function of chemical structure relied on the principles of thermodynamics. The free-energy terms E, H and S were represented by a series of parameters which could be derived for a given molecule.

Electronic effects such as electron donating and withdrawing tendencies, partial atomic charges and electrostatic field densities were defined by Hammett sigma values, resonance parameters (R values), inductive parameters (F values) and Taft substituent values (*, *, E_s). Steric effects such as molecular volume and surface area were represented by values calculated for Molar Refractivity and the Taft steric parameter. Enthalpic effects were calculated using partition coefficients (LogP) or the hydrophobic parameter, which was derived from the partition coefficient. In addition, assortments of structural indices were used to describe the presence of specific functional groups at positions within the molecule. The linear equation which described the relationship between activity and this parameter set was the Hansch equation

$$\log 1/[C] = A(\log P) - B(\log P)^2 + C(E_s) + D() + E + ...$$

Multiple linear regression analysis was used to derive the values of the coefficients. In general, Hansch type studies were performed on compounds which contained a common template (usually a rigid one such as an aromatic ring) with structural variation limited to functional group changes at specific sites.

In 1972, John Topliss published a paper which detailed methodology to automate the Hansch approach[2]. The method assumed that the lead compound of interest contained at least one phenyl ring which could serve as the template for functional group modifications. The first modification to the template was preparation of the para-chloro derivative to examine lipophilicity. Additional substitution patterns were then made sequentially in an attempt to explore and optimize the relationship between activity and the hydrophobic and electronic character of the molecule. While the Topliss approach is easy to follow, it has several drawbacks. The primary problems are that the procedure is not applicable to all types of studies and that there is a high degree of risk associated with its use (it essentially ignores the possibility of interactions between substituents as it changes one substituent at a time).

The use of classical QSAR was expanded during the 1960's as a means of correlating observed activity to chemical properties. However, there are many areas where these techniques could not be used or where they failed to provide useful correlations. These included situations in which activity was found to be determined by 3-dimensional geometry, where poor training sets of compounds were used or the set of compounds were too small or insufficiently diverse and cases where biological activity could not be well quantified. Many of these problems were addressed by extensions to the Hansch method and the development of alternative approaches to QSAR.

There are cases where biological activity values cannot be determined accurately for a variety of reasons, e.g. lack of sensitivity of a particular test system. Alternative statistical techniques can be used in these cases; the problem is simplified to a classification scheme in which compounds are labeled as active, partially active, inactive, etc. The resulting data set is then searched for patterns which predict these categories. The methods which have been used for this type of analysis include SIMCA (Soft, Independent Modeling of Class Analogy)[13], ADAPT (Automated Data Analysis by Pattern recognition Techniques)[14], CASE (Computer Automated Structure Evaluation)[15] and CSA (Cluster Significance Analysis)[16].

Pattern recognition methods[17] attempt to define the set of parameter values which will result in clustering compounds of similar activity into regions of n-dimensional space. The methods used to accomplish this goal can be parametric or nonparametric. Parametric methods search the n-dimensional space for clusters of compounds based on their calculated properties. These methods do not use derived values (e.g., mean vectors and covariance matrices), but instead use the original data to find clustering definitions and apply iterative procedures to find the linear set of parameters which best define the classification scheme.

Where the methods described above develop discriminant functions, SIMCA methods use Principal Component Analysis (PCA) to describe the data set. The objective of PCA is to create a reduced number variables which describe biological activity or chemical properties into a relatively few independent ones. This is accomplished through an analysis of the correlation matrix of biological or chemical properties.

Principle component analysis can be used to create derived variables for each class (e.g., active and inactive) separately by decomposing the correlation matrix; this method is useful to point out redundancies or interrelationships among the variables. PCA seeks to find simplified relationships in data by transforming the original parameters into a new set of uncorrelated variables which are termed principal components. The symmetric correlation matrix is decomposed by eigenvalue decomposition. The largest eigenvalue and its

eigenvector are used to form a linear combination of the original variables with maximum variance. Successively smaller eigenvalues and vectors produce linear combinations of the original variables with diminishing variance. Successive eigenvectors are independent of one another. The simplification is derived by disregarding eigenvectors associated with small eigenvalues. In summary, the procedure finds the set of orthogonal axes for the data which decompose variance in the data.

Another approach to examining the effects of chemical structure on activity was developed by the Jurs' group. Rather than rely on multivariant statistics to highlight these relationships, Jurs used the combination of cluster analysis and pattern recognition techniques as a tool to develop these correlations. The ADAPT program generated a data set of molecular descriptors (topological, geometrical and physicochemical) derived from three dimensional model building, projected these data points onto an n-dimensional surface and analyzed them using pattern recognition methods. The goal of this analysis was to discriminate between active and inactive compounds in a series.

Jurs has reported several applications of the methodology contained in ADAPT. In one study of chemical carcinogens[18], a linear discriminant function was derived from a set of 28 calculated structure features including fragment descriptors, substructure descriptors, environment descriptors, molecular connectivity descriptors and geometric descriptors. Two hundred and nine compounds from twelve structural classes (130 carcinogens, 79 noncarcinogens) were selected for this study. The program was used to identify a training set of 192 compounds which was used to find the best set of descriptors and analyze the entire data set. A predictive success of 90% for carcinogenic compounds and 78% for noncarcinogenic compounds was obtained in randomized testing.

The CASE program extended the techniques in ADAPT by using topological methods to define substructural fragments which were essential for activity. CASE was able to differentiate between positional isomers. Both CASE and ADAPT are limited to analyzing structurally similar data sets.

The analysis methods described to this point have not explicitly incorporated the contribution of three dimensional shapes in the analysis of the activity of a molecule. While the use of chemical graph indexes[18], intermolecular binding distances[19], molecular surface areas[20] and electrostatic potentials[21] contain some information about the 3-D shape of molecules, the Hopfinger[22] and Marshall[23] groups were the first to exhaustively analyze these effects.

In 1979, Marshall extended the 2-D approach to QSAR by explicitly considering the conformational flexibility of a series as reflected by their 3-D shape[23]. The first step of the Active Analog Approach was to exhaustively search the conformations of a compound which was highly active in a particular biological assay. The result of the search was a map of interatomic distances which was used to filter the conformational searches of subsequent molecules in the series. The implicit assumption of the method was that all compounds which display similar activity profiles were able to adopt similar conformations. Once the "active conformation" was determined, molecular volumes for each molecule were calculated and superimposed. Regression analysis of the volumes was used to establish a relationship to biological activity. Marshall and co-workers commercialized the Active Analog Approach and a suite of other drug design techniques in the SYBYL molecular modeling program.

Hopfinger and co-workers also used 3-D shape in QSAR. In molecular shape analysis[24] of the Baker Triazines, the common space shared by all molecules of a series and the differences in

their potential energy fields were computed. When these calculations were combined with a set of rules for overlapping the series, comparative indicies of the shape of different molecules were obtained. Inclusion of these shape descriptors in standard Hansch analysis schemes lead to improved descriptions relating computed parameters to biological activity such that no compounds in the original data set had to be eliminated from the calculations. The techniques developed by Hopfinger and co-workers were made available in the CAMSEQ, CAMSEQ-II, CHEMLAB and CAMSEQ-M computer programs.

In 1988, Richard Cramer proposed that biological activity could be analyzed by relating the shape-dependent steric and electrostatic fields for molecules to their biological activity[25]. Additionally, rather than limiting the analysis to fitting data to a regression line, CoMFA (Comparative Molecular Field Analysis) utilized new methods of data analysis, PLS (Partial Least Squares) and cross-validation, to develop models for activity predictions.

The approach used in the CoMFA procedure requires that the scientist define alignment rules for the series which overlap the putative pharmacophore for each molecule; the active conformation and alignment rule must be specified. Once aligned, each molecule is fixed into a three-dimensional grid by the program and the electrostatic and steric components of the molecular mechanics force field, arising from interaction with a probe atom (e.g., an SP^3 C atom), are calculated at intersecting lattice points within the 3-D grid. The equations which result from this exercise have the form

$$Act_1 = Const_1 + a_1(steric_{xyz}) + b_1(steric_{xyz}) + \ldots + a'_1(estatic_{xyz}) + b'_1(estatic_{xyz}) + \ldots$$
$$Act_2 = Const_2 + a_2(steric_{xyz}) + b_2(steric_{xyz}) + \ldots + a'_2(estatic_{xyz}) + b'_2(estatic_{xyz}) + \ldots$$
$$Act_n = Const_n + a_n(steric_{xyz}) + b_n(steric_{xyz}) + \ldots + a'_n(estatic_{xyz}) + b'_n(estatic_{xyz}) + \ldots$$

Traditional regression methods require that the number of parameters must be considerably smaller than the number of compounds in the data set (or the number of degrees of freedom in the data). The data tables which result from CoMFA analysis have far more parameters than compounds. PLS, which removes this limitation, is used to derive the coefficients for all of the steric and electrostatic terms. PLS essentially relies upon the fact that the correlations among near parts of a molecule are similar so that the real dimensionality is smaller than the number of grid points. Since these coefficients are position dependant, substituent patterns for the series are elucidated which define regions of steric bulk and electrostatic charge associated with increased or decreased activity. The size of the model (the number of components[27] needed for the best model) and the validity of the model as a predictive tool are assessed using cross-validation.

As opposed to traditional regression methods, cross-validation evaluates the validity of a model by how well it predicts data rather than how well it fits data. The analysis uses a "leave-one-out" scheme; a model is built with N-1 compounds and the N^{th} compound is predicted. Each compound is left out of the model derivation and predicted in turn. An indication of the performance of the model is obtained from the cross-validated (or predictive) r^2 which is defined as

$$r^2 \text{ (cross-validated)} = (SD - Press)/SD$$

SD is the Sum-of-Squares deviation for each activity from the mean. Press (or Predictive Sum of Squares) is the sum of the squared differences between the actual and that predicted when the compound is omitted from the fitting process.

As we have discussed, values for conventional r^2 range from 0 to 1. Values for the cross-validated r^2 are reported by the method to range from -1 to 1. Negative values indicate that biological activity values are estimated by the mean of the activity values better than they are

by the model (i.e, the predictions derived from the model are worse than no model). Once a model is developed which has the highest cross-validated r^2, this model is used to derive the conventional QSAR equation and conventional r^2 and s values. The results of the final model are then visualized as contour maps of the coefficients.

Apex-3D is an automated pharmacophore identification system which can identify possible pharmacophores from a set of biologically active molecules using statistical techniques and 3-D pattern matching algorithms. The program classifies molecular structures using three methods: the agreement inductive method identifies common structural patterns in compounds having similar activity; the difference inductive method identifies structural patterns which differentiate active and inactive compounds and the concomitant variations inductive method highlights variations in structural features that explain changes in biological activity for sets of compounds.

The methods defined above follow logic similar to that used by a practicing medicinal chemist: What pharmacophoric patterns are present in the active molecule which is not present in the inactive ones? Pharmacophores are defined by different chemical centers (atom centered functional groups) and the distances between these centers. These descriptor centers can include such things as aromatic ring centers, electron donor ability, and hydrogen bonding sites, lipophilic regions, and partial atomic charge. The information for each molecule is stored in a knowledge base in the form of rules which can be used to predict the activity of novel structures.

Apex-3D contains an expert system which automatically selects the best conformation and alignment for structures based on identified pharmacophores. When quantitative biological data is available, a 3D-QSAR model can be developed for any of the possible identified pharmacophores. Depending on the type of biological activity available, it is possible to identify pharmacophores for different binding orientations, receptor subtypes, or agonist versus antagonist activity.

We have used ChemSketch program of ACD lab[28] and Hyperchem[29]Software for the calculation of various parameters.

At the heart of the additive-constitutive calculation algorithm of all physicochemical properties in ChemSketch lies the presumption that these properties can be estimated using additive atomic or group increments. Apart from molecular weight (MW), which is trivial to calculate, the algorithms may be divided into three general groups:

Basic macroscopic properties: Molar Volume (MV), Molar Refractivity (MR) and Parachor (P_r);

Derived macroscopic properties: density (d), refractive index (n) and surface tension (g).

Basic macroscopic properties such as Molar Volume (MV), Molar Refractivity (MR) and the Parachor (Pr) are calculated first for the input structure. The atomic additive increments in such an algorithm depend on the bonds (single, double, aromatic, etc.) of this atom and on neighboring atoms. ChemSketch rapidly analyzes the input structure to determine the class of each atom, i.e., whether it is cyclic, aromatic, aliphatic, etc.

The prediction algorithms for density (d), refractive index (n) and surface tension (g) are founded on well-known physicochemical formula which can be found in most physical chemistry textbooks. These express d, n and as functions of MV, MR or Pr. Once the MV, MR or P_r, have been predicted by additive means, it is straightforward to predict d, n and g using these formula.

Table-1:
Compounds with their ln S values

Comp. No.	Compound Name	lnS	Comp. No.	Compound Name	lnS
1.	n-Butanol	0.006	32	2,4-Dimethyl-3-pentanol	-2.801
2.	2-Methylpropanol	0.023	33	2,2-Dimethyl-3-pentanol	-2.643
3.	2-Butanol	0.066	34	3-Heptanol	-3.194
4.	n-Pentanol	-1.347	35	4-Heptanol	-3.196
5.	3-Methylbutanol	-1.167	36	n-Octanol	-5.401
6.	2-Methylbutanol	-1.058	37	2,2,3-Trimethyl-3-	
7.	2-Pentanol	-0.635		pentanol	-2.931
8.	3-Pentanol	-0.486	38	2-Octanol	-4.755
9	3-Methyl-2-butanol	-0.405	39	2-Ethylhexanol	-4.996
10	2-Methyl-2-butanol	0.339	40	n-Nonanol	-6.907
11	n-Hexanol	-2.790	41	2-Nonanol	-6.319
12	2-Hexanol	-1.995	42	3-Nonanol	-6.119
13	3-Hexanol	-1.832	43	4-Nonanol	-5.952
14	3-Methyl-3-pentanol	-0.830	44	5-Nonanol	-5.744
15	2-Methyl-2-pentanol	-1.117	45	2,6-Dimethyl-3-heptanol	-5.776
16	2-Methyl-3-pentanol	-1.609	46	3,5-Dimethyl-4-heptanol	-5.298
17	3-Methyl-2-pentanol	-1.639	47	1,1-Diethylpentanol	-5.572
18	2,3-Dimethyl-2-butanol	-0.851	48	7-Methyloctanol	-5.744
19	3,3-Dimethylbutanol	-2.590	49	3,5,5-Trimethylhexanol	-5.769
20	3,3-Dimethyl-2-butanol	-1.410	50	n-Decanol	-8.517
21	4-Methylpentanol	-2.282	51	n-Tetradecanol	-12.772
22	4-Methyl-2-pentanol	-1.814	52	n-Pentadecanol	-13.796
23	2-Ethylbutanol	-2.787	53	n-Hexadecanol	-14.603
24	Cyclohexanol	-0.960	54	2,2-Dimethyl propanol	-0.889
25	n-Heptanol	-4.166	55	4-Penten-1-ol	-0.355
26	2-Methyl-2-hexanol	-2.473	56	3-Penten-2-ol	0.127
27	3-Methyl-3-hexanol	-2.263	57	1-Penten-3-ol	0.035
28	3-Ethyl-3-pentanol	-1.917	58	1-Hexen-3-ol	-1.354
29	2,3-Dimethyl-2-pentanol	-2.002	59	2-Hexen-4-ol	-0.939
30	2,3-Dimethyl-3-pentanol	-1.937	60	2-Methyl-4-penten-3-ol	-1.156
31	2,4-Dimethyl-2-pentanol	-2.145			

The determination of the additive-constitutive atomic increments for MV, MR and P_r were obtained internally by ACD using large experimental data bases relating structure to density, refractive index and surface tension. The MV, MR and Pr were recalculated from d, n andg. These parameters are proprietary information of Advanced Chemistry Development.

II-1 Molar Volume, MV

By definition,

$$MV = \frac{MW}{d}$$

ChemSketch calculates molar volume from additive increments. The additive atomic increments were obtained using a Data Base of density and calculated MW.

Table-2:
Observed and estimated ln S values.

Comp. No.	Observed	Calculated	Residual	Comp. No.	Observed	Calculated	Residual
1.	0.006	-0.088	0.094	31	-2.145	-5.383	3.238
2.	0.023	-0.795	0.818	32	-2.801	-5.575	2.774
3.	0.066	-0.795	0.861	33	-2.643	-5.383	2.74
4.	-1.347	-1.238	-0.109	34	-3.194	-4.233	1.039
5.	-1.167	-1.933	0.766	35	-3.196	-4.233	1.037
6.	-1.058	-1.933	0.875	36	-5.401	-4.677	-0.724
7.	-0.635	-1.933	1.298	37	-2.931	-6.985	4.054
8.	-0.486	-1.933	1.447	38	-4.755	-5.348	0.593
9	-0.405	-2.64	2.235	39	-4.996	-5.348	0.352
10	0.339	-2.424	2.763	40	-6.907	-5.854	-1.053
11	-2.79	-2.413	-0.377	41	-6.319	-6.498	0.179
12	-1.995	-3.083	1.088	42	-6.119	-6.498	0.379
13	-1.832	-3.083	1.251	43	-5.952	-6.498	0.546
14	-0.83	-3.563	2.733	44	-5.744	-6.498	0.754
15	-1.117	-3.563	2.446	45	-5.776	-7.899	2.123
16	-1.609	-3.754	2.145	46	-5.298	-7.899	2.601
17	-1.639	-3.754	2.115	47	-5.572	-6.977	1.405
18	-0.851	-4.269	3.418	48	-5.744	-6.498	0.754
19	-2.59	-3.563	0.973	49	-5.769	-7.683	1.914
20	-1.41	-4.269	2.859	50	-8.517	-6.968	-1.549
21	-2.282	-3.083	0.801	51	-12.772	-11.557	-1.215
22	-1.814	-3.754	1.94	52	-13.796	-12.671	-1.125
23	-2.787	-3.083	0.296	53	-14.603	-13.821	-0.782
24	-0.96	-3.435	2.475	54	-0.889	-2.424	1.535
25	-4.166	-3.527	-0.639	55	-0.355	-2.226	1.871
26	-2.473	-4.713	2.24	56	0.127	-3.429	3.556
27	-2.263	-4.713	2.45	57	0.035	-2.921	2.956
28	-1.917	-4.713	2.796	58	-1.354	-4.071	2.717
29	-2.002	-5.383	3.381	59	-0.939	-4.567	3.628
30	-1.937	-5.383	3.446	60	-1.156	-4.742	3.586

II-2 Molar Refractivity, MR

The Lorentz-Lorenz equation relates refractive index, density, and refractive index:

$$MR = \frac{n^2 - 1}{n^2 + 2} \cdot \frac{MW}{d}$$

ChemSketch calculates molar refractivity from additive increments. The additive atomic increments were obtained using a Data Base of density, refractive index and calculated MW.

II-3 Parachor, P_r By definition,

$$P_r = \left(\frac{MW}{d} \right) \gamma^{1/4}$$

ChemSketch calculates the parachor from additive increments. The additive atomic

increments were obtained using a Data Base of density, surface tension, and calculated MW.

II-4 Density, d

By definition,

$$d = \frac{MW}{MV}$$

ChemSketch calculates the density from MW and the calculated molar volume (see above).

II-5 Refractive Index, n

By the Lorentz-Lorenz equation,

$$n = \sqrt{\frac{2.MR + MV}{MV - MR}}$$

ChemSketch calculates the refractive index from the molar volume and molar refractivity, both of which are calculated as above.

II-6 Surface Tension, g

By definition,

$$\gamma = \left(\frac{P_r}{MV} \right)^4$$

ChemSketch calculates the surface tension from calculated MV (see above) and calculated P_r (see above).

II-7 Polarizability, a

This property is calculated from the Molar Refractivity (MR) as follows:

Polarizability = 0.3964308. MR

2. Results and Discussion

Modeling of ln S of 60 Alcohols:

First attempt has been made to model the solubuility of 60 alcohols which have been reported in Table 1 and Table 2. The physico chemical parameters found effective in modeling lnS are PC, MV,, MR and ST. On the basis of R and Q values it can be inferred that the four-parametric model is the best model for modeling lnS of entire data set. The only parameters which shows its positive role is MW and other three parameters POL, MV and PC have negative roles towards exhibition of ln S of 193 compounds.

The mdodels have been tested by plotting graph between the observed and estimated ln S of compounds using the model. Such graph is demonstrated in Figure 1. The R^2 values in the graph depict the variance or the predictive power of the respective models.

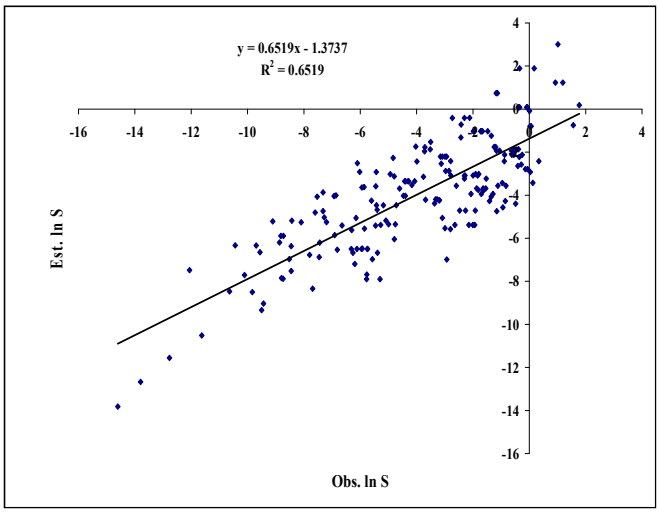

Figure 1: Correlation betweenobserved and estimated ln S.

Conclusion

In conclusion, first attempt has been made to model the solubuility of 60 alcohols which have been reported. The physico chemical parameters found effective in modeling lnS are PC, MV,, MR and ST. Identify all possible binding interaction centers for each compound in the data set; Generate topological (2D) or topographical (3D) distance matrices based on the set of descriptors; Identify possible pharmacophores from all pairs of molecules using clique selection algorithms; Classify these pharmacophores based upon their occurrences in compounds in each activity class using Bayesian statistics and their nonchance occurrence; Set thresholds for probability and reliability statistics associated with a pharmacophore so that all training set molecules are properly classified by the pharmacophore rules; Align compounds containing high probability pharmacophores on the pharmacophore. Once the knowledge base has been constructed, the scientist can use it to predict biological activity for compounds not included in the training set.

References

1. S.J. Christopher Walpole, Roger Wrigglerworth, Stuart Bevan, Elizabeth A. Campbell, Andy Dray, Iain F. James, Kay J. Masdin, Martin N. Perkins and Janet Winter, *J. Med. Chem.*, **36**, 2381 (1993).
2. G. John Topliss, Utilization of Operational Schemes for Analog Synthesis in Drug Design, *J. Med. Chem.*, **15**, 1006 (1972).
3. C. Yvonne Martin, A Practitioner's Perspective of the Role of Quantitative Structure Activity Analysis in Medicinal Chemistry, *J. Med. Chem.*, **24**, 229 (1981).
4. G. John Topliss, Quantitative Structure-Activity Relationships of Drugs, Academic Press, New York, (1983).
5. Franke, R., Theoretical Drug Design Methods, Elsevier, Amsterdam, (1984).
6. Seydel, J.K., QSAR and Strategies in the Design of Bioactive Compounds, VCH, Weinheim, (1985).
7. Yvonne C. Martin,*Accounts of Chem. Res.*, **19**, 392 (1986).
8. Corwin Hansch,*Accounts of Chem. Res.*, **2**, 232 (1969).
9. Corwin Hansch and Carlo Silipo, *J. Amer. Chem. Soc.*, **97**, 6849 (1975).
10. Robert F. Gould (ed.), Biological Correlations — The Hansch Approach, Advances in Chemistry Series, **No. 114**, American Chemical Society, Washington, D.C., 1972.
11. Yvonne C. Martin, Quantitative Drug Design, Marcel Dekker, New York, (1978).
12. Spencer M. Free and James W. Wilson, A Mathematical Contribution to Structure-Activity Studies, *J. Med. Chem.*, **7**, 395 (1964).
13. SvanteWold, Pattern Recognition by Means of Disjoint Principal Components Models, *Pattern Recognition*, **8**, 127 (1976).
14. A.J. Stuper, W.E. Brugger and P.C. Jurs, Chemometrics: Theory and Application, B.R. Kowalski (Ed.), American Chemical Society, Washington, D.C., 1977.
15. G. Klopman and M.I. Dimayuga, *J. Comput.-Aided Mol. Design*, **4**, 117 (1990).
16. J. McFarland, *J. Med. Chem.*,
17. Peter C. Jurs, Chemometrics and Multivariate Analysis in Analytical Chemistry, in Reviews in Computational Chemistry, Volume 1, K.B. Lipkowitz and B.B. Boyd (Eds.), VCH Publishers, Inc., New York, 1990.
18. Peter C. Jurs, J. T. Chou and M. Yuan, *J. Med. Chem.*, **22**, 476 (1979).
19. Lemont B. Kier, Molecular Orbital Theory in Drug Research, Academic Press, New York, NY, 1971.
20. Gordon M. Crippen, *J. Med. Chem.*, **22**, 988 (1979).
21. M. Mabilia, R.A. Pearlstein and A.J. Hopfinger, Computer Graphics in Molecular Shape Analysis, in Molecular Graphics and Drug Design, A.S.V. Burgen, G.C.K. Roberts and M.S. Tute (Eds.), Elsevier Science Publishers, Amsterdam, 1986.

22. H. Weinstein, in Chemical Applications of Molecular Electrostatic Potentials, Peter Politzer and Donald G. Truhlar (Eds.), Plenum Press, New York, NY, 1981.
23. R. Potenzone, E. Cavicchi, H.J.R. Weintraub and A.J. Hopfinger, *Comput. Chem.*, **1**, 187 (1977).
24. G.R. Marshall, C.D. Barry, H.E. Bosshard, R.A..Dammkoehler and D.A. Dunn, The Conformational Parameter in Drug Design: The Active Analog Approach, in Computer Assisted Drug Design, ACS Symposia, **112**, E.C. Olson and R.E. Christofferson (Eds.), American Chemical Society, Washington D.C., 1979
25. Anton J. Hopfinger, *J. Amer. Chem. Soc.*, **102**, 7196 (1980).
26. Richard D. Cramer III, David E. Patterson and Jeffrey D. Bunce, *J. Amer. Chem. Soc.*, **110**, 5959 (1988).
27. Components are defined as the linear combination of all parameters (the independent data) which are used to fit the shape of the parameters to the shape of the activity data (the dependent data).
28. ACD-lab Software for calculating the molecular modeling parameters, Chem Sketch 10.0, www.acdlabs.com
29. Hyperchem-7 Software for calculating the molecular modeling parameters, www.hyper.com