COMPUTER AIDED DRUG DESIGN METHODS & QUANTITATIVE STRUCTURE- ACTIVITY/PROPERTY RELATIONSHIPS

SUMAN SIRIMULLA

Department of Chemistry

APPROVED BY:

William C. Herndon, Ph.D., Chair

James Salvador, Ph.D.

Mahesh Narayan, Ph.D.

Felicia Manciu, Ph.D.

Patricia D. Witherspoon, Ph.D.
Dean of the Graduate School

DEDICATION

To my Family

Bhaskar sirimulla , Nirmala sirimulla, Phani Sirimulla , and Manga Sirimulla

and my love

Ammu

COMPUTER AIDED DRUG DESIGN METHODS & QUANTITATIVE STRUCTURE- ACTIVITY/PROPERTY RELATIONSHIPS

by

SUMAN SIRIMULLA, M.S

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

Department of Chemistry

THE UNIVERSITY OF TEXAS AT EL PASO

May 2010

ACKNOWLDEGEMENTS

With a deep sense of gratitude, I thank my family members and few people for their continuous support and encouragement throughout my career, and without their support this thesis would not have been a possibility. Firstly I would like to express my sincere thanks to Dr. William C. Herndon for being my research mentor, choosing me as his student, guiding in my research work, and his suggestions in the preparation of this thesis. I am also grateful to him for sending me for two major conferences 3rd Pharmaceutical Scientists world congress, Amsterdam 21-25, 2007and 62nd Regional American chemical society meeting, Houston, October 19-22, 2006. I am proud to be one of his students at UTEP. Finally, I would like to sincerely thank him for all the support that he offered me during the course of this program.

I would like to express my deepest gratitude to my parents, my girl friend mrudula raparla, my brother Phani Sirimulla, and my sister Manga Sirimulla who inculcated the art of learning in any domain with pure, unselfish and honest passion - this fundamental quality made me grow and appreciate the world in the way I see.

I would also like to thank my friend Mr. Bharaneeshwar Renukuntla for guiding me to get an admission in to UTEP and giving me suggestions in every step of my life. I wish to thank Dr. Mahesh Narayan for his helpful suggestions concerning my research, my career and also for being my research committee member. I would like to thank Dr Marion L. Ellezy and Dr Gardea Torresedy for supporting me financially with a Teaching Assistantship in the Department of Chemistry during this course.

I also want to thanks Ms. Armenta Lucema for her prompt and timely help in executing official work. I also would like to thank, Mr. Marco Olguin and other colleagues for creating a friendly

ambience in my laboratory. I also want to thanks the Department of chemistry faculty at The University of Texas at El Paso for guiding me through out my master's program and giving me the right knowledge and experience to excel in my areas of interest.

I would also like to extend my gratitude to my thesis committee members, Dr. James Salvador-Gyan, and Dr. Felicia Manciu for the dedication of their time and participation in the thesis.

Finally, I thank the *Almighty* for all his blessings in the successful completion of my thesis work.

ABSTRACT

The first part of dissertation consists of development of a QSAR model for 229 mutagenic aromatic amines and a QSPR model of partial molar volumes of amino acids. A common procedure for QSAR analysis consist of data selection (generally sets of homologous series of compounds and their corresponding biological activities), tabulation of trial physicochemical or molecular structural descriptors, followed by a multilinear statistical analysis to derive a statistically valid QSAR correlation of the activity data making use of a subset of the trial descriptors. A final important step is cross-validation to assess the putative predictive (rather than just correlative) capabilities of the derived QSAR model equationThe results of a very successful elementary QSA(/P)R studies using substituent indicator variables, coupled with calculated theoretical parameters for the compounds in the work outlined above are presented.

The second part of the dissertation illustrates that betalactoglobulin and human serum albumin can be used as a vehicle to improve the bioavailability of curcumin and it's derivatives. Curcumin a major component of Indian spice turmeric (Curcuma longa), possesses diverse anti-inflammatory, anti-tumour and antioxidant properties. Several studies have confirmed that curcumin can reduce the oxidative/nitrosative stress and there by decrease the neuronal attrition. But the bioavailability of curcumin is poor and has raised several concerns regarding limited clinical impact. The aim of this study was to find molecules similar to curcumin which can assist in decreasing nitrosative stress and possess enhanced bioavailability. Here, we examined the use of beta-lactoglobulin as a vehicle to transport molecules to the gut. Curcumin analogs were searched from Zinc database and 6457 compounds were selected for the study. These compounds were docked to betalactoglobulin using Glide to find the best fit ligands. Our findings indicated four compounds that have better binding to betalactoglobulin and efficient NOx (free radical) scavenging activity.

TABLE OF CONTENTS

QSAR OF BIOLOGICALLY ACTIVE COMPOUNDS

Introduction to Quantitative Structure-Activity Relationships

INTRODUCTION

Quantitative Structure-Activity Relationships shortly abbreviated as QSAR is defined as process by which chemical structures are quantitatively correlated with their biological/toxicological activity. If the chemical structures are correlated to their physical properties then the process is called Quantitative Structure-Property Relationships (QSPR).

A common procedure for QSAR analysis consist of data selection (generally sets of homologous series of compounds and their corresponding biological activities), tabulation of trial physicochemical or molecular structural descriptors, followed by a multilinear statistical analysis to derive a statistically valid QSAR correlation of the activity data making use of a subset of the trial descriptors. A final important step is cross-validation to assess the putative predictive (rather than just correlative) capabilities of the derived QSAR model equation. 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

$$\text{Biological Activity} = \text{Const} + (A1 \bullet X1) + (A2 \bullet X2) + (A3 \bullet X3) + \ldots$$

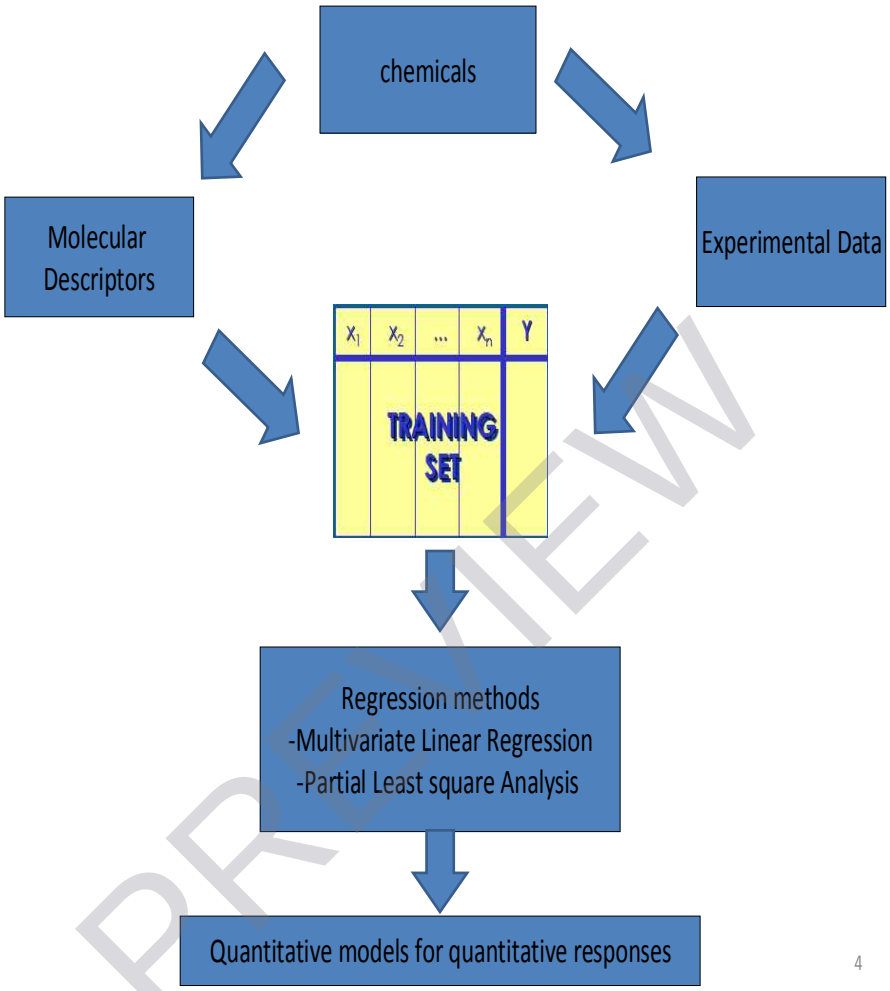
where the parameters X1 through Xn are computed for each molecule in the series and the coefficients A1 through An are calculated by fitting variations in the parameters and the biological activity. Since these relationships are generally discovered through the application of statistical techniques.

There are simple rules to this approach

- Choose well-defined activity endpoints.
- Choose plausible molecular descriptors.
- Explore the data with statistical methods.
- Test hypotheses with new data.

In a hierarchical approach molecular structural descriptors are classified into different levels of complexity. The hierarchical QSAR analysis starts by correlating biological activities with simple low level descriptors like enumeration of different types of atoms, then advances to the next level of parameters like the number of functional groups and ring types, through larger structural fragments, and ends with the highest level descriptors such as experimental properties or the results of semi empirical or *ab initio* molecular orbital calculations (usually Hansch analysis uses this kind of descriptors). The levels of hierarchical structural descriptors are augmented and tested sequentially to obtain information regarding the lowest levels of descriptors that are crucial for a statistically significant rectification of a particular dependent variable property. Rather than using single descriptor level, a combination of significant descriptors from the different descriptor levels can be also be used to obtain a high quality QSAR. The quality of a QSAR may increase assuming that the essential descriptors from different levels incorporate high amounts of variance into the dependent variable. The studies detailed in this chapter will lead to predict the activity of other similar compounds.

Flowchart to describe the steps involved in QSAR



4

METHODS

1 Free-wilson method

In 1967 Free and Wilson have proposed a model. In a homologous series of drug molecules, each molecule can be divided into various segments and the activity of the whole molecule is determined by the summation of activity of each segment. Therefore, when a new segment with a known activity is added to an existing molecule the activity of the newer analogue formed could be estimated by using this method. This procedure involves the segmentation of a molecule into two parts namely, a constant segment (basic nucleus) and a variable segment (the various substituents on the basic nucleus).

One can assign a uniform code for each variable segment by numbering all substituents positions and all the substituents. Usually a two digit code is used which is denoted by jk where 'j' refers to the substitution position and 'k' refers to the substituent.

The structure of any compound of the homologous series can be described in simpler terms by means of a vector of structural parameters, $b_{i\,jk}$ which have the form of a logical quantity and are assigned either a value of 0 or 1. The structural parameter $b_{i\,jk}$ indicates whether in the i^{th} compound of the series the variable segment, jk, i.e. the k^{th} substituent at the j^{th} position, is present ($b_{i\,jk} = 1$) or absent ($b_{i\,jk} = 0$).

When the contributions of the parent structure and of the variable segments activity are designated, respectively, by μ and z_{jk}, it follows from the basic percept of the FREE- WILSON model that biological activity of a given compound in a series can be expressed as the sum obtained from μ and the sum of the z_{jk} for the variable segments occurring in that compound. The latter sum is equivalent to the scalar product of the vector of the structural parameters with the vector of the z_{jk} so that

$$\text{Log } A_i = \mu + b_{i\,jk} . z_{jk}$$

Where A_i = the biological activity of the i^{th} compound.

When the molecule of the series have p substitution sites (j =1,….,p) and m_j substituents may be present at the j^{th} substitution site (k = 1,…, m_j) , it then follows that ;

$$\text{Log } A_i = \mu + \Sigma_{jk}\, b_{i\,jk}\, z_{jk}$$

2 Hansch method

Hansch QSAR uses Hammett's relationship where electronic properties are used as the descriptors of structures. But, when investigators tried to apply Hammett-type relationships to biological systems, many difficulties were encountered which indicated that other structural descriptors were necessary.

When studying the biological activity of plant growth regulators, Indoleacetic acid and phenoxyacetic acid analogues, Robert Muir, a botanist at Pomona College, attempted to correlate the structures of these compounds with their activities, he consulted his colleague in chemistry, Corwin Hansch.

Hansch Used Hammett sigma parameters to account for the electronic effect of substituents but it did not lead to meaningful QSAR. However, Hansch identified the significance of the lipophilicity (expressed as the octanol -water partition coefficient) on biological activity [14]. From then, we recognize lipophilicity as a parameter to provide a measure of the bioavailability of various compounds, which partially determines the amount of the compound that reaches the target site.

From then, various relationships were developed to correlate a structural parameter (i.e., lipophilicity)

with activity. A relationship with one variable correlating structure and activity was sufficient in some cases. The form of the equation is:

$$\log\left(\frac{1}{C}\right) = a \log P + b$$

where C is the molar concentration of compound that produces a standard response (e.g., LD_{50}, ED_{50}). With their data, it was observed that correlations could further be improved by combining Hammett's electronic parameters and Hansch's measure of lipophilicity using an equation such as

$$\log\left(\frac{1}{C}\right) = k_1\pi + k_2\sigma + k_3$$

where σ is the Hammett substituent parameter and pi is defined analogously to σ. That is,

$$\pi = \log\left(\frac{P_x}{P_H}\right)$$

3 Hierarchical approach

In a hierarchical approach molecular structural descriptors are classified into different levels of complexity. The hierarchical QSAR analysis starts by correlating biological activities with simple low level descriptors like enumeration of different types of atoms, then advances to the next level of parameters like the number of functional groups and ring types, through larger structural fragments, and ends with the highest level descriptors such as experimental properties or the results of semi empirical or *ab initio* molecular orbital calculations (usually Hansch analysis uses this kind of descriptors). The levels of hierarchical structural descriptors are augmented and tested sequentially to obtain information regarding the lowest levels of descriptors that are crucial for a statistically significant rectification of a particular dependent variable property. Rather than using single descriptor level, a combination of significant descriptors from the different descriptor levels can be also be used to obtain a high quality

QSAR. The quality of a QSAR may increase assuming that the essential descriptors from different levels incorporate high amounts of variance into the dependent variable [42].

This table comprises of the list of different level descriptors with examples that are used in Hierarchical approach of Quantitative Structural Activity Relationship analysis

Level of descriptors	**Type**	**examples**
Level 1	**Enumeration of atoms and atom types**	**C, N, O, S**
Level 2	**a.specific substructures and groups** **b. specific substituents' positions on aromatic ring**	**ArMe, ArEt, ArPr, ArOMe, ArOH, ArSMe** **6-CH_3, 8-CH_3, 4'-OCH_3 8-CF_3, 4'-Cl**
Level 3	**Experimental properties, semiempirical or *ab initio* molecular orbital calculations**	**Log P, heats of formation, HOMO, LUMO, surface area, volume, electron densities, hammet constant,**

Different level descriptors of the Hierarchical approach

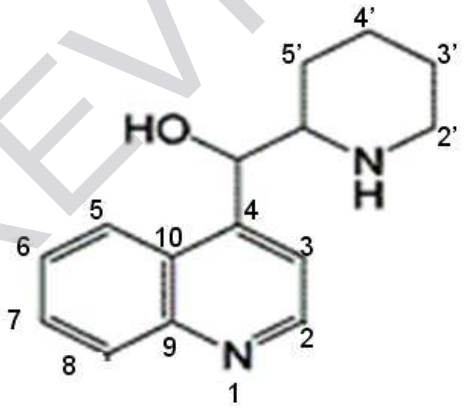
A. Level 1 Descriptors:

The descriptors C, N, O, S represents number of carbon, nitrogen, oxygen and Sulphur atoms respectively that are present in the molecule.

B. Level 2 Descriptors:

B.1 The descriptors ArMe, ArEt, ArPr, ArOMe, ArOH, ArSMe represents the number of methyl, ethyl , propyl , oxymethyl , hydroxyl and thiomethyl groups on the aromatic ring respectively.

B.2 The descriptors 6-CH_3, 8-CH_3, 4'-OCH_3 8-CF_3, 4'-Cl represents there is a methyl group attached at the 6th position in the structure, methyl group is attached at the 8th position, oxymethyl group at the 4' position, trifluromethyl at the 8th postion and a chlorine at the 4' postion respectively in the structure of the molecule. The basic structure of the molecule and its numbering used by Rode *et. al* is depicted in the following figure



Basic nucleus of mefloquine derivates used by Rode et al.,

The value for Descriptor 6-CH_3 is given one for the compounds which have methyl group at 6 position and the value becomes zero for the compounds which do not have methyl group at 6 position. Similarly, for the compounds having methyl group at 8 position the value of descriptor 8-ch3 takes one and if it doesn't then it takes value of zero.

C. Level 3 descriptors

Most of these descriptors are calculated using semiempirical molecular orbital theory which solves the Schrodinger equation of the molecule. Different types of semiempirical methods have been developed mainly by two research groups. In 1960s Pople's group developed CNDO, INDO and NNDO methods and In 1970s Dewar's group developed MNDO/3 and MNDO methods. To improve the predictive power of the molecular system the MNDO method is further optimized and parameterized to get advanced versions like AM1 and PM3 [10-13]. Log P is calculated using PCMODEL and heats of formation, HOMO, LUMO, surface area, surface volume, weight, and dipole movement are calculated using TITAN program.

MULTILINEAR REGRESSION

The basic method for QSPR analysis is essentially the solution of a multilinear regression problem. This can be expressed compactly and conveniently using matrix notation.[1, 2, 3] Suppose that there are *n* property values in **Y** and *n* associated calculated values for each *k* molecular descriptor in **X** columns. [52,53,54,55] Then Y_i, X_{ik}, and e_i can represent the *i*th value of the **Y** variable (property), the *i*th value of each of the **X** descriptors, and the *i*th unknown residual value, respectively. Collecting these terms into matrices we have:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ \vdots \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & X_{11} & \cdots & \cdots & \cdots & X_{1k} \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ 1 & X_{n1} & \cdots & \cdots & \cdots & X_{nk} \end{bmatrix}, \quad \mathbf{e} = \begin{bmatrix} e_1 \\ \vdots \\ \vdots \\ \vdots \\ e_n \end{bmatrix}$$

The multiple regression model in matrix notation then can be expressed as

$$\mathbf{Y} = \mathbf{Xb} + \mathbf{e}$$

where **b** is a column vector of coefficients (b_1 is for the intercept) and *k* is the number unknown regression coefficients for the descriptors. We recall that the goal of multiple regression is to minimize the sum of the squared residuals:

$$\min_{\mathbf{b}} \|\mathbf{e}\|_2$$

Regression coefficients that satisfy this criterion are found by solving the system of linear equations (multiplying both sides by **X**' from left)

$$\mathbf{X'Y} = \mathbf{X'Xb}$$

When the **X**variables are linearly independent (an **X'X** matrix which is of full rank), there is a unique solution to the system of linear equations. One of the ways for solving the system above is to premultiply both sides of the matrix formula for the normal equations by the inverse matrix **X'X** to give

$$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$$

The other way is to solve directly the system above using LS (underdetermined, n < k) or QR factorization for the overdetermined (n > k) system. This method is more general and does not require time-consuming matrix inversion. Singular value decomposition methods can also be used, but usually such methods are significantly more time-consuming and only advantageous when a strong linear dependence exists that would diminish quality of models.

The third way to solve the problem of linear dependency of variables (determinant of the **X'X** matrix is above zero) is by general matrix inversion, but this is usually outside the sphere of QSPR.

A fundamental principle of least squares methods, the multiple linear regression in particular, is that variance of the dependent variable can be partitioned (divided into parts) according to the source. Suppose that a dependent variable (property) is regressed on one or more descriptors and, for

convenience, the dependent variable is scaled so that its mean is 0. Next, a basic least squares identity is calculated in which the total sum of squared values on the dependent variable equals the sum of squared predicted values plus the sum of squared residual values. Stated more generally,

$$\sum(y-\bar{y})^2 = \sum(\hat{y}-\bar{y})^2 + \sum(y-\hat{y})^2$$

where the term on the left is the total sum of squared deviations of the observed values on the dependent variable from the dependent variable mean, and the terms on the right are:

(i) the sum of the squared deviations of the predicted values for the dependent variable from the dependent variable mean and

(ii) the sum of the squared deviations of the observed values on the dependent variable from the predicted values, that is, the sum of the squared residuals.

Stated yet another way,

$$SS_{Total} = SS_{Model} + SS_{Error}$$

Note that the SS_{Total} is always the same for any particular data set, but SS_{Model} and the SS_{Error} vary with the regression equation. Assuming again that the dependent variable is scaled so that its mean is 0, the $SS_{Model}$and SS_{Error} can be computed using

$$SS_{Model} = \mathbf{b}'\,\mathbf{X}'\,\mathbf{Y}'$$

$$SS_{Error} = \mathbf{Y}'\,\mathbf{Y} - \mathbf{b}'\,\mathbf{X}'\mathbf{Y}$$

Assuming that $X'X$is full-rank,

$$r^2 = 1 - \frac{SS_{Error}}{SS_{Total}}$$

$$s^2 = \frac{SS_{Error}}{n-k-1}$$

$$F(k, n-k-1) = \frac{SS_{Model}}{ks^2}$$

where $r^2$is squared correlation coefficient which is the measure of the quality of model fitnessto the property, s^2 is an unbiased estimate of the residual or error variance, and F is Fisher criteria of (k, n - k - 1) degrees of freedom. If **X'X**is not full rank, *rank(***X'X***)* + 1is substituted for k.

STATISTICAL TERMS

Correlation

Covariance is an indicator of the magnitude and direction of the linear relationship between two variables, X and Y. However, the magnitude of covariance is influenced by the units of measurement. This can be taken care of by another measure called correlation coefficient. Correlation coefficient gets derived from covariance when working with standardized data. Mathematically, correlation coefficient, r_{by}, is

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

For purpose of computation of correlation coefficient, the following expression is recommended, where s_x and s_y are the standard deviations of X and Y:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\,\sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

Correlation coefficient, like covariance, indicates the strength and direction of a linear relationship between the two variables. If X and Y are perfectly related, then $r_{xy} = 1$ when the relationship is positive and -1 when the relationship is negative. If $r_{xy} > 0$ then X and Y are positively related, and closer the value is to +1, stronger is the relation. Similarly, if $r_{xy} < 0$ then X and Y are negatively related, and closer the value is to -1, stronger is the relation

Dependent Variable

In QSAR modeling, the endpoint or the activity/property of interest that we are trying to model is our dependent variable whose value is assumed to be influenced by the independent variables that happen to be descriptors in this case.

Independent variable

Independent variables s are those variables that are assumed to have some statistical relationship with the dependent variable .One of the aims of data modeling is to capture this relation in a mathematical form.

In QSAR modeling, descriptors are the independent variables that are believed to have some influence on the endpoint or the activity/property of our interest, which we are trying to model.

Variance

When computing measures to describe the dispersion or variability of a distribution, we may look at 'spread' or 'deviation' of values from the mean value. We can then represent this 'spread' by reporting an index like 'mean deviation from the mean'. However, sum of deviation about the mean is 0 and thus 'mean deviation from the mean' turns out to be 0, irrespective of the spread in the distribution. This can