2010-01-01

Computer Aided Drug Design Methods & Quantitative Structure-Activity/Property Relationships

Suman Sirimulla
University of Texas at El Paso, ssirimulla@miners.utep.edu

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

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

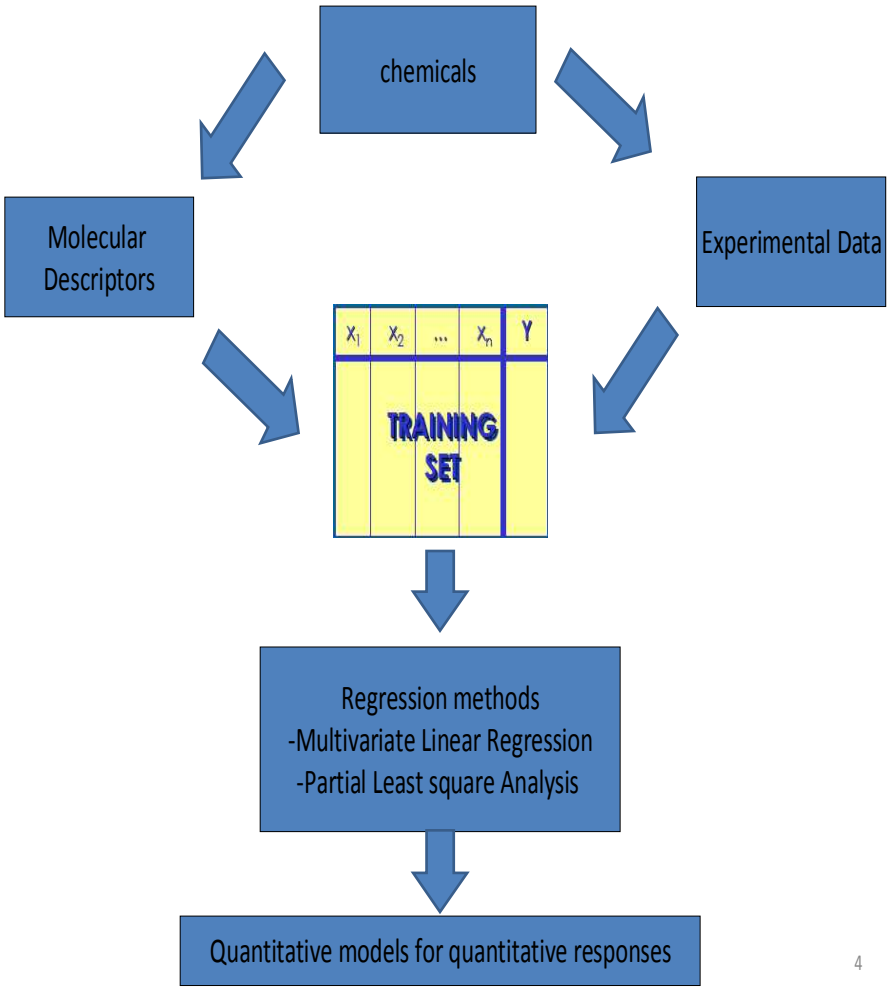
where the parameters X1 through Xn are computed for each molecule in the series and the coefficients A1 through An are calculated by fitting variations in the parameters and the biological activity. Since these relationships are generally discovered through the application of statistical techniques.

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



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

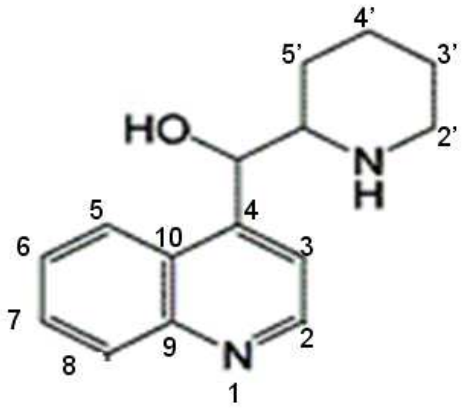
A. Level 1 Descriptors:

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 6^{th} position in the structure, methyl group is attached at the 8^{th} position, oxymethyl group at the 4' position, trifluromethyl at the 8^{th} 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 ith value of the **Y** variable (property), the ith value of each of the **X** descriptors, and the ith 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

be taken care of by squaring the deviation values and then summating them. Thus, the variance of N observations is the 'mean of squared deviations' and is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2$$

Algebraically, variance is square of standard deviation, s^2. However, standard deviation is used instead of variance because the units of the calculated variance don't make sense as they are to the power of 2 and not in the same units as the data itself.

Standard Deviation

Standard Deviation is an indicator of the 'spread' of the data (dispersion of a distribution). If mean is taken as the measure of central tendency of distribution, standard deviation tells us how much each value on an average is 'away' from the mean value (square-root of mean of squared deviations from the mean).

Thus, standard deviation is

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

In context of QSAR, descriptors having low standard deviation are not particularly relevant for regression modeling. Such descriptors don't have the required variance to be able to capture the variance in the property of interest that needs to be modeled. For classification scenario, however, descriptors with low standard deviation may capture some class or sub-class particularly well.

Residuals

Residual is simply the difference between the observed value and the predicted or estimated value.

$$Y_{resid} = Y - Y_{pred}$$

Dogra, Shaillay K., "Residuals" From QSARWorld--A Strand Life Sciences Web Resource.

http://www.qsarworld.com/qsar-statistics-residuals.php

QSAR of Mutagenic Aromatic Amines

INTRODUCTION

Aromatic amines represent one of the most important classes of industrial and environmental chemicals. Many aromatic amines have been reported to be powerful carcinogens and mutagens, and/or hemotoxicants. Exposure to aromatic amines occurs in different industrial and agricultural activities as well as in tobacco smoking. Substantial worker exposure to aromatic amines with subsequent induction of bladder cancer occurred before preventive measures were instituted. Owing to their hazard potential, aromatic amines have been the subject of many in vivo and in vitro experimental studies, as well as biomonitoring investigations [43]. Given their importance and the large amount of data available, the toxicity of the aromatic amines has been studied also with methods based on structure–activity relationship (SAR) and quantitative SAR (QSAR) concepts.1 The foundation of the modern QSAR science came about in the 1960s, after many attempts essentially qualitative. At present QSAR is one of the basic tools of modern drug and pesticide design [44 and 45], and has an increasing role in environmental sciences [46, 47, 48, **49** and 50]. Its strength derives from the fact that it permits the identification of the molecular determinants of the biological action of the chemicals, and provides mathematical models to predict the activity of chemicals not tested experimentally provided that data on similar chemicals, acting with the same mechanism, are available.

This chapter presents: (a) data of carcinogenic aromatic amines; (b) hierarchical QSAR model development of the aromatic amines; (c) the implications of our results for the risk assessment of the aromatic amines.

DATA SET AND MOLECULAR MODELING

Table I reports mutagenicity data relative to the Salmonella typhimurium TA98 strains, together with names and CAS numbers. TA98pot is the mutagenic potencies [expressed as log of (revertants/nmol)]. For nonmutagenic compounds, the potency is arbitrarily coded as -100. The chemicals with Codes 1–144 were used as training set in the present QSAR analyses. The data were retrieved from [Debnath et al., 1992], where the original sources of experimental results is also given. The chemicals with codes higher than 144 were collected by Benigini et al. and were used as external test set.

Code	Chemical name	CAS	TA98pot
1	1-Aminofluoranthene	13177-25-8	3.35
2	1,7-Diaminophenazine	28124-29-0	0.75
3	1,9-Diaminophenazine	102877-14-5	0.04
4	1-Amino-2-methylnaphthalene	2246-44-8	
5	1-Amino-4-nitronaphthalene	776-34-1	−1.77
6	1-Aminoanthracene	610-49-1	1.18
7	1-Aminonaphthalene	134-32-7	−0.6
8	1-Aminocarbazole	18992-86-4	−1.04
9	1-Aminofluorene	6344-63-4	0.43
10	1-Aminophenanthrene	4176-53-8	2.38
11	1-Aminophenazine	2876-22-4	−0.01
12	1-Aminopyrene	1606-67-3	1.43
13	2,2′-Diaminobiphenyl	1454-80-4	−1.52
14	2,3-Dimethylaniline(2,3-xylidine)	87-59-2	
15	2,4,5-Trimethylaniline	137-17-7	−1.32
16	2,4-Diamino-*n*-butylbenzene	63921-07-3	−2.7
17	2,4-Diaminoisopropylbenzene	14235-45-1	−3
18	2,4-Difluoroaniline	367-25-9	−2.7
19	2,4-Dimethylanilineb(2,4-xylidine)	95-68-1	−2.22
20	2,4-Dinitroaniline	97-02-9	−2
21	2,4′-Diaminobiphenyl	492-17-1	−0.92
22	2,5-Dimethylaniline(2,5-xylidine)	95-78-3	−2.4
23	2,6-Dichloro-1,4-phenylenediamine	609-20-1	−0.69
24	2,7-Diaminofluorene	525-64-4	0.48
25	2,7-Diaminophenazine	120209-97-4	3.97
26	2,8-Diaminophenazine	7704-40-7	1.12
27	2-Amino-1-methylnaphthalene	771-13-1	
28	2-Amino-1-nitronaphthalene	606-57-5	−1.17
29	2-Amino-3-methylnaphthalene	10546-24-4	
30	2-Amino-3′-nitrobiphenyl	34862-87-8	−0.89
31	2-Amino-4-chlorophenol	95-85-2	−3
32	2-Amino-4-methylphenol	95-84-1	−2.1
33	2-Amino-4′-nitrobiphenyl	6272-52-2	−0.62
34	2-Amino-5-nitrophenol	121-88-0	−2.52
35	2-Amino-7-acetamidofluorene	6957-50-2	1.18
36	2-Amino-7-nitrofluorene	1214-32-0	3
37	2-Aminoanthracene	613-13-8	2.62
38	2-Aminobiphenyl	90-41-5	−1.49
39	2-Aminocarbazole	4539-51-9	0.6
40	2-Aminofluoranthene	13177-26-9	3.23
41	2-Aminofluorene	153-78-6	1.93
42	2-Aminonaphthalene	91-59-8	−0.67
43	2-Aminophenanthrene	3366-65-2	2.46
44	2-Aminophenazine	2876-23-5	0.55
45	2-Aminopyrene	1732-23-6	3.5
46	2-Bromo-4,6-dinitroaniline	1817-73-8	−0.54
47	2-Bromo-7-aminofluorene	6638-60-4	2.62
48	2-Chloroaniline	95-51-2	−3
49	2-Ethyl-4-chloroaniline	30273-39-3	
50	2-Hydroxy-7-aminofluorene	1953-38-4	0.41
51	2-Methoxy-5-methylaniline(*p*-cresidine)	120-71-8	−2.05
52	2-Methyl-4-bromoaniline	583-75-5	
53	2-Methyl-4-chloroaniline	95-69-2	−100
54	3,3′-Diaminobiphenyl	2050-89-7	−1.3
55	3,3′-Dichlorobenzidine	91-94-1	0.81
56	3,3′-Dimethoxybenzidine	119-90-4	0.15
57	3,3′-Dimethylbenzidine	612-82-8	0.01
58	3,4-Dimethylaniline(3,4-xylidine)	95-64-7	
59	3,4′-Diaminobiphenyl	32316-90-8	0.2
60	3-Amino-2′-nitrobiphenyl	96187-18-7	−1.3

Code	Chemical name	CAS	TA98pot
61	3-Amino-3′-nitrobiphenyl	31835-64-0	−0.55
62	3-Amino-4-methylbiphenyl	80938-67-6	
63	3-Amino-4′-nitrobiphenyl	53059-29-3	0.69
64	3-Trifluoromethylaniline	98-16-8	−0.8
65	3-Aminocarbazole	1635530	−0.48
66	3-Aminofluoranthene	2693-46-1	3.31
67	3-Aminofluorene	6344-66-7	0.89
68	3-Aminophenanthrene	1892-54-2	3.77
69	3-Aminoquinoline	580-17-6	−3.14
70	3-Methoxy-4-methylaniline(o-cresidine)	16452-01-0	−1.96
71	4,4′-Ethylenebis(aniline)	621-95-4	−2.15
72	4,4′-Methylenebis(*o*-ethylaniline)	19900-65-3	−0.99
73	4,4′-Methylenebis(*o*-fluoroaniline)	13824-23-2	0.23
74	4,4′-Methylenebis(*o*-isopropyl-aniline)	19900-66-4	−1.77
75	4,4′-Methylenedianiline	13552-44-8	−1.6
76	4-Aminophenyldisulfide	722-27-0	−1.03
77	4-Amino-2′-nitrobiphenyl	1140-28-9	−0.92
78	4-Amino-3-methylbiphenyl	63019-98-7	
79	4-Amino-3′-nitrobiphenyl	1141-29-3	1.02
80	4-Amino-4′-nitrobiphenyl	1211-40-1	1.04
81	4-Aminobiphenyl	92-67-1	−0.14
82	4-Aminocarbazole	18992-64-8	−1.42
83	4-Aminofluorene	7083-63-8	1.13
84	4-Aminophenanthrene	17423-48-2	−100
85	4-Aminophenylether	101-80-4	−1.14
86	4-Aminophenylsulfide	139-65-1	0.31
87	4-Aminopyrene	17075-03-5	3.16
88	4-Bromoaniline	106-40-1	−2.7
89	4-Chloro-1,2-phenylenediamine	95-83-0	−0.49
90	4-Chloro-2-nitroaniline	89-63-4	−2.22
91	4-Chloroaniline	106-47-8	−2.52
92	4-Cyclohexylaniline	6373-50-8	−1.24
93	4-Ethoxyaniline(*p*-phenetidine)	156-43-4	−2.3
94	4-Fluoroaniline	371-40-4	−3.32
95	4-Methoxy-2-methylaniline(*m*-cresidine)	102-50-1	−3
96	4-Methyl-2-bromoaniline	583-68-6	
97	2-Chloro-4-methylaniline	615-65-6	−100
98	4-Methyl-2-chloroaniline	615-65-6	
99	4-Phenoxyaniline	139-59-3	0.38
100	5-Aminoquinoline	611-34-7	−2
101	6-Aminochrysene	2642-98-0	1.83
102	6-Aminoquinoline	580-15-4	−2.67
103	7-Aminofluoranthene	13177-27-0	2.88
104	8-Aminofluoranthene	5869-25-0	3.8
105	8-Aminoquinoline	578-66-5	−1.14
106	9-Aminoanthraceneb	779-03-3	0.87
107	9-Aminophenanthrene	947-73-9	2.98
108	Benzidine	92-87-5	−0.39
109	4-Methoxyaniline	20265-97-8	−100
110	3-Methoxyaniline	536-90-3	−100
111	Aniline	142-04-1	−100
112	3-Chloroaniline	108-42-9	−100
113	3-Ethoxyaniline	621-33-0	−100
114	2-Ethoxyaniline	94-70-2	−100
115	4-Aminophenol	123-30-8	−100
116	3-Aminophenol	591-27-5	−100
117	4,4′ Methylenebis(2,6-diisopropylaniline)	19900-69-7	−100
118	4,4′ Methylenebis(2,6-diethylaniline)	13680-35-8	−100
119	4,4′ Methylenebis(2-methyl-6-t-butylaniline)	13680-36-9	−100
120	4,4′ Methylenebis(2-methyl-6-isopropylaniline)	16298-38-7	−100
121	4,4′ Methylenebis(2-methyl-6-ethylaniline)	19900-72-2	−100

Code	Chemical name	CAS	TA98pot
122	4,4′ Methylenebis(2,6-dimethylaniline)	4073-98-7	−100
123	3-Aminobiphenyl	2243-47-2	−100
124	2,3-Diaminobiphenyl	ND	−100
125	2-Methoxyaniline	134-29-0	−100
126	2-Aminophenol	95-55-6	−100
127	2,4,6-Trimethylaniline	1619-79-8	
128	2,4,6-Tribromoaniline	147-82-0	
129	2,4,6-Trichloroaniline	634-93-5	
130	2,6-Diethylaniline	579-66-8	
131	3,5-Dimethylaniline	108-69-0	
132	2,6-Dimethylaniline	87-62-7	
133	2,4-Dibromoaniline	615-57-6	
134	2,4-Dichloroaniline	554-00-7	
135	4-Iodoaniline	540-37-4	
136	2-Iodoaniline	615-43-0	
137	2-Fluoroaniline	348-54-9	
138	2-Bromoaniline	615-36-1	
139	4-Ethylaniline	589-16-2	
140	2-Ethylaniline	578-54-1	
141	4-Methylaniline	540-23-8	
142	3-Methylaniline	638-03-9	
143	2-Methylaniline	636-21-5	
144	9-Aminofluorene	525-03-1	
148	2,4-Diaminophenol*2HCl	137-09-7	0.13
149	2,4-Diaminotoluene	95-80-7	−2.68
151	2,5-Toluenediaminesulfate	6369-59-1	−1.76
152	2,6-Toluenediamine*2HCl	15481-70-6	−1.34
153	2-Acetylaminofluorene	53-96-3	0.87
154	2-Amino-4-nitrophenol	99-57-0	−100
156	2-Nitro-*p*-phenylenediamine	5307-14-2	0.01
162	4,4′-Methylenebis(2-chloroaniline)	101-14-4	−0.53
165	4,4′-Sulphonylbisacetanilide	77-46-3	−100
166	4-(Chloroacetyl)-acetylanilide	140-49-8	−0.61
169	4-Amino-2-nitrophenol	119-34-6	−100
172	4-Nitroanthranilicacid	619-17-0	−0.64
173	4-Nitroo-phenylenediamine	99-56-9	1.16
178	Dapsone(4,4-sulfonyldianiline)	80-08-0	−100
179	Fluometuron	2164-17-2	−100
180	H.C.RedN3	354-65-6	0.73
181	HCBlueN1	2784-94-3	−0.99
185	Monuron	150-68-5	−100
186	*N,N*-Dimethylaniline	121-69-7	−100
191	*N*-Phenyl-2-naphthylamine	135-88-6	−100
194	Phenacetin	62-44-2	−100
196	Trifuralin	1582-09-8	−100
197	*o*-Anthranilicacid	118-92-3	−100
199	*p*-Nitroaniline	100-01-6	−2
200	*p*-Phenylenediamine*2HCl	624-18-0	−1.97
201	1-Ethyl-2-aminonaphthalene	389104-53-4	−0.01
202	1-*i*Propyl-2-aminonaphthalene	389104-54-5	−0.63
203	1-*n*Butyl-2-aminonaphthalene	ND	0.01
204	1-*t*Butyl-2-aminonaphthalene	ND	−100
205	1-Ethyl-2-aminofluorene	ND	−0.19
206	1-*i*Propyl-2-aminofluorene	ND	0.19
207	1-*n*Butyl-2-aminofluorene	ND	0.82
208	1-*t*Butyl-2-aminofluorene	ND	−100
209	3-Ethyl-4-aminobiphenyl	389104-60-3	0.17
210	3-*i*Propyl-4-aminobiphenyl	ND	−0.01
211	3-*n*Butyl-4-aminobiphenyl	ND	−0.17
212	3-*t*Butyl-4-aminobiphenyl	ND	−0.41
213	3,5-Dimethyl-4-aminobiphenyl	ND	−1.43

Code	Chemical name	CAS	TA98pot
214	3,5-Diethyl-4-aminobiphenyl	ND	−100
215	3,5-Diipropyl-4-aminobiphenyl	ND	−100
216	4′-Methyl-4-aminobiphenyl	1204-78-0	0.64
217	4′-Ethyl-4-aminobiphenyl	ND	−0.14
218	4′-*i*Propyl-4-aminobiphenyl	ND	−100
219	4′-*n*Butyl-4-aminobiphenyl	ND	−100
220	4′-*t*Butyl-4-aminobiphenyl	ND	−100
221	7-Methyl-2-aminofluorene	ND	1.47
222	7-tButyl-2-aminofluorene	ND	−0.2
223	7-Adamantyl-2aminofluorene	ND	−0.94
224	7-Trifluoromethyl-2-aminofluorene	ND	0.76
225	3′-Methyl-4-aminobiphenyl	ND	0.74
226	3′,5′-Dimethyl-4-aminobiphenyl	ND	0.35
227	4′-Trifluoromethyl-4-aminobiphenyl	ND	−100
228	3′-Trifluoromethyl-4-aminobiphenyl	397-28-4	−100
229	3′,5′-Ditrifluoromethyl-4-aminobiphenyl	ND	−100

MOLECULAR DESCRIPTORS/INDEPENDENT VARIABLES

Level-1

In Level-1 we used enumeration of different type of atoms as indicator variables.

CARBON represents number of carbons in a compound

HYDROGEN represents number of hydrogens in a compound

NITROGEN represents number of nitrogens in a compound

OXYGEN represents number of oxygens in a compound

CHLORINE represents number of chlorines in a compound

BROMINE represents number of bromines in a compound

SULPHUR represents number of sulphurs in a compound

Level-2

In level-2 we used both functional group indicator variables and topological descriptors

NRING represents number of aromatic rings in a compound

FIVERING represents number of five membered rings in a compound

NO2 represents number of nitro groups in a compound

NCH3N represents number of alkyl chains between rings in a compound

METHYL represents number of methyl in a compound

AMINES represents number of amino groups in a compound

Level-3

In level-3 we used physical properties such a as hydrophobicity, molecular orbital energies of HOMO and LUMO, molecular refractivities of substituents on the functional aniline rings of the compounds.

LogP represents hydrophobicity

HOMO represents molecular orbital energy of Highest occupied molecular orbital in eV.

LUMO represents molecular orbital energy of Lowest unoccupied molecular orbital in eV.

MR3 represents molar refractivity of substituents in the meta position of aniline ring.

MR5 represents molar refractivity of substituents in the para position of aniline ring.

MR6 represents molar refractivity of substituents in the ortho position of aniline ring.

RESULTS

A simple QSAR method has been developed by using enumeration of different types of atoms, Functional group indicator variable model, coupled with LogP, HOMO, LUMO and molecular refractivities. We omitted the compound which has no activity. The Multi Linear Regression (MLR) was carried out by different methods such as including all parameters, stepwise forward method, and stepwise backward method using SYSTAT (statistical software). The results presented in this thesis will consist of an analysis of structural descriptors and mutagenic activities of 229 aromatic amines. Both Free-Wilson analysis and Hierarchical approach outlined in the above chapter are discussed here with detailed results. Several possible QSAR model equations are developed for each analysis by conducting the regression including all the parameters and then conducting stepwise forward method followed by stepwise backward method. The best fit equation is taken into consideration for the prediction of the activities of new molecules.

Several multilinear regression analyses were carried out for this data, starting from the level-1 descriptors, then advanced with level-2 descriptors and ends with level-3 descriptors. The regression models for each level are shown below.

Multilinear regression analysis of aromatic amines using counting of different types of atoms (level 1)

Level-1

The atom level is also the first and simplest level for descriptors in our preferred hierarchical variant of QSAR methodology. Model-1 descriptors are defined as the numbers of atoms by element as descriptors (C, H, N, O, and S). The statistical results of Model-1 regression are tabulated in Table 2.2. A squared multiple square of 0.6 is achieved using level-1 descriptors. This suggests there is lot of scope for improvement of the model.

Table 2.2 Regression model for 92 Aromatic amines using different types of atoms (level 1)

ASSUMING MIXTURE MODEL.

DEP VAR: TA98 N: 92 MULTIPLE R: 0.776 SQUARED MULTIPLE R: 0.602
ADJUSTED SQUARED MULTIPLE R: .574 STANDARD ERROR OF ESTIMATE: 1.244

VARIABLE	COEFFICIENT	STD ERROR	STD COEF	TOLERANCE	T	P(2 TAIL)
CARBON	0.529	0.050	3.314	0.047	10.548	0.000
HYDROGEN	-0.459	0.053	-2.680	0.048	-8.603	0.000
NITROGEN	-0.640	0.176	-0.605	0.169	-3.630	0.000
OXYGEN	-0.287	0.152	-0.160	0.653	-1.891	0.062
CHLORINE	-0.668	0.298	-0.156	0.971	-2.244	0.027
BROMINE	-0.537	0.360	-0.106	0.920	-1.493	0.139
SULPHUR	-0.105	0.741	-0.010	0.939	-0.142	0.887

ANALYSIS OF VARIANCE

SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
REGRESSION	198.997	6	33.166	21.430	0.000
RESIDUAL	131.552	85	1.548		

Multilinear regression analysis of aromatic amines using level 1 & 2 descriptors

Level-2

This analysis was carried out by incorporating functional group indicator variables and substituents to level-1 descriptors. The substituents on the basic structure are taken into account and they are correlated to its biological activity. According to free-wilson method, each substituent on the compound has its own effect on the total activity of that particular compound. This theory assumes that the total activity of the compound equals to the sum of activities of all substituents including the basic structure. The statistical results of Model-2 regression are tabulated in Table 2.3. There is no significant improvement of statistical results from Model-1 and still there is lot of scope for improvement of the model.

Table 2.3 Regression model for 92 Aromatic amines using all level 1 & 2 descriptors

MODEL CONTAINS NO CONSTANT.

DEP VAR: TA98 N: 92 MULTIPLE R: 0.784 SQUARED MULTIPLE R: 0.615
ADJUSTED SQUARED MULTIPLE R: .556 STANDARD ERROR OF ESTIMATE: 1.280

VARIABLE	COEFFICIENT	STD ERROR	STD COEF	TOLERANCE	T	P(2 TAIL)
CARBON	0.528	0.248	3.278	0.002	2.128	0.036
HYDROGEN	-0.472	0.142	-2.732	0.007	-3.325	0.001
NITROGEN	-0.594	0.250	-0.557	0.089	-2.380	0.020
OXYGEN	-0.311	0.228	-0.172	0.306	-1.364	0.176
CHLORINE	-0.610	0.336	-0.141	0.808	-1.819	0.073
BROMINE	-0.431	0.405	-0.085	0.767	-1.064	0.291
SULPHUR	-0.117	0.774	-0.011	0.911	-0.151	0.881
NRING	0.105	0.686	0.131	0.007	0.153	0.879
FIVERING	-0.113	0.509	-0.021	0.575	-0.223	0.824
NO2	0.085	0.249	0.033	0.528	0.339	0.735
NCH3N	0.397	0.445	0.078	0.636	0.892	0.375
METHYL	0.122	0.270	0.045	0.500	0.452	0.652
AMINES	-0.194	0.512	-0.127	0.044	-0.379	0.705

ANALYSIS OF VARIANCE

SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
REGRESSION	206.555	13	15.889	9.696	0.000
RESIDUAL	129.457	79	1.639		

Multilinear regression analysis of aromatic amines using level 1, 2 & 3 descriptors

Level-3

In Level-3 we added LogP, HOMO, LUMO and molecular refractivities of substituents at different positions of aniline ring to level 1 & 2 descriptors. The Multi Linear Regression (MLR) was carried out by different methods such as including all parameters, stepwise forward method, and stepwise backward method using SYSTAT (statistical software). Several iterations were carried out using all the descriptors to get a good model.

Run # 1 **Multilinear regression analysis of aromatic amines using all level 1, 2 & 3 descriptors**

The below shown table consists of all the descriptors used in the study and mixture model is used in this regression. Though the p value is 0 which implies the regression is statistically valid, it has high standard error of estimate.

Table 2.4 Regression model for 93 Aromatic amines using all level 1, 2 & 3 descriptors

```
>MODEL TA98 = NRING+FIVERING+NO2+NCH3N+METHYL+AMINES+CARBON+HYDROGEN+,
>NITROGEN+OXYGEN+CHLORINE+BROMINE+SULPHUR+LOGP+HOMO+LUMO+MR5+MR3+MR2+,
>MR6
>ESTIMATE/MIX

 ASSUMING MIXTURE MODEL.

 DEP VAR:    TA98      N:      93  MULTIPLE R: 0.791  SQUARED MULTIPLE R: 0.626
 ADJUSTED SQUARED MULTIPLE R:  .529    STANDARD ERROR OF ESTIMATE:        1.335

 VARIABLE      COEFFICIENT    STD ERROR     STD COEF TOLERANCE    T   P(2 TAIL)

 NRING              -0.246        0.850       -0.305     0.005   -0.289    0.773
 FIVERING           -0.397        0.594       -0.071     0.459   -0.669    0.506
 NO2                -0.051        0.273       -0.019     0.477   -0.185    0.854
 NCH3N               0.303        0.471        0.059     0.618    0.643    0.522
 METHYL              0.123        0.291        0.044     0.468    0.423    0.674
 AMINES             -1.130        0.589       -0.727     0.036   -1.919    0.059
 CARBON              0.744        0.297        4.569     0.002    2.505    0.014
 HYDROGEN           -0.500        0.156       -2.857     0.006   -3.203    0.002
 NITROGEN           -0.117        0.305       -0.111     0.062   -0.384    0.702
 OXYGEN             -0.302        0.295       -0.164     0.198   -1.023    0.310
 CHLORINE           -0.598        0.391       -0.136     0.647   -1.528    0.131
 BROMINE            -0.264        0.556       -0.051     0.443   -0.475    0.636
 SULPHUR             0.145        1.073        0.013     0.516    0.135    0.893
 LOGP               -0.066        0.357       -0.088     0.022   -0.184    0.855
 HOMO               -0.007        0.013       -0.126     0.093   -0.538    0.592
 LUMO               -0.009        0.025       -0.074     0.111   -0.344    0.732
 MR5                -0.562        0.832       -0.079     0.377   -0.675    0.502
 MR3                -0.339        0.235       -0.151     0.467   -1.442    0.153
 MR2                -0.612        0.240       -0.277     0.433   -2.545    0.013
 MR6                 0.635        1.753        0.056     0.215    0.362    0.718


                        ANALYSIS OF VARIANCE

 SOURCE       SUM-OF-SQUARES   DF  MEAN-SQUARE     F-RATIO       P

 REGRESSION         218.033    19       11.475       6.440    0.000
 RESIDUAL           130.080    73        1.782
```

Stepwise backward with constant

93 compounds were considered for the study from the first 121 compounds. Compounds with no activity and outliers were removed from the study. The statistical results were improved from each level. This result consists of descriptors from all the three levels and stepwise backward regression method is carried out.

DEP VAR: TA98 N: 93 MULTIPLE R: 0.875 SQUARED MULTIPLE R: 0.765
ADJUSTED SQUARED MULTIPLE R: .743 STANDARD ERROR OF ESTIMATE: 0.986

VARIABLE	COEFFICIENT	STD ERROR	STD COEF	TOLERANCE	T	P(2 TAIL)
CONSTANT	-5.471	0.517	0.000	.	-10.580	0.000
NRING	1.777	0.226	0.850	0.238	7.846	0.000
FIVERING	-0.590	0.360	-0.098	0.774	-1.639	0.105
CARBON	0.109	0.064	0.178	0.255	1.701	0.093
OXYGEN	0.292	0.122	0.139	0.821	2.391	0.019
CHLORINE	0.470	0.254	0.105	0.869	1.847	0.068
BROMINE	0.921	0.314	0.172	0.815	2.936	0.004
SULPHUR	1.093	0.601	0.100	0.927	1.818	0.073
MR2	-0.293	0.156	-0.104	0.909	-1.874	0.064

ANALYSIS OF VARIANCE

SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
REGRESSION	266.400	8	33.300	34.232	0.000
RESIDUAL	81.713	84	0.973		

Stepwise forward with constant

The best statistical results obtained for the above data set using stepwise forward multilinear regression analysis. Statistical results were significantly improved when compared to previous results. A Squared multiple R is of 0.807 and a multiple R of 0.898 is obtained using this equation. This equation suggests that the mutagenicity is proportional to increase in number of aromatic rings, carbons, oxygens, chlorine, bromine and sulphur.

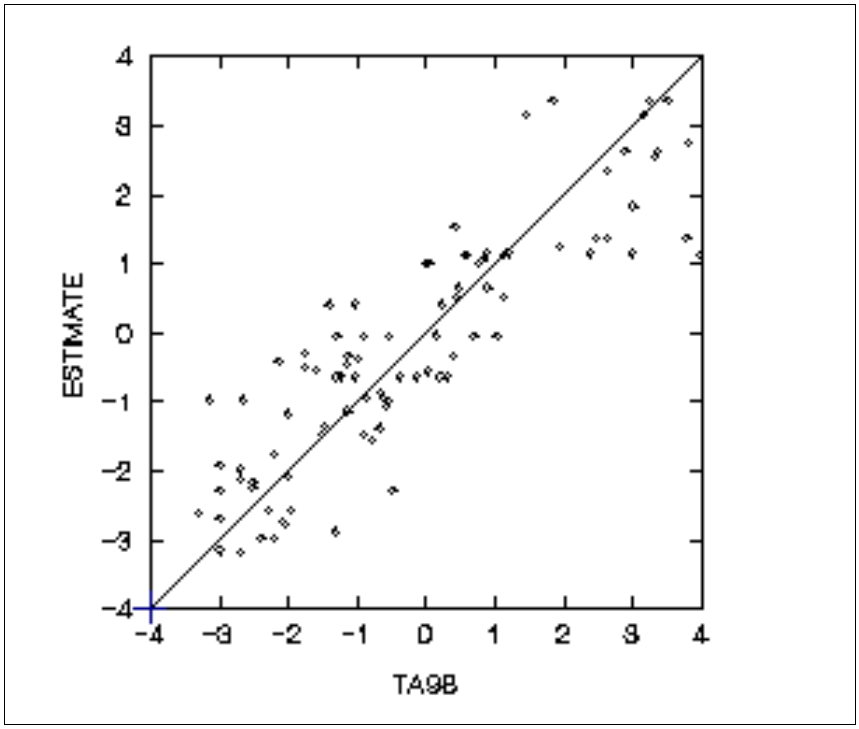
DEP VAR: TA98 N: 90 MULTIPLE R: 0.898 SQUARED MULTIPLE R: 0.807
ADJUSTED SQUARED MULTIPLE R: .786 STANDARD ERROR OF ESTIMATE: 0.859

VARIABLE	COEFFICIENT	STD ERROR	STD COEF	TOLERANCE	T	P(2 TAIL)
CONSTANT	-4.631	0.498	0.000	.	-9.910	0.000
NRING	1.729	0.201	0.875	0.233	8.611	0.000
FIVERING	-0.535	0.320	-0.095	0.745	-1.672	0.098
CARBON	0.105	0.057	0.181	0.250	1.847	0.069
NITROGEN	-0.294	0.133	-0.115	0.891	-2.205	0.030
OXYGEN	0.356	0.111	0.180	0.765	3.215	0.002
CHLORINE	0.433	0.223	0.103	0.862	1.944	0.055
BROMINE	0.915	0.275	0.182	0.808	3.334	0.001
SULPHUR	1.057	0.525	0.103	0.922	2.013	0.047
MR2	-0.243	0.137	-0.091	0.909	-1.777	0.079

ANALYSIS OF VARIANCE

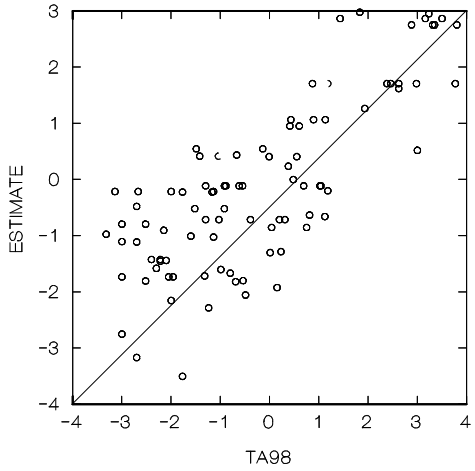
SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
REGRESSION	247.033	9	27.448	37.216	0.000
RESIDUAL	59.003	80	0.738		

Figure illustrates Estimate vs Experimenta l(TA98) mutagenic activities of best model equation.



Crossvalidation

Crossvalidation is an important step in the process of QSAR development. Crossvalidation is a process of verifying the predictive ability of the developed model. Chemical codes from 144 to 229 are used as the test set. The crossvalidation results can be viewed in the following graph as Experimental results vs predicted results. The R2 for the crossvalidation results is 0.76



Conclusion

A QSAR was developed for mutagenic aromatic amines using using simple descriptors such as enumeration of atoms, substitutents, log P, HOMO LUMO, and Molar refractivity of substituents on aniline rings. The statistical results shows that these results are satisfactory and can be used to predict the mutagenicity of other similar aromatic amines.

REFERENCES

1) Rainer Franke; "Theoretical Drug Design Methods" 1984.
2) Polman, S.; Kokpol, S. K.; Hannongbua, S. V.; Rode, B. M.; *Analytical Sciences (1989), 5, 641-644.*
3) Nguyen-Cong V; Rode B M Institute for General, Inorganic and Theoretical Chemistry, University of Innsburck, Austria *Journal of chemical information and computer sciences (1996), 36(1), 114-7.*
4) Kokpol, S. K.; Hannongbua, S. V.; Thongrit, N.; Polman, S.; Rode, B. M.; Schwendinger, M. G. Analysis of structure-activity relation for primaquine antimalarial drugs by a quantum pharmacological approach**.** Analytical Sciences (1988), 4(6), 565-8.
5) Journal of medicinal chemistry (1979), 22(4), 366-91.
6) C. J. Ohnmacht, A. R. Patel and R. E. Lutz, *J. Med. chem., 14, 926 (1971).*
7) D. W. Boykin, A. R. Patel and R. E. Lutz, *J. Med. Chem., 3, 273 (1968).*
8) R. M. Pinder and A. Burger, *J. Med. Chem.,* 11, 267 (1968
9) The pharmacological basis of therapeutics, seventh Edition, Goodman and Gilman
10) Pople et al., *J. chem.. phy (1965) 43, S129*
11) Pople et al., J. Chem . phys (1965) 43, S136
12) Pople et al., J. Chem . phys (1967) 47, 2026.\
13) R.C. Bingham, M.J.S Dewar and D. H. Lo, J. Amer. Chem. Soc (1975) 97, 1285.
14) Hansch, C. (1969) A Quantitative Approach to Biochemical Structure-Activity Relationships. *Acct. Chem. Res. 2: 232-239*
15) Hansch, C., Leo, A., and Taft, R.W. (1991) A Survey of Hammett Substituent Constants and Resonance and Field Parameters. Chem. Rev. 91: 165-195.
16) Hansch, C., Leo, A., and Hoekman, D. (1995) *Exploring QSAR - Hydrophobic, Electronic, and Steric Constants.* American Chemical Society, Washington, D.C.

17) Seydel, J.K. (1966) Prediction of *in Vitro* Activity of Sulfonamides, Using Hammett Constants or Spectrophotometric Data of the Basic Amines for Calculation. Mol. Pharmacol. 2: 259-265.

18) Hansch, C. (1974) Drug Research or the Luck of the Draw. J. Chem. Ed. 51: 360-365.

19) Alexander golbraikh, Alexander Tropsha, Journal of Molecular Graphics and Modeling 20 (2002) 269-276.

20) D.C. Warhurst, Infection 27 (1999), p. S55

21) J.F Trape, *Am.J Trop. Med. Hyg.* 64 (2001), p 12

22) W.Marquino, M Huilca, C. Calampa, E. Falconi and C. Cabezas, *Am. J. Trop. Med. Hyg*. 68 (2003), p, 608

23) Mankil Jung, Hanjo Kim, Ki Youp Nam and Kyoung Tai No, *Bioorganic & medicinal Chemistry Letters vol 15, Issue 12, June 2005, pages 2994-2995.*

24) Hien, T. T.; White, N. J.; Qinghaosu. *The Lancet* **1993**, *341*, 603-608

25) Luo, X. D.; Shen, C. C. The chemistry, pharmacology, and clinical applications of Qinghaosu (artemisinin) and its derivatives. *Med. Res. Rev.* **1987**, *7*, 29-52.

26) People's Health Publishing: Lanzhou, China, 1991; pp 944-946

27) *J. Med. Chem.,* **45** (19), 4321 -4335, 2002.

28) *J. Med. Chem.,* **46**, 4244 -4258, 2003.

29) Desowitz, R. S. The Malaria Capers (More Tales of People, Research and Reality); W. W. Norton & Company: New York, 1991.

30) 2. Van Agtmael, M. A.; Eggelte, T. A.; Van Boxtel, C. J. Artemisinin drugs in the treatment of malaria: from medicinal herb to registered medication. *Trends Pharmacol. Sci.* **1999**, *20*, 199-205.

31) S. Wold, L. Eriksson, Statistical validation of QSAR results, in: H. van de Waterbeemd (Ed.), Chemometrics Methods in Molecular Design, VCH, Weinheim, 1995, pp. 309–318.

32) Journal of molecular graphics and modeling 20 (2002) 269-276.

33) Breiman, L.; Friedman, J.J. Estimating Optimal Transformations for Multiple Regression and Correlation (with discussion) *J.Amer. statist. Assoc.* 1985, 80, 580-619.

34) Bruce-Chwatt LJ et al. *chemotherapy of malaria*, Geneva, World Health organization, 1986.

35) http://www.cdc.gov/malaria/drug_resistance.htm

36) http://www.cdc.gov/malaria/pdf/bloland_WHO2001.pdf

37) S. Clementi, S. Wold, How to choose the proper statistical method, in: H. van de Waterbeemd (Ed.), Chemometrics Methods in Molecular Design, VCH, Weinheim, 1995, pp. 319–338.

38) S. Wold, PLS for multivariate linear modeling, in: H. van de Waterbeemd (Ed.), Chemometrics Methods in Molecular Design, VCH, Weinheim, 1995, pp. 195–218.

39) B. Hoffman, S.J. Cho, W. Zheng, S. Wyrick, D.E. Nichols, R.B. Mailman and A. Tropsha, Quantitative structure–activity relationship modeling of dopamine D(1) antagonists using comparative molecular field analysis, genetic algorithms-partial least-squares, and *k* nearest neighbor methods. *J. Med. Chem.* **42** (1999), pp. 3217–3226

40) Ajay, A unified framework for using neural networks to build QSARs, J. Med. Chem. 36 (1993) 3565–3571.

41) http://www.infomed.org/100drugs/meflphar.html

42) Hung-Ta Chen, M.S thesis, Department of chemistry, University of Texas at El Paso, Dec 1995. page41-46.

43) Y.T. Woo, D.Y. Lai, Aromatic amino and nitro-amino compounds and their halogenated derivatives, in: E. Bingham, B. Cohrssen, C.H. Powell (Eds.), Patty's Toxicology, Wiley, New York, 2001, pp. 969–1105.

44) C. Hansch, A. Leo, Exploring QSAR. Part 1. Fundamentals and Applications in Chemistry and Biology, American Chemical Society, Washington, DC, 1995.

45) T. Fujita , Recent success stories leading to commercializable bioactive compounds with the aid of traditional QSAR procedures. Q. Struct. Act. Relat. **16** (1997), pp. 107–112.

46) C. Hansch, D. Kim, A.J. Leo, E. Novellino, C. Silipo and A. Vittoria , Toward a quantitative comparative toxicology of organic compounds. Crit. Rev. Toxicol. **19** (1989), pp. 185–226.

47) H.J.M. Verhaar, J. Solbe, J. Speksnijder, C.J. van Leeuwen and J.L.M. Hermens , Classifying environmental pollutants. Part 3. External validation of the classification system. Chemosphere **40** (2000), pp. 875–883.

48) H. Hong, W. Tong, H. Fang, L.M. Shi, Q. Xie, J. Wu, R. Perkins, J.D. Walker, W. Branham and D.M. Sheehan , Prediction of estrogen-receptor binding for 58,000 chemicals using an integrated system of a tree-based model with structural alerts. Environ. Health Perspect. **110** (2002), pp. 29–36.

49) R. Benigni and A. Giuliani , Quantitative structure–activity relationship (QSAR) studies of mutagens and carcinogens. Med. Res. Rev. **16** (1996), pp. 267–284.

50) A.M. Richard, R. Benigni, AI and SAR approaches for predicting chemical carcinogenicity: survey and status report, SAR QSAR Environ. Res. 13 (1) (2002) 1–1.

51) Benigni, Romualdo; Bossa, Cecilia; Netzeva, Tatiana; Rodomonte, Andrea; Tsakovska, Ivanka. Mechanistic QSAR of aromatic amines: new models for discriminating between homocyclic mutagens and nonmutagens, and validation of models for carcinogens. Environmental and Molecular Mutagenesis (2007), 48(9), 754-771.

52) http://www.statsoft.com/textbook/stathome.html

53) Darlington, R. B. *Regression and linear models*. New York: McGraw-Hill, 1990.

54) Neter, J.; Wasserman, W.; Kutner, M. H. *Applied linear regression models* (2nd ed.).

55) www.codessa-pro.com/methods/MR.htm

Prediction of partial molar volumes of amino acids

1. INTRODUCTION

The research presented in this chapter is an extension of a master's thesis written by Ms. Mericarmen Lerma. Ms.Lerma developed a QSPR to predict partial molar volume's of amino acids and small pepdtides by just counting the atoms. In this part of my research I tried not only to reproduce her results but added 3-dimensional descriptors and also adjusted the amino acids pH to their Isoelectric PH.

The term "amino acids" refers in relation with the twenty genetically encoded amino acids, either in their neutral molecular forms or their charged zwitterionic structures. Amino acid partial structures called residues, found in peptides and proteins, are also identified using the name of the parent acid[1]. These names, common abbreviations, used to identify the residual structures, and linear diagrams for the neutral molecular structures of each amino acid compound are given in Table I. The word "peptide" generally denotes a small polymer made up of two or larger numbers of amino acid residues, linked by peptide bonds. Peptides can possess linear, cyclic, or branched molecular structures[1].

Partial molar volumes (PMVs) of the zwitterionic ά-amino acids and peptides in neutral aqueous solution are experimental properties, related to the thermodynamic properties by the equations of statistical mechanics (PMV thermochemical symbol, V2º, units, cm3/mol). Experimental PMVs of aqueous amino acids and peptides to be considered in this paper are generally known to be very precise and accurate, derived from measurements of solution densities in a series of decreasing concentration, with extrapolation to infinite dilution.

A great deal of the interest pertaining to amino acids and larger peptides is coupled with their characterization as the basic building blocks of proteins. Proteins are macromolecules (polymers) that are constructed from one or more unbranched chains of amino acids. A typical protein contains 200- 300

amino acids, but some are smaller (often called peptides) and some are much larger. The largest to date is titin, a protein found in skeleton and cardiac muscle; containing 26,926 amino acids in a single chain. Every function of the living cell depends in some way on proteins. In fact, the structure of cells, and the extracellular matrix in which they are embedded, is largely made of protein. Furthermore, the catalysis of a majority of biochemical reactions is carried out by enzymes, in which the essential components are mainly proteins; the receptors for hormones and other signaling molecules are also proteins. From a perusal of apropos protein literature, one gains the distinct impression that it is axiomatically assumed that experimental, aqueous solution, infinite dilution partial molar volumes of the zwitterionic ά-amino acids and peptides provide useful and important fundamental reference information related to studies of protein denaturation, hydration, intermolecular association, and other significant physical properties.

The research to be described in this chapter primarily consists of development and evaluation of a fundamentally new and innovative method for predicting high accuracy values (compared with experimental data) of the partial molar volumes of amino acids and peptides.

2. DISCUSSION

2.1 Partial Molar Volumes of Amino Acids

The aqueous solution, infinite dilution partial molar volumes of all twenty encoded amino acids have been known for some time[27, 28, 33, and 34]. They are included in a valuable key reference published in 1997 by Kharakoz et al.[16], that summarizes and extensively evaluates all of the then existing experimental data, finally tabulating a set of recommended PMV values for 24 amino acids and 13 dipeptides. A table of this data, augmented with additional contemporaneous experimental data and the analyzed results is presented. Kharakoz et al. found that the precision of the experimental data was generally excellent for amino acids, in every case actually better than 1.0 cm3/mol., about one percent of the largest values of measured amino acid PMVs. One also notes that since the 1930's [4] the PMVs for many of the compounds have been reinvestigated in several different laboratories with good agreement between results, consistently better than 0.5 cm3/mol.

2.2 Previous Correlation Studies

The initial impetus for the research described in this paper is a QSPR (Quantitative/Structure/Property/Relationship) of amino acid PMVs, which was previously carried out by Randic, Mills and Basak[26], published in 2000. They defined a new type of mathematical descriptor, a so-called "generalized topological index". The development of the index descriptor for each amino acid is carried out by fairly complex arithmetical calculations, and the full paper, cited above, should be consulted for a detailed explanation. The topological indices were then used as the independent variable parameters in a statistical QSPR study to obtain putative high quality correlations of PMVs for 16 and 17 compound subsets of the 20 natural amino acids.

One of the major presuppositions used by Dr. Herndon research group at UTEP in numerous QSAR studies carried out during the last 20 years has been that optimum descriptors for analyzing molecular structure-property and structure-activity relationships are normally simple indicator variables. For these types of organic chemical systems, effective choices of indicator variables generally include direct counts of atoms by element, often extended to make use of the atom types sorted by hybridization and/or types of substituents, the bond-types, or more complex functional groups and structural feature, all descriptors evaluated in order of increasing complexity. This additivity methodology, with the exception of the hierarchical extension is, of course, one of the prosaic, common standards for legions of successful QSPR and QSAR studies[2,31,35] . A significant advantage is that the additive parameters of this type are local structural descriptors, bearing one-to-one correspondence to the conventional representations of molecular structure. As such, they normally allow precise structural interpretations of importance in determining the value of a physical chemical or biochemical property. In addition, the design of new molecules with altered desirable structures and properties is facilitated, and the additive predictive capabilities of derived QSPR and QSAR relationships are easily tested.

In addition a previously published study in which Herndon and Radhakrishnan[13] discovered an exact mathematical algebraic equivalence between group additivity and topological index QSPR equations for the PMVs of saturated normal alkanes strongly suggested that a similar equivalence might be expected when these two different approaches were separately used to correlate the PMVs of the amino acids. However, we were interested in examining the capabilities of calculated methods to treat larger structures related to the amino acids such as polypeptides and possibly proteins. Extensions of the topological index methodology to these more complicated systems appeared to be much more difficult to implement than the simple additivity analysis. An improved level of accuracy was also desirable, and this did not seem to be attainable using the topological matrix method. Of course, the initial step in an extended investigation of the additivity approach required a detailed comparison of the two methodologies and a critical comparison of the two sets of correlation results. The differences between the two methods and the critical comparison of correlation results are outlined below.

2.3 General Topological Index Vs Group Additivity

During the last thirty years, in addition to group additivity, the topological index[13] approach has evolved as a general protocol for developing quantitative structure-property and structure-additivity relationships, with large numbers of successful studies[4,9-11,24]. One of the adduced advantages is succinctness. Numerous topological index QSPR applications make use of only one topological structure index as an independent variable, exemplified in the first ά-amino acid/ PMV study of this type[17]. This may be the reason that the number of the descriptors in the Randic, et al., topological index study[26] is emphasized to be unity, i.e., a single, numerical generalized connectivity index value for each one of the ά-amino acids. However, one should note that such emphasis is very misleading. In actuality, the number of parameters that must be optimized to establish the required individual generalized topological indices used in a final QSPR linear regression equation is a total of six: first, after several optimizations (by hand), four different weighting factors for C, N, O and S variables, used in calculating the individual numerical values of the connectivity index for each compound(Table II)[26], second, after

regression analysis at every stage of optimization, the final independent variable index coefficient and a constant term (Table IV)[26].

Other recent topological index studies on molecular volumes of the ά-amino acids have used a like number of adjustable parameters[23,24] and for the research described in this paper, this number of the optimized parameters in the topological index study turned out to be useful, since it allowed direct comparison to the results in the present paper, and to several older examples of structure-PMV amino acid studies. Contrary to the first sentence of the Randic, et al., paper, there are numerous such published additivity studies for ά-amino acids. A 1934 ACS monograph edited by Cohn and Edsal[4], gives summaries of the earlier work, including the PMV work of Traube, J published in 1899. In general, both earlier and later experimental data and various model studies are uniformly characterized by concordant high-precision precision results,[4,8] and there are also numerous extensions to peptides and proteins[7,12,14,15,19] .

In contrast, many, more recent additivity studies on peptides use a multitude of descriptors[20,21,25,30], counting structural variables like the numbers of each of the 20 different residues, the types of end-groups, the number of amide linkages, etc. In the additivity results to be described later, we attempt a close comparison with the Randic, et al., results by limiting a descriptor pool for regression to six elementary parameters, atom counts by (C, H, O, N, and S) and one additional descriptor, identified in the course of the investigation. Linear regression equations containing a constant were eschewed after finding that the value of an included constant was statistically insignificant in every attempted analysis, and also because of the ambiguity in defining a molecular structural interpretation for such a constant.

The two methodologies can be evaluated by comparison of the usual statistical parameters for correlated results (correlation coefficient, standard and mean deviations, and F-ratios). However, we were not able to formulate more stringent strategies for comparison of the two methods. This was principally because the several iterations for the non-linear hand calculations of the generalized

topological index, required before optimization in each case, precluded using cross-validation procedures.

3. METHODS AND RESULTS

3.1 PMVs of Encoded Amino Acids

The information to be analyzed and evaluated in this part of the paper consists primarily of results for the linear regression correlation of encoded amino acid experimental PMVs data, based on an atomistic additivity model. Additional data for 13 amino acids with unusual substituents or side-chain functionality[22] and ten straight-chain aliphatic alpha-omega amino acids[3] have been added to the set of amino acids recommended by Kharakoz et al.[16] to compile our amino acid PMV data set, listed in Table II. The original journal citation for each of the listed compounds was examined, and the PMV values that are given for each compound have been compared to the data in the original articles.

As far as can be ascertained, these 43 amino acids comprise the complete set of amino acids with known experimental PMVs, where the PMVs were determined at 25 ºC and several appropriate series of low concentrations. The partition of the data in the table (Table II) into a reference set comprising the twenty uncoded amino acids, used for development of parameters to correlate their experimental PMVs, and a 23 compound test set to evaluate predictions, is an obvious natural consequence of the composition of the table.

3.2 PMV Data Analysis: Model-1

The starting point for the data analysis is the PMV data listed in the second column of Table II. A single parameter, the molecular weights of the twenty coded amino acids, gives what might be considered as an acceptable correlation with the aqueous solution PMVs (correlation coefficient 0.919). Taking the numbers of atoms by element as descriptors (C, H, N, O, and S), defined as Model-1, must necessarily lead to an improved correlation because of the additional flexibility in the independent variable descriptor set. In addition, the reader may recall that the atom level is also the first and simplest

level for descriptors in our preferred hierarchical variant of QSPR methodology. The statistical results of Model-1 regression are tabulated in Table 3, and the expectation of obtaining an improved correlation is unquestionably fulfilled, generating the astonishing results illustrated in Figure 1.

Calculated PMVs and residual errors for Model-1 and the topological model results for the 17 and 16 compounds subsets studied by Randic, et al.,[26] are compared in Table IV. The 16 compound Randic, et al., results were obtained after disregarding the largest outlier (Glutamine) in the 17 compounds model. Any benefits arising from this modification are not readily apparent.

Randic, et al., in their amino acid PMV study, justified the removal of outliers with an error greater than twice the standard error, on the supposition that the ensuing higher quality regression would have greater usefulness for the data set in hand.

However, in the present study, it is preferable to retain all the experimental data, since the ultimate objective in this present research is to test the capability of the amino acid atom count parameters to actually predict accurate values for partial molar volumes of related compounds, that is, molecular systems with distinctly different structural motifs, but composed of atoms of the same five elements. Examples include noncoded amino acids, whose PMVs are given in Table II (Test Set # 1) and a sizable number of small polypeptides whose PMVs will be treated and discussed later. Additional examples (not treated in this paper) with various types of known thermodynamic properties include a few large polypeptides and proteins, and over 100 other known compounds of biological interest containing one or more of each of the specified elemental atoms.

3.3 PMV Data Analysis: Model-2

The Model-1 correlation of the amino acid data is certainly remarkable, particularly if compared to the generalized topological index results, in which a larger number of adjustable parameters is employed, applied to a fewer number of molecules. However, it should be noted that the mean deviation for Model-1 (1.09 cm3/mol), although small, is still more than twice as large as the usual reported precision of amino acid PMV experimental data. However, it is readily apparent that approximately 20% of the total

mean error for Model-1 is due to the single compound, Lysine, with a negative 4.545 cm3/mol error, approximately four times the mean error, and 2.7 times the calculated standard error (see Figure 1 and Table II).

Of course, the statistical regression results can be substantially improved by simply omitting lysine from the analysis, as was done in the Randic, et al, topological index study. However, a better, structure related approach is suggested by the fact that Lysine happens to belong to a small subgroup of the encoded ά-amino acids in which the predominating structures at neutral pH in aqueous solution do not possess the customary ά-ammonium carboxylate zwitterion. Instead, ionized side chain functional groups are partially involved, aliphatic ammonium group in the cases of Lysine, Arginine and Histidine, and the side chain carboxylate for Aspartic and Glutamic Acids.

The most simplistic way to model these side chains without exceeding six parameters is to define a single indicator-type variable; +1 for side-chain amino group structures (which thus must include Tryptophan), -1 for side-chain carboxyl group, and a value of zero otherwise. A physical interpretation of this construct could be that one is postulating that effects of anionic and cationic side chains on the measured PMVs are close to equal in magnitude and opposite in sign. The results of this atom type with side-chain parameter regression model (Model-2), tabulated in Tables V and VI, provide a practical justification for this six parameter model. There are dramatic improvements in regression statistics, including a 20 % improvement in the important F-ratio statistic, and a significant reduction in standard error of estimate from 1.6 to 1.2. The self-evident high quality of the correlation is illustrated below in Figure II.

However, the practical usefulness of a QSPR analysis doesn't really lie in the ability of a model equation to correlate data for one specific property, even for a large set of dissimilar compounds. The ultimate goals of QSAR activity and QSPR property studies are prediction of biological potencies or physical properties of yet uninvestigated compounds. Obtaining a really excellent QSPR correlation

equation, as is found for Model-2, should not end an investigation, and the correlation results should not be considered as proof that a particular methodology will have general predictive usefulness.

Thus the additivity methodology presented in this model, and the correlation equation represented by Figure II need to be evaluated for the ability to predict PMVs, rather than just correlate already known experimental values. For the additivity model, this predictive capability will be demonstrated and discussed next, but such a corresponding detailed analysis is not feasible for the topological index model. However, one comparison of results from the atom count and topological models is possible which does address the question of prediction accuracy, at least for three of the twenty encoded amino acids. The results of this comparison, in the next paragraph, will terminate both our criticisms of the generalized topological index, and also bring this Section to a germane conclusion.

Randic, et al.[26] used their best QSPR regression (six optimized parameters, N=16) to, in fact, predict values for the PMVs of Isoleucine, Threonine, and Lysine. As it was mentioned previously, these compounds were not used in developing their QSPR equations. We used Model-2 (also six optimized parameters) to develop a QSPR equation, leaving out the experimental data for the same three amino acids. The Model-2 errors are -1.22, -0.84, and -5.35 cm3/mol, respectively, to be compared with the Randic, et al. respective errors, +9.9,+10.05, and -13.7 cm3/mol. Thus, the true predictions are comparable with those obtained for correlations, very good for the atom additivity model and unacceptable large errors for the generalized topological index model.

3.4 PMV Data Analysis: Model-3 & 4 (effect of pH and 3D solvent accessible surface area)

As all the amino acids comprises of an acidic carboxylic group and a basic amine group, there is always an exchange hydrogen ion from carboxylic group to amine group which makes the amino acids to have both positive charge and negative charge. The charge of amino acids is highly dependent on pH. In order to make the comparisons more meaningful, all the amino acids are adjusted to their neutral form

and their pH at this point, which is also called as Isoelectric pH (abbreviated as PI) is calculated and introduced as a descriptor in the equation. Table IX and Figure V summarizes the statistical results of this model (Model-3). Although there is not a significant change in the statistical parameters from the model-2, the residuals were decreased to small extent.

To further examine the effect of 3-dimensional descriptors to this equation, we calculated 3-dimensional solvent accessible surface area (ASA) of amino acids present at their Isoelectric pH. Both Isoelectric pH and their ASA are calculated using JChem[18] molecular modeling software. The major 3D conformer of each amino acid at their Isoelectric pH is considered for the solvent accessible surface area calculation and water is used as the solvent (solvent radius: 1.4 Å). The statistical results are shown in Table X and the fitting graph is shown in Figure VI. The results look equally good as model-3 except the F-Ratio is decreased from 1207 to 1040. All calculated PMV values from this model regression are now within 1% to 3% of the experimental values in every case, close to the limits of precision of the experimental measurements.

4. CROSS VALIDATION FOR PMVS OF NONCODED AMINO ACIDS AND DIPEPTIDES.

4.1 Pragmatic Aspects of QSAR/QSPR Predictions

The work to be described in this Section concerns the capability of the additivity model correlation calculations to provide easily obtained, and most importantly, verifiable predictions, not just correlations of the PMV data under consideration. One of the most common of the contemporary procedures used to validate the capability to predict in medicinal chemistry QSPR and QSAR studies has been to exclude a small random sample of the compounds to be used as a test set for prediction[32]. As far as can be determined, no significant failures of this validation procedure have ever been reported in the scientific literature, particularly for the sets of similar, related, so-called congeneric, molecules normally considered in biochemical or biophysical studies. One always seems to obtain acceptable “predicted” values of the target property for the compounds of the test set. Of course, however, one has to suspend

any belief in the value of random sampling as an effective statistical analysis tool in order to interpret such results as bona fide predictions.

In general, regarding prediction, Herndon and coworkers have previously proposed that the quality and usefulness of QSAR/QSPR equations can only be truly appraised by requiring demonstrations of predictive capabilities with concrete verifiable examples. For a sizable large fraction of the studies carried out in biophysical and physical chemical systems, this requirement must be met by using the limited amounts of data in hand. In such cases, it is necessary to treat the available data so as to allow one or both of two different types of predictive results to be generated and verified: (1) predictions for compounds with biological activities or physical properties lying outside of the range of such properties for corresponding training sets, or (2) predictions for properties of compounds with nontrivial molecular structural features, not existing in the training structures. These precepts are illustrated and enforced in the two examples given below.

4.2 Verifiable Predictions of PMVs for Noncoded Amino Acids.

The sources and general accuracy of the data for the work to be described in this section was discussed in section 3.1 of the preceding Section. The data consists of experimentally obtained PMVs for the twenty encoded ά-amino acids, and corresponding experimental results for 23 additional noncoded amino acids. Thus, the number of the noncoded amino acids with experimentally measured PMV values is somewhat larger than the number of the natural ά- amino acids. In addition several types of functionality are present in this test set, not found in the coded acids, and ten of the noncoded compounds are aliphatic alpha-omega amino acids in which only one of the ten compounds has the basic structure of an alpha amino acid.

The Model-2 regression analysis of the PMVs of the twenty coded amino acids provides the optimized coefficients of the six additivity parameters, given in Table V. These coefficients are used to obtain predicted values of the PMVs for the 23 noncoded amino acids. The results, of course can then be verified by comparison with the actual experimental data. These data and the predicted PMV values are

listed in columns two and three of Table VII, respectively, with the prediction errors given in column four of the table. Each of the compounds listed in Table VII has a molecular structure with distinct individual structural differences, differing from the molecular structures of the twenty encoded amino acids. Thus, accurate predictions of the PMVs for the compounds in Table VII using the optimized Model-2 atom parameters (Table V) present a stringent validation challenge for the atomistic additive protocol.

However, as one can see from the tabulated results, the challenge is simply met by a straightforward application of an additivity procedure. In fact, the mean and standard deviations for the prediction errors are only ±1.252 and ± 1.393 cm3/mol, respectively and the correlation coefficient comparing experimental and predicted values is larger than 0.9999. In addiction, only one compound of this test set of non-coded compounds has a prediction error larger than 1 cm3/mol, a 3% error for 5-aminopentanoic acid (C5H11NO2). The upper limit of error, observed for the encoded amino acid correlation is about the same, a 2% discrepancy for Lysine (C6H14N2O2).

The results of this study are indisputable and accurate predictions of PMVs compared to actual precise experimental results. It is possible that one might think that graphs illustrating these types of results might be considered to be superfluous. Even so, such a graph illustrating the results summarized in Table VIII is presented below in Figure III. This figure simply adds graphical emphasis to the overall aspects of the extraordinary high quality of this prediction study. It is also interesting to note the distinct similarity of Figure III to the previous results presented in Figures I and II in which optimized results of regression correlations have been illustrated.

4.3 Verifiable Predictions of PMVs for Dipeptides.

The next logical step in the development of the research described in this paper is an extension of the successful amino acid methodology to a consideration of the PMVs of peptides, beginning, of course, with dipeptides. The data that will be examined consists of PMVs for the 18 dipeptides listed in Table VII. Thirteen of the dipeptides are taken from the Kharakoz compilation[16] and the other five are from

other sources[5, 9, 22]. Three of these latter compounds[9] are diketopiperazine derivatives, i.e., neutral cyclic dipeptides, for which chemical intuition projects a systematic error in predicted PMVs if calculated with the open zwitterionic amino acids parameters.

The data for PMVs of dipeptides are obviously not extensive, nor as diverse as the amino acid data. Only four of the 18 known values are for compounds that are not derivatives of glycine, and three of the 18 are derivatives of 2,5-diketopiperazine (the cyclic GlyGly peptide). Be that as it may each of the dipeptides is distinctly different from the amino acids from section 4.2, containing either one or two unionized peptide linkages, in addition to possessing an increased distance separating the normal zwitterionic charges. Thus, the calculated dipeptide PMVs are significant extrapolations from the Model-2 correlation of PMVs for the monomeric amino acids.

The over all result clearly identifies the cyclic dipeptides as outliers, and small overall decreased accuracy of the predicted PMV values is found, demonstrated in the graph shown in Figure 4. A graph of predicted versus experimental values, excluding the three cyclic dipeptides, is actually very similar to the parameterization graph (Figure II), and to the Model-2 prediction graph (Figure III), i.e., quite uninformative, just data points lying close to a straight line drawn with close to unit slope. However, such a comparison does emphasize the fact that the errors for predicted monomeric amino acids and for open chain dipeptide compounds are the same order of precision as the errors of the correlation in the investigation of the parent monomeric acids.

The large errors in the predicted values for the three cyclic dipeptides were not unexpected. These compounds are not dissimilar to the other dipeptides in possessing small cyclic structures, but the structures of the three molecules also differ by not existing as zwitterionic at the pH of neutral water. Since they are neutral organic compounds, the surrounding water structure is anticipated to be highly dissimilar to that of the normal ionic open-chain amino acids and dipeptides. Of course it is possible to mitigate a large fraction of the errors for the three cyclic dipeptides if the independent variable list for

the statistical regression analysis made use of a parameter denoting the presence or absence of the cyclic peptide structures. Similar indicator variables are commonly used in QSAR and QSPR studies and, in fact, our Model-2 parameters include just such a descriptor for the three types of amino acid side chains. The parameterization of this type of descriptor for cyclic dipeptides in the present study is not justifiable due to the absence of a set of experimental data that could be used to evaluate both correlative and, more importantly, the predictive utility of the cyclicity parameter.

5. SUMMARY AND CONCLUDING REMARKS

The water solubility of an amino acid or peptide is one of its most important physical properties, for both practical and scientific reasons. Some of the main scientific uses involve aqueous measurements of partial molar quantities such as the partial molar heat capacity or the partial molar volume, key parameters in understanding the thermodynamics of aqueous solutions. The most important of these partial molar quantities is the partial molar free energy or chemical potential. However, as explained in introduction of this paper, the easiest such property to measure precisely is the partial molar volume. This is actually a very fortunate circumstance since it can be coupled with the fact that that experimental, aqueous solution, infinite dilution partial molar volumes of amino acids and peptides provide useful and important fundamental reference information related to studies of protein denaturation, hydration, intermolecular association, and other significant physical properties.

Generally, quantitative structure/property studies (QSPR) of amino acids, peptides and even proteins have made use of the encoded ά-amino acid residues as the basic units for understanding the physical properties of such biomolecules. The main initial goal of the research for this paper was to examine the possibility that experimental partial molar volumes of amino acids might be understandable on an atomistic basis (five parameters for C, H, N, O, S) rather than twenty parameters (for the twenty different residues).

All of the most significant findings and conclusions related to the work reported in this paper are linked to the initial, completely unexpected results, strongly supporting this hypothesis. These results are

presented earlier, tabulated in Tables III, IV, V and VI, and illustrated in Figure I and Figure II. The Tables and Figures, based on multilinear regression studies, demonstrate that coded amino acid PMV data is precisely determined by the number of atoms of each type, with one additional indicator variable designating types of side chains. Some might consider the ensuing studies and results, reported in Section 4, to be even more intriguing and potentially useful. In this Section the atom parameters from the 20 amino acid correlation are used to provide accurate, verifiable predictions of the PMVs in a dipeptide data set. (See tables VII and VIII for the relevant data and the graphs in Figure III and Figure IV). Although the incorporation of pH and 3-dimensional solvent accessible surface area parameters does not change the statistics drastically, they improve the predictability to small extent.

Future research along the lines described in the present paper could include studies incorporating PMV data for additional amino acids, tripeptides, tetrapeptides, and perhaps, even larger peptides. PMVs for several larger peptides are known, and it might be possible to bracket the peptide size where conformational effects make a linear model ineffective. Extending this idea, it is possible to imagine that the difference between the calculated atomistic PMVs of a series of denatured proteins (compared to the experimental PMV values) might be a useful measure of conformational and folding properties.

6. GENERALIZATION OF METHOD

Finally, it should be noted that an atomistic model is easily applicable to nearly all types of molecular chemical systems and can be used to correlate and predict nearly all types of physicochemical properties and biological properties. In the Herndon group we believe that the atomistic hierarchical methodology for QSAR and QSPR studies provides an optimum starting point for such studies.

TABLES

Table I. Names, abbreviations and linear structures of 20 amino acids.

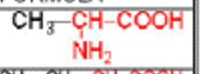
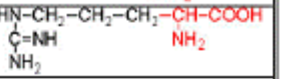
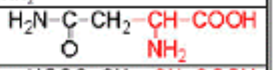
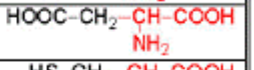
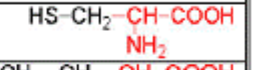
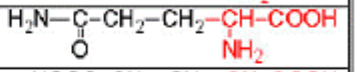
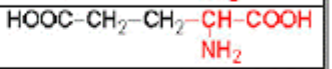
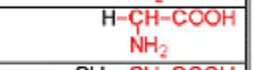
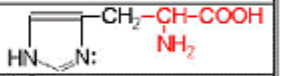
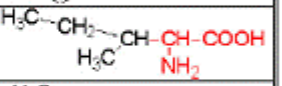
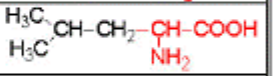
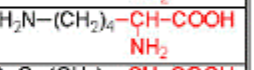
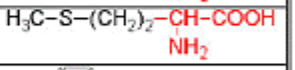
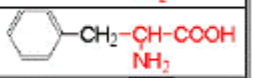
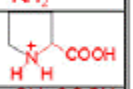
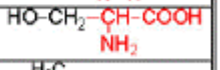
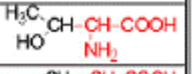
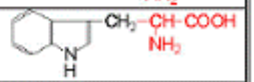
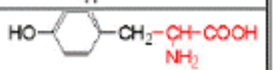
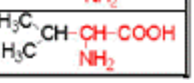
NAME	ABBREVATIONS		LINEAR STRUCTURE FORMULA
Alanine	Ala	A	
Arginine	Arg	R	
Asparagine	Asn	N	
Aspartic acid	Asp	D	
Cysteine	Cys	C	
Glutamine	Gln	Q	
Glutamic Acid	Glu	E	
Glycine	Gly	G	
Histidine	His	H	
Isoleucine	Ile	I	
Leucine	Leu	L	
Lysine	Lys	K	
Methionine	Met	M	
Phenylalanine	Phe	F	
Proline	Pro	P	
Serine	Ser	S	
Threonine	Thr	T	
Tryptophan	Trp	W	
Tyrosine	Tyr	Y	
Valine	Val	V	

Table II. Experimental partial molar volumes of amino acids (cm3/ mol).

Encoded Amino acids	PMV EXPERIMENTAL	Noncoded Amino acids Test set # 1	PMV EXPERIMENTAL
Glycine	43.24	Hydroxyproline	84.2
Alanine	60.44	Norleucine	107.75
Valine	90.79	Norvaline	91.7
Leucine	107.66	B-alanine	58.5
Isoleucine	105.6	Isoserine	59.07
Methionine	105.36	Allothreonine	76.87
Proline	82.2	beta-phenylserine	124.69
Phenylalanine	121.8	S-ethylcystein	104.2
Tryptophan	143.9	Ethionine	119.58
Serine	60.69	m-tyrosine	123.11
Threonine	76.86	3-aminotyrosine	129.61
Asparagine	77.3	3,4-dihyroxyphenylalanine	125.76
Glutamine	93.9	Citrulline	115.94
Tyrosine	123.7	alpha-aminobutyric acid	75.56
Cysteine	73.45	beta-aminobutyric acid	76.21
Lysine	108.7	gamma-aminobutyric acid	73.23
Arginine	123.8	5-aminopentanoic acid	87.65
Histidine	98.8	6-aminohexanoic acid	104.09
Aspartic acid	74.3	7-aminoheptanoic acid	120
Glutamic acid	89.5	8-aminooctanoic acid	136.03
		9-aminononanoic acid	151.3
		10-aminodecanoic acid	167.3
		11-aminoundecanoic acid	183

Table III. Model-1 multiple linear regression analysis (PMV vs. C, H, O, N and S).

DEP VAR: PMV N: 20 MULTIPLE R: 0.998 SQUARED MULTIPLE R: 0.996
ADJUSTED SQUARED MULTIPLE R: .995 STANDARD ERROR OF ESTIMATE: 1.692

VARIABLE	COEFFICIENT	STD ERROR	STD COEF	TOLERANCE	T	P(2 TAIL)
CARBON	7.783	0.238	1.830	0.076	32.673	0.000
HYDROGEN	3.736	0.181	1.546	0.042	20.669	0.000
NITROGEN	4.220	0.513	0.285	0.198	8.230	0.000
OXYGEN	2.904	0.364	0.300	0.168	7.986	0.000
SULPHUR	14.624	1.282	0.188	0.871	11.404	0.000

ANALYSIS OF VARIANCE

SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
REGRESSION	12028.416	4	3007.104	1050.095	0.000
RESIDUAL	42.955	15	2.864		

Table IV. Comparison of results from Model-1 and Generalized Topological Index Models (cm3/ mol).

Amino acid	PMV EXP	Model- 1 Calculated	Model-1 Residuals	Topolog-17 residuals	Topolog-16 residuals
Glycine	43.24	44.272	-1.032	-0.24	-0.13
Alanine	60.44	59.527	+0.913	-0.46	-0.70
Valine	90.79	90.035	+0.755	-3.48	-4.39
Leucine	107.66	105.290	+2.370	-1.73	-2.94
Isoleucine[a]	105.60	105.290	+0.310	No data	No data
Methionine	105.36	104.659	+0.701	+4.87	+3.84
Proline	82.20	82.564	-0.364	+1.72	+1.08
Phenylalanine	121.80	121.169	+0.631	+9.42	+8.15
Tryptophan	143.90	144.690	-0.790	-0.70	-2.62
Serine	60.69	62.431	-1.741	+3.99	+3.84
Threonine[a]	76.86	77.685	-0.825	No data	No data
Asparagine	77.30	78.169	-0.869	-5.76	-6.45
Glutamine[b]	93.90	93.424	+0.476	-13.65	No data
Tyrosine	123.70	124.073	-0.373	-0.47	-1.98
Cysteine	73.45	74.151	-0.701	-6.56	-7.18
Lysine[a]	108.70	113.245	-4.545	No data	No data
Arginine	123.80	121.684	+2.116	+8.51	+7.11
Histidine	98.80	98.787	+0.013	-1.90	-2.94
Aspartic acid	74.30	73.118	+1.182	+5.28	+4.89
Glutamic acid	89.50	88.372	+1.128	+1.17	+0.45

[a]Not used in Randic, et al.[26], [b]defined as an outliner in Randic, et al.[26] to be ignored.

Table V. Model-2 Multiple Linear Regression Analysis (PMV vs. C,H,O,N, S, SIDE-CHN).

DEP VAR: PMV N: 20 MULTIPLE R: 0.999 SQUARED MULTIPLE R: 0.998
ADJUSTED SQUARED MULTIPLE R: .997 STANDARD ERROR OF ESTIMATE: 1.413

VARIABLE	COEFFICIENT	STD ERROR	STD COEF	TOLERANCE	T	P(2 TAIL)
CARBON	7.867	0.201	1.849	0.074	39.106	0.000
HYDROGEN	3.722	0.151	1.540	0.042	24.654	0.000
NITROGEN	5.514	0.637	0.372	0.089	8.656	0.000
OXYGEN	2.219	0.393	0.229	0.100	5.645	0.000
SULPHUR	14.492	1.072	0.187	0.869	13.524	0.000
SCHAIN	-2.541	0.926	-0.065	0.291	-2.743	0.016

ANALYSIS OF VARIANCE

SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
REGRESSION	12043.434	5	2408.687	1207.059	0.000
RESIDUAL	27.937	14	1.996		

Table IX. Model-3 Multiple Linear Regression Analysis (PMV vs. C,H,O,N, S, SIDE-CHN, PI).

DEP VAR: PMV N: 20 MULTIPLE R: 0.999 SQUARED MULTIPLE R: 0.998
ADJUSTED SQUARED MULTIPLE R: .997 STANDARD ERROR OF ESTIMATE: 1.281

VARIABLE	COEFFICIENT	STD ERROR	STD COEF	TOLERANCE	T	P(2 TAIL)
CARBON	7.730	0.195	1.817	0.065	39.730	0.000
HYDROGEN	4.014	0.200	1.661	0.020	20.091	0.000
NITROGEN	6.058	0.638	0.409	0.073	9.494	0.000
OXYGEN	2.245	0.357	0.232	0.100	6.296	0.000
SULPHUR	14.780	0.982	0.190	0.850	15.049	0.000
SCHAIN	-2.316	0.847	-0.060	0.286	-2.733	0.017
PI	-0.509	0.253	-0.133	0.031	-2.007	0.066

ANALYSIS OF VARIANCE

SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
REGRESSION	12050.043	6	2008.341	1224.163	0.000
RESIDUAL	21.328	13	1.641		

Table X. Model-2 Multiple Linear Regression Analysis (PMV vs. C,H,O,N, S, SIDE-CHN, PI, ASA).

DEP VAR: PMV N: 20 MULTIPLE R: 0.999 SQUARED MULTIPLE R: 0.998
ADJUSTED SQUARED MULTIPLE R: .997 STANDARD ERROR OF ESTIMATE: 1.286

VARIABLE	COEFFICIENT	STD ERROR	STD COEF	TOLERANCE	T	P(2 TAIL)
CARBON	7.448	0.357	1.751	0.019	20.842	0.000
HYDROGEN	3.768	0.328	1.559	0.007	11.490	0.000
NITROGEN	5.992	0.645	0.404	0.072	9.295	0.000
OXYGEN	1.670	0.707	0.173	0.026	2.363	0.036
SULPHUR	13.350	1.806	0.172	0.254	7.391	0.000
SCHAIN	-2.372	0.853	-0.061	0.284	-2.781	0.017
PI	-0.617	0.279	-0.161	0.026	-2.211	0.047
ASA	0.023	0.025	0.259	0.002	0.945	0.363

ANALYSIS OF VARIANCE

SOURCE	SUM-OF-SQUARES	DF	MEAN-SQUARE	F-RATIO	P
REGRESSION	12051.521	7	1721.646	1040.821	0.000
RESIDUAL	19.849	12	1.654		

Table VI. Comparisons of Experimental PMVs vs. calculated PMVs of Model-1, Model-2, Model-3 and Model-4

Amino acids	PMV EXP	Model-1 predicted	Model-1 residuals	Model-2 predicted	Model-2 residuals	Model-3 predicted	Model-3 residuals	Model-4 predicted	Model-4 residuals
alanine	60.44	59.527	0.913	59.603	0.837	58.796	1.644	59.646	0.794
arginine	123.8	121.684	2.116	123.255	0.545	123.506	0.294	123.364	0.436
asparagine	77.3	78.169	-0.869	76.383	0.917	76.919	0.381	76.911	0.389
asparticacid	74.3	73.118	1.182	74.449	-0.149	74.635	-0.335	74.685	-0.385
cysteine	73.45	74.151	-0.701	74.095	-0.645	73.734	-0.284	73.423	0.027
glutamine	93.9	93.424	0.476	91.693	2.207	92.412	1.488	92.431	1.469
glutamicacid	89.5	88.372	1.128	89.759	-0.259	90.708	-1.208	90.691	-1.191
glycine	43.24	44.272	-1.032	44.293	-1.053	43.139	0.101	42.548	0.692
histidine	98.8	98.787	0.013	99.133	-0.333	98.778	0.022	98.997	-0.197
isoleucine	105.6	105.29	0.31	105.532	0.068	105.962	-0.362	105.814	-0.214
leucine	107.66	105.29	2.37	105.532	2.128	105.988	1.672	105.925	1.735
lysine	108.7	113.245	-4.545	112.227	-3.527	111.872	-3.172	111.866	-3.166
methionine	105.36	104.659	0.701	104.715	0.645	105.076	0.284	105.387	-0.027
phenyalanine	121.8	121.169	0.631	121.689	0.111	121.243	0.557	121.392	0.408
proline	82.2	82.564	-0.364	82.779	-0.579	81.709	0.491	81.6	0.6
serine	60.69	62.431	-1.741	61.822	-1.132	61.188	-0.498	61.355	-0.665
threonine	76.86	77.685	-0.825	77.132	-0.272	76.992	-0.132	76.874	-0.014
tryptophan	143.9	144.69	-0.79	144.116	-0.216	144.455	-0.555	144.407	-0.507
tyrosine	123.7	124.073	-0.373	123.908	-0.208	123.718	-0.018	123.529	0.171
valine	90.79	90.035	0.755	90.223	0.567	90.225	0.565	90.242	0.548

Table VII. Model-2 predictions of PMVs for 23 non-coded amino acids (cm3/ mol).

Noncoded Amino acids	PMV Experimental	PMV Calculated	Calc errors
Hydroxyproline	84.2	83.846	0.354
Norleucine	107.75	106.043	1.707
Norvaline	91.7	90.4	1.3
B-alanine	58.5	59.114	-0.614
Isoserine	59.07	60.169	-1.099
Allothreonine	76.87	75.812	1.058
beta-phenylserine	124.69	123.59	1.1
S-ethylcystein	104.2	105.048	-0.848
Ethionine	119.58	120.691	-1.111
m-tyrosine	123.11	123.59	-0.48
3-aminotyrosine	129.61	128.454	1.156
3,4-dihyroxyphenylalanine	125.76	124.644	1.116
Citrulline	115.94	114.429	1.511
alpha-aminobutyric acid	75.56	74.757	0.803
beta-aminobutyric acid	76.21	74.757	1.453
gamma-aminobutyric acid	73.23	74.757	-1.527
5-aminopentanoic acid	87.65	90.4	-2.75
6-aminohexanoic acid	104.09	106.043	-1.953
7-aminoheptanoic acid	120	121.686	-1.686
8-aminooctanoic acid	136.03	137.329	-1.299
9-aminononanoic acid	151.3	152.972	-1.672
10-aminodecanoic acid	167.3	168.615	-1.315
11-aminoundecanoic acid	183	184.258	-1.258

Table VIII. Experimental and prediction data of PMVs of Dipeptides (cm3/ mol).

Test Set # 2 Dipeptides	PMV experimental	PMV Model-2 Predicted	PMV Error
GlyGly	76.30[a]	78.279	-1.979
GlyLeu	139.30[a]	140.851	-1.551
LeuGly	143.70[a]	140.851	2.849
GlyAla	92.80[a]	93.922	-1.122
AlaGly	94.80[a]	93.922	0.878
GlyPhe	155.54[a]	157.343	-1.803
PheGly	160.30[d]	157.343	2.657
GlyVal	122.30[a]	125.208	-2.908
ValGly	126.00[a]	125.208	0.792
GlySer	92.93[a]	94.977	-2.047
GlyThr	108.50[a]	110.620	-2.120
GlyAsn	110.11[a]	113.087	-2.977
AlaAla	110.60[a]	109.565	1.035
SerSer	111.80[a]	111.674	0.126
Gly(α-aminobutane)	107.81[b]	109.565	-1.755
DIKETOPIPERAZINES			
c-GlyGly	76.85[c]	69.616	7.234
c-AlaAla	112.58[c]	100.902	11.678
c-SarSar	113.36[c]	100.902	12.458

[a]Kharakoz, ref. 16, [b]Mishra and Ahluwalia, ref. 22, [c] Hakin, et al., ref. 9, and [d] Greenstein and Wyman, ref. 5

FIGURES

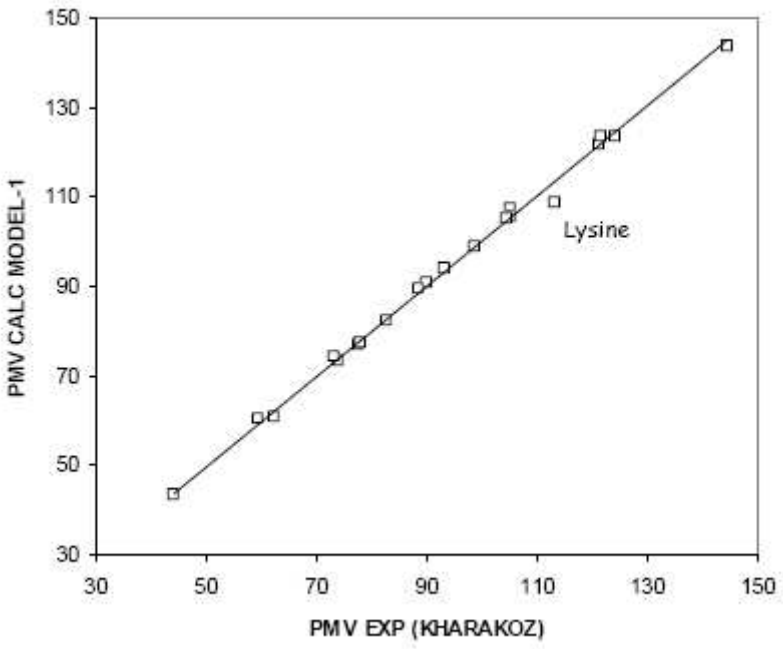


Figure I. Model-1 PMVs for encoded amino acids: calculated versus experimental (cm3/ mol).

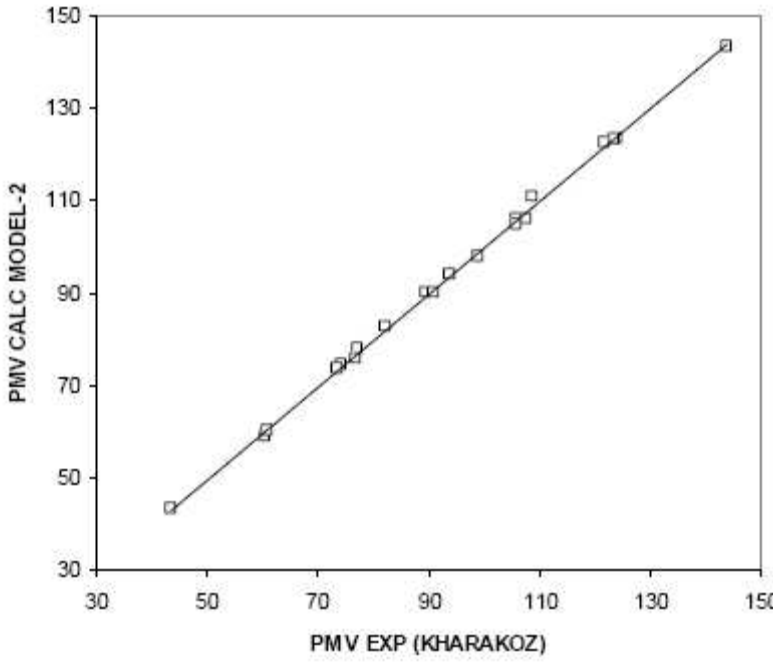


Figure II. Model-2 linear regression: calculated PMV values (abscissa) vs. experimental values (ordinate) (cm3/ mol).

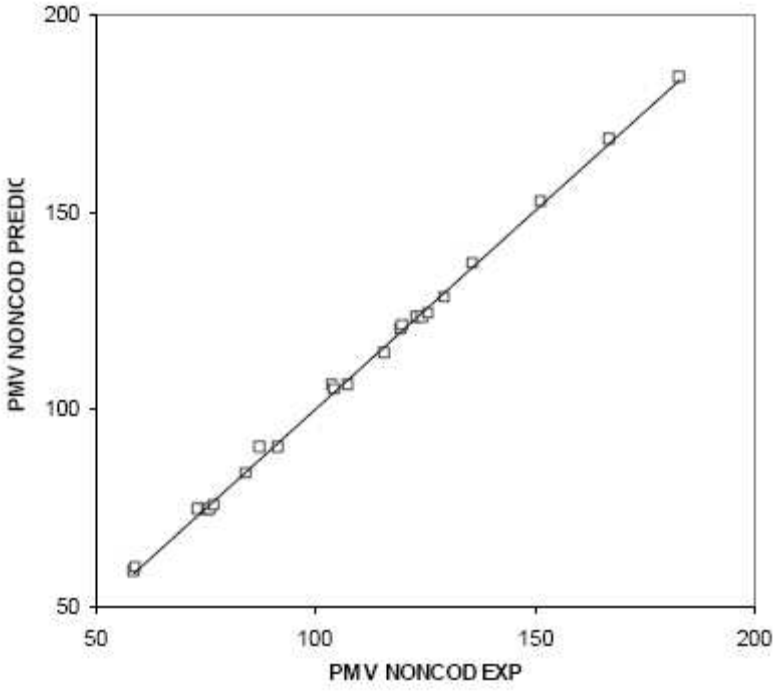


Figure III. Predicted vs. Experimental PMVs for noncoded amino acids. The data points are superposed on a single drawn with unit slope(cm3/ mol).

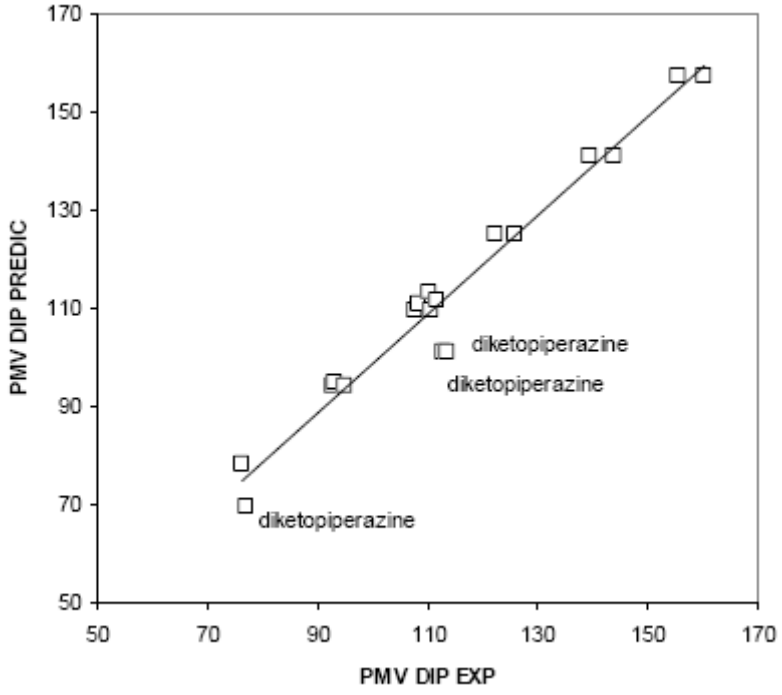


Figure IV. PMVs for Dipeptides: predicted data versus experimental data (cm3/ mol).

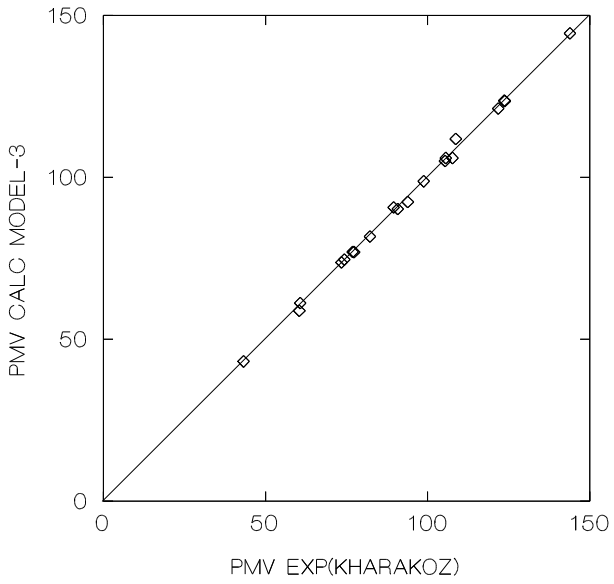


Figure V. Model-3 linear regression: calculated PMV values (abscissa) vs. experimental values (ordinate) (cm3/ mol).

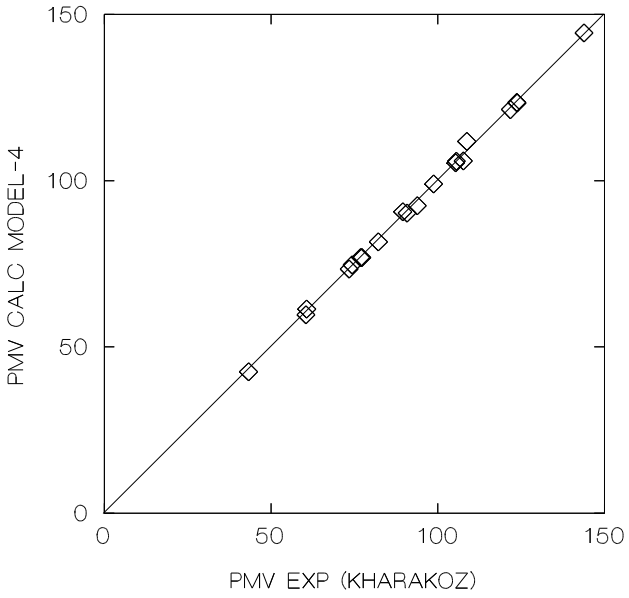


Figure VI. Model-4 linear regression: calculated PMV values (abscissa) vs. experimental values (ordinate) (cm3/ mol).

REFERENCES

1. Barret, G. C.; Elmore D.T.; Introduction. In *Amino Acids and Peptides.* First ed.; Press syndicate of University of Cambridge, Cambridge, UK, 1998: pp 1-3.

2. Zefirov, N. S.; Palyulin, V. A. QSAR for boiling points of "small" sulfides. Are the "high-quality-structure-property-activity regressions" the real high quality QSAR models? *J. Chem. Inf. Comput. Sci.* **2001,** 41, 1022-1027.

3. Chalikian, T. V.; Sarvazyan, A. P.; Breslauer, K, J. Partial molar volumes, expansibilities, and comprenssibilities of alpha.,omega.,.-aminocarboxylic acids in aqueous solutions between 18 and 55 ºC. *J. Phys. Chem.* **1993,** 97, 13017-13026.

4. Cohn, E.J; McMeekin, T. L; Edsall, J. T; Weare J. H; Studies in the Physical Chemistry of Amino Acids, Peptides and Related Substances. II. The Solubility of α-Amino Acids in Water and in Alcohol—Water Mixtures, *J. Am. Chem. Soc.* **1934,** 56, pp 2270–2282

5. Greenstein, J. P.; Wyman, J.; Cohn, E. J. Studies of multivalent amino acids and peptides. III. The dielectric constants and electrostriction of the solvent in solutions of tetrapoles. *J. Am. Chem. Soc.* **1935,** 57, 637-642.

6. Rum, G; Herndon, W. C; Three- Dimensional Topological Descriptors and Similarity of Molecular Structures: Binding Affinities of Corticosteroids. In *QSAR and Molecular Modeling: Concepts, Computational Tools and Biological Application;* Proceedings of the 10th European Symposium on Structure-Activity Relationships, QSAR and Molecular Modeling, Barcelona, Spain, Sep 4-9, 1994.

7. Häckel, M.; Hinz, H.; Hedwig, G. R. Partial molar volumes of proteins: amino acid side-chain contributions derived from the partial molar volumes of some tripeptides over the temperature range 10-90 ºC. *Biophysical Chemistry.* **1999,** 82, 35-50.

8. Hakin, A. W.; Hedwig, G. R. Group additivity calculations of the thermodynamic properties of unfolded proteins in aqueous solution: a critical comparison of peptide-based and HFK models. *Biophysical Chemistry.* **2001,** 89, 253-264.

9. Hakin, A. W.; Kowalchuck, M.G.; Liu, J. L.; Marriott, R.A. Thermodynamics of protein model compounds: apparent and partial molar heat capacities and volumes of several cyclic dipeptides in water. *Journal of Solution Chemistry.* **2000,** 29, 131-151.

10. Hall, Lowell H.; Dailey, Robert S.; Kier, Lemont B. Design of molecules from quantitative structure-activity relationship models. 3. Role of higher order path counts: path 3. *Journal of Chemical Information and Computer Sciences.* **1993,** 33, 598-603.

11. Hall, Lowell H.; Kier, Lemont B. Determination of topological equivalence in molecular graphs from the topological state. Quantitative Structure-Activity Relationships. 1990, 9, 115-131

12. Harano, Y., Imai, T. & Kovalenko, A., Kinoshita, M., Hirata, F. Theoretical study for partial molar volume of amino acids and polypeptides by the three-dimensional reference site model. *Journal of Chemical Physics.* **2001,** 114, 9506-9511.

13. Herndon, W. C.; Radhakrishnan, T. P.; Zivkovic, T. P. Characteristic and matching polynomials of chemical graphs. *Chemical Physics Letters.* **1998,** 152, 233-238.

14. Imai, T.; Kinoshita, M.; Hirata, F. Theoretical study for partial molar volume of amino acids in aqueous solution: Implication of ideal fluctuation volume. *Journal of Chemical Physics*. **2000,** 112, 9469-9478.

15. Amend, J. P.; Helgeson, H. C. Calculation of the standard molal thermodynamic properties of aqueous biomolecules at elevated temperatures and pressures II. Unfolded proteins. *Biophysical Chemistry.* **2000,** 84, 105-136.

16. Kharakoz, D.P. Partial volumes and compressibilities of extended polypeptide chains in aqueous solution: additivity scheme and implication of protein unfolding at normal and high pressure. *Biochemistry* **1997,** 36, 10276-10285.

17. Kier, L. B.; Hall, L. H. Deviation and significance of valence molecular connectivity. *J. Pharm. Sci.* **1981,** 70, 583-589.

18. JChem, version 5.2, ChemAxon: Budapest, Hungary, 2009.

19. Makhatadze, G. I.; Medvedkin, V. N.; Privalov, P. L. Partial molar volumes of polypeptides and their constituent groups in aqueous solution over the broad temperature range. *Biopolymers.* **1990,** 30, 1001-1010.

20. Matta, C. F.; Bader, R. F. W. Atoms-in-molecules study of the genetically encoded amino acids. II. Computational study of molecular geometries. *Proteins: Structure, Function and Genetics.* **2002,** 48, 519-538.

21. Matta, C. F.; Bader, R. F. W. Atoms-in-molecules study of the genetically encoded amino acids III. Bond and atomic properties and their correlations with experiment including mutation-induced changes in protein stability and genetic coding. *Proteins: Structure, Function and Genetics.* **2003,** 52, 360-399.

22. Mishra, A. K.; Ahluwalia, J. C. Apparent molal volumes of amino acids, N-acetylamino acids, and peptides in aqueous solutions. *J. Phys. Chem.* **1984**, 88, 86-92.

23. Pogliani, L. Molecular connectivity model for determination of physicochemical properties of alpha-amino acids. *J. Phys. Chem.* **1993,** 97, 6731-6736.

24. Pogliani, L. Modeling with special descriptors derived from a medium-sized set of connectivity indices. *J. Phys. Chem.* **1996,** 100, 18065 -18077

25. Popelier, P. L. A.; Aicken, F. M. (2003) Atomic properties of selected biomolecules: quantum topological atom types of hydrogen, oxygen, nitrogen and sulfur occurring in natural amino acids and their derivatives. *Chem. Eur. J.* **2003,** 9, 1207-1216.

26. Randic, M.; Mills, D; Basak, S. C. On characterization of physical properties of amino acids. *Int. J. Quant. Chem*. **2000,** 80, 1199-1209.

27. Rao, M. V. R.; Atreyi, M.; Rajeswari, M. R. Partial molar volumes of ά-amino acids with ionogenic side chains in water. *J. Phys. Chem.* **1984,** 88, 3129-3131.

28. Rellick, L. M.; Beckel, W.J. Comparison of van der Waals and semiempirical calculations of the molecular volumes of small molecules and proteins. *Biopolymers.* **1997,** 42, 191-202.

29. Shahidi, F. A.; Farrell, P. G. Partial Molar Volumes of Some ά-aminocarboxylic Acids in Water. *J.Chem. Soc. Faraday Trans. I.* **1981,** 77, 963-968.

30. Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J. Med. Chem.* **1998,** 41, 2481-2491.

31. Sotomatsu-Niwa, T.; Ogino, A. Evaluation of the hydrophobic parameters of the amino acid chains of peptides and their application in QSAR and conformational studies. *Journal of Molecular Structure.* **1997,** 392, 43-54.

32. Tropsha, A.; Gramatica, P.; Gombar, V. K. (2003) The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* **2003**, 22, 69-77.

33. Yasuda, Y.; Tochio, N.; Sakurai, M.; Nitta, K. Partial molar volumes and isentropic comprenssibilities of amino acids in dilute aqueous solutions. *J. Chem. Eng. Data.* **1998,** 43, 205-214.

34. Wang, J.; Yan, Z.; Zhuo, K.; Lu, J. Partial molar volumes of some ά-amino acids in aqueous sodium acetate solutions at 308.15 K. *Biophysical Chemistry*, **1999,** 80, 179-188.

35. Zefirov, N. S.; Palyulin, V. A. Fragmental Approach in QSPR. *J. Chem. Inf. Comput. Sci*. **2002,** 42, 1112-1122.

COMPUTER AIDED-DRUG DESIGN METHODS

Docking of curcumin in to beta-lactoglobulin and human serum albumin

MOLECULAR DOCKING

Molecular modeling is a powerful tool in drug design which generously assists in investigating, interpreting, explaining and identifying of molecular properties using n-dimensional molecular structures. Docking is frequently used to predict the binding orientation of small molecule drug candidates to their protein targets in order to in turn predict the affinity and activity of the small molecule. Hence docking plays an important role in the rational design of drugs.[2] Given the biological and pharmaceutical significance of molecular docking, considerable efforts have been directed towards improving the methods used to predict docking. The focus of molecular docking is to computationally stimulate the molecular recognition process. The aim of molecular docking is to achieve an optimized conformation for both the protein and ligand and relative orientation between protein and ligand such that the free energy of the overall system is minimized.

Molecular docking can be thought as a *"lock-and-key"* problem, where one is interested in finding the correct relative orientation of the *"key"* (ligand) which will open up the *"lock"* (protein). Molecular docking may be defined as an optimization problem, which would describe the "best-fit" orientation of a ligand that binds to a particular protein of interest. During the course of the process, the ligand and the protein adjust their conformation to achieve an overall "best-fit" and this kind of conformational adjustments resulting in the overall binding is referred to as "induced-fit".[4]

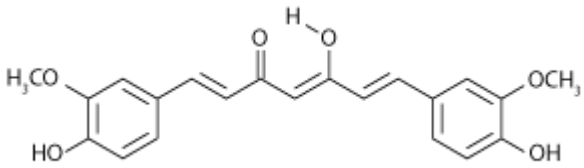
CURCUMIN AND IT'S BIOLOGICAL EFFECTS

Curcumin is known for its antitumor,[2][3] antioxidant, antiarthritic, anti-amyloid, anti-ischemic[4]

and anti-inflammatory properties. The low bioavailability of curcumin in both humans and animals has raised several concerns that this may limit its clinical impact [15]. Curcumin associates with serum albumin by hydrophobic interactions and transported to appropriate cells and elicits its pharmacological actions [30]. As the bioavailability of curcumin is low, there is a need to develop new similar molecules which have same or more biological activity and more bioavailability. In this project I would like to use AutoDock to quantify possible interactions between small molecules (curcumin and its derivatives) and protein (human serum albumin).

For the last few decades, extensive work has been carried out to establish the biological activities and pharmacological actions of curcumin. Curcumin is known for its antitumor,[2][3] antioxidant, antiarthritic, anti-amyloid, anti-ischemic[4] and anti-inflammatory properties.[5] Anti-inflammatory properties may be due to inhibition of ecosinoids biosynthesis.[6] In addition it may be effective in treating malaria, prevention of cervical cancer, and may interfere with the replication of the HIV virus.[7] In HIV, it appears to act by interfering with P300/CREB-binding protein (CBP).It is also hepatoprotective.[8] A 2008 study at Michigan State University showed that low concentrations of curcumin interfere with Herpes simplex virus-1 (HSV-1) replication.[9] The same study showed that curcumin inhibited the recruitment of RNA polymerase II to viral DNA, thus inhibiting the transcription of the viral DNA.[9] This effect was shown to be independent of effect on histone acetyltransferase activities of p300/CBP.[9] A previous (1999) study performed at University of Cincinnati indicated that curcumin is significantly associated with protection from infection by HSV-2 in animal models of intravaginal infections.[10] Curcumin acts as a free radical scavenger and antioxidant, inhibiting lipid peroxidation [11] and oxidative DNA damage. Curcuminoids induce glutathione S-transferase and are potent inhibitors of cytochrome P450. Its anticancer effects stem from its ability to induce apoptosis in cancer cells without cytotoxic effects on healthy cells. Curcumin can interfere with the activity of the transcription factor NF-κB, which has been linked to a number of inflammatory diseases such as cancer.[12] Indeed, when 0.2% curcumin is added to diet given to rats or mice previously given a

carcinogen, it significantly reduces colon carcinogenesis (Data from sixteen scientific articles reported in the Chemoprevention Database). A 2007 report indicates that curcumin may suppress MDM2, an oncogene involved in mechanisms of malignant tumor formation.[13]. A 2004 UCLA-Veterans Affairs study involving genetically altered mice suggests that curcumin might inhibit the accumulation of destructive beta-amyloid in the brains of Alzheimer's disease patients and also break up existing plaques associated with the disease.[14]



Structure of curcumin

ROLE OF CURCUMIN IN REDUCING NITROSATIVE STRESS

Curcumin, a phenolic compound is a natural antioxidant known to possess therapeutic properties and has been reported to scavenge free radicals [36-37]. It has already been used clinically and is approved by the FDA as a safe food additive [1].

Protein disulfide isomerase (PDI), **[Fig.1A]** the chief endoplasmic reticulum (ER) resident oxidoreductase chaperone that catalyzes maturation of disulfide-bond-containing proteins is involved in the pathogenesis of both Parkinson's (PD) and Alzheimer's (AD) diseases. S-nitrosylation of PDI cysteines due to nitrosative stress is associated with cytosolic debris accumulation and Lewy-body aggregates in PD and AD brains. Polyphenolic phytochemical, curcumin can rescue PDI from becoming S-nitrosylated and maintain its catalytic function under conditions mimicking nitrosative stress by forming stable NOx adducts **[38]**. Furthermore, curcumin intervenes to prevent the formation of PDI-resistant polymeric misfolded protein forms that accumulate upon exposure to oxidative stress. Curcumin can serve as lead-candidate prophylactics for reactive oxygen species induced chaperone

damage, protein misfolding and neurodegenerative disease; importantly, they can play a vital role in sustaining traffic along the ER's secretory pathway by preserving functional integrity of PDI.**[38]**

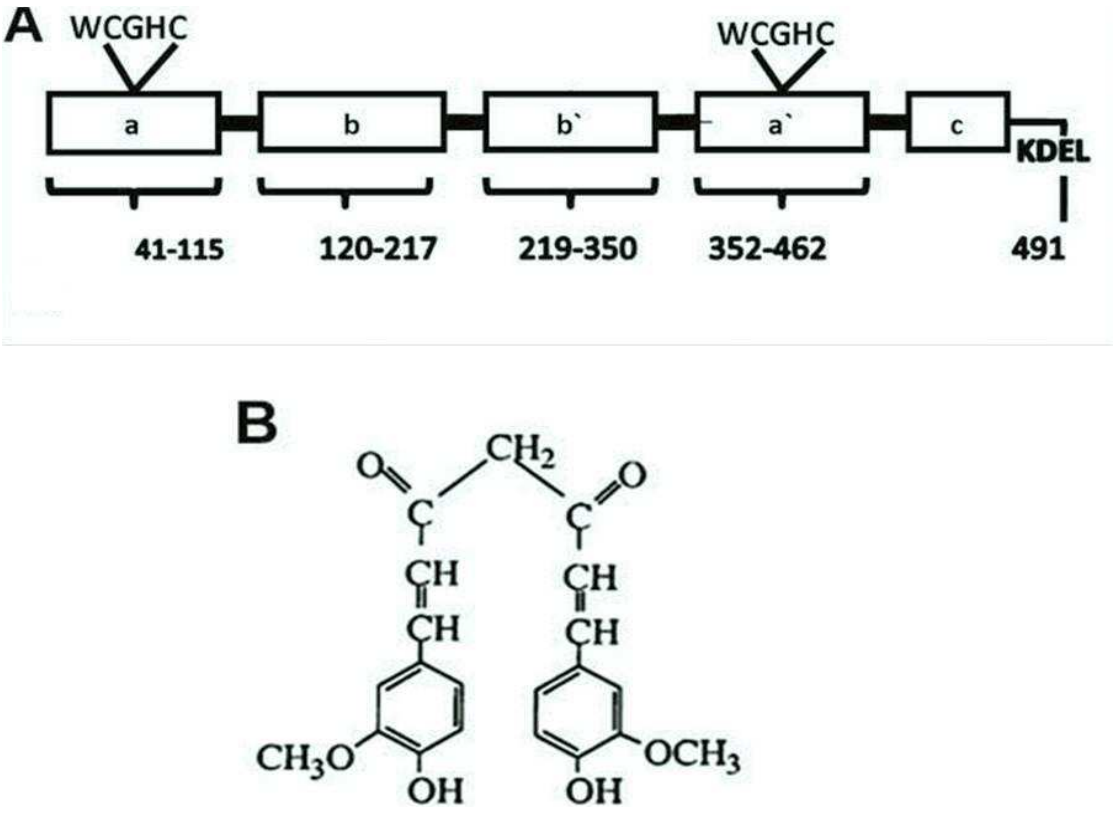


Fig. 1. (A) Schematic of protein disulfide isomerase (PDI) and **(B)** curcumin.

CURCUMIN AND IT'S BIOAVAILABILITY

The poor oral absorption of curcumin in both humans and animals has raised several concerns that this may limit its clinical impact [15]Curcumin is a biphenolic compound with hydroxyl groups at the *ortho*-position on the two aromatic rings that are connected by a β-diketone bridge, containing two double bonds (dienone), which can undergo Michael addition, critical for some of the effects of

curcumin [16]but contributing to chemical instability in aqueous solution [17] . Pharmacokinetic studies of curcumin demonstrate extensive intestinal sulfation and glucuronidation [16], [17], [19], [21]. Typically, clinical trials show negligible unconjugated curcumin plasma levels with oral dosing [22] , leading to the suggestion that in vivo efficacy may come from a more bioavailable and/or potent metabolite [14]. Potential metabolites, some of which may be active, include tetrahydrocurcumin (TC), hexahydrocurcumin, hexahydrocurcuminol, vanillin, vanillic acid, and ferulic acid. However, detectable levels of these metabolites in active unconjugated forms after administration of the parent compound have not been reported, presumably due to their low concentrations, or vulnerability to sulfation, glucuronidation, or hydrolysis.

BETA-LACTOGLOBULIN AS VEHICLE FOR IMPROVING CURCUMIN BIOAVAILABILITY

Beta-lactoglobulin also abbreviated as 'β-Lg' (18.4 kDa, 162 amino acid residues), the major whey protein, belongs to the lipocalin protein-family [39]many of whose members are known to bind small hydrophobic molecules [17]. It has recently been shown to present some antioxidant activity, apparently thanks to its free thiol group [40]. Beta-Lg folds up into an eight-stranded, antiparallel β-barrel with a three-turn α-helix on the outer surface and a ninth β-strand flanking the first strand[41]. At physiological pH β-Lg is mostly found as dimers, and at pH values below 3.5 and above 7.5 the protein tends to be monomeric [42]. The solvent accessible-conical β-barrel, or calyx, forms the main ligand binding site, although there are indications for the existence of a second ligand binding site in a crevice near the α-helix on the external surface of the β-barrel [43]. A possible third binding site was suggested to be located at the dimer interface [44]. Since the finding of the retinol–β-Lg complex [45] , many other hydrophobic ligands have been observed to bind to β-Lg, such as retinoic acid [46][47] , cholesterol [48], vitamin D [49][50] and [51] , and various aromatic compounds [52] and fatty[47] [53] and [54]. Our hypothesis is to use betalactoglobulin as a vehicle for polyphenolic compounds to improve their bioavailability. This thesis explains the interaction of betalactoglobulin interaction with different polyphenolic compounds.

DOCKING OF CURCUMIN AND ITS ANALOGS IN TO BETA-LACTOGLOBULIN

1. Dataset

Curcumin analogs were searched from Zinc database (http://zinc.docking.org). 6457 compounds similar to curcumin were selected from the database and obtained as an sdf file. Later Ligprep was used to convert the sdf file in to 3-dimensional structures. 2D structures of the compound library selected for this study are listed in Appendix B.

2. Protein preparation

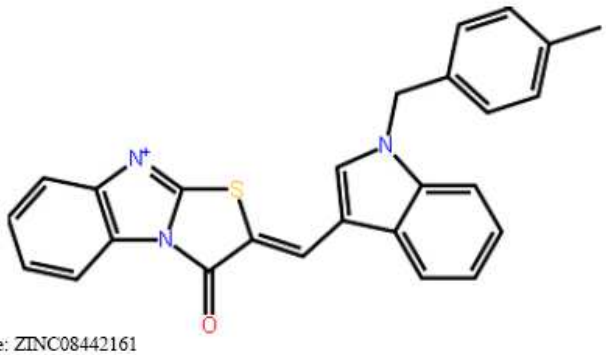
PDB file of Betalactoglobulin (PDB ID: 2Q2M) was obtained from Protein Data bank (www.rcsb.org). Betalactoglobulin was prepared using protein preparation wizard. We preprocessed and analyzed the protein structure using default settings, generated Het states, optimized H-bond assignment using sample water orientations and minimized using OPLS2005 force field. Glide (molecular docking program from Schrodinger, Inc) was used to perform virtual screening of Curcumin analogs in to betalactoglobulin.

3. Docking and virtual screening

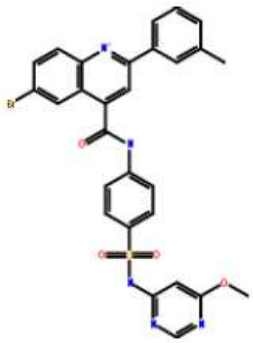
The present work describes the identification of the 4 potent nitrosative stress inhibitors that are structurally related to curcumin. In this approach, we have taken advantage of the availability of the crystal structure of betalactoglobulin and have used computational docking to screen libraries of commercially available compounds for predicted binding to betalactoglobulin. Experimental tests of a limited selection of candidate compounds verified four potent nitrosative stress inhibitors.

Virtual screening of 6457 curcumin like compounds to betalactoglobulin, followed by visual inspection of the docked poses of 50 of the best scoring ligands, has led to the identification of ten potent nitrosative stress inhibitors (nine superimposed structures are illustrated schematically and in their predicted binding conformation in Figure y). Ten compounds were selected for experimental testing based on both their predicted affinity (i.e., their Docking Scores) and the predicted structures of their betalactoglobulin-bound complexes (Figure 2), however only nine of them were available. Freshly prepared stock solutions of tetranitromethane (TNM) in acetonitrile is added to these nine compounds to check their ability to undergo nitration followed by evaluation using mass spectroscopy. Only four out of nine showed ability to undergo nitration and these results were discussed in the next section. The docking scores of the ten compounds are given in the following table.

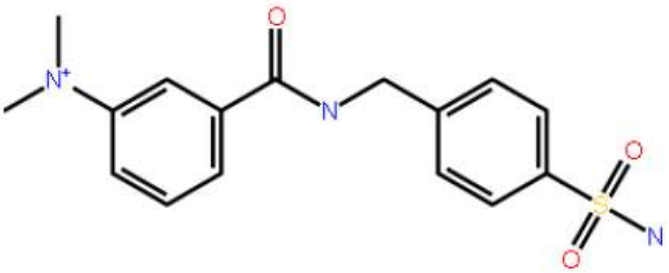
Compound #	Supplier code	ZINC code	Docking Score
1	AJ-292/14925212	ZINC00030345	-5.836
2	AM-807/41932130	ZINC05919492	-6.026
3	AM-807/14146032	ZINC08442014	-5.91
4	AM-879/14887007	ZINC08442161	-5.867
5	AM-879/42011358	ZINC08442191	-5.997
6	AM-879/42012288	ZINC08442197	-6.272
7	AM-879/42012312	ZINC08442202	-5.988
8	AK-968/41170301	ZINC08442285	-6.037
9	AK-968/41170344	ZINC08442286	-5.954
10	AK-968/41171569	ZINC08442287	-5.847



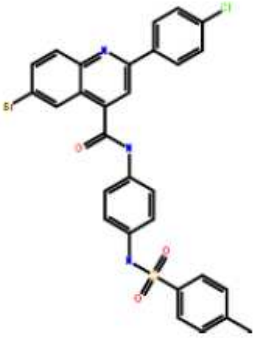
title: ZINC08442161
docking score: -5.85666143468



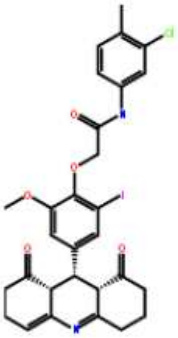
title: ZINC08442285
docking score: -6.03682700906



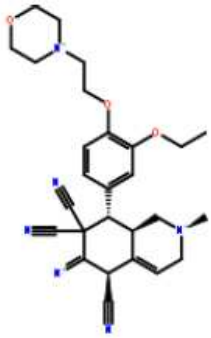
title: ZINC00030345
docking score: -5.83569164865



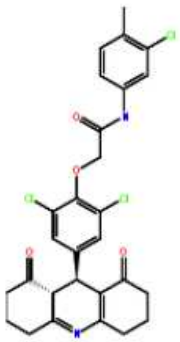
title: ZINC08442287
docking score: -5.8468376046



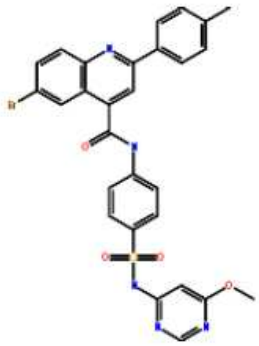
title: ZINC08442197
docking score: -6.27217023475



title: ZINC08442014
docking score: -5.91025819861

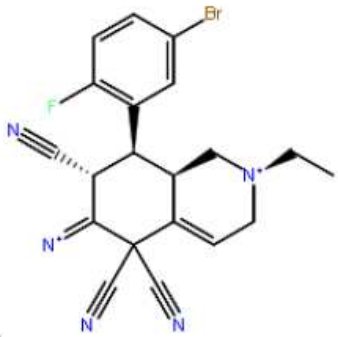


title: ZINC08442191
docking score: -5.99748065947



title: ZINC08442286
docking score: -5.95390016812

title: ZINC08442202
docking score: -5.98764582758



title: ZINC05919492
docking score: -6.02450980487

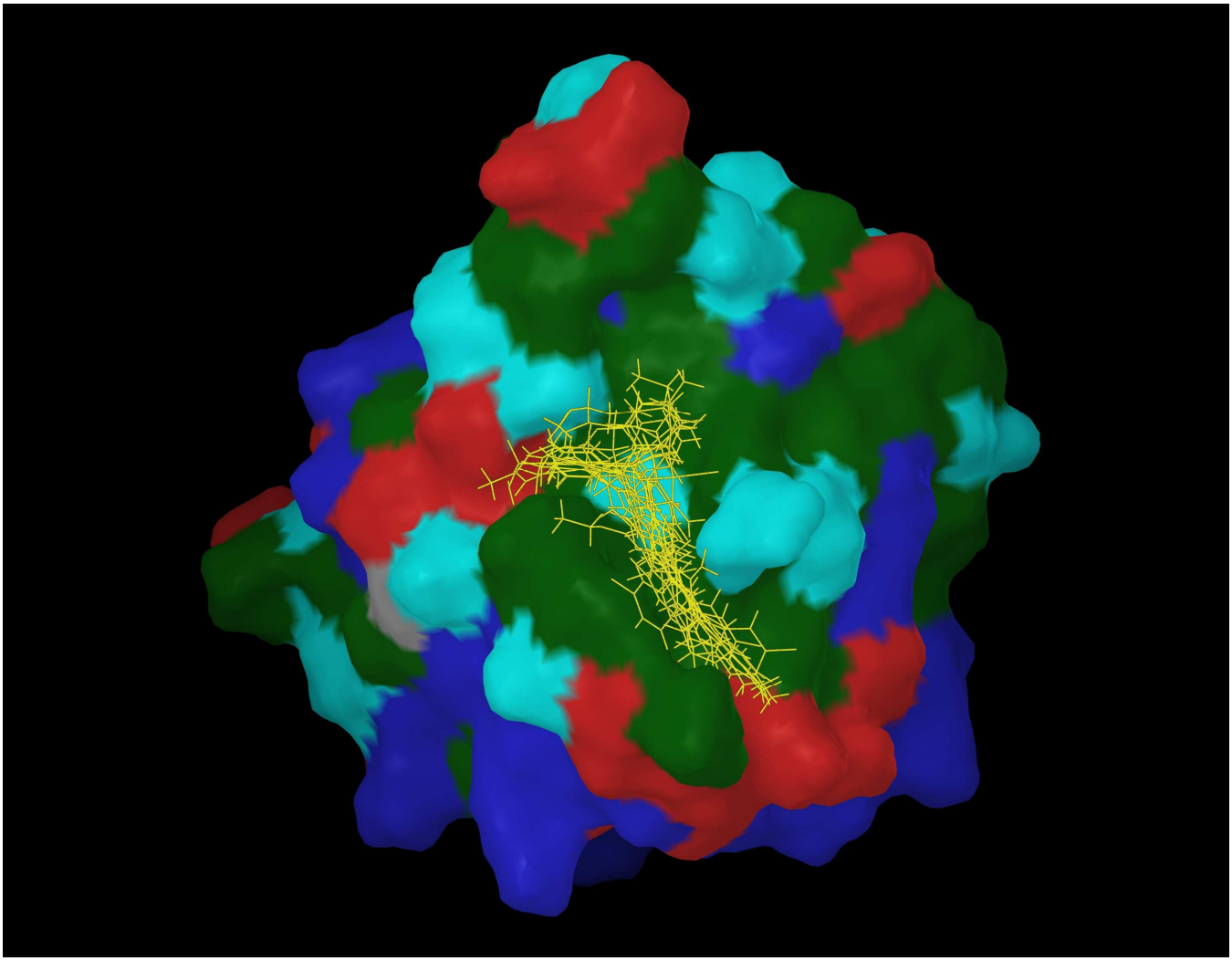


Fig Y: Superimposed structures of best docking score compounds in to binding site of betalactoglobulin.

Determining residue properties relies on PDB residue names. If an atom does not have a PDB residue name, then it will be colored gray.

Color Residue Property
---
gray: GLY and atoms without PDB Residue names
green: hydrophobic (ALA, CYS, ILE, LEU, MET, PHE, TRP, TYR, VAL, PRO)
cyan: polar uncharged (SER, THR, HIS, GLN, ASN)
blue: Positives (LYS, ARG)
red: Negatives(ASP, GLU)

---

4. Experimental Evaluation

The Mass spec and UV-Vis experiments were done by Mr.Rituraj pal, a doctoral student in Dr.Naryan's lab.

Procedure

Peaks corresponding to the polyphenols (identified using wavelengths of 360 and 418 nm) were collected from UV–Vis analysis (Perkin–Elmer) and subjected to ESI-FTMS (LTQXL, Thermo Fisher Scientific, San Jose, CA). Solutions of curcumin (5–200 lM) and compounds 4, 5, 7 and 8 (5–200) lM in acetonitrile were obtained by dilution from stock solutions that were prepared by weight. Freshly prepared stock solutions of tetranitromethane (TNM) in acetonitrile were separately added to curcumin and masoprocol. Added tetranitromethane concentrations varied between 0–10 lM. Samples were immediately analyzed using UV–Vis spectroscopy and by reversed-phase HPLC (Supelco Discovery_ BIO Wide Pore C18, 5 lm, 15 cm _ 100.0 mm) using an acetonitrile gradient (1%/min). Fractions were collected and analyzed using mass spectrometry.

Results

Mass spectrum of curcumin and other polyphenols in a comparative study suggests that a mass addition of 45 Da to curcumin and other polyphenols is indicative of nitration. Analysis of the mass chromatogram suggested that mono-nitrated polyphenols were formed with highest abundance relative to other adducts (data not shown). Tetra-nitro adducts of compound 8 was found but could not be detected in all other polyphenols (Table 1). Mass analysis of curcumin and other polyphenols after exposure to nitrosative stress conditions revealed the formation of mono- and poly-nitrosylated polyphenols. The presence of polyphenol dimers was also detected. (Table1). The presence of the bisphenol mitigated polymer formation; mass analysis revealed the presence of mono- and poly

nitrosylated polyphenol in addition to a polyphenol dimer, suggesting that the bisphenol was capable of scavenging NOx radicals by multiple mechanisms

Adduct formation between polyphenols and NOx (pH 8, 100 mM Tris–HCl, 25 ◦C).

Experimental conditions	Mass(Da) + identity	Mass (Da) + identity	Mass (Da) + identity	Mass (Da) + identity
Curcumin (control)	369.16 (C)	□	□	□
NOx + curcumin	369.16 (C)	415.2551 (NitroC)	437.2384 (NitroC)	453.2196 (NitroC)
Compound 4 (control)	425.1045 (M1)	□	□	□
NOx + compound 4	425.1045 (M1)	471.0886 (NitroM1)	□	□
Compound 5 (control)	489.2668 (M2)	□	□	□
NOx + compound 5	489.2668 (M2)	534.2813 (NitroM2)	□	□
Compound 7 (control)	608.0721 (M3)	□	□	□
NOx + compound 7	608.0432 (M3)	654.0763 (Nitro M3)	□	□
Compound 8 (control)	334.1407 (M4)	□	667.2886 (M4 Dimmer)	□
NOx + compound 8	334.1407 (M4)	500.2264 (Tetranitro M4)	665.2855 (M4 Dimmer)	712.2666 (Nitro M4 dimmer)

C, curcumin.
M1, compound 4
M2, compound 5
M3, compound 7
M4, compound8
*** n.o., not observed. [R. Pal. et,al.]

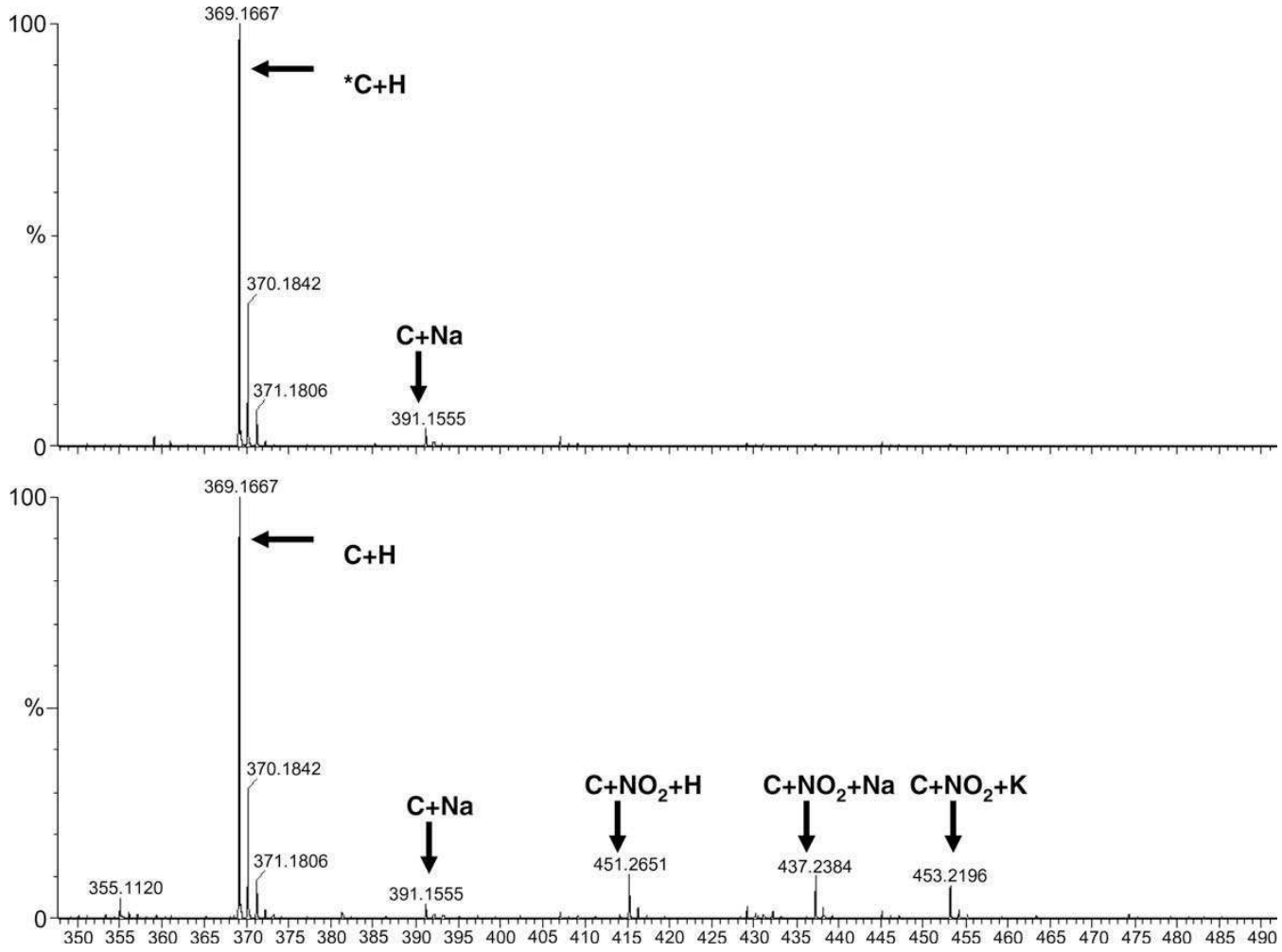


Fig. 2. ESI–FTMS of curcumin obtained by exposing to 4 lM tetranitromethane and 30 lM *C, curcumin (pH 8, 100 mM Tris–HCl).[R. Pal. et.al]

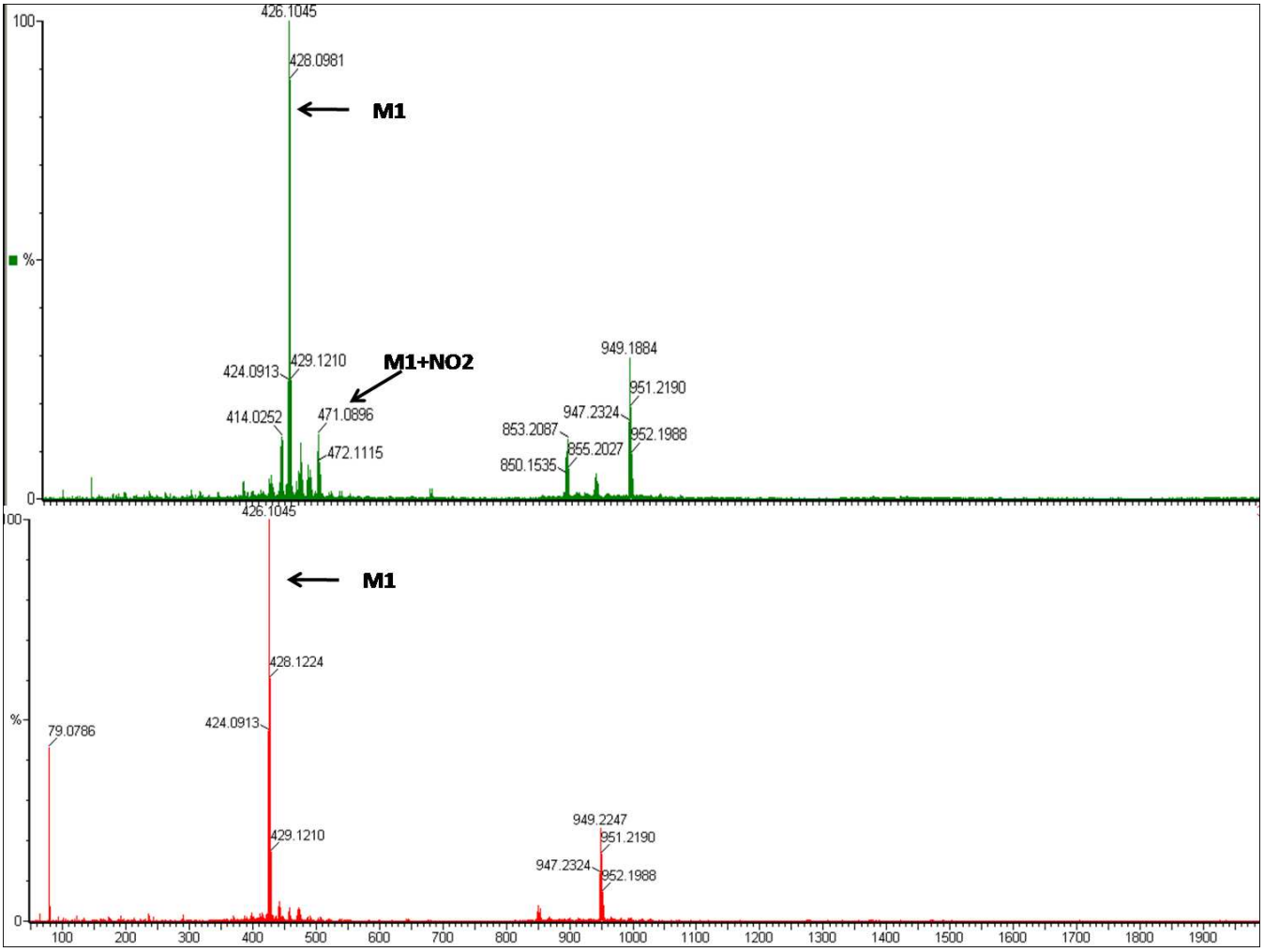


Fig. 3. ESI–FTMS of compound 4 obtained by exposing to 4 microM tetranitromethane and 30 microM compound 4 (pH 8, 100 mM Tris–HCl).

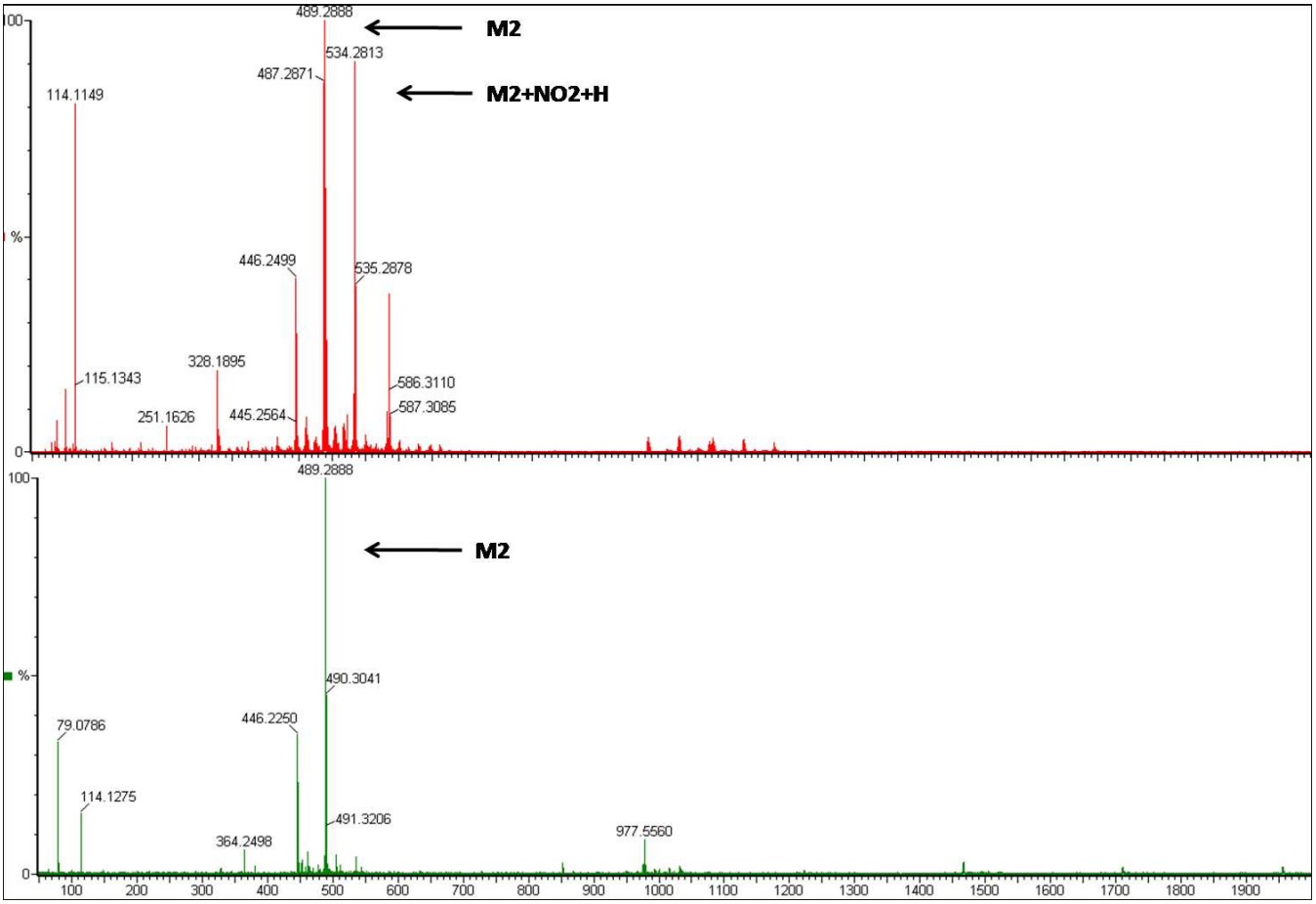


Fig. 4. ESI–FTMS of compound 5 obtained by exposing to 4 microM tetranitromethane and 30 microM compound 5 (pH 8, 100 mM Tris–HCl).

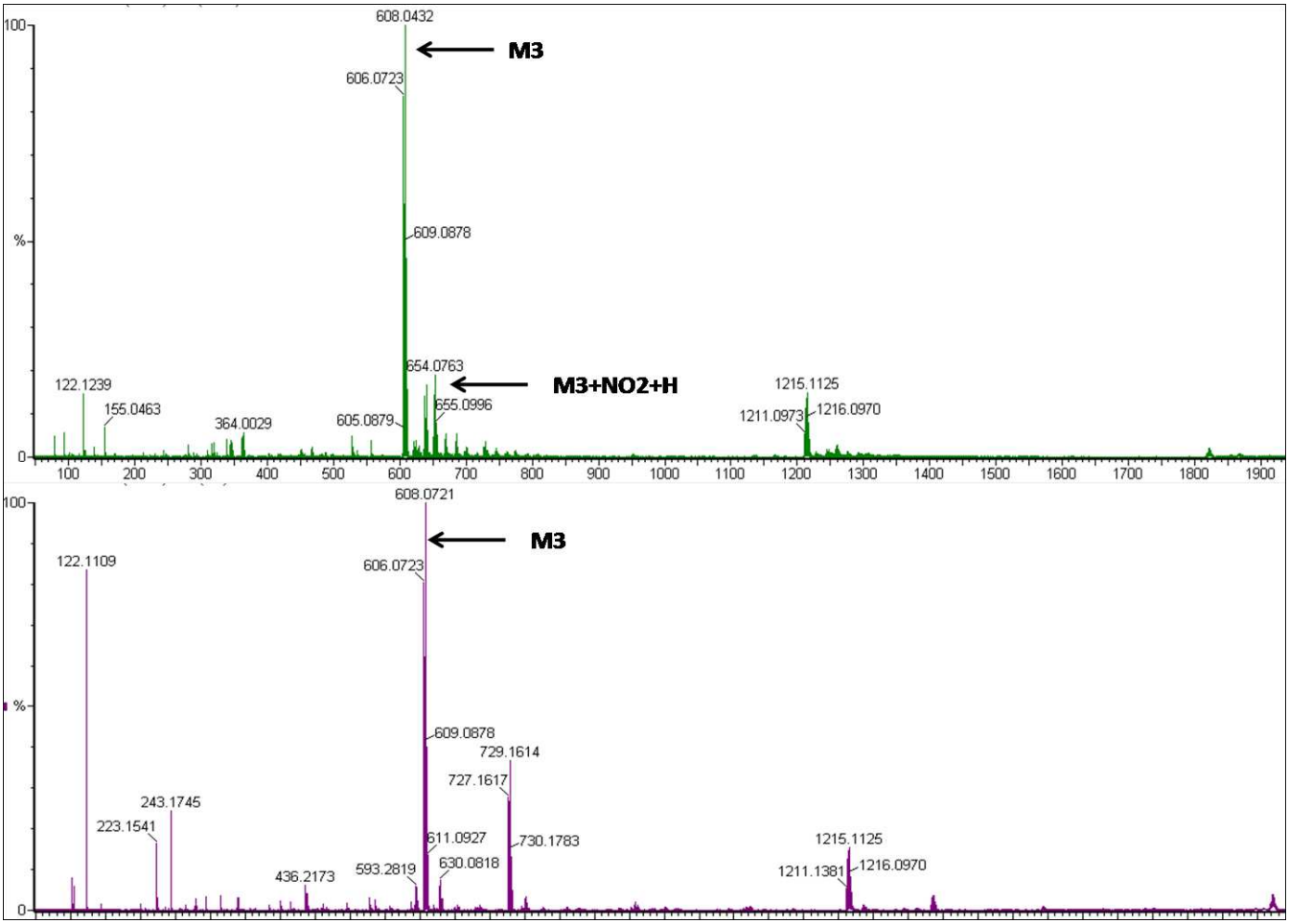


Fig. 5. ESI–FTMS of compound 7 obtained by exposing to 4 microM tetranitromethane and 30 microM compound 7 (pH 8, 100 mM Tris–HCl).

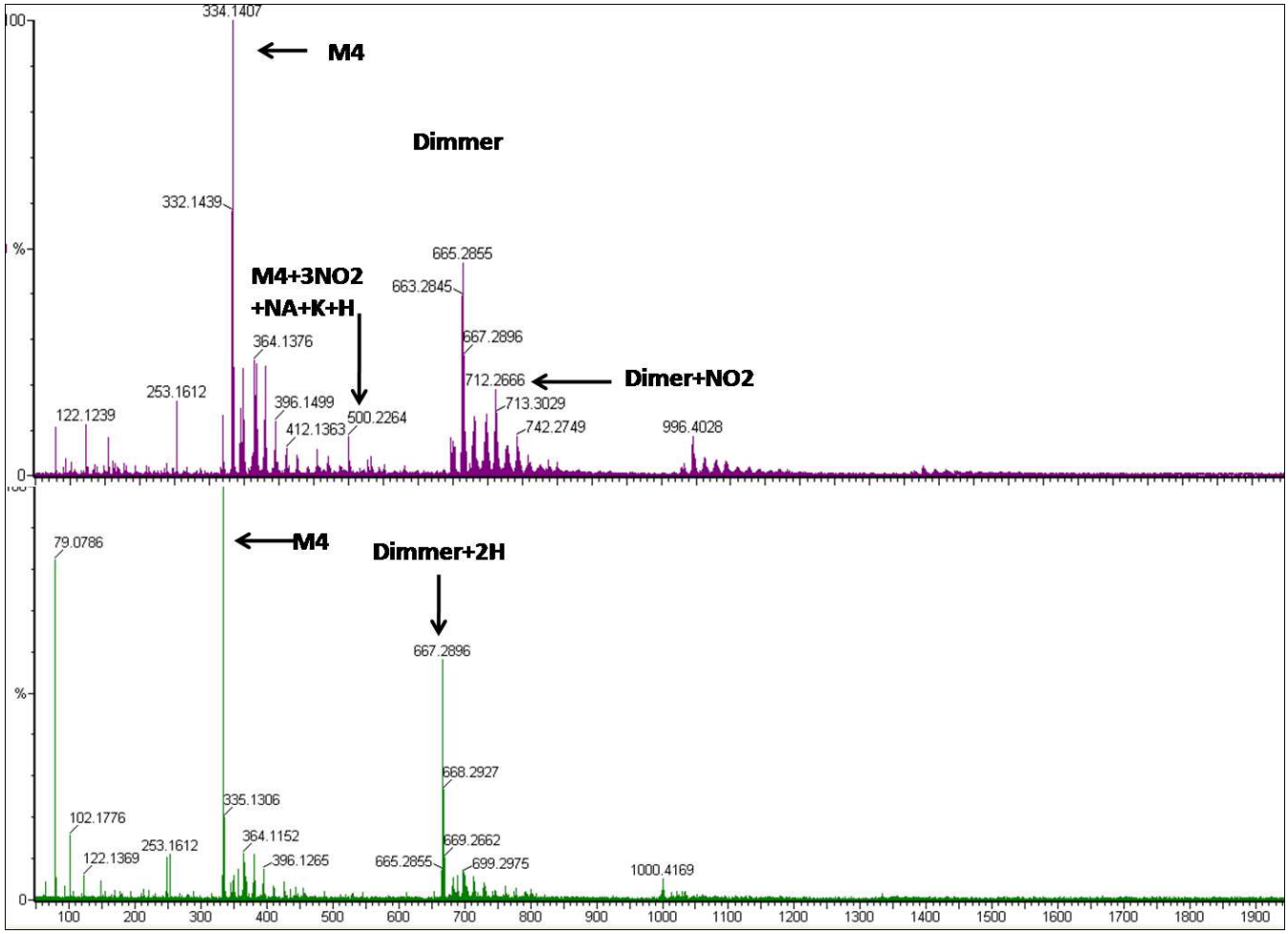


Fig. 5. ESI–FTMS of compound 8 obtained by exposing to 4 microM tetranitromethane and 30 microM compound 8 (pH 8, 100 mM Tris–HCl).[R. Pal. et.al]

CONCLUSION

Four potent nitrosative stress inhibitors are identified by virtual screening and experimental evaluation. The experimental results show the evidence of nitration on these compounds and Molecular docking results explain that these compounds can bind to betalactoglobulin. Hence betalactoglobulin can be used as vehicle for these compounds to improve their bioavailability.

CURCUMIN INTERACTION WITH HUMAN SERUM ALBUMIN

1. Introduction

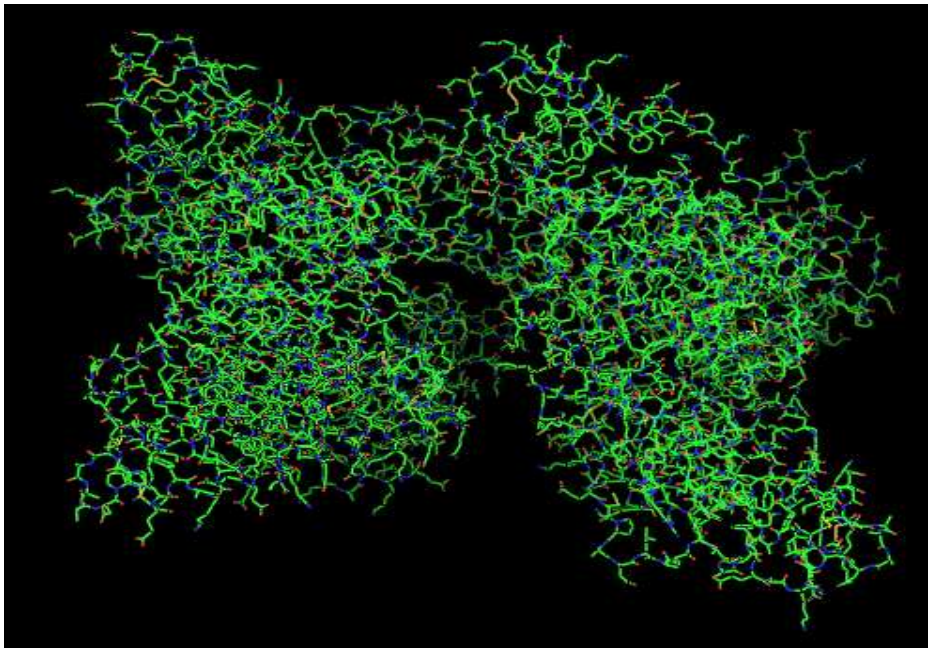
Curcumin associates with serum albumin by hydrophobic interactions and transported to appropriate cells and elicits its pharmacological actions. As the bioavailability of curcumin is low, there is a need to develop new similar molecules which have same or more biological activity and more bioavailability. In this project we would like to dock curcumin and its derivatives to human serum albumin.

Human serum albumin [33] is the most abundant protein in human blood plasma. It is produced in the liver. Albumin comprises about half of the blood serum protein. It is soluble and monomeric. The gene for albumin is located on chromosome 4 and mutations in this gene can result in various anomalous proteins. The human albumin gene is 16,961 nucleotides long from the putative 'cap' site to the first poly (A) addition site. It is split into 15 exons which are symmetrically placed within the 3 domains that are thought to have arisen by triplication of a single primordial domain. Albumin is synthesized in the liver as preproalbumin which has an N-terminal peptide that is removed before the nascent protein is released from the rough endoplasmic reticulum. The product, proalbumin, is in turn cleaved in the Golgi vesicles to produce the secreted albumin. The reference range for albumin concentrations in blood is 30 to 50 g/L. It has a serum half-life of approximately 20 days. It has a molecular mass of 67 kDa.

The approximate amino acid sequence of human serum albumin is:

*MKWVTFISLL FLFSSAYSRG VFRR*DAHKSE VAHRFKDLGE ENFKALVLIA FAQYLQQCPF EDHVKLVNEV TEFAKTCVAD ESAENCDKSL HTLFGDKLCT VATLRETYGE MADCCAKQEP ERNECFLQHK DDNPNLPRLV RPEVDVMCTA FHDNEETFLK KYLYEIARRH PYFYAPELLF FAKRYKAAFT ECCQAADKAA CLLPKLDELR DEGKASSAKQ RLKCASLQKF GERAFKAWAV ARLSQRFPKA EFAEVSKLVT DLTKVHTECC HGDLLECADD RADLAKYICE NQDSISSKLK ECCEKPLLEK SHCIAEVEND EMPADLPSLA ADFVESKDVC KNYAEAKDVF LGMFLYEYAR RHPDYSVVLL LRLAKTYETT LEKCCAAADP HECYAKVFDE FKPLVEEPQN LIKQNCELFE QLGEYKFQNA LLVRYTKKVP QVSTPTLVEV SRNLGKVGSK CCKHPEAKRM PCAEDYLSVV LNQLCVLHEK TPVSDRVTKC CTESLVNRRP CFSALEVDET YVPKEFNAET FTFHADICTL SEKERQIKKQ TALVELVKHK PKATKEQLKA VMDDFAAFVE KCCKADDKET CFAEEGKKLV AASQAALGL

Where the italicized first 24 amino acids are signal and propeptide portions not observed in the transcribed, translated and transported protein but present in the gene. There are 609 amino acids in this sequence with only 585 amino acids in the final product observed in the blood.



Crystal structure of human serum albumin

2. Docking of curcumin in to human serum albumin

In this study, we used Autodock 4.0, molecular modeling software to build and optimize molecular models of the curcumin with Human serum albumin. Autodock 4.0 generated a series of docking poses and ranked them using energy-based criterion. Based on this ranking, the lowest energy pose of the ligand–receptor complex can be selected.

- The first thing to check was that the ligand (curcumin or its derivatives) is docking into some kind of pocket on the receptor (human serum albumin).
- The second was that there should be a chemical match between the atoms in the ligand and those in the receptor. For example, check that carbon atoms in the ligand are near hydrophobic atoms in the receptor while nitrogen's and oxygen's in the ligand are near similar atoms in the binding pocket.
- Check for charge complementarity.
- Check whatever else you may know about your particular system: for instance, if you know that the enzymatic action of your protein involves a particular residue, examine how the ligand binds to that residue.

The following steps illustrate how the docking procedure should be carried out using AutoDock 4.0 [31] [35]

Step 1: Editing a PDB file

Protein Data Bank (PDB) files can have a variety of potential problems that need to be corrected before they can be used in AutoDock. These potential problems include missing atoms, added waters, more than one molecule, chain breaks, alternate locations etc. In this step, the PDB file of human serum albumin is prepared (to remove waters, how to add the polar hydrogen's) and which will be kept fixed during the dockings.

Step 2: Preparing a ligand file for AutoDock

In this step ligand file is converted from PDB to PDBQ format, detection of ROOT and selections of

torsions are done. The keywords ROOT, ENDROOT, BRANCH, and ENDBRANCH establish a "torsion tree" object or tor Tree that has a root and branches. The root is a rigid set of atoms, while the branches are rotatable groups of atoms connected to the rigid root. The keyword TORSDOF signals the number of torsional degrees of freedom in the ligand. The TORSDOF for a ligand is the total number of possible torsions in the ligand minus the number of torsions that only rotate hydrogens. TORSDOF is used in calculating the change in free energy caused by the loss of torsional degrees of freedom upon binding.

Step 3: Preparing the macromolecule file

The protein file should be converted to pdbqs format which is pdb plus 'q' charge and 's' salvation parameters.

Step 4: Preparing the grid parameter file

The grid parameter file (gpf) is prepared by sticking ligand file near tryptophan-214 of human serum albumin and creating a grid around it. The gpf tells AutoGrid the types of maps to compute, the location and extent of those maps and specifies pair-wise potential energy parameters.

Step 5: Running AutoGrid

Run the AutoGrid by typing the following commands in cygwin shell. The results should be always saved in a file with glg extension.

"Autogrid4 –p example.gpf –l example.glg & "

Step 6: preparing the docking parameter file

The docking parameter file tells AutoDock which map files to use, the ligand molecule to move, what its center and number of torsions are, where to start the ligand, which docking algorithm to use and how many runs to do. It usually has the file extension, ".dpf". Four different docking algorithms are currently available in AutoDock: SA, the original Monte Carlo simulated annealing; GA, a traditional Darwinian genetic algorithm; LS, local search; and GALS, which is a hybrid genetic algorithm with local search. The GALS is also known as a Larmarckian genetic algorithm, or LGA, because children are allowed to

inherit the local search adaptations of their parents. Each search method has its own set of parameters, and these must be set before running the docking experiment itself. These parameters include what kind of random number generator to use, step sizes, etc. The most important parameters affect how long each docking will run. In simulated annealing; the number of temperature cycles, the number of accepted moves and the number of rejected moves determine how long a docking will take. In the GA and GALS, the number of energy evaluations and the number of generations affect how long a docking will run. ADT lets you change all of these parameters, and others not mentioned here.

Step 7: Running Autodock

Run the AutoDock by typing the following commands in cygwin shell. The results should be always saved in a file with dlg extension.

"Autogrid4 –p example.dpf –l example.dlg &"

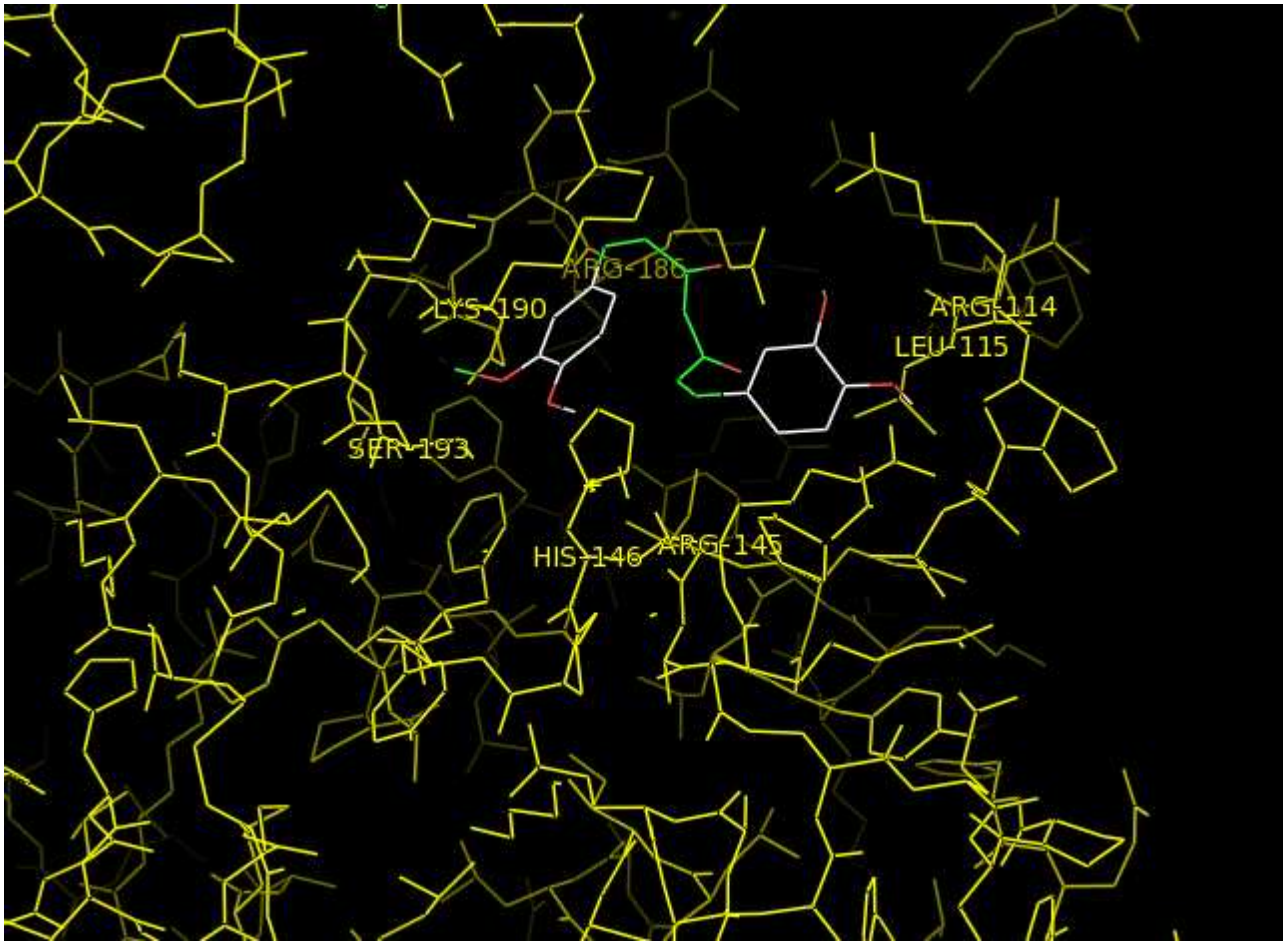
Step 8: Analyzing results

Reading a docking log or a set of docking logs is the first step in analyzing the results of docking experiments. (By convention, these results files have the extension ".dlg".) During its automated docking procedure, AutoDock outputs a detailed record to the file specified after the –l parameter. The output includes many details about the docking which are output as AutoDock parses the input files and reports what it finds. For example, for each AutoGrid map, it reports opening the map file and how many data points it read in. When it parses the input ligand file, it reports building various internal data structures. After the input phase, AutoDock begins the specified number of runs. It reports which run number it is starting; it may report specifics about each generation. After completing the runs, AutoDock begins an analysis phase and records details of that process. At the very end, it reports a summary of the amount of time taken and the words 'Successful Completion'. The key results in a docking log are the docked structures found at the end of each run, the energies of these docked structures and their similarities to each other. The similarity of docked structures is measured by computing the root-mean-square-deviation, rmsd, between the coordinates of the atoms. The docking results consist of the PDBQ

of the Cartesian coordinates of the atoms in the docked molecule, along with the state variables that describe this docked conformation and position.

3. Results and Discussion

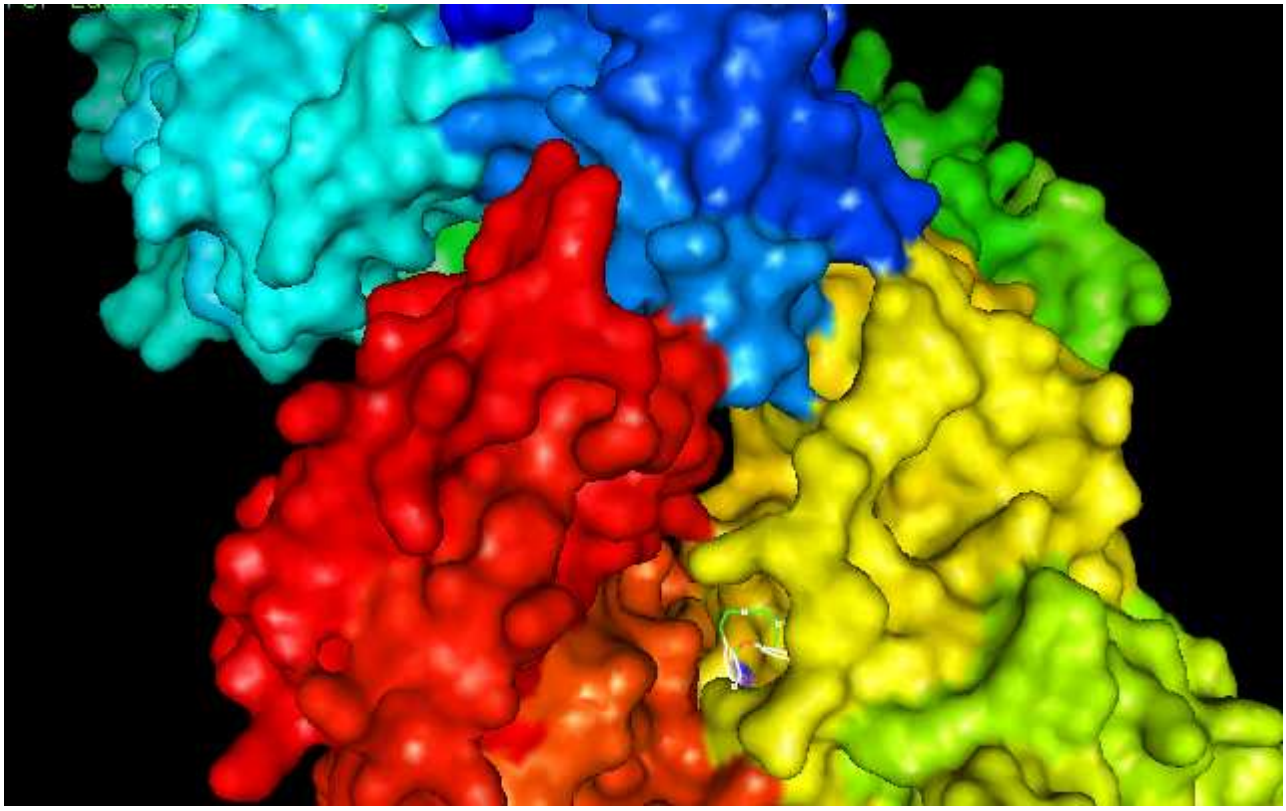
In the present study we used Autodock 4.0 (molecular docking software) to predict the binding site of curcumin in human serum albumin. Docking was performed placing grid box on entire protein molecule. The docking results propose four possible binding pockets for curcumin in human serum albumin, two in each chain. Depending upon the binding affinity best 9 docking poses were taken in to further consideration. The best docking pose was near amino acids, ARG 186, LYS 190, LEU 115, ARG 114. The binding affinities of these poses range from -6.9 to -6.0 kcal/mol. These results indicate that human serum albumin may serve as a viable agent to improve the bioavailability of curcumin



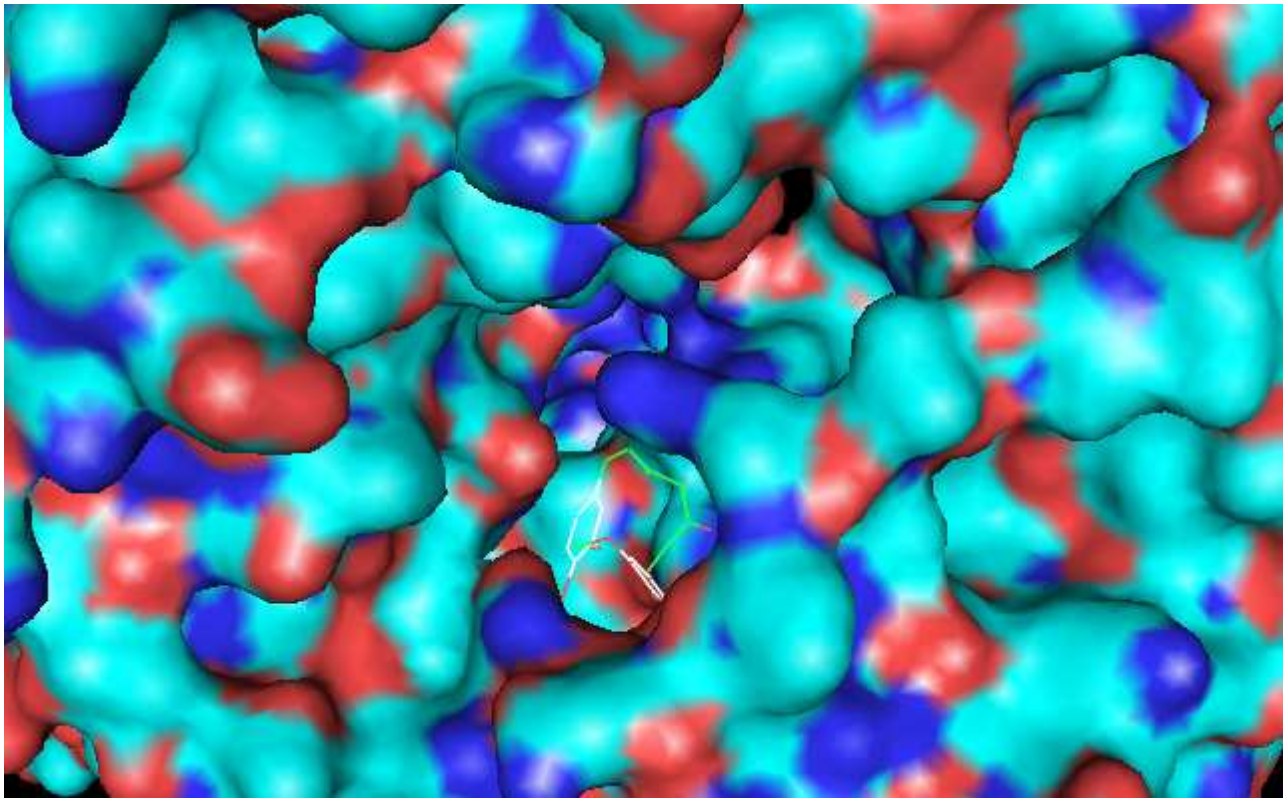
Amino acids of human serum albumin near best docking pose of curcumin.

Binding Affinities of different poses of curcumin with human serum albumin is as follows

mode	affinity (kcal/mol)	dist from best mode rmsd l.b.	rmsd u.b.
1	-6.9	0.000	0.000
2	-6.8	34.081	37.180
3	-6.7	16.407	21.619
4	-6.7	43.545	46.637
5	-6.4	36.669	39.184
6	-6.3	12.915	15.718
7	-6.2	17.451	21.940
8	-6.1	39.484	41.110
9	-6.0	21.852	25.518



Best docking pose of curcumin in human serum albumin



Best docking pose of curcumin in human serum albumin

REFERENCES

1. Kolev, Tsonko M.; et al. (2005). "DFT and Experimental Studies of the Structure and Vibrational Spectra of Curcumin". *International Journal of Quantum Chemistry* (Wiley Periodicals) **102** (6): 1069–79.

2. Aggarwal, BB.; Shishodia S. (May 2006). "Molecular targets of dietary agents for prevention and therapy of cancer". *Biochemical Pharmacology* (Elsevier) **71** (10): 1397–421.

3. Choi, Hyunsung; et al. (July 2006). "Curcumin Inhibits Hypoxia-Inducible Factor-1 by Degrading Aryl Hydrocarbon Receptor Nuclear Translocator: A Mechanism of Tumor Growth Inhibition". *Molecular Pharmacology* (American Society for Pharmacology and Experimental Therapeutics) **70**: 1664–71.

4. Shukla PK, Khanna VK, Ali MM, Khan MY, Srimal RC.Anti-ischemic effect of curcumin in rat brain, Neurochem Res. 2008 Jun;33(6):1036-43. Epub 2008 Jan 18, PMID: 18204970.

5. Stix, Gary (February 2007). "Spice Healer". *Scientific American*, http://sciam.com/article.cfm?chanID=sa006&articleID=131CED4F-E7F2-99DF-3C84BB412D1D3B51&pageNumber=1&catID=2.

6. Srivastava, KC; Bordia A; Verma SK (April 1995). "Curcumin, a major component of the food spice turmeric (Curcuma longa), inhibits aggregation and alters eicosanoid metabolism in human blood platelets". *Prostaglandins Leukot Essent Fatty Acids* **52** (4): 223–7

7. Padma, TV (2005-03-11). "Turmeric can combat malaria, cancer virus and HIV", *SciDev.net*.

8. Marotta, F.; et al. (October 2003). "Hepatoprotective effect of a curcumin/absinthium compound in experimental severe liver injury". *Chinese Journal of Digestive Diseases* (Blackwell Publishing) **4** (3): 122–7.

9. Kutluay SB, Doroghazi J, Roemer ME, Triezenberg SJ (January 2008). "Curcumin inhibits herpes simplex virus immediate-early gene expression by a mechanism independent of p300/CBP histone acetyltransferase activity". *Virology* **373**: 239.

10. Bourne KZ, Bourne N, Reising SF, Stanberry LR (1999 July). "Plant products as topical microbicide candidates: assessment of in vitro and in vivo activity against herpes simplex virus type 2". *Antiviral research* **42** (3): 219–26.

11. Shukla PK, Khanna VK, Khan MY, Srimal RC.Protective effect of curcumin against lead neurotoxicity in rat.Hum Exp Toxicol. 2003 Dec;22(12):653-8.PMID: 14992327>

12. Aggarwal BB, Shishodia S. Suppression of the nuclear factor-kappaB activation pathway by spice-derived phytochemicals: reasoning for seasoning. Ann N Y Acad Sci. 2004 Dec;1030:434-41.

13. "Curcumin's anti-cancer mechanism proposed", *nutraingredients-usa.com* (2007-04-13).

14. Yang, F; Lim GP; Begum AN; Ubeda OJ; Simmons MR; Ambegaokar SS; Chen PP; Kayed R; Glabe CG; Frautschy SA; Cole GM (February 2005). "Curcumin inhibits formation of amyloid beta oligomers and fibrils, binds plaques, and reduces amyloid in vivo". *Journal of Biological Chemistry* (American Society for Biochemistry and Molecular Biology) **280** (7): 5892–901.

15. Kelloff GJ, Crowell JA, Hawk ET, Steele VE, Lubet RA, Boone CW, Covey JM, Doody LA, Omenn GS, Greenwald P, et al. (1996) Strategy and planning for chemopreventive drug development: clinical development plan: curcumin. *J Cell Biochem Suppl* **26:** 72–85.

16. Weber WM, Hunsaker LA, Gonzales AM, Heynekamp JJ, Orlando RA, Deck LM, and Vander Jagt DL (2006) TPA-induced up-regulation of activator protein-1 can be inhibited or enhanced by analogs of the natural product curcumin. *Biochem Pharmacol* **72:** 928–940.

17. Pan MH, Huang TM, and Lin JK (1999) Biotransformation of curcumin through reduction and glucuronidation in mice. *Drug Metab Dispos* **27:** 486–494.

18. Pan MH, Lin-Shiau SY, and Lin JK (2000) Comparative studies on the suppression of nitric oxide synthase by curcumin and its hydrogenated metabolites through down-regulation of IkappaB kinase and NFkappa B in macrophages. *Biochem Pharmacol* **60:** 1665–1676.

19. Pan MH, Huang TM, and Lin JK (1999) Biotransformation of curcumin through reduction and glucuronidation in mice. *Drug Metab Dispos* **27:** 486–494.

20. Pan MH, Lin-Shiau SY, and Lin JK (2000) Comparative studies on the suppression of nitric oxide synthase by curcumin and its hydrogenated metabolites through down-regulation of IkappaB kinase and NFkappa B in macrophages. *Biochem Pharmacol* **60:** 1665–1676.

21. Sharma RA, Euden SA, Platton SL, Cooke DN, Shafayat A, Hewitt HR, Marczylo TH, Morgan B, Hemingway D, Plummer SM, et al. (2004) Phase I clinical trial of oral curcumin: biomarkers of systemic activity and compliance. *Clin Cancer Res* **10:** 6847–6854.

22. Lao CD, Ruffin MT 4th, Normolle D, Heath DD, Murray SI, Bailey JM, Boggs ME, Crowell J, Rock CL, and Brenner DE (2006) Dose escalation of a curcuminoid formulation. *BMC Complement Altern Med* **6:** 10.

23. G.M. Morris, D.S. Goodsell, R.S. Halliday, R. Huey, W.E. Hart, R.K. Belew and A.J. Olson, *J. Comput. Chem.* **19** (1998), p. 1639.

24. C.M. Venkatachalam, X. Jiang, T. Oldfield and M. Waldman, *J. Mol. Graph. Model.* **21** (2003), p. 289.

25. R. Abagyan, M. Totrov and D. Kuznetsov, *J. Comput. Chem.* **15** (1994), p. 488.

26. M. Rarey, B. Kramer, T. Lengauer and G. Klebe, *J. Mol. Biol.* **261** (1996), p. 470.

27. G. Jones, P. Willett, R.C. Glen, A.R. Leach and R. Taylor, *J. Mol. Biol.* **267** (1997), p. 727.

28. T.A. Halgren, R.B. Murphy, R.A. Friesner, H.S. Beard, L.L. Frye, W.T. Pollard and J.L. Banks, *J. Med. Chem.* **47** (2004), p. 1750.

29. T.J.A. Ewing, S. Makino, A.G. Skillman and I.D. Kuntz, *J. Comput. Aided Mol. Des.* **15** (2001), p. 411.

30. A. Ch Pulla Reddy[1], E. Sudharshan[2], A. G. Appu Rao[2] and B. R. Lokesh[1] Interaction of curcumin with human serum albumin—A spectroscopic study, Volume 34, Number 10 / October, 1999, pg 1025-1029.

31. Using AutoDock 4 for Virtual Screening Written by William Lindstrom, Garrett M. Morris, Christoph Weber and Ruth Huey, The Scripps Research Institute Molecular Graphics Laboratory.

32. http://en.wikipedia.org/wiki/Curcumin

33. http://en.wikipedia.org/wiki/Human_serum_albumin

34. Structural relationship of curcumin derivatives binding to the BRCT domain of human DNA polymerase λ, Genes Cells, Mar 2006; 11: 223 - 235.

35. Using AutoDock with AutoDock Tools: A tutorial; written by Ruth Huey and Garrett M. Morris The Scripps Research Institute Molecular Graphics Laboratory.

36. Kunchandy E, Rao MNA. Oxygen radical scavenging activity of curcumin. Int J Pharm 1990; 58:237–240.

37. Subramaniam M, Sreejayan N, Rao MNA, Davasagayam TP, Sigh BB. Diminuition of singlet oxygen- induced DNA damage by curcumin and related antioxidants. Mutation Res 1994;311:249–255.

38. Pal R, Cristan EA, Schnittker K, Narayan M., Rescue of ER oxidoreductase function through polyphenolic phytochemical intervention: implications for subcellular traffic and neurodegenerative disorders, Biochem Biophys Res Commun. 2010 Feb 19; 392(4):567-71.

39.

40. G. Kontopidis, C. Holt and L. Sawyer, Invited review: Beta-lactoglobulin: Binding properties, structure, and function, *Journal of Dairy Science* **87** (4) (2004), pp. 785–796.

41. H.C. Liu, W.L. Chen and S.J.T. Mao, Antioxidant nature of bovine milk beta-lactoglobulin, *Journal of Dairy Science* **90** (2) (2007), pp. 547–555.

42. S. Brownlow, J.H.M. Cabral, R. Cooper, D.R. Flower, S.J. Yewdall and I. Polikarpov et al., Bovine beta-lactoglobulin at 1.8 angstrom resolution – still an enigmatic lipocalin, Structure **5** (4) (1997), pp. 481–495.

43. McKenzie, H. A. (1971). β-Lactoglobulins. In Milk proteins, (pp. 2257–2330).

44. G. Kontopidis, C. Holt and L. Sawyer, Invited review: Beta-lactoglobulin: Binding properties, structure, and function, *Journal of Dairy Science* **87** (4) (2004), pp. 785–796.

45. B.J. Harvey, E. Bell and L. Brancaleon, A tryptophan rotamer located in a polar environment probes ph-dependent conformational changes in bovine β-lactoglobulin A, *Journal of Physical Chemistry B* **111** (10) (2007), pp. 2610–2620.

46. S. Futterman and J. Heller, The enhancement of fluorescence and the decreased susceptibility to enzymatic oxidation of retinol complexed with bovine serum albumin, lactoglobulin, and the retinol-binding protein of human plasma, *Journal of Biological Chemistry* **247** (16) (1972), pp. 5168–5172.

47. D.C. Lange, R. Kothari, R.C. Patel and S.C. Patel, Retinol and retinoic acid bind to a surface cleft in bovine beta-lactoglobulin: a method of binding site determination using fluorescence resonance energy transfer, *Biophysical Chemistry* **74** (1) (1998), pp. 45–51.

48. M. Narayan and L.J. Berliner, Fatty acids and retinoids bind independently and simultaneously to beta-lactoglobulin, *Biochemistry* **36** (7) (1997), pp. 1906–1911.

49. M.D. Perez and M. Calvo, Interaction of beta-lactoglobulin with retinol and fatty-acids and its role as a possible biological function for this protein – a review, *Journal of Dairy Science* **78** (5) (1995), pp. 978–988.

50. Q.W. Wang, J.C. Allen and H.E. Swaisgood, Binding of vitamin D and cholesterol to beta-lactoglobulin, *Journal of Dairy Science* **80** (6) (1997), pp. 1054–1059

51. Q.W. Wang, J.C. Allen and H.E. Swaisgood, Binding of lipophilic nutrients to beta-lactoglobulin prepared by bioselective adsorption, *Journal of Dairy Science* **82** (2) (1999), pp. 257–264.

52. Liu, Y. (2003). Beta-lactoglobulin complexed vitamins A and D in skim milk: Shelf life and bioavailability, PhD dissertation, advisor: Dr. Jonathan C. Allen, North Carolina State University, NC, USA.

53. H.M. Farrell Jr., M.J. Behe and J.A. Enyeart, Binding of p-nitrophenyl phosphate and other aromatic compounds by beta-lactoglobulin, *Journal of Dairy Science* **70** (2) (1987), pp. 252–258.

54. M.D. Perez and M. Calvo, Interaction of beta-lactoglobulin with retinol and fatty-acids and its role as a possible biological function for this protein – a review, *Journal of Dairy Science* **78** (5) (1995), pp. 978–988.

55. A.A. Spector and J.E. Fletcher, Binding of long chain fatty acids to beta-lactoglobulin, *Lipids* **5** (4) (1970), pp. 403–411

APPENDIX A

Level-3 descriptors of aromatic amines

Code	logP	HOMO	LUMO	MR5	MR3	MR2	MR6
1	3.72	8.159	0.816	0.1	0.8	0.56	0.1
2	1.64	8.089	1.045	0.1	0.54	0.54	0.1
3	1.64	7.997	1.04	0.1	0.54	0.54	0.1
4	2.59	8.034	0.176	0.8	0.1	0.56	0.8
5	2.48	8.663	1.035	0.1	0.8	0.9	0.1
6	3.69	7.82	0.804	0.1	0.8	0.8	0.1
7	2.25	8.108	0.195	0.1	0.8	0.8	0.1
8	2.3	8.035	0.115	0.1	0.56	0.54	0.1
9	3.18	8.467	0.191	0.1	0.56	0.56	0.1
10	3.26	8.132	0.357	0.1	0.8	0.8	0.1
11	2.18	8.23	1.164	0.1	0.54	0.54	0.1
12	4.31	7.58	20.7	0.1	0.8	0.8	0.1
13	1.58	8.193	0.267	0.1	0.1	2.98	0.1
14	1.91	8.403	0.581	0.1	0.56	0.56	0.1
15	2.41	8.244	0.581	0.56	0.1	0.56	0.1
16	1.77	8.13	0.702	0.54	0.1	1.96	0.1
17	1.12	8.121	0.722	0.54	0.1	1.5	0.1
18	1.54	8.695	0.085	0.1	0.1	0.09	0.1
19	1.68	8.288	0.605	0.1	0.1	0.56	0.1
20	1.84	9.931	1.491	0.1	0.1	0.74	0.1

21	1.58	8.144	0.215	0.1	0.1	2.98	0.1
22	1.83	8.404	0.581	0.56	0.1	0.56	0.1
23	1.79	8.18	0.029	0.1	0.1	0.6	0.6
24	1.47	7.753	0	0.1	0.56	0.1	0.1
25	1.64	8.11	1.069	0.1	0.54	0.1	0.1
26	1.64	8.369	1.039	0.1	0.54	0.1	0.1
27	2.59	8.142	0.203	0.1	0.8	0.56	0.1
28	2.95	8.8	0.998	0.1	0.8	0.74	0.1
29	2.59	8.214	0.199	0.8	0.1	0.56	0.1
30	2.68	8.746	1.029	0.1	0.1	3.17	0.1
31	1.81	8.508	0.154	0.6	0.1	0.28	0.1
32	1.16	8.274	0.478	0.56	0.1	0.28	0.1
33	2.68	8.852	1.148	0.1	0.1	3.17	0.1
34	1.36	9.052	0.923	0.1	0.1	0.28	0.1
35	1.72	7.876	0.143	0.1	0.56	0.1	0.1
36	3.06	8.595	1.164	0.1	0.56	0.1	0.1
37	3.26	7.869	0.771	0.1	0.8	0.1	0.1
38	2.84	8.407	0.107	0.1	0.1	2.54	0.1
39	2.3	8.023	0.025	0.1	0.54	0.1	0.1
40	3.72	8.273	0.887	0.8	0.56	0.1	0.1
41	3.14	8.106	0.108	0.1	0.56	0.1	0.1
42	2.28	8.227	0.198	0.1	0.8	0.1	0.1
43	3.26	8.233	0.365	0.1	0.8	0.1	0.1
44	2.18	8.495	1.165	0.1	0.54	0.1	0.1

45	3.72	8.144	0.865	0.8	0.8	0.1	0.1
46	2.78	10.02	1.691	0.1	0.1	0.74	0.89
47	3.92	8.236	0.405	0.1	0.56	0.1	0.1
48	1.9	8.632	0.268	0.1	0.1	0.6	0.1
49	2.96	8.502	0.305	0.1	0.1	1.03	0.1
50	2.03	7.998	0.127	0.1	0.56	0.1	0.1
51	1.74	8.449	0.419	0.56	0.1	0.79	0.1
52	2.58	8.562	0.226	0.1	0.1	0.56	0.1
53	2.43	8.508	0.289	0.1	0.1	0.56	0.1
54	1.58	8.414	0.056	0.1	2.98	0.1	0.1
55	3.51	8.125	0.142	0.1	0.1	0.6	0.1
56	1.81	7.927	0.005	0.1	0.1	0.79	0.1
57	2.34	7.873	0.185	0.1	0.1	0.56	0.1
58	1.91	8.318	0.602	0.1	0.56	0.1	0.1
59	1.58	8.187	0.103	0.1	2.98	0.1	0.1
60	2.68	8.665	0.648	0.1	3.17	0.1	0.1
61	2.68	8.801	1.034	0.1	3.17	0.1	0.1
62	3.3	8.398	0.007	2.54	0.1	0.56	0.1
63	2.68	8.879	1.178	0.1	3.17	0.1	0.1
64	2.29	9.032	0.104	0.1	0.5	0.1	0.1
65	2.3	7.877	0.107	0.1	0.56	0.1	0.1
66	4.2	8.033	0.818	0.1	0.8	0.8	0.1
67	2.7	8.333	0.182	0.1	0.56	0.1	0.1
68	3.26	8.122	0.203	0.1	0.8	0.1	0.1

69	1.63	8.534	0.376	0	0.8	0.1	0.1
70	1.52	8.274	0.583	0.1	0.79	0.1	0.1
71	2.13	8.318	0.589	0.1	0.1	0.1	0.1
72	3.66	8.181	0.605	0.1	0.1	1.03	0.1
73	2.5	8.467	0.186	0.1	0.1	0.09	0.1
74	4.46	8.165	0.61	0.1	0.1	1.5	0.1
75	1.59	8.274	0.576	0.1	0.1	0.1	0.1
76	1.99	8.209	21.33	0.1	0.1	0.1	0.1
77	2.68	8.597	0.554	0.1	0.1	0.1	0.1
78	3.3	8.208	0.069	0.1	0.1	0.56	0.1
79	2.68	8.63	0.959	0.1	0.1	0.1	0.1
80	2.68	8.707	1.102	0.1	0.1	0.1	0.1
81	2.86	8.263	0.048	0.1	0.1	0.1	0.1
82	2.3	7.993	0.029	0.1	0.54	0.56	0.1
83	2.7	8.256	0.157	0.1	0.56	0.56	0.1
84	3.26	8.524	0.411	0.1	0.8	0.8	0.1
85	1.36	8.167	0.317	0.1	0.1	0.1	0.1
86	2.18	7.528	0.047	0.1	0.1	0.1	0.1
87	3.72	7.837	20.85	0.8	0.8	0.8	0.1
88	2.26	8.62	0.212	0.1	0.1	0.1	0.1
89	1.28	8.325	0.241	0.1	0.1	0.54	0.1
90	2.72	9.11	1.019	0.1	0.1	0.74	0.1
91	1.88	8.568	0.284	0.1	0.1	0.1	0.1
92	3.65	8.38	0.62	0.1	0.1	0.1	0.1

93	1.24	8.182	0.513	0.1	0.1	0.1	0.1
94	1.15	8.56	0.275	0.1	0.1	0.1	0.1
95	1.23	8.152	0.474	0.1	0.1	0.56	0.1
96	2.58	8.498	0.228	0.1	0.1	0.89	0.1
97	2.43	8.559	0.274	0.1	0.1	0.6	0.1
98	2.43	8.477	0.293	0.1	0.1	0.6	0.1
99	2.96	8.598	0.255	0.1	0.1	0.1	0.1
100	1.16	8.395	0.395	0.1	0.65	0.8	0.1
101	4.98	7.926	0.592	0.8	0.8	0.8	0.1
102	1.28	8.443	0.383	0.1	0.8	0.1	0.1
103	3.72	8.186	0.895	0.1	0.56	0.56	0.1
104	3.72	8.166	0.853	0.1	0.56	0.1	0.1
105	1.79	8.134	0.294	0.1	0.8	0.65	0.1
106	3.26	7.6	0.712	0.8	0.8	0.8	0.8
107	3.56	8.099	0.364	0.8	0.8	0.8	0.1
108	1.34	7.931	0.147	0.1	0.1	0.1	0.1
109	0.95	8.212	0.492	0.1	0.1	0.1	0.1
110	0.93	8.527	0.597	0.1	0.79	0.1	0.1
111	0.9	8.53	0.615	0.1	0.1	0.1	0.1
112	1.88	8.739	0.246	0.1	0.6	0.1	0.1
113	1.55	8.504	0.621	0.1	1.25	0.1	0.1
114	1.55	8.344	0.525	0.1	0.1	1.25	0.1
115	0.04	8.269	0.408	0.1	0.1	0.1	0.1
116	0.17	8.523	0.504	0.1	0.28	0.1	0.1

117	7.31	8.059	0.666	0.1	0.1	1.5	1.5
118	5.72	8.111	0.643	0.1	0.1	1.03	1.03
119	6.26	8.058	0.673	0.1	0.1	1.96	0.56
120	5.46	8.097	0.641	0.1	0.1	1.5	0.56
121	4.66	8.113	0.636	0.1	0.1	1.03	0.56
122	3.6	8.12	0.624	0.1	0.1	0.56	0.56
123	2.8	8.498	0.003	0.1	2.54	0.1	0.1
124	1.58	8.305	0.167	0.1	2.54	0.54	0.1
125	1.18	8.331	0.525	0.1	0.1	0.79	0.1
126	0.62	8.356	0.448	0.1	0.1	0.28	0.1
127	2.41	8.222	0.616	0.1	0.1	0.56	0.56
128	4.03	8.803	0.319	0.1	0.1	0.89	0.89
129	3.69	8.767	0.258	0.1	0.1	0.6	0.6
130	2.97	8.359	0.594	0.1	0.1	1.03	1.03
131	1.91	8.44	0.597	0.56	0.56	0.1	0.1
132	1.91	8.37	0.582	0.1	0.1	0.56	0.56
133	3.05	8.741	0.069	0.1	0.1	0.89	0.1
134	2.91	8.682	0.001	0.1	0.1	0.6	0.1
135	2.34	8.685	0.211	0.1	0.1	0.1	0.1
136	2.32	8.684	0.204	0.1	0.1	1.39	0.1
137	1.26	8.56	0.275	0.1	0.1	0.09	0.1
138	2.11	8.656	0.203	0.1	0.1	0.89	0.1
139	1.96	8.389	0.613	0.1	0.1	0.1	0.1
140	1.74	8.438	0.593	0.1	0.1	1.03	0.1

141	1.39	8.362	0.606	0.1	0.1	0.1	0.1
142	1.4	8.486	0.589	0.1	0.56	0.1	0.1
143	1.32	8.443	0.585	0.1	0.1	0.6	0.1
148	0.609	8.246	0.526	0.54	0.1	0.28	0.1
149	0.137	8.121	0.706	0.54	0.1	0.56	0.1
151	0.137	7.861	0.609	0.1	0.1	0.56	0.1
152	0.087	8.247	0.684	0.1	0.54	0.56	0.1
153	3.094	8.282	0.246	0.1	0.56	0	0.1
154	1.177	9.474	1.076	0.74	0.1	0.28	0.1
156	0.751	8.472	0.802	0.1	0.1	0.74	0.1
162	3.741	8.441	0.24	0.1	0.1	0.6	0.1
165	1.305	9.244	0.583	0.1	0.1	0.1	0.1
166	1.182	9.209	0.59	0.1	0.1	0.1	0.1
169	0.807	8.959	1.119	0.1	0.74	0.1	0.1
172	1.406	9.312	1.484	0.74	0.1	0.69	0.1
173	0.611	8.945	0.85	0.1	0.1	0.54	0.1
178	0.886	8.888	0.254	0.1	0.1	0.1	0.1
179	2.389	9.284	0.224	0.1	0.5	0.1	0.1
180	0.371	8.343	0.728	0.1	0.74	0.1	0.1
181	0.976	8.075	0.706	0.1	0.1	0.74	0.1
185	1.992	8.757	0.112	0.1	0.1	0.1	0.1
186	2.307	8.277	0.708	0.1	0.1	0.1	0.1
191	4.794	8.14	0.271	0.1	0.8	0.1	0.1
194	1.765	8.353	0.353	0.1	0.1	0.1	0.1

196	5.323	10.211	2.028	0.1	0.1	0.74	0.74
197	1.208	8.66	0.239	0.1	0.1	0.69	0.1
199	1.258	9.159	0.705	0.1	0.1	0.1	0.1
200	0.312	7.915	0.64	0.1	0.1	0.1	0.1
201	3.29	8.127	0.169	0.1	0.8	1.03	0.1
202	3.71	8.102	0.172	0.1	0.8	1.5	0.1
203	4.27	8.124	0.171	0.1	0.8	1.96	0.1
204	4.16	8.052	0.179	0.1	0.8	1.96	0.1
205	4.14	8.06	0.073	0.1	0.56	1.03	0.1
206	4.55	8.023	0.051	0.1	0.56	1.5	0.1
207	5.12	8.045	0.065	0.1	0.56	1.96	0.1
208	5.01	7.997	0.041	0.1	0.56	1.96	0.1
209	3.88	8.214	0.073	0.1	0.1	1.03	0.1
210	4.3	8.21	0.076	0.1	0.1	1.5	0.1
211	4.86	8.212	0.072	0.1	0.1	1.96	0.1
212	4.75	8.17	0.108	0.1	0.1	1.96	0.1
213	3.93	8.166	0.088	0.1	0.1	0.56	0.56
214	4.92	8.15	0.096	0.1	0.1	1.03	1.03
215	5.75	8.144	0.114	0.1	0.1	1.5	1.5
216	3.39	8.208	0.04	0.1	0.1	0.1	0.1
217	3.88	8.216	0.043	0.1	0.1	0.1	0.1
218	4.3	8.211	0.056	0.1	0.1	0.1	0.1
219	4.86	8.217	0.041	0.1	0.1	0.1	0.1
220	4.75	8.205	0.059	0.1	0.1	0.1	0.1

221	3.65	8.035	0.1	0.1	0.56	0.1	0.1
222	5.01	8.036	0.078	0.1	0.56	0.1	0.1
223	6.71	8.034	0.082	0.1	0.56	0.1	0.1
224	4.1	8.45	0.66	0.1	0.56	0.1	0.1
225	3.39	8.256	0.073	0.1	0.1	0.1	0.1
226	3.93	8.239	0.1	0.1	0.1	0.1	0.1
227	3.8	8.584	0.549	0.1	0.1	0.1	0.1
228	3.8	8.535	0.401	0.1	0.1	0.1	0.1
229	4.77	8.754	0.825	0.1	0.1	0.1	0.1

CURRICULUM VITTAE

Suman Sirimulla was born on August 15, 1984 in karimnagar, state of Andhra Pradesh, India. Suman Sirimulla was the youngest child of three of Bhaskar sirimulla and Nirmala sirimulla. He entered J.S.S college of Pharmacy, Mysore, India in December 2001 and graduated with Bachelor of Pharmacy (B. Pharm) degree in the November 2005. In the spring 2006, he entered the graduate school at The University of Texas at El Paso. He was awarded a Teaching Assistantship from spring 2006 to summer 2007. He joined Dr William C. Herndon's lab in spring 2006 to do his thesis work in the area of medicinal chemistry. After finishing masters degree he continued his research in Dr. Herndon's lab as PhD student in fall 2007 and graduated with a PhD degree in spring 2010. Later he joined El Paso Community college as chemistry lecturer.

Permanent address: H.No. 8-6-188/6/1

Alakapuri, Kothirampur

Karimnagar, Andhra Pradesh

India. 505001

This thesis was typed by Suman Sirimulla.