QSAR: Quantitative structure activity relationship

CADD.....

CADD is capable of increasing the hit rate of novel drug compounds because it uses a much more targeted search than traditional HTS and combinatorial chemistry.

It not only aims to explain the molecular basis of therapeutic activity but also to predict possible derivatives that would improve activity. In a drug discovery campaign, *CADD is usually used for three major purposes*:

(1) filter large compound libraries into smaller sets of predicted active compounds that can be tested experimentally;

(2) guide the optimization of lead compounds, whether to increase its affinity or optimize drug metabolism and pharmacokinetics (DMPK) properties including ADMET;

(3) Design novel compounds, either by "growing" starting molecules one functional group at a time or by piecing together fragments into novel chemotypes.

In silico screening……

VIRTUAL SCREENING (CADD) OF DATABASES

ONE OF THE WIDELY USED APPROACHES FOR REDUCING THE SIZE OF HAYSTACK IS TO FILTER OUT UNDESIRED MOLECULES USING COMPUTATIONAL APPROACHES

In silico screening……

Methods for virtual Screening of compounds:

It is seen as complementary approach to experimental screening (HTS), and when coupled with structural biology, promises to increase the number, and enhance the success of projects in the lead identification of stage of the drug discovery process.

•Ligand-based screening: The strategy uses the information provided by a compound or set of compounds that are known to be active and to use this to identify other compounds in any databases. This can be performed by one of the followings: Similarity and substructure searching, phamacophore matching or 3D shape matching.

•Structure-based screening: When the structure of the target protein is known, receptor-based computational methods can be employed. These involve explicit molecular docking of each ligand into the binding site of the target, producing a predicted binding mode for each data base compound together with a measure of the quality of the fit in the binding site. This information is then used to rank the compounds with a view to compounds for biological testing.

Databases comprise millions of individual molecular entries

- Biomolecular databases: genes and proteins
- Databases of organic molecules
- Databases of biological/therapeutic/disease targets
- Data bases of natural products

Scientists are now trying to integrate both experimental and computational approaches towards addressing two major challenges of pharmaceutical research, that is, discovery of drugs (or leads) and their targets. (A drug discovery program may or may not be guided by a biological target but when it is without a target then one has to establish its mechanism.)

Two options that are generally used : Either virtual screening of compounds or targets

VS of compounds

Several million compounds (Databases)

↓ In-silico screening against a specified biological target

Few compounds

VS of targets

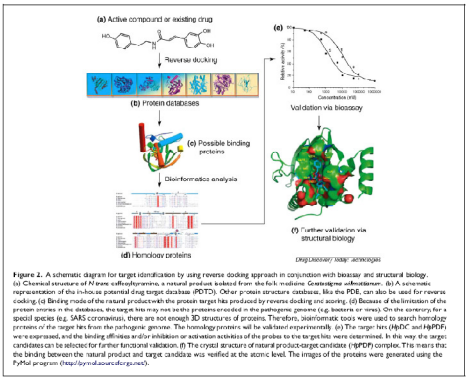


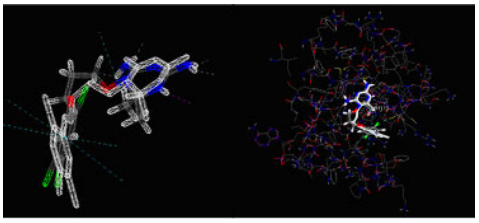
Figure 2. A schematic diagram for target identification by using reverse docking approach in conjunction with bioassay and structural biology. (a) Chemical structure of N-trans-caffeoylnormine, a natural product isolated from the folk medicine Gnetum gnemon var. tenerum. (b) A schematic representation of the in-house potential drug target database (PDTD). Other protein structure databases, like the PDB, can also be used for reverse docking. (c) Binding mode of the natural product with the protein target has produced by reverse docking and scoring. (d) Because of the limitation of the protein entries in the databases, the target hits may not be the proteins encoded in the pathogenic genome (e.g. bacteria or virus). On the contrary, for a special species (e.g. SARS coronavirus), there are not enough 3D structures of proteins. Therefore, bioinformatic tools were used to search homology proteins of the target hits from the pathogenic genome. The homology proteins will be validated experimentally. (e) The target hits (HpDC and HpPDF) were overexpressed, and the binding affinities and/or inhibition or activation activities of the probes to the target hits were determined. In this way, the target candidates can be selected for further functional validation. (f) The crystal structure of natural product-target candidate (-HpPDF) complex. This means that the binding between the natural product and target candidate was verified at the atomic level. The images of the proteins were generated using the PyMol program (http://pymol.sourceforge.net/).

World-wide In Silico Docking On Malaria (WISDOM)

WISDOM simulated the docking of 31 million molecules against target proteins on the malarial parasite, using the equivalent of 80 years CPU time in six weeks.

•testing up to 150,000 docked compounds per hour on 3,000 computers around the world in 15 countries
•identified three preclinical molecules that inhibit the haemoglobin breakdown

The strategy demonstrated how grid computing can be used to accelerate drug discovery research, by speeding up the virtual screening process and reducing the cost of developing new drugs.



Left: Structure of a potential antimalarial drug. Right: A simulation of the drug binding to a protein from the malaria parasite.

Rational approach in Drug design

Quantitative Structure Activity Relationships (QSAR)

QSAR concept introduced in 1964 was implemented in computers and constituted first generation rational approach to drug design

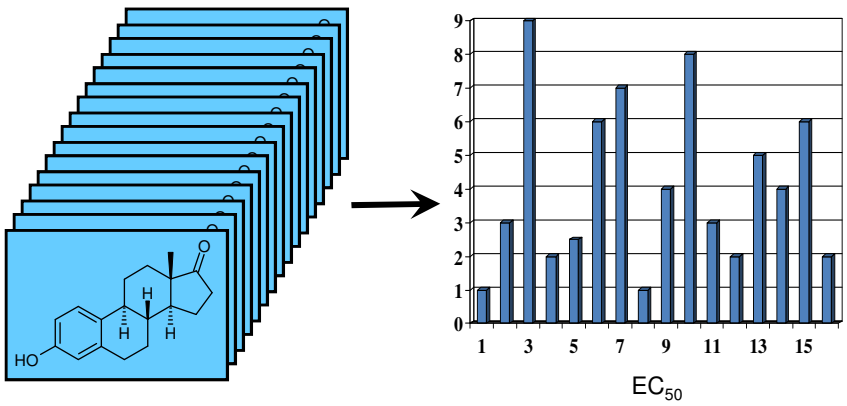
If we can understand how a molecular structure brings about a particular effect in a biological system, we have a key to unlocking the relationship and using that information to our advantage.

If we take a series of chemicals and attempt to form a quantitative relationship between the biological effect (i.e. the activity) and the chemistry (i.e. the structure) of each of the chemicals, then we are able to form a quantitative structure-activity relationship or QSAR.

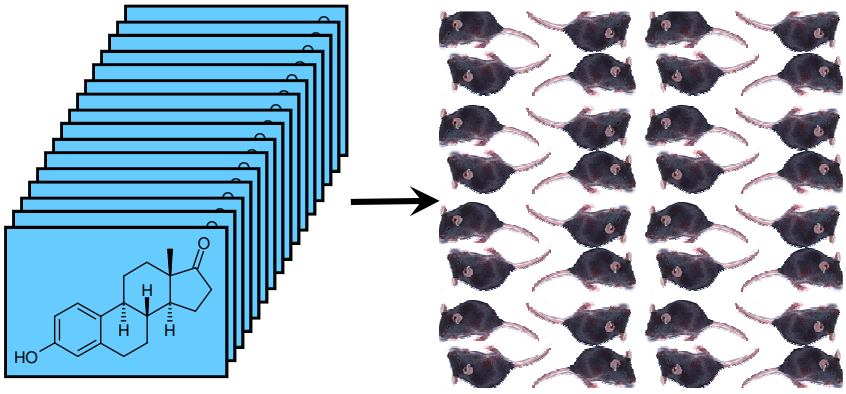
QSAR: The Setting

Quantitative structure-activity relationships are used when there is little or no receptor information, but there are measured activities of (many) compounds

From Structure to Property



From Structure to Property



LD_{50}

QSAR: Which Relationship?

Quantitative structure-activity relationships* correlate chemical/biological activities with structural features or atomic, group or molecular (physico-chemical) properties.

**(within a range of structurally similar compounds)*

Rationale for QSAR

In drug design, in vitro potency addresses only part of the need; a successful drug must also be able to reach its target in the body while still in its active form.

The in vivo activity for substance is a composite of many factors:

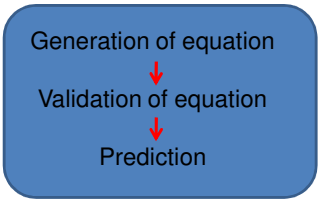
- Intrinsic reactivity of the drug
- Its solubility in water
- Its ability to cross blood brian barrier
- Its nonreactivity with nontarget molecules
- others

QSAR and mathematics

A quantitative SAR correlates measurable or calculable physical or molecular properties to some specific biological activity in terms of equation.

Once a valid QSAR has been determined, it should be possible to predict the biological activity of related drug candidates before they are put through expensive and time-consuming biological testing.

In some cases only computed values need to be known to make an assessment.



QSAR equations have been used to describe thousands of biological activities within different series of drugs and drug candidates.

Especially enzyme inhibitions data have been successfully correlated with physico-chemical properties of the ligands.

In certain cases, where X-ray structure of proteins became available, the results of QSAR regression models could be interpreted with the additional information from the three-dimensional (3D) structures.

Free Energy of Binding and Equilibrium Constants

The free energy of binding is related to the reaction constants of ligand-receptor complex formation:

$$\Delta G_{binding} = -2.303\ RT \log K$$

$$= -2.303\ RT \log (k_{on} / k_{off})$$

Equilibrium constant K

Rate constants k_{on} (association) and k_{off} (dissociation)

Basic Assumption in QSAR

The structural properties of a compound contribute in a linearly additive way to its biological activity provided there are no non-linear dependencies of transport or binding on some properties

Free Energy of Binding

$$\Delta G_{binding} = \Delta G_0 + \Delta G_{hb} + \Delta G_{ionic} + \Delta G_{lipo} + \Delta G_{rot}$$

ΔG_0	entropy loss (translat. + rotat.)	+5.4
ΔG_{hb}	ideal hydrogen bond	–4.7
ΔG_{ionic}	ideal ionic interaction	–8.3
ΔG_{lipo}	lipophilic contact	–0.17
ΔG_{rot}	entropy loss (rotat. bonds)	+1.4

(Energies in kJ/mol per unit feature)

History of QSAR

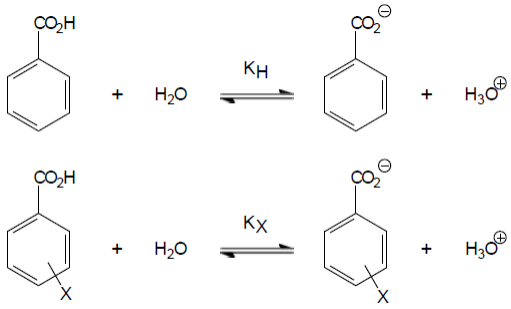
- Free- Wilson Analysis
- Hansch Analysis

- QSARs are the mathematical relationships linking chemical structures with biological activity using physicochemical or any other derived property as an interface.
- Crum-Brown and Fraser **in (1868)** expressed the idea that the physiological action of a substance in a certain biological system(Φ) was a function (f) of its chemical composition and constitution (C).

  $\Phi = f\ C$ Equation
- Thus, an alteration in chemical constitution, ΔC, would be reflected by an alteration in biological activity ΔΦ.
- Richardson **(1868)** expressed the chemical structure as a function of solubility.
- Richet **(1893) Correlated** toxicities of a set of alcohols, ethers and ketones with aqueous solubility and showed that their cytotoxicities are inversely related to their corresponding water solubilities.

Hammett equation

The seminal work of Hammett **(1935, 1937) gave rise to the σ-ρ** culture correlated the effect of the addition of a substituent on benzoic acid with the ionization constant, postulated electronic sigma-rho constants and established the linear free energy relationship (LFER) principle.



$$\log \frac{K_X}{K_H} = \rho \log \frac{K_X}{K_H} = \rho\,\sigma$$

Thus Hammett equation relates to

- the relative magnitude of the equilibrium constants to a reaction constant ρ and a substituent constant σ.
- Observed **changes in equilibrium or rate constants** to systematic changes in substituents that govern electron donating/withdrawing ability.
- It is an example of a linear free-energy relationship as changes in log K (log k) are linear with substituent effects.
- Effect of substitutent on acidity (ΔpKa) of benzoic acid in a similar manner as it would effect other reactions.
- Quantification of effect of substituents on any reaction by defining an empirical electronic substituent parameter(σ), which is derived from the acidity constants, Ka's of substituted benzoic acids

For the chemical equilibrium;

$$R\text{-}CO_2H + H_2O \text{ ==== } H_3O^+ + R\text{-}CO_2^-$$

$$K_a = \frac{[H_3O^+][RCO_2^-]}{[RCO_2H]}$$

Thus, when $[RCO_2H] = [RCO_2^-]$;

$$K_a = [H_3O^+]$$

and $pK_a = pH$

e.g. For the ionization of benzoic acid in pure water at 25oC (the reference reaction), the constant ρ is defined as 1.00. Thus, the electronic substituent parameter (σ) is defined as:

$$\sigma = \log (K_X / K_H)$$

Reaction constant ρ

The reaction constant is a measure of how sensitive a particular reaction is to changes in **electronic effects** of substituent groups . The reaction constant depends on the

- nature of the chemical reaction
- reaction conditions (solvent, temperature, etc)

Both the sign and magnitude of the reaction constant are indicative of the extent of charge build up during the reaction progress.

Reactions with $\rho > 0$ are favoured by electron withdrawing groups (i.e., the stabilization of negative charge).

Those with $\rho < 0$ are favoured by electron donating groups (i.e., the stabilization of positive charge).

The greater the magnitude of ρ, the more sensitive the reaction is to electronic substituent effects.

Substituent constant σ

This pertains to the observed electronic **(inductive and resonance)** effect that a particular substituent imparts to a molecule. E.g the rate of reaction is 10^5 times slower when NO_2 than when CH_3

- Electron withdrawing substituents will have a positive σ value
- Electron donating substituents will have a negative substituent constant:

(e.g., $\sigma_{para}(NO_2) = 0.78$, $\sigma_{para}(OCH_3) = -0.27$)

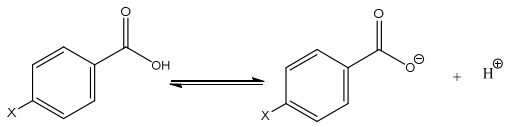
Resonance contributions can only occur for ortho and para substituents, ortho substituents are excluded from the Hammett treatment because steric effects play a complicating role.

Meta substituents will have a negligible resonance contribution (σR=0) and are almost entirely due to inductive effects (σmeta=σI).

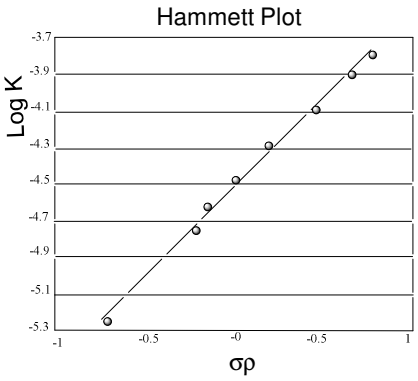
Inductive effects arise as a result of electronegativity differences and diminish with the distance between the substituent and the reactive centre.

Thus, σI(meta) >> σI(para) substitution, because of their closer proximity.

Example of Hammett plot

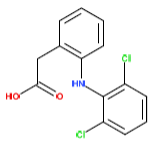


Substituent	σρ	Log K
NH2	-0.66	-5.25649
OMe	-0.27	-4.82391
Me	-0.17	-4.63827
H	0.00	-4.46852
Cl	0.23	-4.25964
COCH3	0.5	-4.05552
CN	0.66	-3.89279
NO2	0.78	-3.77989

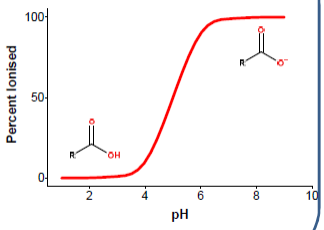


Effect of pKa on Aqueous Solubility

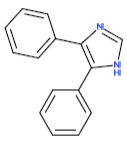
Diclofenac – pK_a 4.08 (acidic)



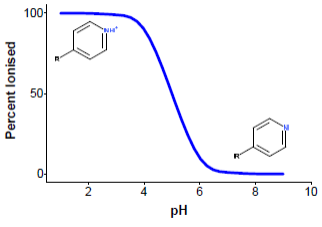
pH 3	pH 5	pH 7.4	pH 9
17 uM	>350 uM	>350 uM	>350 uM



4,5-diphenylimidazole – pK_a 5.88 (basic)



pH 3	pH 5	pH 7.4	pH 9
> 350 uM	> 350 uM	84 uM	63 uM



Although (Bell and Roblin 1942) studies examining effects of ionization/electron distribution on biological (antibacterial) activities were carried out on a series of synthetic compounds (sulfanilamides) in terms of their ionizations, following were drawbacks: Steric and hydrophobcity factors.

Taft **(1952) Postulated a method for separating polar, steric, and resonance** effects and introduced the first steric parameter, ES and subsequently Hansch and Muir **(1962) Correlated the** biological activities of plant growth regulators with Hammett constants and **hydrophobicity**.

Hansch Analysis

Thus, yet another parameter hydrophobicity as a new scale was combined with Hammett equation.

Using the octanol/water system, a whole series of partition coefficients were measured, and thus a new hydrophobic scale was introduced. The parameter π, which is the relative hydrophobicity of a substituent, was defined in a manner analogous to the definition of sigma.

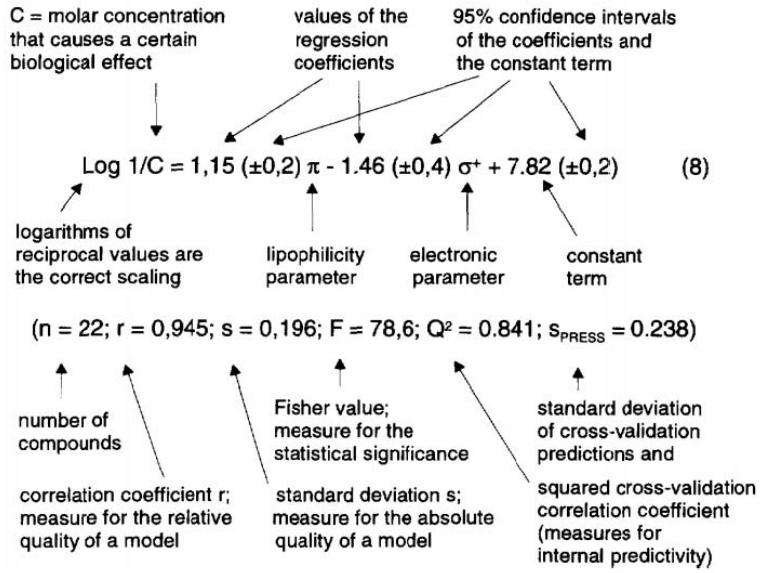
$$\Pi_X = \log P_X - \log P_H$$

PX and PH represent the partition coefficients of a derivative and the parent molecule, respectively.

This laid the basis for the development of the QSAR paradigm by Hansch and Fujita **(1964), which combined the hydrophobic constants with** Hammett's electronic constants to yield the linear Hansch equation and its many extended forms.

$$\text{Log } 1/C = a\,\sigma + b\,\pi + ck$$

QSAR Equation



Due to the curvilinear, or bilinear, relationship between log1/C50 and hydrophobicity normally found in single dose tests the quadratic π2 term was later introduced to the model. Hansch **(1969) Developed the parabolic** Hansch equation for dealing with extended hydrophobicity ranges.

$$\text{Log } 1/C = -\,a\,(\log P)^2 + b.\log P + c\,\sigma + k$$

C=minimum effective dose, P= n-octanol/water partition coefficient; σ = Hammett electronic parameter;
a, b, c = regression coefficient; k =constant term.

In summary Hansch analysis comprises affinities of ligands to their binding sites, inhibition constants, rate constants, and other biological end points, with atomic, group or molecular properties such as lipophilicity, polarizability, electronic and steric properties.

Drug transport *and* binding affinity depend nonlinearly on lipophilicity

Log P

Log P is a measure of the drug's hydrophobicity, which was selected as a measure of its ability to pass through cell membranes.

The log P value reflects the relative solubility of the drug in octanol (representing the lipid bilayer of a cell membrane) and water (the fluid within cell and in blood).Thus partition coefficient is the ratio of concentrations of a compound in the two phases of ammixture of two immiscible solvents at equilibrium

$$P = \frac{\text{Concentration of drug in organic phase}}{\text{Concentration of drug in aqueous phase}}$$

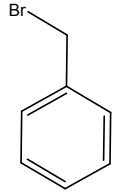
Hydrophobic compounds will have a high P value, whereas hydrophilic compounds will have a low P value. The substituent hydrophobicity constant is a measure of how hydrophobic a substituent is, relative to hydrogen. A positive value of π indicates that the substituent is more hydrophobic then hydrogen. A negative value indicates that the substituent is less hydrophobic.

Calculating Log P

$$\text{Log P} = \text{Log } K_{(o/w)} = \text{Log})[X]_{octanol}/[X]_{water}$$

Most programs use a group additivity approach:

1 Aromatic	0.780
7 H's on C	1.589
1C-Br	-0.120
1 alkyl C	0.195
Sum 2.924 = calc Log P	



Some use more complicated algorithms, including factros such as the dipole moment, molecular size and shape.

Log P values can be measures experimentally or more commonly calculated.

Hansch Analysis

+ Fewer regression coefficients needed for correlation
+ Interpretation in physicochemical terms
+ Predictions for other substituents possible

Free Wilson analyses

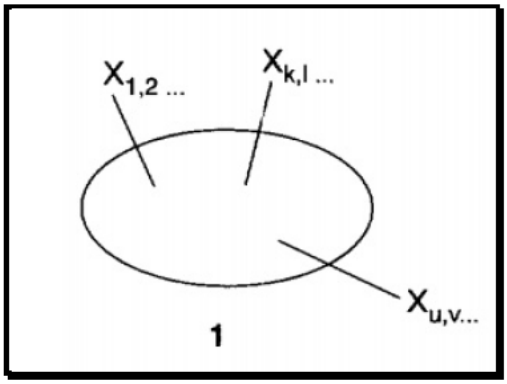
In 1964, Free and Wilson derived a mathematical model that describes the presence and absence of certain structural features i.e. those groups that are chemical modified, coded by values of 1 and 0 and correlates the resulting structural matrix with biological activity values.

- $\text{Log } 1/C = \Sigma a_i + \mu$

  where C=predicted activity,

  a_i= contribution per group, and

  μ=biological activity of reference compound

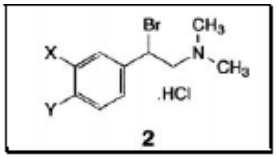


A common skeleton bears substituents X_i in different position p; the presence or absence of these substituents is coded by the values 1 and 0 respectively.

Equation below describes the antiadrenergic activities for 22 different *m-, p- and m,p-disubstituted analogs of the N,Ndimethyl-* α-bromophenylamine 2, where C is the concentration that causes a 50% reduction of the adrenergic effect of a certain epinephrine dose.

log 1/C = – 0.301 (±0.50) [*m-F] + 0.207 (±0.29) [m-Cl]* +0.434 (±0.27) [*m-Br] + 0.579 (±0.50) [m-I] + 0.454 (±0.27)*[*m-Me] +0.340 (±0.30) [p-F] + 0.768 (±0.30) [p-Cl]* + 1.020 (±0.30) [*p-Br] +1.429 (±0.50) [p-I] + 1.256 (±0.33)*[*p-Me] + 7.821 (±0.27)*

*(n = 22; r = 0.969; s = 0.194;*F = 16.99)



N,N-dimethyl-α-bromophenylamines (X, Y = H, F, Cl, Br, I, Me).

Where n = number of compounds;
r = correlation coefficient, measure for the relative quality of a model;
s = standard deviation, measure for the absolute quality of a model;
F = fisher value, measure for statistical significance;
C = molar concentration that causes a certain biological effect.

The main advantage of Free Wilson analysis: only the biological activity values and the chemical structure of the compounds need to be known to derive a QSAR model.

Nevertheless, Free Wilson analysis is often used to see at a glance which physicochemical properties might be important for the biological activity. In this data set, it can be easily concluded from equation that:

•Biological activities increase with increasing lipophilicity (F to Cl, Br and I);

•Biological activities increase with electron donor properties (methyl has larger group contributions than the equi-lipophilic Cl);

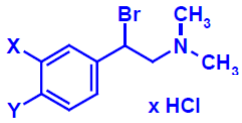
meta-substituents have lower group contributions than para-substituents.

Free-Wilson Analysis

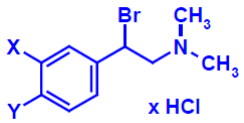
+ Computationally straightforward
– Predictions only for substituents already included
– predictions those are too far outside the range of investigated parameters, such as for *tert-Bu or -OH or* -SO2NH2 will most probably fail because of narrow chemical relationship among the investigated substituents and the very different nature of these chemical groups, in size or in their hydrogen bond donor and acceptor properties.
– Requires large number of compounds

Hansch Equations

log 1/C = 1.151 (±0.19) π - 1.464 (±0.19)
(n = 22; r = 0.945; s = 0.19



log 1/C = 1.259 (±0.19) π - 1.460 (
+ 0.208 (±0.17) E_s^{meta} + 7.6
(n = 22; r = 0.959; s = 0.173; F = 69.24)



Free Wilson Equation

log 1/C = - 0.301 (±0.50) [m-F] + 0.207 (±0.29) [m-Cl]
+ 0.434 (±0.27) [m-Br] + 0.579 (±0.50) [m-I]
+ 0.454 (±0.27) [m-Me] + 0.340 (±0.30) [p-F]
+ 0.768 (±0.30) [p-Cl] + 1.020 (±0.30) [p-Br]
+ 1.429 (±0.50) [p-I] + 1.256 (±0.33) [p-Me]
+ 7.821 (±0.27)
(n = 22; r = 0.969; s = 0.194; F = 16.99)

The Free Wilson model is a simple and efficient method for the quantitative description of structure activity relationships. It is the only numerical method which directly relates structural features with biological properties,

in contrast to Hansch analysis, where physicochemical properties are correlated with biological activity values.

Nevertheless both approaches are closely interrelated, not only from a theoretical point of view, but also in their practical applicability.

In many cases both models can be combined to a mixed approach which includes Free Wilson type parameters to describe the activity contributions of certain structural modifications and physicochemical parameters to describe the effect of some other substituents on the biological activity.

$$\log 1/C = a\,(\log P)^2 + b \log P + c\sigma + \ldots + \sum a_i + k \quad \ldots..(1$$

General Scheme of a QSAR Study

The chemoinformatics methods used in building QSAR models can be divided into three groups:

- Extracting descriptors from molecular structure,
- Choosing those informative in the context of analyzed activity,
- Finally using the values of the descriptors as independent variables to define a mapping that correlates them with the activity in question.

Descriptors

Physicochemical or any other property used for generating QSARs is termed as Descriptors and treated as independent variable.

Descriptors (Molecular properties)

1. density
2. Ionization energy
3. $H_{vaporization}$
4. Molecular weight
5. $H_{Hydration}$
6. Log P (Lipophilicity)
7. pKa
8. Dipole moment
9. Reduction potential
10. molecular volume
11. surface area
12. Polarizability
13. LUMO/HOMO energy

Selection of Descriptors

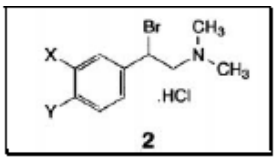
1. What is particularly relevant to the therapeutic target?
2. What variation is relevant to the compound series?
3. What property data can be readily measured?
4. What can be readily calculated?

QSAR methodology

Often it is found that several descriptors are correlated

Statistical analysis is used to determine the best descriptors that correlates with biological activity.

The final QSAR involves only the most important 3-5 descriptors



2

QSAR equation using 2 variables:
$Log(1/C) = 1.151\pi - 1.464\sigma + 7.817$
(n=22, r = 0.945)

QSAR equation using 3 variables:
$Log(1/C) = 1.259\pi - 1.460\sigma + 0.208\ E_{s(meta)} + 7.817$
(n=22, r = 0.959)

Types of QSARs

Two Dimensional QSAR

- Classical Hansh Analysis
- Two dimensional molecular properties

Three Dimensional QSAR

- Three dimensional molecular properties
- Molecular Field Analysis
- Molecular Shape Analysis
- Distance Geometry
- Receptor Surface Analysis

QSAR Generation Process

1. Selection of training set
2. Enter biological activity data
3. Generate conformations
4. Calculate descriptors
5. Selection of statistical method
6. Generate a QSAR equation
7. Validation of QSAR equation
8. Predict for Unknown

		Receptor Structure	
		Unknown	Known
Ligand structre	Unknown	Generate 3D structures, similarity /dissimilarity Homology model Screening/synthesis	Active site search, Receptor based DD, 3D searching
Ligand structure	**Known**	**Indirect DD Ligand based DD Analogs design 2D/3D pharmacophore**	**Structure based drug design**

PHARMACOPHORE APPROCH

Pharmacophore:

The Spatial orientation of various functional groups or features in 3D necessary to show biological activity.

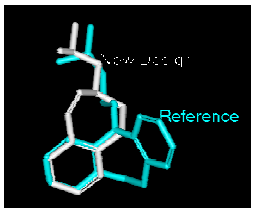
Types of Pharmacophore Models

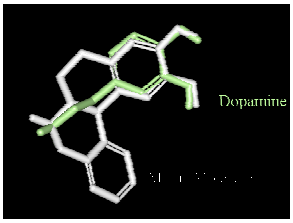
Distance Geometry based Qualitative Common Feature Hypothesis.

Quantitative Predictive Pharmacophores from a training set with known biological activities.

Pharmacophore-based Drug Design

•Examine features of *inactive* small molecules (ligands) and the features of *active* small molecules.
•Generate a hypothesis about what chemical groups on the ligand are necessary for biological function; what chemical groups suppress biological function.
•Generate new ligands which have the same necessary chemical groups in the same 3D locations. ("Mimic" the active groups)





Advantage: Don't need to know the biological target structure

Pharmacophore Generation Process

Five Steps

Training set selection.

Features selection

Conformation Generation

Common feature Alignments

Validation

Pharmacophore Features

- **HB Acceptor & HB Donor**
- **Hydrophobic**
- **Hydrophobic aliphatic**
- **Hydrophobic aromatic**
- **Positive charge/Pos. Ionizable**
- **Negative charge/Neg. Ionizable**
- **Ring Aromatic**

Each feature consists of four parts:

1. **Chemical function**
2. **Location and orientation in 3D space**
3. **Tolerance in location**
4. **Weight**

Pharmacophore Generation