QSAR and 3D QSAR in drug design Part 2: applications and problems

Hugo Kubinyi

QSAR already plays an important role in lead structure optimization and it can be predicted that QSAR methods will become essential for handling the huge amount of data associated with combinatorial chemistry. 3D QSAR has already been successfully applied to many data sets of enzyme and receptor ligands. The theory and methodology of these approaches were outlined in Part 1 of this article, published in the November issue. In the second part of this two-part review, the author explains the applications of these methods and addresses the associated problems.

Many successful applications of QSAR and 3D QSAR methods[1] prove the usefulness of these approaches in drug research (for reviews see Refs 2–12). Typical examples from literature are presented and discussed in this review, to illustrate the proper use of QSAR and 3D QSAR in lead structure optimization. Not all published QSAR studies fulfill all statistical criteria that are nowadays applied in a more rigid manner. Some of these problems are also discussed and recommendations for the validation of QSAR studies are given. Results from QSAR studies should always be considered as working hypotheses, on the one hand supported by some statistical parameters, on the other hand to be justified by the design and testing of new analogs, in order to prove or disprove the working hypotheses. If QSAR is applied in such a manner, it will be most helpful in the structural optimization of drugs.

QSAR and 3D QSAR applications

Absorption and distribution of drugs

If relatively polar compounds or only a narrow range of lipophilicity values are considered, linear lipophilicity dependences result. The blood–brain barrier penetration of various H_2-receptor antagonists and some CNS-active drugs has been correlated with van der Waals volumes V_M and a hydrogen-bonding capability parameter Λ_{alk} (Eqn 1; n = number of compounds; r = correlation coefficient; s = standard deviation; F = Fisher value; a more extensive explanation of these and other statistical parameters was given in Figure 3 of Part 1 of this review[1]), which together describe the lipophilicity and desolvation behavior of these compounds[3,13].

$$\log (C_{brain}/C_{blood}) = 0.007\ (\pm0.001)\ V_M - 0.34\ (\pm0.03)\ \Lambda_{alk} + 1.73\ (\pm0.30) \qquad (1)$$
$$(n = 20;\ r = 0.934;\ s = 0.290;\ F = 58)$$

As already discussed in the preceding review[1], the absorption and distribution of drugs most often show nonlinear lipophilicity relationships. Typical examples for drug transport and drug absorption are given in Figure 1a, whereas Figure 1b describes the nonlinear lipophilicity dependence of blood–brain barrier and blood–placenta barrier penetration. In all cases, the experimental biological data were correlated with calculated or experimental lipophilicity values, using the bilinear model[1]. Many similar nonlinear absorption/distribution relationships have been described in the literature[3,14].

Eqn 2 (P_{app} = apparent partition coefficient at a certain pH value; the coefficient β is a nonlinear term that must be estimated by an iterative procedure[1]) correlates the buccal absorption rate constants of an acid (*p*-hexylphenylacetic

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

Copyright © Elsevier Science Ltd. All rights reserved. 1359-6446/97/$17.00. PII: S1359-6446(97)01084-2

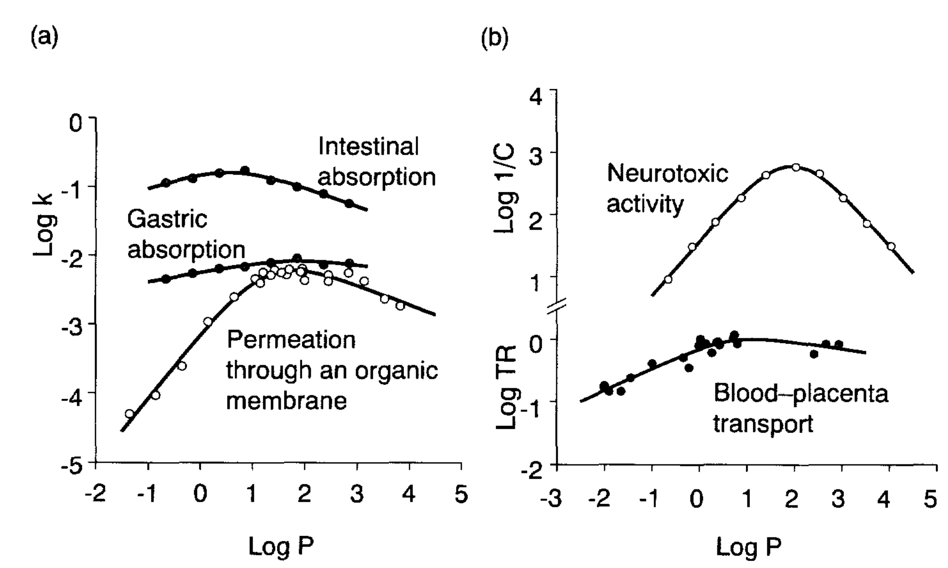


Figure 1. *(a) The permeation of barbiturates from an aqueous phase through an organic membrane into another organic phase follows a nonlinear dependence on the lipophilicity of the compounds. Similar dependences are observed for the gastric and intestinal absorption rates of carbamates in the rat and for other absorption processes. (b) The blood–brain barrier is permeable only for compounds with a certain lipophilicity. The neurotoxic activities of a series of homologous alcohols shows a nonlinear lipophilicity dependence, with a maximum near log P = 2. The blood–placenta barrier shows a similar but slightly less pronounced lipophilicity dependence for a group of chemically different drugs.*

acid, $pK_a = 4.36$) and a base (propranolol, $pK_a = 9.45$) at different pH values with log P_{app} values[3].

$$\begin{aligned}\log k_{abs} &= 0.448\ (\pm 0.05)\log P_{app} - 0.448\ (\pm 0.05)\log(\beta P_{app} + 1) - 1.689 \qquad (2)\\ \log\beta &= -2.792\\ (n &= 12;\ r = 0.988;\ s = 0.102)\end{aligned}$$

If only transport is considered, lipophilic drugs have the advantage that they are easily absorbed and readily distributed. However, lipophilic drugs have also some disadvantages. Most often such compounds show, despite their good absorption, a low bioavailability because they are directly eliminated at their first liver passage, via the bile ('first pass effect'). Undesired central nervous system side effects may result when nonpolar compounds enter the brain. In addition, lipophilic xenobiotics sometimes are metabolized by cytochromes to chemically reactive, toxic oxidation products. Thus, a recommendation in drug design is to make drugs only as lipophilic as absolutely necessary[15].

Box 1. QSARs of enzyme ligands (substrates and inhibitors)

Hydrolases
Chymotrypsin[2–4]
Trypsin[2–4]
Subtilisin[4]
Cholinesterases[4]
Emulsin[4]
Papain[2–4]
Other cystein proteases[2,4]

Oxidoreductases
Alcohol dehydrogenase[2,4]
Dihydrofolate reductase[2–4]
Malate dehydrogenase[4]
Xanthine oxidase[4]
Monoamine oxidase[4]
Prostaglandin synthase[4]
Thymidilate synthase[4]
Lipoxygenase[4]
Inosinic acid dehydrogenase[4]

Transferases
Acetyl transferases[2,4]
Catechol-*O*-methyl transferase[4]

Lyases
Carbonic anhydrase[4]

The theory of passive drug absorption models has recently been reviewed by Camenisch, Folkers and van de Waterbeemd[16]. Design principles for orally bioavailable drugs were discussed by Navia and Chaturvedi[17].

Enzyme inhibitors

The quantitative description of structure–activity relationships of the binding affinities of enzyme substrates and inhibitors is the domain of classical QSAR, especially Hansch analysis (Box 1), and of comparative molecular field analysis (CoMFA) (Box 2). In cases where 3D structures of the enzymes are known, the structure-derived ligand–receptor interactions correlate well with the results of the QSAR analyses. This has been investigated, for example, for dihydrofolate reductase (DHFR), papain and other cysteine proteases, trypsin and related serine proteases, acetylcholinesterase, renin, HIV protease and thermolysin[2,3,6,18].

Two closely related QSAR models were derived for the inhibition of *E. coli* DHFR and *L. casei* DHFR by

Box 2. 3D QSARs of enzyme ligands (substrates and inhibitors)[a]

Hydrolases	
Acetylcholinesterase	(225/96)
Angiotensin-converting enzyme (ACE)	(62/94, 64/94)
Chymotrypsin	(305/95)
Dipeptidyl peptidase	(338/96)
Glycogen phosphorylase	(306/95)
HIV protease	(361/94, 219/95)
Papain	(364/94, 491/94)
Renin	(326/93)
Thermitase	(226/96)
Thermolysin	(326/93, 62/94, 64/94, 304/95)
Oxidoreductases	
Aromatase	(410/96)
Cytochrome P450	(144/96)
Dihydrofolate reductase	(70/95, 222/96)
Lanosterol-14α-demethylase	(411/96)
Monoaminoxidases	(143/96,492/96)
Ligases	
Phenylethanolamine *N*-methyltransferase	(502/94)
Other enzymes	
HIV integrase	(543/95)
Topoisomerase	(491/96)

[a] The numbers in brackets refer to number/year of an abstract published in *Quant. Struct.-Act. Relatsh.* (years 1993–1996 only)

benzylpyrimidines, (Eqns 3 and 4; π_X and MR'_X = lipophilicity and molar refractivity parameters of substituent X)[3,4,18].

Escherichia coli DHFR

$$\log 1/K_{i\,app} = 0.75\ (\pm 0.26)\ \pi_{3,4,5} - 1.07\ (\pm 0.34)\ \log(\beta \cdot 10^{\pi_{3,4,5}} + 1) + 1.36\ (\pm 0.24)\ MR'_{3,5} + 0.88\ (\pm 0.29)\ MR'_4 + 6.20 \qquad (3)$$

$$\log \beta = 0.12 \qquad \text{optimum } \pi = 0.25$$

$$(n = 43;\ r = 0.903;\ s = 0.290)$$

Lactobacillus casei DHFR

$$\log 1/K_{i\,app} = 0.31\ (\pm 0.11)\ \pi_{3,4} - 0.88\ (\pm 0.24)\ \log(\beta \cdot 10^{\pi_{3,4}} + 1) + 0.95\ (\pm 0.21)\ MR'_{3,4} + 5.32 \qquad (4)$$

$$\log \beta = -1.33 \qquad \text{optimum } \pi = 1.05$$

$$(n = 42;\ r = 0.876;\ s = 0.222)$$

The only difference between these equations is that 5-substituents of the benzyl group contribute to biological activities in the case of *E. coli* DHFR, whereas they have no effect in the case of *L. casei* DHFR. An explanation could be given as soon as the X-ray structures of the enzymes were available. Both have about the same geometry of the binding site but a rigid leucine side chain in the *L. casei* DHFR forms a much narrower cleft than a more flexible methionine side chain in *E. coli* DHFR[3,4,18].

The inhibition of monoamine oxidase by amines (at different pH values) and alcohols (Eqn 5; I = 0 for amines, I = 1 for alcohols) corresponds to an equilibrium system, where the biological data have to be corrected for the concentration of the unionized form[3].

$$\log 1/K_i + \log(1 + 10^{pK_a - pH}) = 3.130\ (\pm 0.17)\ \log P - 3.797\ (\pm 0.32)\ \log(\beta P + 1) - 3.507\ (\pm 0.12)\ I + 3.379 \qquad (5)$$

$$\log \beta = -1.781 \qquad \text{optimum } \log P = 2.45$$

$$(n = 21;\ r = 0.999;\ s = 0.118)$$

Other biological data

The affinities of receptor, transporter and ion channel ligands have mainly been investigated by 3D QSAR methods.

Box 3. 3D QSARs of receptor, transporter and ion channel ligands[a]

Receptors	
5-HT (serotonin) receptors	(327/93, 215/94, 308/95, 41/96, 223/96)
α_1-adrenergic receptor	(365/94, 495/94)
ATII (angiotensin II) receptor	(397/95)
Benzodiazepine receptor	(328/93, 311/95)
CCK_A (cholecystokinin) receptor	(500/94)
Dioxine (Ah) receptor	(187/93)
Dopamine receptors	(398/95, 337/96)
Estrogen receptor	(339/96)
ET_A (endothelin) receptor	(483/95)
GHRH (gonadotropin hormone release hormone) receptor	(544/95)
Morphine receptors	(312/95, 482/95)
NK_1 (neurokinin) receptor	(142/96)
Purine receptor	(409/96)
Transporters	
Corticosteroid binding globulin	(304/95, 481/95)
Dopamine transporter	(496/94, 310/95)
Testosterone binding globulin	(304/95, 481/95)
Ion channels	
Calcium channel	(68/95)
Chloride channel	(217/95)

[a] The numbers in brackets refer to number/year of an abstract published in *Quant. Struct.-Act. Relatsh.* (years 1993–1996 only)

Although receptor modelling, based on the 3D structure of bacteriorhodopsin, has advanced in recent years, so far the models built for G-protein-coupled receptors (GPCR) do not allow a structure-based ligand design or the derivation of QSARs[6]. Thus, CoMFA and related field-based approaches offer the only chance to derive such relationships and to predict the affinities of new ligands. Some recent CoMFA applications to receptor, transporter and ion channel ligands are listed in Box 3.

In vivo data should not be correlated by 3D field-based methods because absorption, distribution, metabolism and elimination (the ADME parameters) overlap with the ligand–protein interaction and obscure the results. Such complex structure–activity relationships are better modelled by classical QSAR models, especially Hansch analysis. Often one step in the complex cascade is responsible for the derived QSAR equation, e.g. the concentration in the compartment where the drugs exert their action. QSAR applications in different therapeutic fields have been reviewed (Box 4).

Box 4. Other QSAR applications

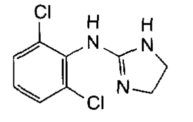
Cardiovascular agents
Antiadrenergic agents[3]
Antihypertensive agents[3]
Calcium antagonists[2,3]

CNS active agents
General anesthetics[4]
Anticonvulsants[4]
Antidepressants[2]
CNS stimulants[4]

Oncology
Mutagenicity[3,4]
Carcinogenicity[3,4]
Cancer chemotherapy[2–4]
Multiple drug resistance[3,4]

Antimicrobial agents
Antiviral agents[4]
Antibacterial agents[4]
Antifungal agents[4]
Antiprotozoal agens[3,4]

Metabolism
Prodrugs[4]
Microsomal oxidation[4]
Glucuronidation and other conjugations[4]
Elimination[4]



1

***Figure 2.** The antihypertensive drug clonidine.*

Activity–activity relationships

An important application of QSAR is the quantitative description of activity–activity relationships[2,3,5,19]. Nowadays, when combinatorial chemistry produces thousands and ten thousands of analogs within certain series that are tested by high-throughput screening methods, such relationships become more and more important. Only with the help of a standard series of analogs, for which *in vitro* and *in vivo* activities have been successfully correlated, can the usefulness and validity of the chosen *in vitro* test model be justified.

The application of quantitative activity–activity relationships in drug development shall be illustrated by a group of analogs of the antihypertensive drug clonidine **1** (Figure 2), for which different biological activities are observed. In normal, anesthetized rats, hypotension results after an intravenous application of the compounds. Their partition coefficients (log P at pH 7.4), *in vitro* activities [binding affinities to α_1 and α_2 receptors, $IC_{50}(\alpha_1)$ and $IC_{50}(\alpha_2)$] and *in vivo* activities (central antihypertensive activities C_{25} and peripheral hypertensive activities C_{60}) were determined. Correlation of the peripheral activities, expressed by log $1/C_{60}$ values, with the α_2 binding affinities, expressed by $IC_{50}(\alpha_2)$, led to Eqn 6 (Refs 3,5).

$$\log 1/C_{60} = 1.163\ (\pm 0.21)\ \log 1/IC_{50}(\alpha_2) - 0.962\ (\pm 0.39) \quad (6)$$
$$(n = 21;\ r = 0.936;\ s = 0.317;\ F = 135.15)$$

In decerebrated rats, without any central nervous system function (so-called pithed rats), the same compounds show an opposite effect, a hypertensive activity. In order to find out whether the two biological activities are related in their mechanism of action and whether the unwanted hypertensive activity can be separated from the therapeutically desired antihypertensive effect, Eqns 7 and 8 were derived from the experimental data.

$$\log 1/C_{25} = 0.805\ (\pm 0.22)\ \log P - 3.373\ (\pm 1.02)\ \log(\beta P + 1) + 1.071\ (\pm 0.20)\ \log 1/IC_{50}(\alpha_2) - 1.164\ (\pm 0.39) \quad (7)$$
$$\log \beta = -1.986 \qquad \text{optimum } \log P = 1.48$$
$$(n = 21;\ r = 0.971;\ s = 0.284;\ F = 65.22)$$

$$\log 1/C_{25} = 0.784\ (\pm 0.26)\ \log P - 3.685\ (\pm 1.39)\ \log(\beta P + 1) + 0.830\ (\pm 0.20)\ \log 1/C_{60} - 0.189\ (\pm 0.30) \quad (8)$$
$$\log \beta = -2.078 \qquad \text{optimum } \log P = 1.51$$
$$(n = 21;\ r = 0.954;\ s = 0.354;\ F = 40.52)$$

Eqns 6–8 clearly indicate that

- hypertensive activities are correlated with α_2-agonistic activities (Eqn 6);

- antihypertensive activities are also correlated with α_2-agonistic activities but there is an additional nonlinear lipophilicity dependence (Eqn 7), because the compounds have to cross the blood–brain barrier on their way from the site of application to the site of action; as a consequence,
- hypertensive and antihypertensive activities can be correlated with each other, including a nonlinear lipophilicity relationship, because both effects result from the interaction of the compounds with the same receptors (Eqn 8); stimulation of α_2 receptors in the brain causes a CNS-mediated blood pressure decrease, whereas stimulation of peripheral, vascular α_2 receptors causes vasoconstriction which results in a blood pressure increase.

The conclusions from Eqn 8 are that the activities cannot be separated, but for optimum antihypertensive activity the lipophilicity of the compounds should be around log P = 1.5. Such analogs can readily cross the blood–brain barrier and their central effect will override the peripheral effect. All other attempts to find chemically related analogs with improved selectivity must fail. Thus, no more syntheses and testing of compounds within this series need to be performed.

Several other examples are known where desired activities have been compared with toxicities, especially for antitumor drugs. In some cases the syntheses of further analogs were stopped because no better therapeutic indices could be expected[3,4].

Current problems in QSAR applications

Despite the fact that thousands of successful QSAR applications and hundreds of CoMFA studies prove the usefulness of these approaches, there are several problems in their proper application. For classical QSAR studies, the most important recommendations were already summarized in 1973, by Unger and Hansch[20]:

- Several independent physicochemical parameters should be tested.
- All 'reasonable' parameters should be validated.
- The simplest model should be accepted.
- A minimum number of compounds should be included for each variable.
- The regression equation should be explained by a qualitative model.

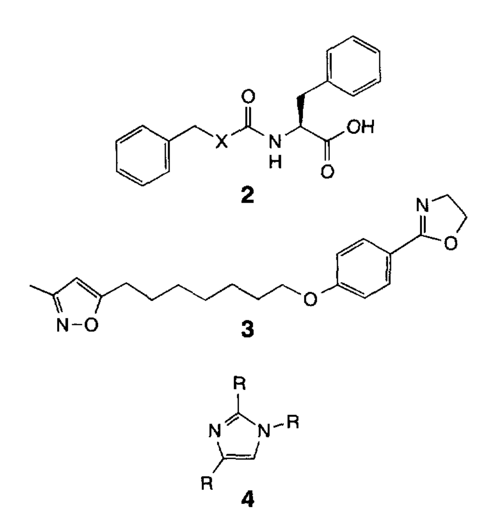
More recently, recommendations for the proper application of 3D QSAR approaches have also been defined[6,21]. In a short version they are summarized below:

- The selection of starting geometries should be rationalized.
- Methods of geometry optimization should be documented.
- Charges used in CoMFA and their calculation method should be defined.
- Alignment criteria and all options (box, grid size, etc.) should be given.
- Scaling and weighting of fields should be documented.
- Cross-validation runs should be performed for every analysis.
- Statististical data for fit and internal predictivity should be given.
- The number of (significant) PLS components must be presented.
- Typical problems in cross-validation should be considered.
- Removed outliers should be mentioned and discussed.
- Variable selection procedures should be used whenever appropriate.
- Contour maps of the final model should be provided or at least discussed.
- Origins of biological data should be documented.
- Standard errors of biological data should be given.
- A table with all observed vs predicted values should be provided.
- Coordinates of the molecules in the used alignments should be available.
- Predictions of biological activity values depend on the training set.

These recommendations should help to avoid the most common errors and pitfalls and should ease the reproduction of CoMFA results by other scientists. Some special problems in QSAR and 3D QSAR studies are discussed in more detail in the following sections.

Unexpected changes in the binding modes

As already mentioned, a most difficult problem in 3D QSAR studies (but implicitly also in classical QSAR) is the derivation of orientation rules for the alignment of protein ligands. Whereas the binding mode of methotrexate, as compared with dihydrofolate[1], is predictable from the different patterns of hydrogen bond donor and acceptor groups, there are many cases where ligands show unexpected binding modes or even multiple binding modes; examples are compounds **2–4** (Figure 3)[3,6,19,22,23].



2

3

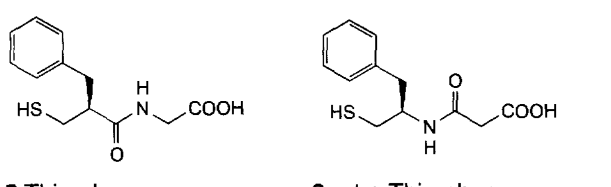
4

Figure 3. *The pair* N*-carbobenzoxy-*L*-phenylalanine (**2**, X = O) and* N*-β phenylpropionyl-*L*-phenylalanine (**2**, X = CH_2) and some rhinovirus coat protein ligands, e.g.* ***3***, *show, despite their chemical similarity, binding modes where some analogs are oriented in a different direction than the other ones. All three isomeric phenylimidazoles* ***4*** *(R = phenyl, H, H) are cytochrome P450 inhibitors, but only 1- and 4-phenylimidazole bind as expected; 2-phenylimidazole shows a completely different binding mode.*

How similar are similar compounds?

Most often similar compounds show similar biological activities. However, there are examples where seemingly analogous compounds have very different activities. Thiorphan **5** and *retro*-thiorphan **6** differ only by the direction of the amide group that connects the benzyl-substituted thiol with the acidic part of the molecule (Figure 4)[3,19]. Can these compounds be considered to be similar or not?

The 3D structures of the complexes of both compounds with the bacterial zinc protease thermolysin have been determined. Both compounds show identical binding modes, which are reflected by their identical thermolysin inhibitor activities; the compounds may be considered to be 'similar'. They also inhibit neutral endopeptidase, NEP 24.11, with even higher affinity constants. However, against angiotensin-converting enzyme (ACE), a zinc protease related to thermolysin and NEP 24.11, they show very different inhibitory activities[24]. The structural reasons for this different behavior will remain unknown, as long as we do not know the exact 3D structure of ACE.



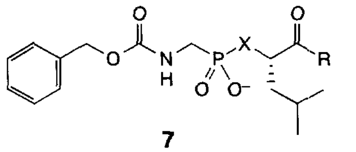
5 Thiorphan		**6** *retro*-Thiorphan	
K_i (thermolysin)	1.8 μM	K_i (thermolysin)	2.3 μM
K_i (NEP 24.11)	0.0019 μM	K_i (NEP 24.11)	0.0023 μM
K_i (ACE)	0.14 μM	K_i (ACE)	>10 μM

Figure 4. *The 'biological similarity' of thiorphan* ***5*** *and* retro*-thiorphan* ***6*** *differs significantly with respect to their biological target.*

Minor structural changes of an active analog may significantly reduce its biological activity. An example is the thermolysin inhibitors **7** (Figure 5)[25].

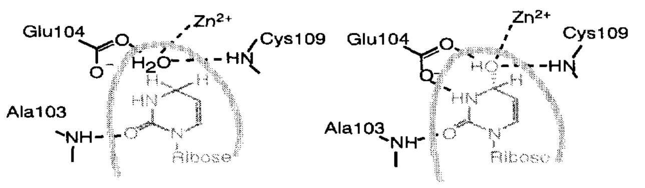
So far the biggest effect of a small structural change is observed for some transition state inhibitors. The explanation for the striking activity differences between the cytosine deaminase inhibitors **8** and **9** can be derived from the X-ray structures of the protein–inhibitor complexes. Compound **9** corresponds to the transition state of the enzymatic reaction. Its hydroxyl group is optimally suited to form several hydrogen bonds and to replace the water molecule that is coordinated to the Zn^{2+} ion and added to cytosine during the deamination. On the other hand, compound **8** does not replace this water molecule and obviously fits less well into the binding site (Figure 6)[26].

All these analogs show the same type of activity with only quantitative differences. But there are also qualititative differences, such as differences in the mechanism of action and in the therapeutic applicability of chemically related compounds, for example the CNS-active drugs **10–12** (Figure 7).



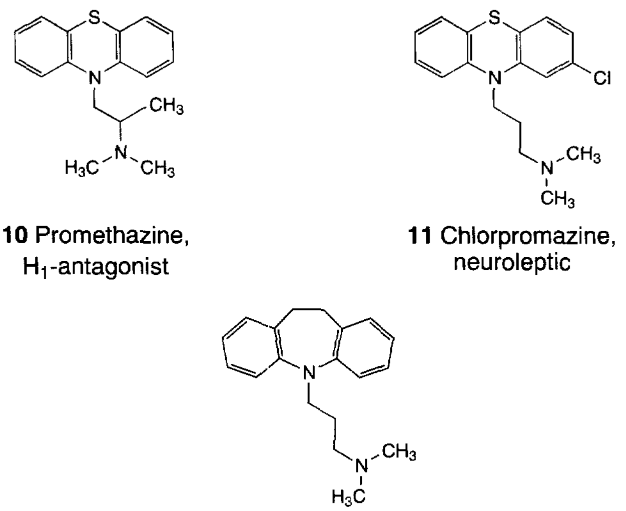
7

Figure 5. *The exchange of X = NH or CH_2 against X = O reduces the inhibitory activities of the thermolysin inhibitors* *7* *by three orders of magnitude.*



8 K_i = 30 μM   9 K_i = 1.2 pM

***Figure 6.** The introduction of a single hydroxy group into the weakly active cytosine deaminase inhibitor **8** leads to analog **9**, which corresponds to the transition state of the enzymatic reaction. Correspondingly, the inhibitor **9** is much more tightly bound; its binding affinity is more than seven orders of magnitude higher than the affinity of compound **8**, where the water molecule cannot be replaced by the ligand.*



10 Promethazine, H_1-antagonist

11 Chlorpromazine, neuroleptic

12 Imipramine, antidepressant

***Figure 7.** The tricyclic compounds promethazine **10** (an H_1-antihistaminic), chlorpromazine **11** (a dopamine-antagonistic neuroleptic) and imipramine **12** (an antidepressant, which acts as a norepinephrine- and serotonine-uptake inhibitor) show significantly different biological actions, despite their close chemical similarity.*

Other illustrative examples are α- and β-adrenergic agonists and antagonists, the male and female steroid hormones, and the gluco- and mineralocorticosteroids.

Variable selection

The proper application of regression analysis requires the formulation of a working hypothesis, the design of experiments (i.e. the compounds to be tested), the selection of a mathematical model, and a test of the statistical significance of the obtained result. However, QSAR studies are most often retrospective studies. Several different variables are tested to find out whether some of them are able to describe the data, alone or in certain combinations. As long as QSAR equations are only used to derive new hypotheses and to design new experiments, the requirements for the application of statistical methods are fulfilled.

Forward selection, backward elimination and stepwise variable selection methods are frequently used to find the most relevant variables. However, these methods work properly only in the case of non-correlated variables and if there are not too many variables. In all other cases they tend to produce suboptimal results which may be far from the global 'best' model. In addition to other approaches, evolutionary and genetic algorithms have proved useful to derive good statistical models[27,28]. In evolutionary algorithms (EA) and genetic algorithms (GA), first one (EA) or several (GA) variable combinations are randomly selected (the start model/s). Then these models are reproduced, performing mutations (EA) or crossovers (GA). The better model (EA) or the best models (GA) are kept, the others are discarded. In the next cycle, again mutation (EA) or crossover (GA) are applied. After repeating these procedures, in most cases a few hundred times, stable models (EA) or stable populations of models (GA) result. Evolutionary algorithms are much faster but many independent runs have to be performed to obtain several good models. Genetic algorithms give a selection of good models; however, because mutation plays a minor role in genetic algorithms, sometimes important variables are missed, they 'die out' during the selection process.

Partial least squares (PLS) analysis is an alternative to variable selection methods, especially in the case of many, highly intercorrelated variables[3,6,29,30]. However, if too many variables not contributing to fit and prediction are involved, PLS analysis fails. Then even in PLS analysis a variable selection procedure has to be applied. Clementi developed the variable selection method GOLPE (generating optimal linear PLS estimations)[6,31]. In GOLPE, first a D-optimal design preselects non-redundant variables and then a fractional factorial design procedure is used to run several PLS analyses with different combinations of these variables. Variables significantly contributing to prediction are selected, whereas all others are discarded. Due to the D-optimal design selection of variables in the first step, many highly interrelated variables are eliminated; accordingly, the resulting CoMFA contour maps are not as smooth as they should (and could) be.

Validation of QSAR models

Already 25 years ago, Topliss pointed out that the risk of chance correlations increases significantly with the number of tested variables and with the number of variables included in the final model compared with the number of compounds for which the model is derived[32]. Validation methods are needed. Obviously the best method is to split the group of compounds into two subgroups, a training set for which the model is derived and a test set for which biological activities are predicted. However, this can only be done in relatively large sets of compounds and only if the structural variations in the training set and in the test set are comparable, as will be shown below. Of course, all data may be described by one model; then predictions for new compounds, not included in the original data set, are made and tested (see below). An alternative to training and test selections is cross-validation in groups[3,6,29,30,33].

Hansch recommended a lateral validation of QSAR results, namely the comparison of models of closely related series of compounds in one biological test system or the comparison of the QSAR models derived for one series of compounds in several related biological test models (for example serine and cysteine proteases)[3,4,34]. If all models are of comparable quality and if they show similar regression coefficients of the physicochemical terms, the results can be accepted. However, in most cases the required effort will be too large to do this routinely. In addition, even closely related enzymes or receptors may have significantly different binding sites.

Predictions: the selection of training and test sets

The most demanding problem in QSAR and 3D QSAR studies is the suitability of a model to derive predictions for new analogs. Thus, it is often recommended to split a data set (if it is large enough) into a training set to derive the model and a test set to check its external predictivity. How critical training and test set selections are, can be illustrated by a simple example. In his first CoMFA study, Cramer separated the data set of 31 steroids into a training set (compounds 1–21) and a test set (compounds 22–31)[35]. The training set can be easily described by a one-parameter free Wilson equation; 4,5-C=C- encodes the presence or absence of a cycloaliphatic 4,5-double bond in ring A of the steroids (Eqn 9; CBG = corticosteroid binding globulin affinities; Q^2 = squared cross-validation correlation coefficient; s_{PRESS} = standard deviation of cross-validation predictions)[36].

$$\log 1/\mathrm{CBG} = 2.022\ (\pm 0.52)\ 4{,}5\text{-C=C-} + 5.186\ (\pm 0.36) \qquad (9)$$
$$(n = 21;\ r = 0.882;\ s = 0.568;\ F = 66.41;\ Q^2 = 0.726;\ s_{PRESS} = 0.630)$$

The external predictivity of this model (test set: compounds 22–31; $n = 10$; $r^2_{pred} = 0.477$, $s_{PRESS} = 0.733$) is not as good as the internal predictivity because some structural features in the test set (compounds 23 and 31) are not covered by the training set. If, for example, compounds 1–12 and 23–31 are selected as the training set, a model with slightly worse fit and internal predictivity ($n = 21$; $r = 0.731$; $s = 0.697$; $F = 21.82$; $Q^2 = 0.454$; $s_{PRESS} = 0.754$) results. However, this model is especially suited for test set prediction (compounds 13–22; $n = 10$; $r^2_{pred} = 0.909$, $s_{PRESS} = 0.406$) which provides striking evidence that obviously all structural features of the test set compounds are well represented by the training set compounds.

The training set compounds should span a parameter space in which all data points are more or less equally distributed. The structures and all relevant properties of the test set compounds should not be too far from the training set compounds. To derive statistical models with reasonable experimental effort, an appropriate design scheme should be used to cover the property space with a small number of objects. Redundancy is minimized by following this recommendation. On the other hand, serious problems may arise if the redundancy of properties is too much reduced: cross-validation is no longer applicable in such series and single point errors may distort the final QSAR model. Especially the latter topic is most often neglected as a possible source of poor test set predictions. Reliable results for the test set predictions can only be expected by including sufficient redundancy in the training set compounds.

Contributions of QSAR and 3D QSAR to drug design

The contributions of QSAR and 3D QSAR to drug design are manifold. The role of lipophilicity as well as dissociation and ionization in drug absorption, transport and distribution could only be understood after correlating the experimental data by appropriate models.

Ligand–protein interactions depend on different interactions:

- The hydrophobic interaction, i.e. the lipophilic contact area; in Hansch analysis it is modelled by different lipophilicity parameters, in 3D QSAR by the steric or hydrophobic fields.
- The formation of neutral and charged hydrogen bonds; in Hansch analysis the relative strength of such hydrogen bonds is modelled by the electronic Hammett parameters, in 3D QSAR by the electrostatic field or by hydrogen bond donor–acceptor fields.
- Steric requirements; in Hansch analysis they are modelled by the MR (molar refractivity) parameter or by certain steric parameters, whereas steric fields are used in 3D QSAR studies.

QSAR and 3D QSAR should be considered as tools to derive hypotheses which should be proven or disproven by further syntheses and biological tests. Understanding a structure–activity relationship is the main goal. Predictions are only a means for the design of new analogs. Despite several difficulties and problems in the proper application of QSAR and 3D QSAR, its value in lead structure optimization is documented by many examples[2–4,6,12]. Several dedicated publications[37–39], including a recent review by Toshio Fujita[40], list success stories leading to commercializable bioactive compounds with the aid of traditional QSAR procedures, in medicinal as well as in agricultural chemistry.

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 to a minimum and to provide the corresponding results in the context of related work in the same field.

The journal *Quantitative Structure–Activity Relationships* publishes, in addition to original contributions, about 500–600 detailed abstracts every year on scientific papers in the fields of QSAR, 3D QSAR and molecular modelling.

REFERENCES

1 Kubinyi, H. (1997) *Drug Discovery Today* 2, 457–467 (Part 1)

2 Ramsden, C.A., ed. (1990) *Quantitative Drug Design* (*Comprehensive Medicinal Chemistry. The Rational Design, Mechanistic Study & Therapeutic Application of Chemical Compounds*) (Vol. 4) (Hansch, C., Sammes, P.G. and Taylor, J.B., eds), Pergamon Press

3 Kubinyi, H. (1993) *QSAR: Hansch Analysis and Related Approaches*, VCH

4 Hansch, C. and Leo, A. (1995) *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology*, American Chemical Society

5 Kubinyi, H. (1995) in *Burger's Medicinal Chemistry* (Vol. 1, 5th edn) (Wolff, M.E., ed.), pp. 497–571, John Wiley & Sons

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

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

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

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

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

11 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, Greenwich, CT, USA

12 Kubinyi, H., Folkers, G. and Martin, Y.C., eds (1997) *3D QSAR in Drug Design. Volume II: Ligand–Protein Interactions and Molecular Similarity* and *Volume III: Recent Advances*, Kluwer Academic Publishers

13 van de Waterbeemd, H. and Kansy, M. (1992) *Chimia* 46, 299–303

14 Pliska, V., Testa, B. and van de Waterbeemd, H., eds (1996) *Lipophilicity in Drug Action and Toxicology*, VCH

15 Hansch, C., Björkroth, J.P. and Leo, A. (1987) *J. Pharm. Sci.* 76, 663–687

16 Camenisch, G., Folkers, G. and van de Waterbeemd, H. (1996) *Pharm. Acta Helv.* 71, 309–327

17 Navia, M.A. and Chaturvedi, P.R. (1996) *Drug Discovery Today* 1, 179–189

18 Blaney, J.M. *et al.* (1984) *Chem. Rev.* 84, 333–407

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

20 Unger, S.H. and Hansch, C. (1973) *J. Med. Chem.* 16, 745–749

21 Thibaut, U. *et al.* (1994) *Quant. Struct.-Act. Relatsh.* 13, 1–3

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

23 Meyer, E.F. (1995) *Perspect. Drug Des. Discovery* 3, 168–195

24 Roques, B.P. *et al.* (1993) *Pharmacol. Rev.* 45, 87–146

25 Morgan, B.P. *et al.* (1991) *J. Am. Chem. Soc.* 113, 297–307

26 Xiang, S. *et al.* (1995) *Biochemistry* 34, 4516–4523

27 Rogers, D. and Hopfinger, A.J. (1994) *J. Chem. Inf. Comput. Sci.* 34, 854–866

28 Kubinyi, H. (1996) *J. Chemomet.* 10, 119–133

29 van de Waterbeemd, H., ed. (1995) *Chemometric Methods in Molecular Design*, VCH

30 van de Waterbeemd, H., ed. (1995) *Advanced Computer-Assisted Techniques in Drug Discovery*, VCH

31 Baroni, M. *et al.* (1992) *Quant. Struct.-Act. Relatsh.* 12, 9–20

32 Topliss, J.G. and Costello, R.J. (1972) *J. Med. Chem.* 15, 1066–1068

33 Cramer, R.D., III *et al.* (1988) *Quant. Struct.-Act. Relatsh.* 7, 18–25, erratum (1988) 7, 91

34 Hansch, C. (1993) *Acc. Chem Res.* 26, 147–153

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

36 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

37 Fujita, T. (1990) in *Quantitative Drug Design* (*Comprehensive Medicinal Chemistry. The Rational Design, Mechanistic Study & Therapeutic Application of Chemical Compounds*) (Vol. 4) (Hansch, C., Sammes, P.G. and Taylor, J.B., eds), pp. 497–560, Pergamon Press

38 Boyd, D.B. (1990), in *Reviews in Computational Chemistry* (Vol. 1) (Lipkowitz, K.B. and Boyd, D.B., eds), pp. 355–371, VCH

39 Fujita, T. (1992) in *QSAR in Design of Bioactive Compounds* (Kuchar, M., ed.), pp. 3–22, Prous Science Publishers

40 Fujita, T. (1997) *Quant. Struct.-Act. Relatsh.* 16, 107–112