QSAR and 3D QSAR in drug design
Part 1: methodology

Hugo Kubinyi

Classical QSAR methods describe structure–activity relationships in terms of physicochemical parameters and steric properties (Hansch analysis, extrathermodynamic approach), or certain structural features (Free Wilson analysis). 3D QSAR methods, especially comparative molecular field analysis, consider the three-dimensional structures and the binding modes of protein ligands. Quantitative similarity–activity relationships derive correlations between the similarities of individual compounds and their biological activities. Theory and methodology of these approaches are described here, together with the proper use of regression and partial least squares analyses for deriving quantitative structure–activity relationships. Part 2, to be published in the December issue, will address applications and problems.

Quantitative structure–activity relationships (QSAR) correlate, within congeneric series of compounds, affinities of ligands to their binding sites, inhibition constants, rate constants, and other biological activities, either with certain structural features (Free Wilson analysis) or with atomic, group or molecular properties, such as lipophilicity, polarizability, electronic and steric properties (Hansch analysis)[1–13]. Although relationships between lipophilicity and unspecific biological properties, such as narcotic, bactericidal, fungicidal, hemolytic and toxic properties, have been known since the turn of our century, the independent publications of the Free Wilson method[14] and of Hansch analysis[15], both in 1964, mark a milestone in the development of QSAR.

Since then, QSAR equations have been used to describe thousands of biological activities within different series of drugs and drug candidates. Especially enzyme inhibition data have been successfully correlated with physicochemical properties of the ligands[1–3,16]. In certain cases, where X-ray structures of the proteins became available, the results of QSAR regression models could be interpreted with the additional information from the three-dimensional (3D) structures[1–3].

After a very slow development of grid-based 3D QSAR approaches, in 1988 the method of comparative molecular field analysis (CoMFA) was published by Cramer *et al.*[17] This molecular field-based method constituted the first real 3D QSAR method. In the past ten years, many successful CoMFA applications proved the value of this method[18–24], especially in cases where classical QSAR methods fail. In contrast to Hansch or Free Wilson analysis, CoMFA is better suited to describe ligand–receptor interactions, because it considers the properties of the ligands in their (supposed) bioactive conformations. As the result of a CoMFA analysis, regions in space are identified that are favorable or unfavorable for the ligand–receptor interaction.

Hugo Kubinyi, Drug Design, ZHV/W – A30, BASF AG, D-67056 Ludwigshafen, Germany. tel: +49 621 60 42115, fax: +49 621 60 20914, e-mail: kubinyi@zhv.basf-ag.de

  1359-6446/97/$17.00. PII: S1359-6446(97)01079-9

Quantitative structure–activity relationships

QSAR theory

All QSAR analyses are based on the assumption of linear additive contributions of the different structural properties or features of a compound to its biological activity, provided that there are no nonlinear dependences of transport or binding on certain physicochemical properties. This simple assumption is proven by some dedicated investigations, for example the scoring function of the *de novo* drug design program LUDI (Eqn 1)[18,25,26]; in addition, the results of many Free Wilson and Hansch analyses support this concept.

$$\Delta G_{binding} = \Delta G_0 + \Delta G_{hb} + \Delta G_{ionic} + \Delta G_{lipo} + \Delta G_{rot} \quad (1)$$

Overall loss of translational and rotational entropy, ΔG_0 = +5.4 kJ mol^{-1}
Ideal neutral hydrogen bond, ΔG_{hb} = –4.7 kJ mol^{-1}
Ideal ionic interaction, ΔG_{ionic} = –8.3 kJ mol^{-1}
Lipophilic contact, ΔG_{lipo} = –0.17 J mol^{-1} Å^{-2}
Entropy loss per rotatable bond of the ligand, ΔG_{rot} = +1.4 kJ mol^{-1}

Eqn 1 correlates the free energy of binding, $\Delta G_{binding}$, with a constant term, ΔG_0, that describes the loss of overall translational and rotational degrees of freedom and ΔG_{hb}, ΔG_{ionic} and ΔG_{lipo}, which are structure-derived energy terms for neutral and charged hydrogen bond interactions and hydrophobic interactions between the ligand and the protein; ΔG_{rot} describes the loss of internal rotational degrees of freedom of the ligand. Eqn 1 holds for a wide range of energy values: the $\Delta G_{binding}$ of 45 different ligand–protein complexes ranges from –9 to –76 kJ mol^{-1}, which corresponds to binding constants between 2.5 × 10^{-2} M and 4 × 10^{-14} M; its standard deviation of 7.9 kJ mol^{-1} corresponds to a mean error of about 1.4 log units in the prediction of ligand binding constants from the mathematical model[18,25,26].

Because of the extrathermodynamic relationship between free energies ΔG and equilibrium constants K (Eqn 2) or rate constants k (k_{on} = association constant, k_{off} = dissociation constant of ligand–receptor complex formation), the logarithms of such values can be correlated with binding affinities.

$$\Delta G = -2.303\ RT \log K = -2.303\ RT \log k_{on}/k_{off} \quad (2)$$

Logarithms of molar concentrations C that produce a certain biological effect can be correlated with molecular features or with physicochemical properties that are also free-energy-related equilibrium constants; normally the logarithms of inverse concentrations, log 1/C, are used to obtain larger values for the more active analogs.

Free Wilson analysis

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 chemically modified, by values of 1 or 0 and correlates the resulting structural matrix with biological activity values, following Eqn 3; the values a_i in Eqn 3 are the biological activity group contributions of the substituents X_1, X_2, ... X_i in the different positions p of compound **1** (Figure 1) and μ is the biological activity value of the reference compound, most often the unsubstituted parent structure of a series[1,2,7,14].

$$\log 1/C = \Sigma\, a_i + \mu \quad (3)$$

Equation 4 describes the antiadrenergic activities for 22 different *m*-, *p*- and *m,p*-disubstituted analogs of the *N,N*-dimethyl-α-bromophenethylamine **2** (Figure 2), where C is the concentration that causes a 50% reduction of the adrenergic effect of a certain epinephrine dose[1,2,7]; for the meaning of r, s, F, and all other terms see Figure 3.

$$\log 1/C = -0.301\ (\pm 0.50)\ [m\text{-F}] + 0.207\ (\pm 0.29)\ [m\text{-Cl}] + 0.434\ (\pm 0.27)\ [m\text{-Br}] + 0.579\ (\pm 0.50)\ [m\text{-I}] + 0.454\ (\pm 0.27)\ [m\text{-Me}] + 0.340\ (\pm 0.30)\ [p\text{-F}] + 0.768\ (\pm 0.30)\ [p\text{-Cl}] + 1.020\ (\pm 0.30)\ [p\text{-Br}] + 1.429\ (\pm 0.50)\ [p\text{-I}] + 1.256\ (\pm 0.33)\ [p\text{-Me}] + 7.821\ (\pm 0.27)$$
$$(n = 22;\ r = 0.969;\ s = 0.194;\ F = 16.99) \quad (4)$$

Figure 1. *Schematic presentation of a molecule for Free Wilson analysis. A common skeleton bears substituents X_i in different positions p; the presence or absence of these substituents is coded by the values 1 and 0, respectively.*

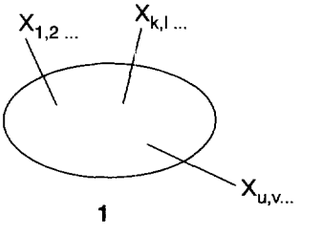
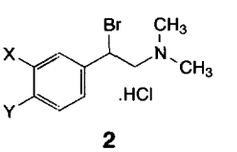


Figure 2. *N,N-dimethyl-α-bromophenethylamines (X, Y = H, F, Cl, Br, I, Me).*



Equation 4 illustrates the main advantage of Free Wilson analysis: only the biological activity values and the chemical structures of the compounds need to be known to derive a QSAR model. On the other hand, Free Wilson analysis has several shortcomings:

- at least two different positions of substitution must be chemically modified;
- predictions can only be made for new combinations of substituents already included in the analysis;
- single point determinations (i.e. the single occurence of a certain structural feature in the whole data set) obscure the statistical results;
- many degrees of freedom are wasted to describe every substituent.

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 easily be concluded from Eqn 4 that

- biological activities increase with increasing lipophilicity (F to Cl, Br, 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.

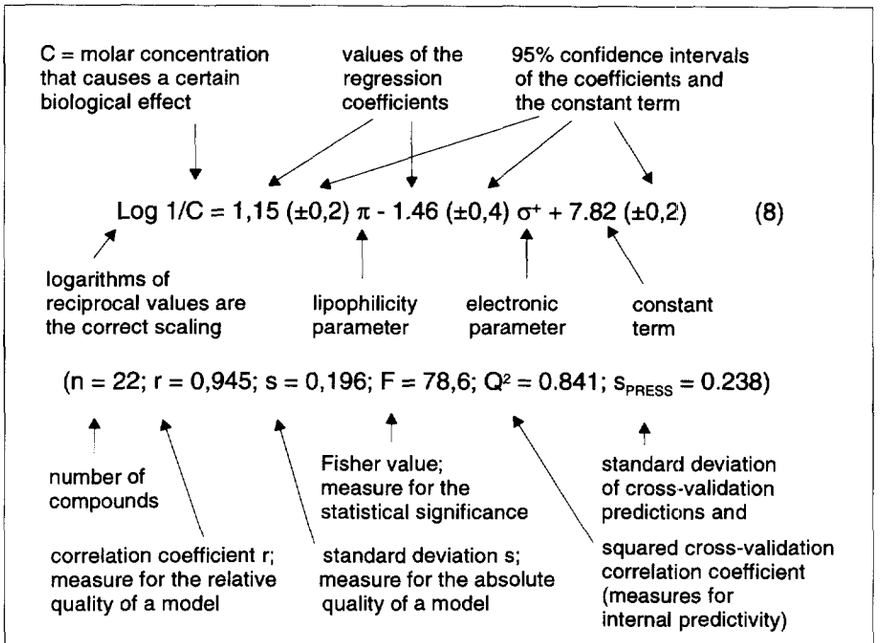


Figure 3. *Equation 8 describes a quantitative relationship between the antiadrenergic activities of compounds* ***2*** *(Figure 2; X, Y = H, F, Cl, Br, I, Me), lipophilicity (expressed by the n-octanol/water-derived parameter π), and electronic properties; σ⁺ is a special Hammett constant for the reactivity of benzyl cations. All relevant statistical parameters are explained.*

Hansch analysis

Also in 1964, the linear free-energy-related Hansch model (sometimes called the 'extrathermodynamic approach') was published (Eqn 5; P = n-octanol/water partition coefficient, σ = Hammett electronic parameter, a, b, c = regression coefficients, k = constant term)[1–3,7,15].

$$\log 1/C = a (\log P)^2 + b \log P + c \sigma + ... + k \qquad (5)$$

Equation 5 was developed from the concept that the transport of a drug from the site of application to its site of action depends in a nonlinear manner on the lipophilicity of the drug, and that the binding affinity to its biological counterpart, such as an enzyme or a receptor, depends on the lipophilicity, the electronic properties and other linear free-energy-related properties. Equation 5 combines the description of both processes in one mathematical model. In addition to the introduction of a parabolic term for the nonlinear lipophilicity dependence and the combination of different physicochemical properties in one equation, Hansch and Fujita defined lipophilicity parameters π of substituents X (Eqn 6), in the same manner as Hammett had defined the electronic parameter σ (Eqn 7), about 30 years earlier. The partition coefficient P in Eqn 6 is an equilibrium constant, similar to the dissociation or reaction constants K in Eqn 7. The absence of a 'reaction term' π in Eqn 6 is explained by the fact that all π values refer to the n-octanol/water system.

$$\pi_X = \log P_{RX} - \log P_{RH} \qquad (6)$$

$$\rho\sigma = \log K_{RX} - \log K_{RH} \qquad (7)$$

With the help of these definitions it was possible to use tabulated values instead of measured values[2,4]. For the data

set described by Eqn 4, Eqns 8 (Figure 3) and 9 (E_s^{meta} = steric parameter for *meta*-substituents) could be derived[2,7]. All parameters that are relevant in a QSAR study are presented and discussed in Figure 3.

$$\log 1/C = 1.259\ (\pm 0.19)\ \pi - 1.460\ (\pm 0.34)\ \sigma^+ + 0.208\ (\pm 0.17)\ E_s^{meta} + 7.619\ (\pm 0.24)$$
$$(n = 22;\ r = 0.959;\ s = 0.173;\ F = 69.24;\ Q^2 = 0.869;\ s_{PRESS} = 0.222) \quad (9)$$

Equations 8 and 9 demonstrate the superiority of Hansch analysis, as compared with Free Wilson analysis. Only a few properties are needed to correlate the biological activities; the model can directly be interpreted in physicochemical terms. The results of the Free Wilson analysis are confirmed in all details but predictions for compounds with other substituents can be made, for example for X = ethyl or CF_3. On the other hand, predictions that are too far outside the range of investigated parameters, such as for *tert*-Bu, –OH or $-SO_2NH_2$, will most probably fail because of the narrow chemical relationship among the investigated substituents and the very different chemical nature of these groups, in size or in their hydrogen bond donor and acceptor properties. For such predictions much more heterogeneous substituents have to be included in the derivation of the QSAR model.

The fact that different models can be derived for the same data set frequently offers a dilemma in Hansch analysis. One can never be sure that a certain QSAR model is the 'correct' one for the data set. On the other hand, different models correspond to different working hypotheses. Proposals for the synthesis of new analogs can be made in the following steps, which allow discrimination between these models.

Free Wilson-type group contributions for every substituent can be derived from Eqns 8 and 9, which clearly indicate the close theoretical relationship between Free Wilson analysis and linear Hansch analysis. Correspondingly, both approaches can be used in one model, the so-called 'mixed approach' (Eqn 10)[2,7].

$$\log 1/C = a\ (\log P)^2 + b \log P + c\ \sigma + \ldots + \Sigma\ a_i + k \quad (10)$$

Equation 10 combines the advantages of Hansch and Free Wilson analyses and widens the applicability of both methods. Physicochemical parameters describe parts of the molecules with broad structural variation, whereas indicator variables a_i (Free Wilson-type variables) encode the effects of structural variations that cannot be described otherwise[1–3,7].

Nonlinear structure–activity relationships

Transport rate constants of organic compounds in simple two-compartment systems, for example n-octanol/water, depend on their lipophilicity. The forward rate constant k_1 describes the transport from the aqueous phase into the organic phase. For polar compounds, k_1 increases linearly with increasing lipophilicity, until the diffusion rate constant from the bulk solution to the aqueous/organic interface and the diffusion through this interface constitute an upper limit; the reverse situation applies to the rate constants k_2, from the organic medium into water. These relationships hold for series of structurally highly diverse compounds (Figure 4); no influence of the size or the shape of molecules could be detected (because of the Einstein–Stokes relationship this influence is expected to be very small, depending on the radius of the particles, which is related to the cubic root of the volumes)[2,7].

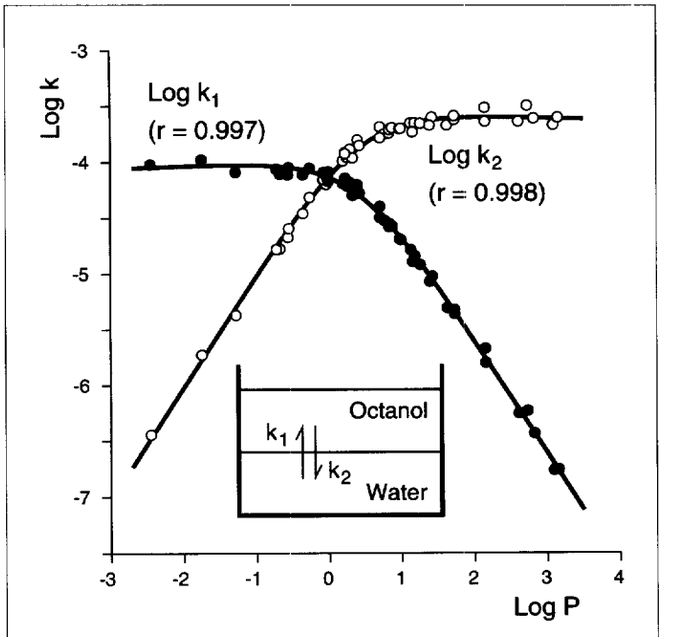


Figure 4. *For a large group of chemically heterogeneous compounds the transport rate constants* k_1*, from an aqueous phase into an n-octanol phase, and* k_2*, in the reverse direction, depend in a nonlinear manner on the lipophilicity. The curves are fitted by Eqns 11 and 12.*

The lipophilicity dependence of the rate constants k_1 and k_2 follows relationships that are expressed by Eqns 11 and 12. Because the transport of a drug from its site of application to its site of action corresponds to a random walk through several aqueous and lipophilic barriers, the bilinear model (Eqn 13) has been derived to describe nonlinear lipophilicity–activity relationships[2,7].

$$\log k_1 = a \log P - a \log (\beta P + 1) + c \quad (11)$$

$$\log k_2 = - a \log (\beta P + 1) + c \quad (12)$$

$$\log 1/C = a \log P - b \log (\beta P + 1) + c \quad (13)$$

The coefficient β in Eqns 11–13 is a nonlinear term that must be estimated by an iterative procedure. Equation 13 correlates a large number of nonlinear lipophilicity–activity relationships in an exact manner[2,3]. However, the simpler parabolic Hansch model (Eqn 5) may be considered as a good approximation to the bilinear model.

Besides nonlinear lipophilicity–activity relationships, also nonlinear dependences of binding affinities or biological activities on the volumes of the substituents, expressed either by molar refraction (MR) or by any steric parameters, are relatively common. Most often such relationships reflect the limited size of a certain binding pocket.

Dissociation and ionization of acids and bases

Acids AH dissociate in aqueous solution and bases B are protonated, according to their dissociation constants (pK_a values). Because ions are more polar than neutral forms, the log P values of ionizable compounds have to be corrected for the amount of ionized and unionized forms (Eqns 14 and 15; P_{app} = apparent partition coefficient at a certain pH value, P_u = partition coefficient of the unionized form, P_i = partition coefficient of the ionized form). Sigmoidal curves are obtained for the pH dependence of the log P_{app} values. The concentrations of A^- and BH^+ in the organic phase may be neglected for compounds that are not too lipophilic, therefore simple approximations can be used to estimate the log P_{app} values of most acids and bases at a given pH value[2,7].

Acids:

$$AH + H_2O = A^- + H_3O^+$$

$$\log P_{app} = \log (P_u.10^{pH} + P_i.10^{pK_a}) - \log (10^{pK_a} + 10^{pH})$$

$$\approx \log P_u - \log (1 + 10^{pH - pK_a}) \quad (14)$$

Bases:

$$B + H_3O^+ = BH^+ + H_2O$$

$$\log P_{app} = \log (P_u.10^{pH} + P_i.10^{pK_a}) - \log (10^{pK_a} + 10^{pH})$$

$$\approx \log P_u - \log (1 + 10^{pK_a - pH}) \quad (15)$$

At pH values where the neutral form predominates (pH < pK_a for acids; pH > pK_a for bases), P_{app} values are identical with P_u values. With increasing ionization, the log P_{app} values decrease linearly with increasing (acids) or decreasing (bases) pH values, till again a constant P_{app} value is obtained, because now only the ionic form contributes to partitioning ($P_{app} = P_i$). Complex pH dependences are obtained in the case of compounds having more than one ionizable group. The proper implementation of Eqns 14 and 15 into QSAR models depends on the system, whether it is equilibrium-controlled (e.g. enzyme inhibition in aqueous buffer) or whether it is a kinetically controlled system (e.g. a whole animal).

The pH-absorption profiles should be parallel to the pH-partition profiles. Deviations from the simple pH partition hypothesis, called pH shifts, are obtained for highly lipophilic compounds; their absorption profiles are shifted to higher (acids) or lower (bases) pH values. The higher the lipophilicity of the neutral species is, the larger is the observed pH shift, which can be explained by the assumption of an 'unstirred' aqueous diffusion layer at the aqueous/organic interface.

3D quantitative structure–activity relationships

In 1979, a new approach was proposed to describe molecular properties by fields, calculated in a regular grid[27]. Vectors were extracted from these fields by principal component analysis and correlated with the biological activities. Later this method was called the DYLOMMS (dynamic lattice-oriented molecular modeling system) approach. But only from 1988 on[17], when partial least squares (PLS) analysis (see below) was used to correlate the field values with biological activities and commercial software became available[28], the method, now called comparative molecular field analysis (CoMFA), became widely used, especially to derive quantitative models for enzyme inhibition constants and other binding affinities[18–24].

Generation of 3D structures and alignment

In CoMFA, first a group of compounds is selected. These compounds should be chemically related and should act via the same mechanism of action; however, in contrast to

classical QSAR methods, they should have a common pharmacophore, not necessarily the same molecular skeleton. As the pharmacophore refers to 3D structures, first the 2D or 2.5D (including configurational information) structures of all molecules are converted to 3D structures. Standard computer programs for this purpose are CONCORD (Ref. 29) and CORINA (Ref. 30). However, both programs create only one low-energy structure per molecule. Most often one or several torsion angles of flexible molecules have to be modified (of course, leading to other low-energy 3D structures) to derive a common pharmacophore for all analogs. The investigation of rigid analogs or analogs with different conformational constraints helps to find or confirm the 'bioactive' conformations of the more flexible molecules.

An example of a pharmacophore, as a spatial arrangement of some structural features in certain distance ranges, is given in Figure 5. In the next step, an alignment is performed by superimposing all molecules, according to orientation rules that follow from their common pharmacophore. This is one of the most critical and difficult steps in a CoMFA study[18,24]. In many 'real world' situations, for example for molecules that are not from a congeneric series or for compounds with a large number of rotatable bonds, a correct alignment is difficult or even impossible. Theoretically, this problem severely limits the applicability of CoMFA; on the other hand, some investigations give evidence that the similarities (and correspondingly the affinities) of molecules are properly described, even if they are studied in geometries that are different from the correct bioactive conformations. In addition, 3D approaches have been developed that do not depend on a common alignment of the molecules[24].

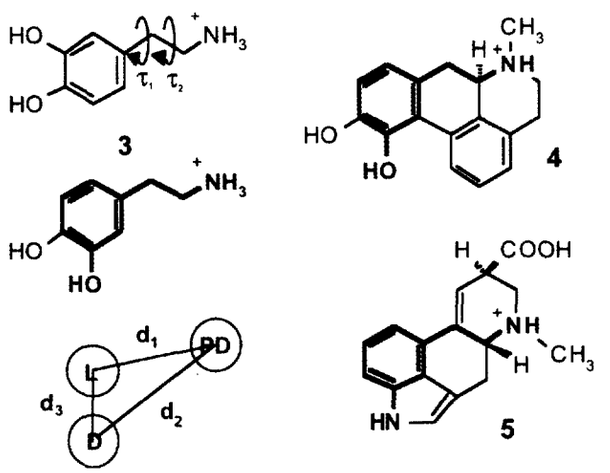


Figure 5. *Dopamine (**3**) is a flexible molecule with two rotatable bonds, characterized by the torsion angles τ_1 and τ_2. Analogs **4** and **5** contain the dopamine pharmacophore, but in frozen conformations. A comparison of the structures **3–5** leads to the derivation of a pharmacophore model (lower left; corresponding structural elements in **3–5** are indicated in red) which is made up from a hydrogen bond donor group D, a lipophilic area L and a protonated hydrogen bond donor group PD.*

In the very first CoMFA study, steroids were superimposed by a rigid body alignment of their skeletons[17]. This group of compounds was a fortunate example, without conformational flexibility of the molecules and without ambiguities regarding the relative orientations in which the molecules should be superimposed. The alignment of dihydrofolate (**6**; DHF) and methotrexate (**7**; MTX) (Figure 6) poses a much more difficult problem. Although 'chemically' related, the different pattern of hydrogen bond donor and acceptor distribution in MTX, when compared with the lead structure DHF, demands a completely different orientation (Figure 6)[12,31]. This different binding mode was predicted from the 3D structure of the DHF/dihydrofolate reductase complex[31] and later experimentally confirmed by the X-ray structure determination of the MTX complex[32].

Computer methods have been described that allow an automated alignment, according to different properties, for example the program SEAL (Refs 33,34). It defines a 'similarity index' A_F between two molecules A and B in any relative orientation to each other (Eqn 16); r_{ij} is the distance between atoms i and j, α is a user-defined value, w_E and w_S are user-attributed values to give different weights for electrostatic and steric overlap, q_i and q_j are the partial charges at atoms i of molecule A and j of molecule B, and v_i and v_j are arbitrary powers (default = 3) of the van der Waals radii of atoms i and j. Any other property, for example hydrophobicity, might be added to the definition of w_{ij}.

$$A_F = \sum_{i=1}^{m}\sum_{j=1}^{n} w_{ij} e^{-\alpha r_{ij}^2}; w_{ij} = w_E q_i q_j + w_S v_i v_j + \ldots \qquad (16)$$

Training, test set selection and calculation of fields

The data set is separated into a training set for which a CoMFA model is derived and a test set that will prove the

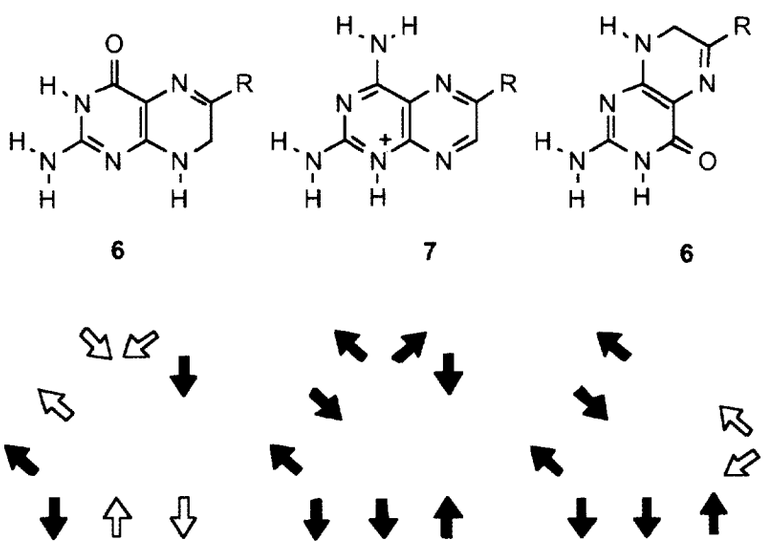


Figure 6. *The chemical structures of dihydrofolate (**6**) and methotrexate (**7**) look relatively similar. However, a closer inspection of the hydrogen bond donor and acceptor patterns of both compounds clearly shows that after a simple atom-by-atom superposition of both molecules only three donor and acceptor functions are in the same place (left; hydrogen bond donors in green, acceptors in red; filled arrows indicate identical positions). In the other binding mode (right) dihydrofolate has six donor and acceptor functions in the same place. This binding mode was predicted and several years later experimentally confirmed.*

external predictivity of the resulting model. A box is placed around the superimposed molecules in such a manner that the box is in all directions several Å larger than the combined volume of all molecules. A regular lattice is laid over the molecules to calculate different molecular fields in each grid point (Figure 7); the default distance between the grid points is 2 Å (Refs 17,18).

Probe atoms or groups, such as a neutral carbon atom (probe for van der Waals interactions), a charged atom (probe for Coulomb interactions) or a hydrogen bond donor or acceptor (probes for hydrogen bond interactions), are used to determine for each molecule the interaction energies in every grid point; the mathematical functions used for van der Waals and Coulomb interactions are given in Eqns 17 (E_{vdW} = van der Waals interaction energy, r_{ij} = distance between atom i of the molecule and the grid point j, where the probe atom is located; A_i and C_i are constants that depend on the van der Waals radii of the corresponding atoms) and 18 (E_C = coulomb interaction energy, q_i = partial charge of atom i of the molecule, q_j = charge of the probe atom, D = dielectric constant, r_{ij} = distance between atom i of the molecule and the grid point j, where the probe atom is located).

$$E_{vdW} = \sum_{i=1}^{n}\left(A_i r_{ij}^{-12} - C_i r_{ij}^{-6}\right) \qquad (17)$$

$$E_c = \sum_{i=1}^{n} \frac{q_i q_j}{D r_{ij}} \qquad (18)$$

The Lennard–Jones potential (Eqn 17) and the Coulomb potential (Eqn 18) create relatively 'hard' fields. They change their values from close to zero to very large numbers within a few tenths of an Ångstrom, i.e. within a small fraction of the most often used grid distance of 2 Å. Thus, cut-off values have to be defined to avoid values that approach infinity at the atom positions. Serious differences in the CoMFA results may be obtained after small shifts in the orientation of a ligand, of the box, or the grid points in which the fields are calculated. Several modifications of the CoMFA procedure have been proposed to avoid such problems[24]:

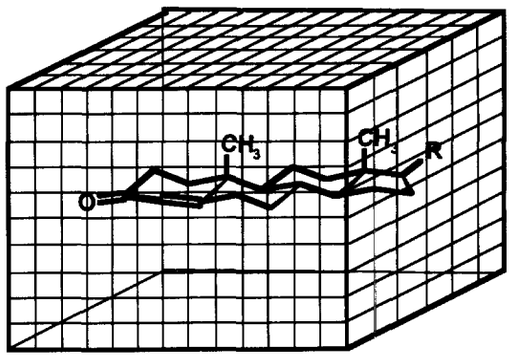


Figure 7. *All molecules of a chemically related series are superimposed in a comparative molecular field analysis (CoMFA) study, according to their common pharmacophore. Then they are imbedded in a box and a regular grid is laid over the molecules. For better visualization, the box is smaller than usual, only the outer lines of the grid are indicated, and only one molecule (instead of all different analogs) is shown.*

- a cross-validated Q^2-guided region selection to achieve consistent results[35],
- reorientations of the molecules for which too low activities are calculated, to obtain an 'improved' (but more subjective) alignment[36],
- single and domain mode variable selection in the PLS analysis[37], and
- a selection of relevant regions by the GOLPE variable selection program[38].

In a recent CoMFA modification, called comparative molecular similarity indices analysis (CoMSIA)[24,39], SEAL functions (Eqn 16) are used to calculate 'similarity fields' to probe atoms and groups, in the same manner as CoMFA fields are calculated. Since the CoMSIA fields are Gaussian potentials, they are much 'softer' than the CoMFA functions. In addition, they are good approximations to the cut-off-corrected Lennard–Jones and Coulomb potentials (Figure 8). As expected, CoMSIA fields have been shown to produce smoother contour maps (see below) than CoMFA fields[39].

Correlation of the molecular fields with the biological data

The resulting matrix of several thousand columns of the individual grid point values for each molecule cannot be correlated by regression analysis. Partial least squares (PLS)

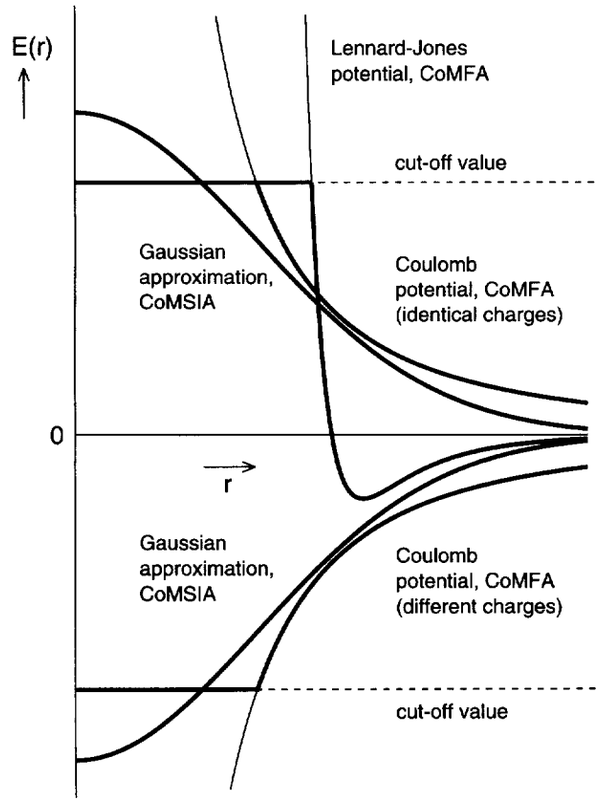


Figure 8. *The Lennard–Jones and Coulomb functions are 'hard' potentials. Within several tenths of an Ångstrom they change their values from close to zero to very large values. Cut-off values have to be defined to avoid such unreasonably large values. The comparative molecular similarity indices analysis (CoMSIA) potentials are much smoother; they may be considered as good approximations for the cut-off-limited Lennard–Jones and Coulomb potentials in CoMFA (comparative molecular field analysis).*

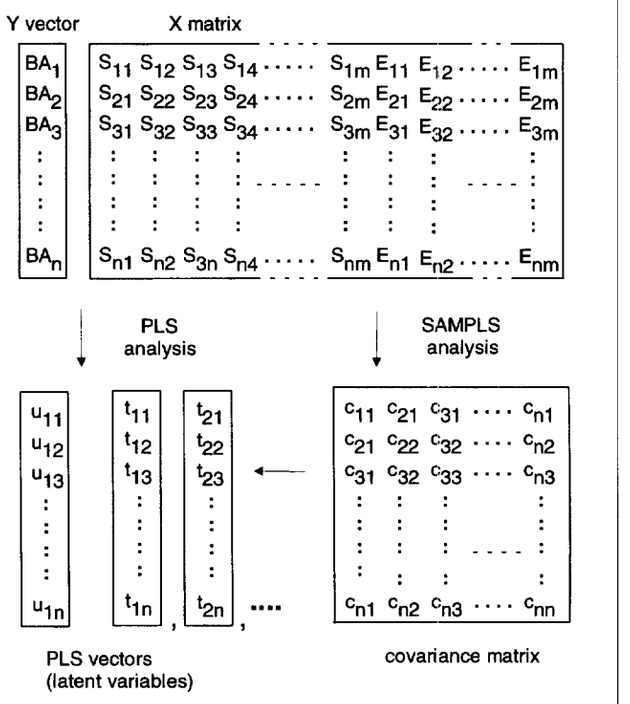
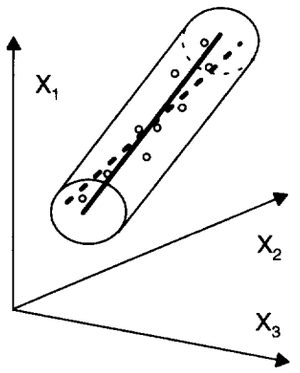


Figure 9. *Partial least squares (PLS) analysis extracts principal component-like vectors (so-called latent variables) from the* ***Y*** *block (biological data; usually only a* ***y*** *vector) and the* ***X*** *block (in comparative molecular field analysis the field variables in the different grid points). SAMPLS analysis is a PLS variant, in which a covariance matrix is calculated from the* ***X*** *block and the latent variables are derived from this matrix. PLS and SAMPLS analysis give absolutely identical results, with the advantage that SAMPLS is some orders of magnitude faster in cross-validation runs.*



From $\mathbf{u} = \mathbf{kt}$ ($\mathbf{k}$ = constants k_1, k_2 ... k_j; j = number of PLS vectors) follows:

$$BA_i = a_1S_{i1} + a_2S_{i2} + a_3S_{i3} + \ldots + a_mS_{im} + b_1E_{i1} + b_2E_{i2} + b_3E_{i3} + \ldots + b_mE_{im}$$

Figure 10. *The main advantage of PLS analysis, as compared with principal component regression analysis is the fact that PLS vectors (dotted line) are shifted against the principal components (bold line), in order to achieve a maximum correlation between the* ***Y*** *and the* ***X*** *block; the tube indicates the confidence volume of the principal component.*

analysis is the method of choice (Figure 9). This method extracts principal component-like vectors, the so-called latent variables, from the **Y** block (usually only a one-column vector **y**) and the **X** block; the latent variables are shifted within the confidence volumes of the principal components, in such a manner that a maximum correlation between these variables, called **u** and **t**, is achieved (Figure 10)[2,5,6,18].

The optimum number of latent variables is determined by cross-validation[2,5,6,18]. In this approach, many models are derived where one or several compounds are excluded from the data set and are predicted by the corresponding model. In the most common 'leave-one-out' cross-validation, every compound is eliminated once. Thus, n models (n = number of compounds) are derived and the n predictions are compared with the experimental values; Q^2, the squared cross-validated correlation coefficient, and s_{PRESS}, the standard deviation of predictions, are calculated from the sum of the squared deviations of the predicted values (PRESS), in the same manner as the squared correlation coefficient r^2 and the standard deviation s are calculated if no cross-validation is performed. Most often, the highest Q^2 or the lowest s_{PRESS} is taken as the criterion for the optimum number of latent variables. For theoretical reasons, Q^2 cannot be larger than r^2 and s_{PRESS} cannot be smaller than s; Q^2 may even be negative, which means that the predictions by the model are worse than taking the average of all biological data as the predictions.

The result of a PLS analysis corresponds to a regression equation with thousands of coefficients, from which predictions for compounds outside the investigated training set can be made. Most often the PLS results are presented as contour plots; favorable and unfavorable regions for certain properties in space are indicated in different colors (Figure 11)[40].

Since 1988, a few hundred publications, several reviews[19–23] and two books[18,24] have appeared on this subject. Unfortunately, several CoMFA studies do not meet all necessary scientific and statistical standards. Critical reviews on the validity of the results and conclusions of CoMFA studies have been published[18,24]. Despite some major problems in its proper application, the method is now generally accepted as a useful tool for deriving 3D QSAR models.

Quantitative similarity–activity relationships (QSiAR)

Similar molecules, having similar structures and similar physicochemical properties, are supposed to have also similar biological properties and comparable activity values. Whereas there are some significant exceptions to this general assumption, it is the basis of all quantitative structure–activity relationships. Quantitative similarity–activity relationships (QSiAR) correlate the degree of mutual similarity of all molecules within a series with their biological activities.

Rum and Herndon, in 1991, used n × n similarity index matrices (n = number of compounds in the data set), derived from 2D topological descriptors, and stepwise regression analysis to correlate several columns of these matrices with biological activities[41]. Two years later, in a more systematic investigation, Good and Richards superimposed molecules, like in CoMFA analyses, and compared their 3D electronic similarity[42,43]. The resulting n × n similarity index matrices were correlated with biological activities, using either neural networks or PLS analysis.

This concept could be extended to many other linear and nonlinear QSiAR relationships, by calculating either n × n distance matrices **D** (especially suited for nonlinear relationships) or n × n covariance matrices **C** as similarity

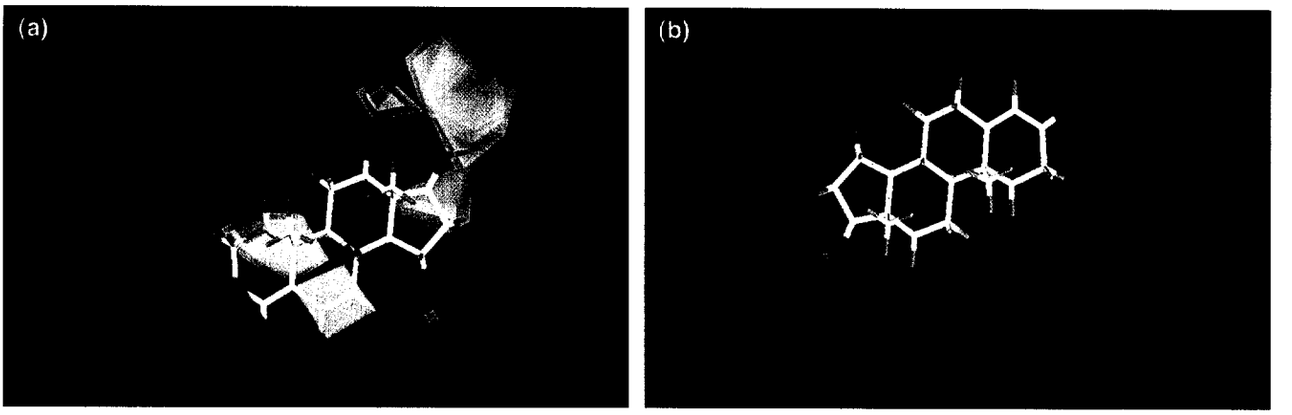


Figure 11. *3D contour maps around testosterone (compare Figure 7, R = OH) resulting from a comparative molecular field analysis (CoMFA) of the testosterone binding globulin (TBG) affinities of different steroids[17]. (a) The color codings indicate regions where substitution enhances (green) or reduces (yellow) the binding affinity. (b) Regions where electronegative substituents enhance (blue) or reduce (red) the binding affinity (reproduced[40] with kind permission from VCH, Weinheim).*

measures[44]. For this purpose, all or only several relevant properties of the compounds are used to derive the corresponding similarity matrices. No superposition of the molecules is necessary. If a distance matrix **D** is derived from the **X** matrix of explanatory physicochemical properties (n rows, m columns), then all x_{ij} values must be normalized before calculating the distances d, i.e. mean-value-centered and standardized, column by column. The great advantage of distance similarity index matrices is that no special models need to be defined in the case of nonlinear relationships[2,44,45]; on the other hand, problems may arise from significant intercorrelations between the different columns of the similarity matrices.

In principle, any QSAR model may be considered as the result of a similarity analysis. As the similarity of molecules can also be described by topological indices, it is not surprising that connectivity indices and related parameters work so well in quantitative structure–activity analyses. On the other hand, the applicability of the similarity concept should not be overestimated. Examples are known, where a minor chemical change of a structure surprisingly leads to a significantly different biological activity, in its quantity differing up to several orders of magnitude[16].

Notes

In most cases reviews and books are cited in this overview instead of the original literature, in order to keep the number of references at a minimum and to provide the corresponding results in their context to related work in the same field.

A discussion forum for QSAR researchers is the WWW home page of *The QSAR and Modelling Society* at http://www.pharma.ethz.ch/qsar. Not only names and e-mail addresses of QSAR colleagues but also tips and tricks, information on recent books and software, and links to related topics can be found there.

REFERENCES

1 Ramsden, C.A., ed. (1990) *Quantitative Drug Design* (*Comprehensive Medicinal Chemistry*) (Vol. 4), Pergamon Press
2 Kubinyi, H. (1993) *QSAR: Hansch Analysis and Related Approaches*, VCH
3 Hansch, C. and Leo, A. (1995) *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology*, American Chemical Society
4 Hansch, C., Leo, A. and Hoekman, D. (1995) *Exploring QSAR. Hydrophobic, Electronic, and Steric Constants*, American Chemical Society
5 van de Waterbeemd, H., ed. (1995) *Chemometric Methods in Molecular Design*, VCH
6 van de Waterbeemd, H., ed. (1995) *Advanced Computer-Assisted Techniques in Drug Discovery*, VCH
7 Kubinyi, H. (1995) in *Burger's Medicinal Chemistry* (Vol. 1, 5th edn) (Wolff, M.E., ed.), pp. 497–571, John Wiley & Sons
8 Sanz, F., Giraldo, J. and Manaut, F., eds (1995) *QSAR and Molecular Modelling: Concepts, Computational Tools and Biological Applications (Proceedings of the 10th European Symposium on Quantitative Structure–Activity Relationships, Barcelona, 1994)*, Prous Science
9 Pliska, V., Testa, B. and van de Waterbeemd, H., eds (1996) *Lipophilicity in Drug Action and Toxicology*, VCH
10 van de Waterbeemd, H., ed. (1996) *Structure–Property Correlations in Drug Research*, Academic Press

11 van de Waterbeemd, H. (1996) in *The Practice of Medicinal Chemistry* (Wermuth, C.G., ed.), pp. 367–389, Academic Press

12 Böhm, H-J., Klebe, G. and Kubinyi, H. (1996) *Wirkstoffdesign*, pp. 363–380 and 399–436, Spektrum Akademischer Verlag

13 van de Waterbeemd, H., Testa, B. and Folkers, G., eds (1997) *Computer-Assisted Lead Finding and Optimization (Proceedings of the 11th European Symposium on Quantitative Structure–Activity Relationships, Lausanne, 1996)*, Verlag Helvetica Chimica Acta and VCH

14 Free, S.M., Jr and Wilson, J.W. (1964) *J. Med. Chem.* 7, 395–399

15 Hansch, C. and Fujita, T. (1964) *J. Am. Chem. Soc.* 86, 1616–1626

16 Kubinyi, H. (1997) *Drug Discovery Today* 2, part 2

17 Cramer, R.D., III, Patterson, D.E. and Bunce, J.D. (1988) *J. Am. Chem. Soc.* 110, 5959–5967

18 Kubinyi, H., ed. (1993) *3D QSAR in Drug Design. Theory, Methods and Applications*, ESCOM Science Publishers

19 Green, S.M. and Marshall, G.R. (1995) *Trends Pharmacol. Sci.* 16, 285–291

20 Kim, K.H. (1995) in *Molecular Similarity in Drug Design* (Dean, P.M., ed.), pp. 291–331, Chapman & Hall

21 Martin, Y.C. and Lin, C.T. (1996) in *The Practice of Medicinal Chemistry* (Wermuth, C.G., ed.), pp. 459–483, Academic Press

22 Blankley, C.J. (1996) in *Structure–Property Correlations in Drug Research* (van de Waterbeemd, H., ed.), pp. 111–177, Academic Press

23 Martin, Y.C., Kim, K-H. and Lin, C.T. (1996) in *Advances in Quantitative Structure Property Relationships* (Vol. 1) (Charton, M., ed.), pp. 1–52, JAI Press

24 Kubinyi, H., Folkers, G. and Martin, Y.C., eds (1997) *3D QSAR in Drug Design. Ligand-Protein Interactions and Molecular Similarity* (Vol. 2) and *Recent Advances* (Vol. 3), Kluwer Academic Publishers

25 Böhm, H-J. (1994) *J. Comput.-Aided Mol. Design* 8, 243–256

26 Böhm, H-J. and Klebe, G. (1996) *Angew. Chem.* 108, 2750–2778, *Angew. Chem., Int. Ed. Engl.* 35, 2588–2614

27 Cramer, R.D., III and Milne, M. (1979) *Am. Chem. Soc. Meeting*, April 1979, Computer Chemistry Section. Abstr. no. 44

28 SYBYL/QSAR, Molecular Modelling Software, Tripos Inc., 1699 S. Hanley Road, St Louis, MO 63944, USA

29 Pearlman, R.S. (1993) *Chem. Design Automation News* 8 (8), 3–15

30 Sadowski, J. and Gasteiger, J. (1993) *Chem. Rev.* 93, 2567–2581

31 Bolin, J.T. *et al.* (1982) *J. Biol. Chem.* 257, 13650–13662

32 Bystroff, C., Oatley, S.J. and Kraut, J. (1990) *Biochemistry* 29, 3263–3277

33 Kearsley, S.K. and Smith, G.M. (1990) *Tetrahedron Comput. Methodol.* 3, 615–633

34 Klebe, G., Mietzner, T. and Weber, F. (1994) *J. Comput.-Aided Mol. Design* 8, 751–778

35 Cho, S.J. and Tropsha, A. (1995) *J. Med. Chem.* 38, 1060–1066

36 Kroemer, R.T. and Hecht, P. (1995) *J. Comput.-Aided Mol. Design* 9, 396–406

37 Norinder, U. (1996) *J. Chemomet.* 10, 95–105

38 Cruciani, G., Pastor, M. and Clementi, C. (1997) in *Computer-Assisted Lead Finding and Optimization (Proceedings of the 11th European Symposium on Quantitative Structure–Activity Relationships, Lausanne, 1996)* (van de Waterbeemd, H., Testa, B. and Folkers, G., eds), pp. 379–395, Verlag Helvetica Chimica Acta and VCH

39 Klebe, G., Abraham, U. and Mietzner, T. (1994) *J. Med. Chem.* 37, 4130–4146

40 Kubinyi, H. (1994) *Chemie in unserer Zeit* 23, 281–290

41 Rum, G. and Herndon, W.C. (1991) *J. Am. Chem. Soc.* 113, 9055–9060

42 Good, A.C., Peterson, S.J. and Richards, W.G. (1993) *J. Med. Chem.* 36, 2929–2937

43 Good, A.C. (1995) in *Molecular Similarity in Drug Design* (Dean, P.M., ed.), pp. 24–56, Chapman & Hall

44 Kubinyi, H. (1997) in *Computer-Assisted Lead Finding and Optimization (Proceedings of the 11th European Symposium on Quantitative Structure–Activity Relationships, Lausanne, 1996)* (van de Waterbeemd, H., Testa, B. and Folkers, G., eds), pp. 7–28, Verlag Helvetica Chimica Acta and VCH

45 Martin, Y.C. *et al.* (1995) *J. Med. Chem.* 38, 3009–3015