QSAR

S. Ravichandran, **Ph.D.**

Advanced Biomedical Computing Center, NCI-Frederick, MD 21702

05/27/2004

Overview of the Talk

- What is QSAR?
  - Lots of data
- Data Reduction
  - Clustering, PCA etc.
- Predicting Activity
  - Linear Regression, MLR, PLS etc.
- Hands-on exercise
  - Cerius2 (QSAR+)

QSAR

- What is QSAR

  Quantitative Structure Activity Relationships

Addresses two questions:

- What feature of a molecule affect its activity?
- What can be modified to enhance properties?

Quantitative in that a mathematical model is used to account for the observed activity

Applications of QSAR

- Drug Design
  - Predictions for new experiments
    - Place 10 substituents on the four open positions of an asymmetrically disubstituted benzene ring
      - Number of compounds required for synthesis: 10,000
  - Correlate different kinds of biological activity
    - *In-vitro*, *in-vivo*
  - Elucidate the mechanism of new drugs
    - Which substituents/functional groups mainly affect the activity

Basic Ideas

Conditions: Y Observations (Dependent variable)

Objective: Correlate Y with X_1, X_2 …

Challenge: Variance is spread over X parameters

Find the QSAR signal in a huge field of variance!

"Variations in X_1,X_2 … that are correlated with changes in biological activity"

Considerations

- Quality of Y:
  "Strength of a model depends on the quality of the dependent variable"

- Choice of X:
  "Improper choice of independent"

- Overfitting:
  "With enough parameters, you can correlate anything with anything!"
  Ideal ratio: 1 independent variables to 5 molecules

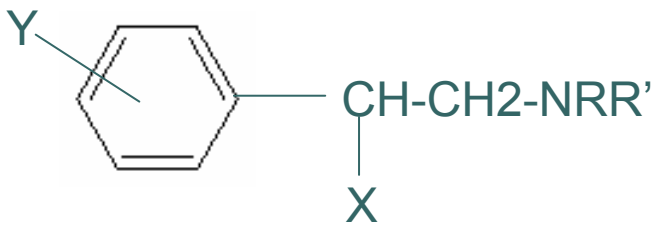
Hansch & Fujita Equation

- Relation of biological activity to more than one parameter
  - log(1/C) = 1.22 π - 1.59σ + 7.89

Electronic factor

Adrenergic blocking activity of β-halo-arylamines

Log(1/C) increases if the substituents has + π value and –σ value (Hydrophobic & Electron donating)



Other models include Free-Wilson, Combined model (Hansch and Free-wilson)

C. Hansch and T. Fujita, *J. Am. Chem. Soc.*, 1964, **86**, 1616

S.M. Free, Jr. and J.W. Wilson, *J. Med. Chem*, 1964, **7**, 395

Physicochemical Properties: Descriptors

- Hydrophobicity, Electronic & Steric
- Quantitative description of hydrophobicity is comparatively easy
  - Partition coefficients (log P) or π (hydrophobic)

$$P = \frac{\text{Concentration of drug in octanol}}{\text{Concentration of drug in aq. soln.}}$$

Hydrophobic (High P); Hydrophilic (Low P)

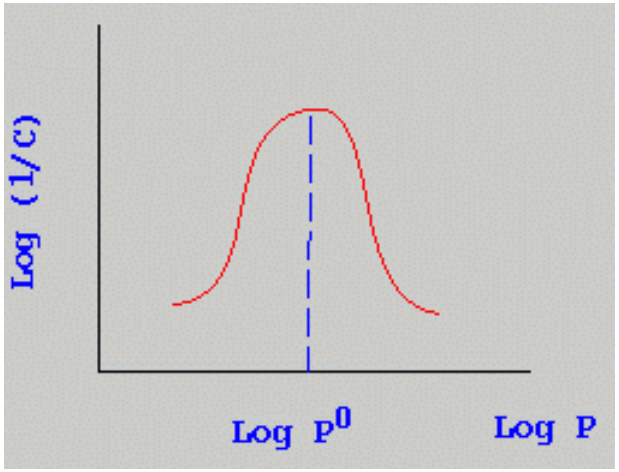
Binding of drugs to Serum Albumin

log(1/c) = 0.75 log P + 2.3 (based on 40 Compunds)

True for small ranges of P (1-4).

log (p)

Parabolic Curve



Does the log (p) increases infinitively?

What happens at higher P values?

Why do we see such as behavior?

Are there drugs which depend only on log P?

Yes, Anaesthetics which enter cell-membranes and affect the CNS activity- No drug-receptor interactions

General Equation;

$Y = (-)\ k_1(\log P)^2 + k_2 \log P + k_3$

$Y = -0.22\ (\log P)^2 + 1.04 \log P + 2.16$

Gl. Eq. for range of anaesthetic ethers

Can we extent it for any anaesthetics?

S. Ravichandran,Ph.D., ABCC, NCI-Frederick

Descriptors- log (P)

- It has been shown that any compound with a log P close to 2 can efficiently enter CNS and act as anaesthetic
- What would you do if you want your drug molecule to get stuck in the CNS?

Substituent Hydrophobicity constant (π)

π_{ali} Aliphatic Compunds

π_{Ar} Aliphatic Compunds

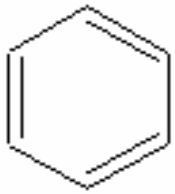
X	CH_3	t-Bu	OH	OMe	CF_3	Cl	Br	F
π_{ali}	0.5	1.68	-1.16	0.47	1.07	0.39	0.60	-0.17
π_{Ar}	0.52	1.68	-0.67	-0.02	1.16	0.71	0.86	0.14

- Log (p) values represent the hydrophobicity for the whole compound. To calculate log(p) we need to synthesize the molecules.
  - Is there anyway to quantify hydrophobic effects of functional groups?
- $\pi = \log P_X - \log P_H$

How does this solve the problem?

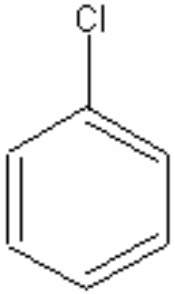
Still the lead compound log P has to be determined experimentally. Once it is known then the analogues can be calculated using π

Substituent Hydrophobicity constant (π)



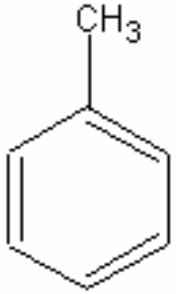
2.13

log(p)



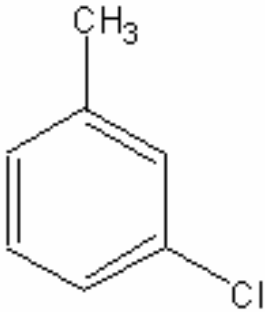
0.71

π_{Cl}



0.56

π_{Me}



3.40

log(p)

Log(p) Experimental value 3.28

Values are also available specifically for meta substituents

Hugo Kumbinyi" QSAR Parameters'

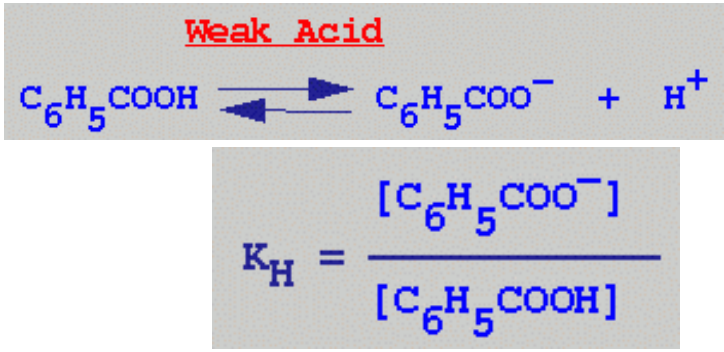
Quantum Mechanical Descriptors:

HOMO (Ionization Potential), LUMO (e-affinity) etc.

Hint about the mechanism

Electronic Effects

- Hammett substitution constant σ
- It is a measure of electron-withdrawing or electron donating ability of a substituent
- For substituent benzoic acids
  - $\sigma_X = K_X/K_H$



n doesn't vary from one Organic molecule to another

MR ∝ V (bulkness)

Steric Factors

Molar Refractivity (MR): Bulkness of the drug molecule

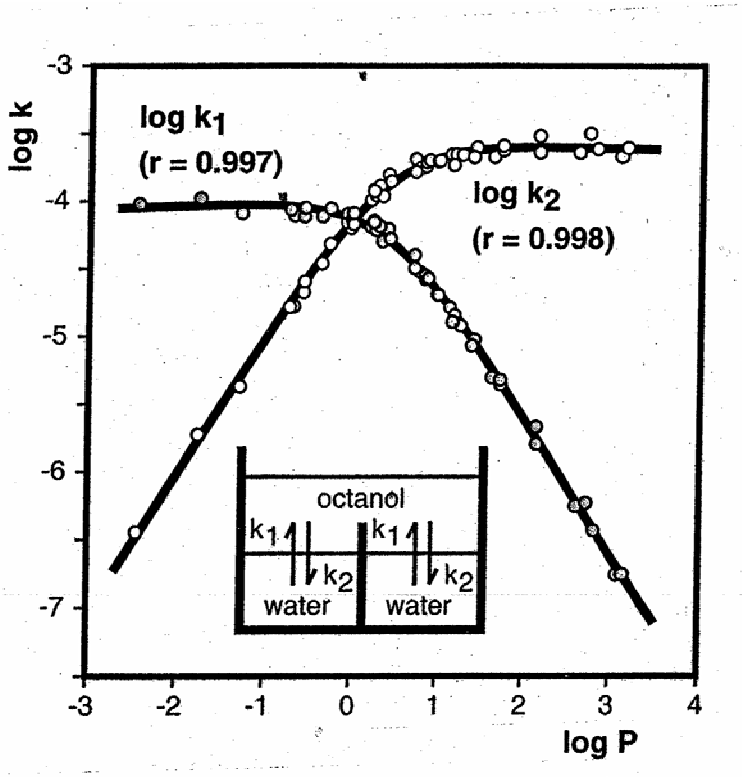
$$MR = [(n^2 - 1)/(n^2+2)] (MW/\rho)$$

n refractive index, MW Molecular Weight, ρ Density

Descriptors

- Topological: Calculated from 2D structure

Concepts of QSAR



Hugo Kubinyi, Encyclopedia of Computational Chemistry

- QSAR models are free-energy related, $\Delta G = -2.303\ RT \log K$
- Additivity of substituent group contributions to biological activity follows from many applications of Hansch analysis
- Activity is usually expressed in log[C] or log1/C. Why?
  - Biological data is often found to be skewed, log transformation makes the data to a normal distribution.

Principal Components Analysis

Case	ht.(x_1)	Wt(x_2)	Age(x_3)	Sbp(x_4)	Heart rate (x_5)
1	153	160	22	113	69
2	181	120	38	128	76
n	198	200	62	145	59

Height (ht) and Weight (wt.) could be related

Heart Rate and Systolic Blood Pressure (sbp) could be again related.

5 variables are reduced (projected) to 2 (PC1 and PC2) without loss of information

We can form two new variables (PC1, PC2), components

$PC1 = k_1x_1 + k_2 x_2$

$PC2 = k_3x_3 + k_4 x_4$

Another Example: 3D real object pictured as 2D objects. Reduction in dimensionality

Based on the lecture Multivariate Statistics, Manchester Metropolitan University

S. Ravichandran,Ph.D., ABCC, NCI-Frederick

Principal Components Analysis (PCA)

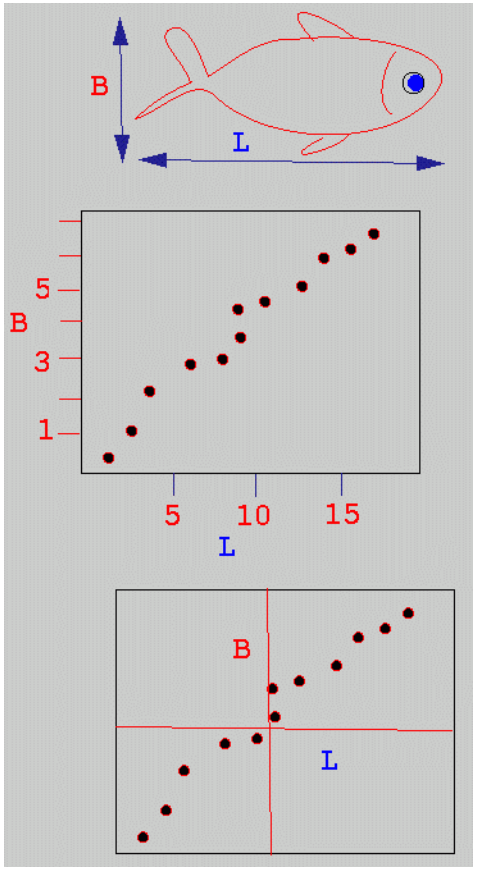
- Data Reduction Technique
  - It is not a regression method
  - Removes redundancies between descriptors
    - Correlation matrix
- What is PCA?
  - Linear combination of original descriptors explaining the maximum variance
  - PC's are orthogonal to each other

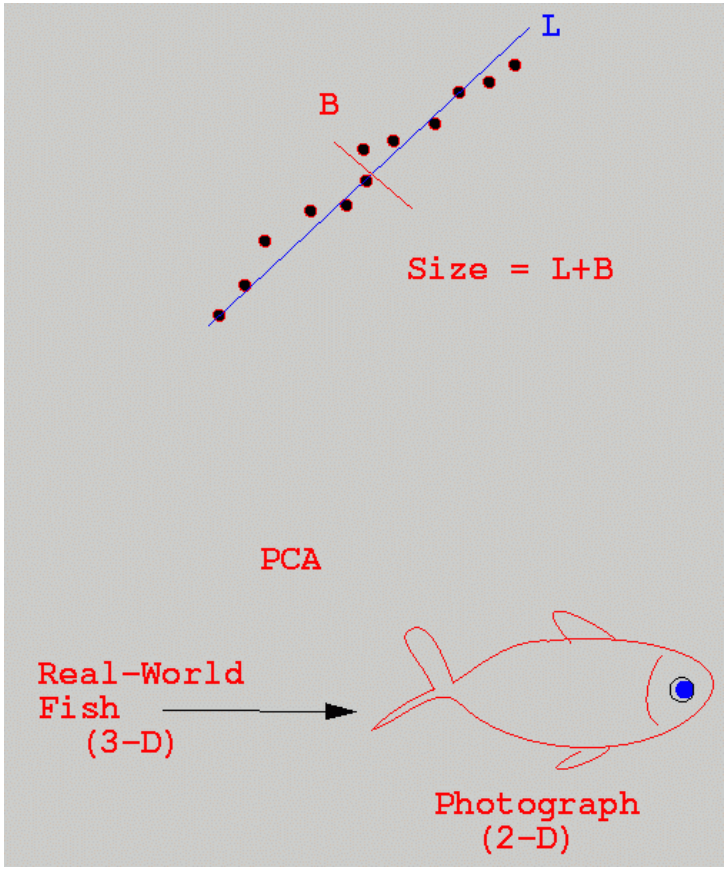
Pearson (1901) Hotelling (1933)

PCA

- $PC_1 = a_{1,1}x_1 + a_{1,2}x_2 + \ldots\ldots a_{1,n}x_n$
  $PC_2 = a_{2,1}x_1 + a_{2,2}x_2 + \ldots\ldots a_{2,n}x_n$
  $PC_3 = a_{3,1}x_1 + a_{3,2}x_2 + \ldots\ldots a_{3,n}x_n$
  Linear combination of original descriptors.

An example of PCA





Based on the Dept. of Biological Sciences, Manchester Metropolitan University

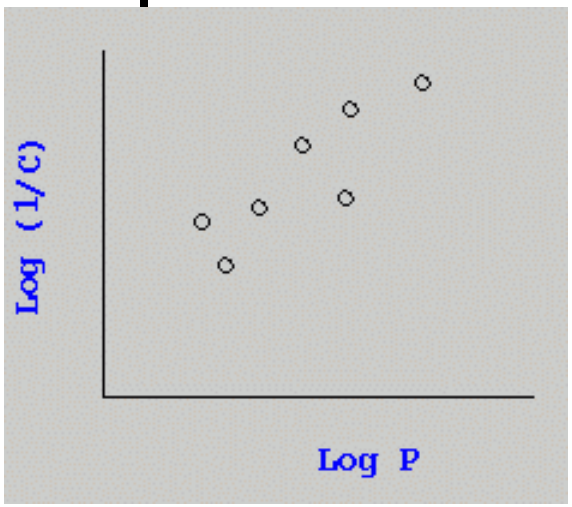
Clustering Algorithms

- Why Clustering?
  - To select meaningful subset of descriptors (Chemical Diversity)
    - There are more than 10^{10} ways to select 10 compounds out of 50

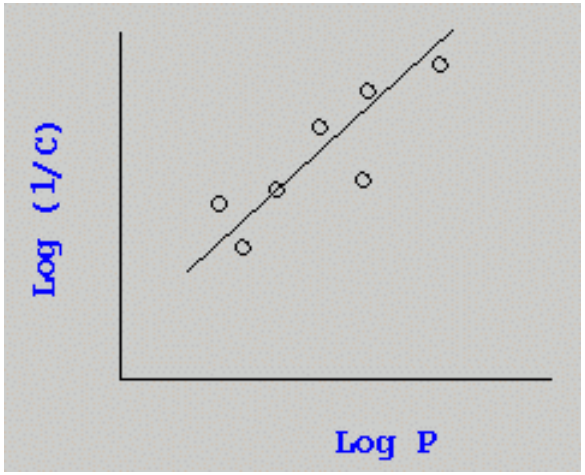
Regression Methods

- Under-Determined: MR, linear
  - Suitable for # molecules >> Descriptors
- Over-Determined: PCR, PLS, GA
  - Descriptors >> # molecules
    - Correlated descriptors

Simple Linear Regression

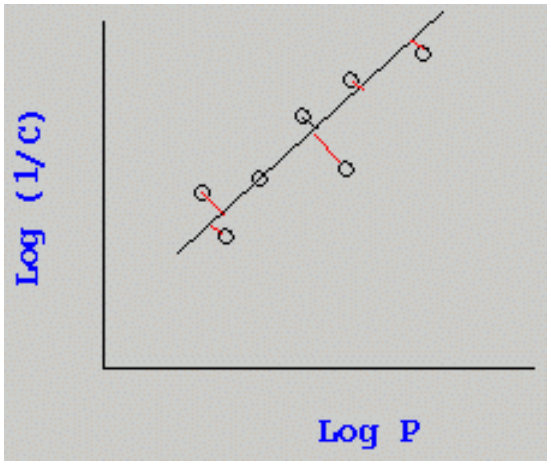


Idea is to see whether there is any relationship between x & y



In reality we can synthesize a series of compounds which can affect only Log (p)

Activity expressed as 1/C
C = concentration of a drug molecule required to produce expected level of biological activity

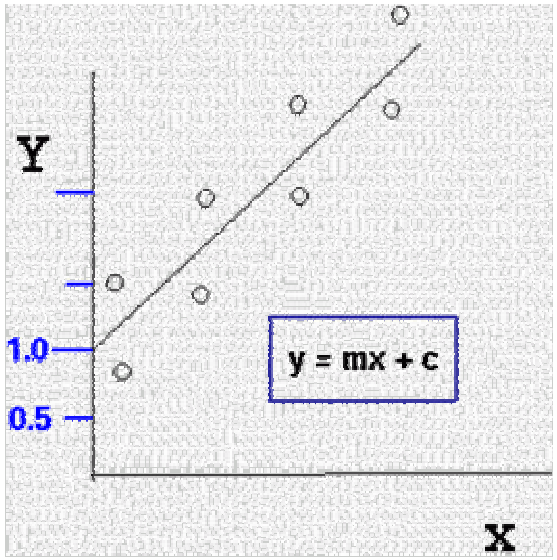


$Y = k_1x + k_2$

k_1 and k_2 are constants

Simple Linear Regression

- Assumes data is linear
- one equation for each descriptors (independent variable)
- Multiple descriptor interaction is ignored



Statistical Analysis

- R-squared (r2)
  - Is the correlation coefficient
  - Goodness of a fit

$$r^2 = 1 - \frac{\sum (Y_{obs} - Y_{pred})^2}{\sum (Y_{obs} - Y_{mean})^2}$$

= 1 perfect fit

= >= 0.95 good model

= <= 0.70 Poor model

Statistical Analysis

Multiple Linear Regression:

$Y = ax_1 + bx_2 + \ldots\ldots + \text{constant}$

- Multiple Linear Regression:
  - Tries to fit data to a multidimensional surface
- Quality of Linear Regression: r^2
- Quality of Multiple Linear Regression: R^2
- Another Quantity reported is F-Values, often called F-Statistics

Which descriptors to use in QSAR

- F-Value: Tells us how significant a variable is-It also controls when a variable will be added or removed from the QSAR equation.
- Forward-Stepping and Reverse-Stepping can help in the choice of the descriptors
- Reverse-Stepping regression starts with a QSAR equation using all descriptors. The one with smallest t statistic is removed.

Statistical Analysis

- Outliers: Residual values greater than 2 times the S.D. of the residuals generated in the validation procedure

- PREicted Sum of Squares:
  - $\sum(Y_i - Y_{Pred})^2$
  - Y_i is the actual value and Y_{pred} is the predicted value (Both are independent variables)
  - Good Model low PRESS values

- Cross-validated r^2 (xv-r2)

Cross-Validation

- Used to check the quality of a regression model (also called Jack-knifing)
- Repeats the regression many times on subsets of data
  - Procedure:
    - Typically starts by leaving one molecule out from the set
    - Build QSAR with the remaining set
    - Predict the activity of the left-out molecule
    - Repeat the procedure for all the molecules in the set

Cross-Validation

- Leave-one-out (LOO) or leave N out

$$PRESS = \sum(Y_i - Y_{Pred})^2$$

$$xv\text{-}r^2 = 1 - \frac{PRESS}{\sum(Y_{obs} - Y_{mean})^2}$$

$xv\text{-}r^2 = 1$ Perfect model

$0.4 < xv\text{-}r^2 < 1$ Predictive Model

$0.0 < xv\text{-}r^2 < 0.4$ Poor Model

Partial Least Squares (PLS)

- PLS expresses y in terms of linear combinations of the x variables

$$y = b_1t_1+b_2t_2+b_3t_3+\ldots..b_mt_m$$

The t's are also orthogonal

Where $t_1 \ldots t_m$ are latent variables

$$t_1 = c_{11}x_1+c_{12}x_2+\ldots.c_{1p}X_p$$
$$t_2 = c_{21}x_1+c_{22}x_2+\ldots.c_{3p}X_p$$
$$t_3 = c_{31}x_1+c_{32}x_2+\ldots.c_{3p}X_p$$

PLS can explain the variation in both X & Y. PCA can only explain the variation in X

PLS using Example: Toxicity of Halogenated Hydrocarbons

LD_{25}: Total toxicity measure, MR: Molar Refractivity, BP: Boiling Point. H_{vap}: Latent enthalpy of vaporization, ….

	LD_{25}	MR	logP	BP	H_{vap}	MW	d_{20}	n_{20}	q_c	q_{cl}	EInC	EInCl
	y	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11
1	0.96	16.56	1.25	40	7.57	85	1.33	1.42	0.097	-0.180	8.88	9.96
2	1.31	23.54	2.30	50	7.11	197	1.48	1.45	0.188	-0.100	9.72	10.04
3	1.45	21.43	1.97	61.7	7.50	119	1.48	1.37	0.180	-0.087	9.69	10.16
4	1.53	26.30	2.83	76.5	8.27	154	1.59	1.46	0.266	-0.066	10.55	10.36
5	2.26	26.05	2.29	86.5	8.01	131	1.46	1.46	0.117	-0.069	9.90	10.33
6	2.26	30.45	2.60	121	9.24	166	1.62	1.51	0.136	-0.068	10.08	10.34
7	2.42	30.92	2.66	146	9.92	168	1.59	1.49	0.137	-0.101	9.27	10.02

Dunn et al. 1984

PLS Using Example

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
1	0.432	0.32	0.442	0.387	0.190	0.326	0.327	-0.103	0.213	0.122	0.210
2	-0.085	0.217	-0.286	-0.176	0.283	0.232	0.048	0.72	0.093	0.415	0.098
3	0.127	0.098	0.034	-0.331	-0.106	-0.141	-0.504	-0.225	0.484	0.235	0.485

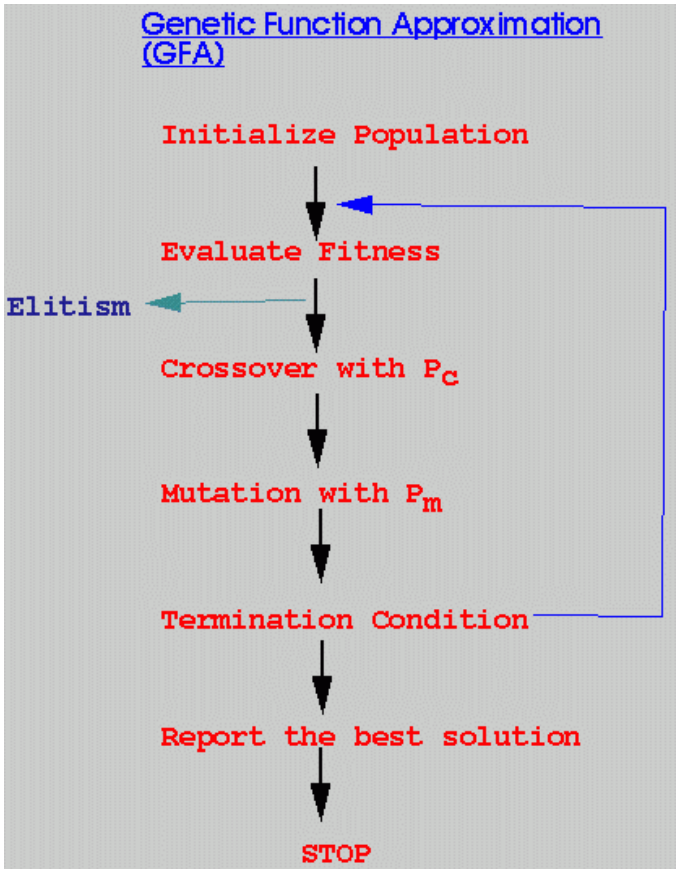
1) All variables contribute to the first component (higher weightings from molar refractivity (x_1), logP(x_2), BP(x_3), Latent Heat of Vaporization(x_4), d20(x_6) and n_{20}(x_7)—Combination of Steric, Hyrdophobic and Electronic Factors
2) Second variable gets contribution from electronic descriptors-Charge on C (x_8) and electro negativity of the Carbon (x_{10})

Applications of PLS

- Comparative Molecular Field Analysis: CoMFA Cramer *et al* 1988

1) Set of conformations of the molecule
2) They are overlaid in the binding site
3) Molecular fields surrounding each molecule are then calculated (probe and lattice method)
4) This is represented in a matrix form (each molecule: row; Each column contains the energy values at the grid points)
    1) N points in the grid, P probe then N X P Columns in the Matrix
5) Relation between Biological Activity (last column) and the field will then be determined

Genetic Algorithm



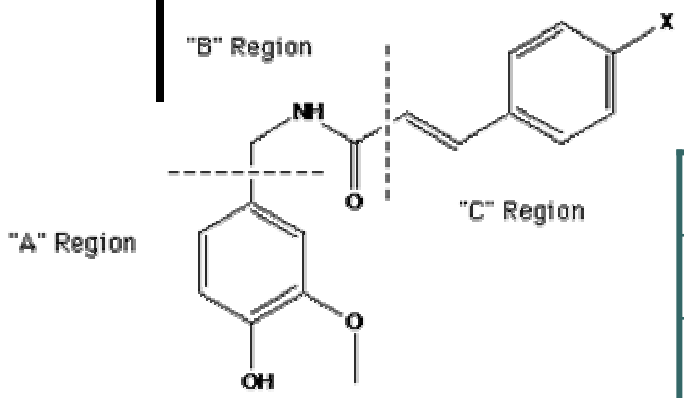
Effective in screening large sample space

QSAR equation will be provided in more than one models

Simple Genetic Algorithm Outline

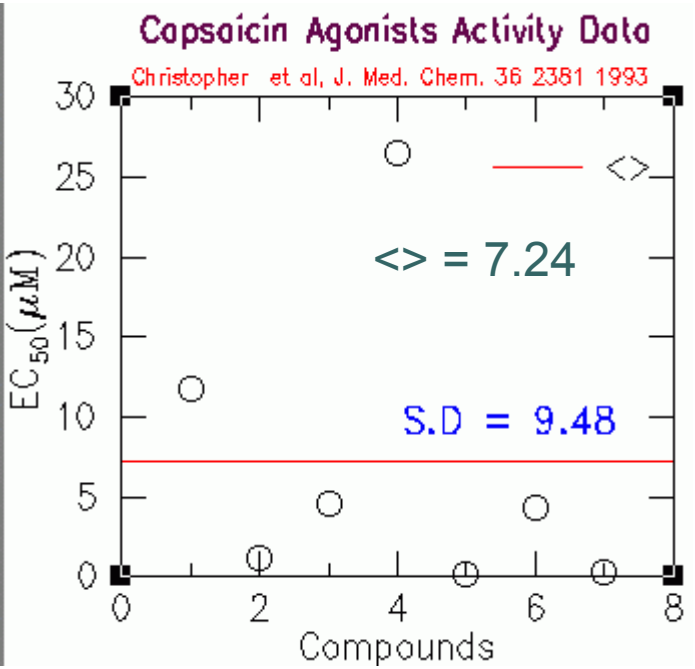
QSAR by example

EC_{50}: Molar concentration of an agonist which produces 50% of the max. possible response



Analogs of Capsaicin & activity tested in vitro assay by measuring Ca2+ influx into dorsal root ganglia neurons

#	Cmpd	X	EC_{50}(μM)
1	6a	H	11.8±1.9
2	6b	Cl	1.24±0.11
3	6d	NO_2	4.58±0.29
4	6e	CN	26.5±5.87
5	6f	C_6H_5	0.24±0.3
6	6g	$N(Me)_2$	4.39±0.67
7	6h	I	0.35±0.05
8	**6i**	**NHCHO**	**??**



Most Active

J. Med. Chem. 36, 2381 (1993)
Christopher *et al*

S. Ravichandran,Ph.D., ABCC, NCI-Frederick

05/27/04

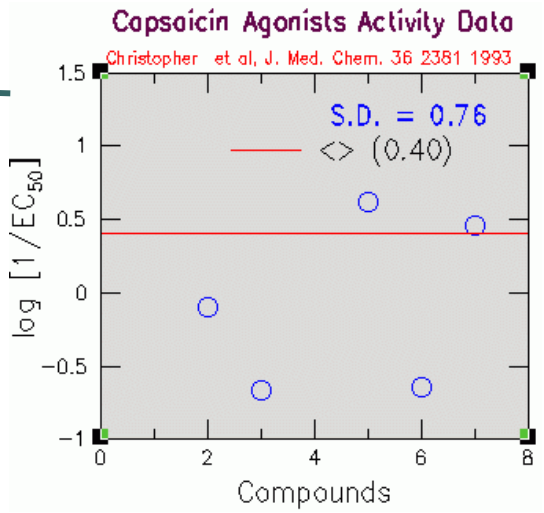
QSAR by example

- EC_{50} values cannot be used as such in QSAR (why? Because QSAR is based on free-energy relationships)
  - $\Delta G_0 = -2.3\ RT\ \log K = \log 1/[S]$ (ΔG_0 can be found proportional to $\log[1/\ EC_{50}]$
- EC_{50} are transformed to $\log[1/\ EC_{50}]$
  - This transformation also helps us in getting back the normal distribution (another assumption of regression analysis)
- Note the transformed data gets uniformly distributed

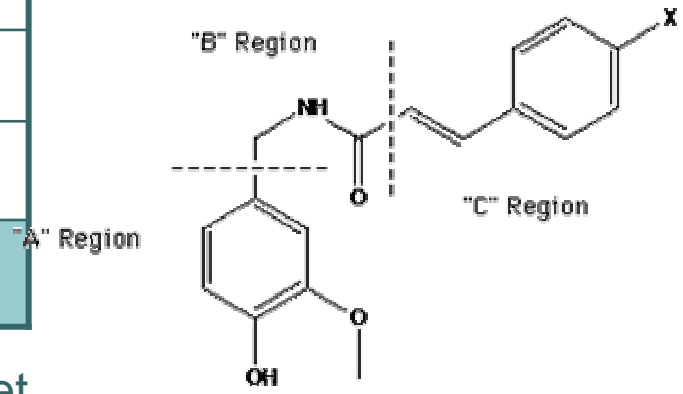
QSAR by example: Data Transformation

Std. Dev. Is error associated with the prediction

#	Cmpd	X	Y=EC_{50}(μM)	Log(y)	Log(1/y)
1	6a	H	11.8±1.9	1.07	-1.07
2	6b	Cl	1.24±0.11	0.09	-0.09
3	6d	NO_2	4.58±0.29	0.66	-0.66
4	6e	CN	26.5±5.87	1.42	-1.42
5	6f	C_6H_5	0.24±0.3	-0.62	-0.62
6	6g	$N(Me)_2$	4.39±0.67	0.64	-0.64
7	6h	I	0.35±0.05	-0.46	0.46
8	**6i**	**NHCHO**	**?? ±??**	**??**	**??**



Uniform Distribution of data



One can guess the activity for 6i as 0.40 <> of the set

S. Ravichandran,Ph.D., ABCC, NCI-Frederick

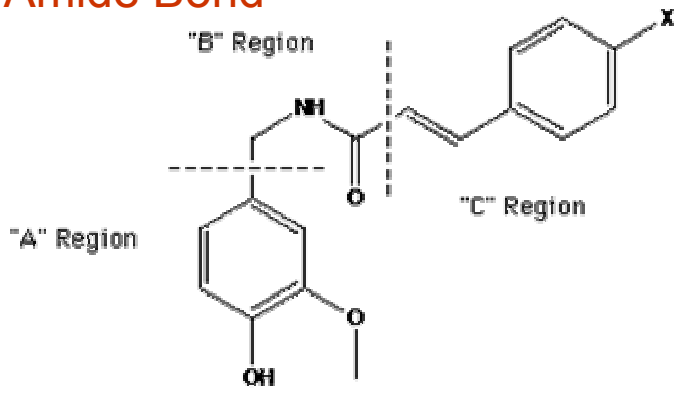
05/27/04

QSAR by example

- In real situations, we will have more compounds & Descriptors
- What is the relation between the descriptor and activity?
  - Why do we care?
    - If the association is strong, then the chances of activity prediction is highly probable.
    - What happens if the descriptor do not show any variation?
- Also relevant descriptors have to be chosen
  - If there reactivity increases with bulky substituents, it is imperative to include descriptors such as MR etc.

QSAR by Example

Amide Bond



Aromatic Ring

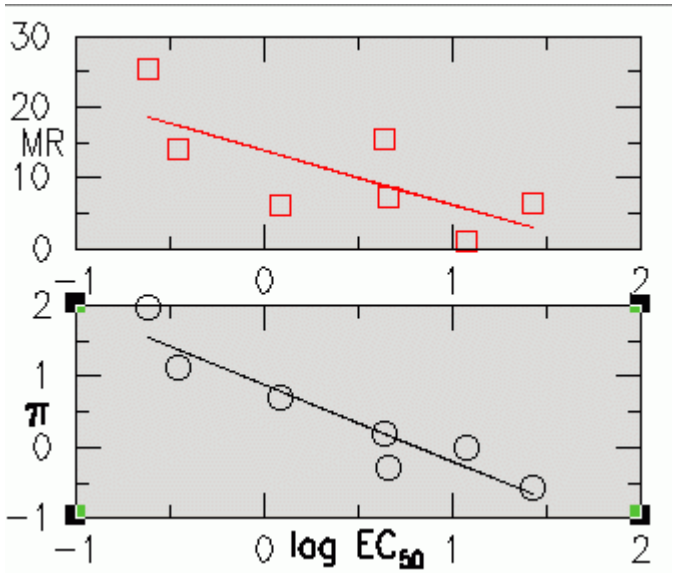
Hypothesis: Small Hydrophobic would increase activity

Sandoz *et al* based on available results assumed that a small hydrophobic substituent would increase activity (ie.)
Ideal choice would be π (hydrophobic) and Size (MR)

QSAR by example

How π and MR related to Activity?

	Cmpd	X	Log(EC_{50})	π	MR
1	6a	H	1.07	0.00	1.03
2	6b	Cl	0.09	0.71	6.03
3	6d	NO_2	0.66	-0.28	7.36
4	6e	CN	1.42	-0.57	6.33
5	6f	C_6H_5	-0.62	1.96	25.36
6	6g	$N(Me)_2$	0,64	0.18	15.55
7	6h	I	-0.46	1.12	13.94
8	**6i**	**NHCHO**	**??**	**??**	**??**



Based on the linear relationships a model was proposed

Log EC_{50} = 0.764 – (0.817) π

QSAR by Example: How good is Model 1?

	Cmpd	X	Log(EC_{50})	π	Log(EC_{50})$_{Calc}$	Residual
1	6a	H	1.07	0.00	0.79	0.28
2	6b	Cl	0.09	0.71	0.21	-0.12
3	6d	NO_2	0.66	-0.28	1.02	-0.36
4	6e	CN	1.42	-0.57	1.26	0.16
5	6f	C_6H_5	-0.62	1.96	-0.81	0.19
6	6g	$N(Me)_2$	0,64	0.18	0.65	-0.01
7	6h	I	-0.46	1.12	-0.12	-0.34
8	**6i**	**NHCHO**	**??**	**??**	**1.60**	**??**

How to quantify the error?

Is this model a improved one?

Yes, because the S.D is less than previous

0.28 < 0.76

$(S.D)^2 = (S)^2 = [0.28^2 + (-0.12)^2 + \ldots\ldots(-0.34)^2]/(7-2)$ and $S = \text{sqrt}[S^2] = 0.28$

S. Ravichandran,Ph.D., ABCC, NCI-Frederick

05/27/04

QSAR by example

- Calculated r^2 value is 0.89

  (regression variance/original variance)

- One cal also calculate F-ratio

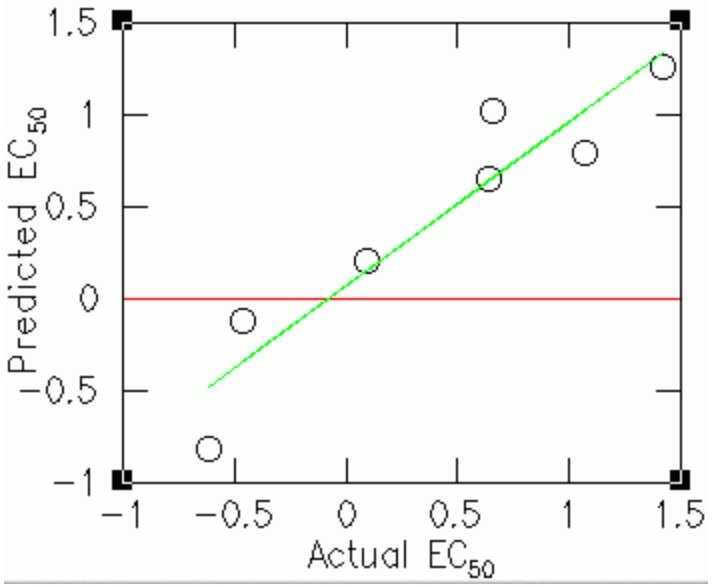
  $F_{1,n} = (n-2)\,[r^2/(1-r^2)] = 40.46$

- One can also form another model by including MR

  Log EC50 = 0.762 – (0.819)π + (0.011) MR

  (obtained by Multiple Regression)

Note: MR coefficient is negligible that shows probably steric bulk is not an important factor in determining the activity



Basic Books/Articles

- **Molecular Modelling: Principles and Applications**, 2nd Edition, Andrew R. Leach
- **Hugo Kumbinyi, QSAR in drug design, Encyclopedia of Computational Chemistry, 5 Volume Set** by Paul Von R. Schleyer (Editor), Paul Von Rague Schleyer
- **An introduction of QSAR Methodology**, Allen B. Richon and Stanley S. Young, Network Science (1997)
- **Data Analysis for Chemists**, David Livingston, Oxford University Press (1995)