Dr David Winkler
is a Senior Principal Research Scientist at CSIRO Molecular Science, and an Honorary Senior Research Fellow at Monash University. His research interests include *de novo* molecular design, and development of novel QSAR, pattern recognition and data mining methods.

The role of quantitative structure–activity relationships (QSAR) in biomolecular discovery

David A. Winkler

Received (in revised form): 22nd November 2001

Abstract

Empirical methods for building predictive models of the relationships between molecular structure and useful properties are becoming increasingly important. This has arisen because drug discovery and development have become more complex. A large amount of biological target information is becoming available through molecular biology. Automation of chemical synthesis and pharmacological screening has also provided a vast amount of experimental data. Tools for designing libraries and extracting information from molecular databases and high-throughput screening experiments robustly and quickly enable leads to be discovered more effectively. As drug leads progress down the development pipeline, the ability to predict physicochemical, pharmacokinetic and toxicological properties of these leads is becoming increasingly important in reducing the number of expensive, late development failures. Quantitative structure–activity relationship (QSAR) methods have much to offer in these areas. However, QSAR analysis has many traps for unwary practitioners. This review introduces the concepts behind QSAR, points out problems that may be encountered, suggests ways of avoiding the pitfalls and introduces several exciting, new QSAR methods discovered during the last decade.

Keywords: *QSAR, molecular descriptors, drug design, structure–property relationships, neural networks*

INTRODUCTION

Many important properties of matter are dependent on the microscopic properties of the molecules from which it is made up. This is particularly true when biomolecular properties are considered. For bioactive molecules, sometimes a single residue change in a peptide sequence, or a single atomic change in a small organic molecule, will profoundly affect the biological properties of these species. At an even more subtle level, chemically identical but isomerically distinct compounds, such as optical isomers or enantiomers, can exhibit very different biological properties. An important, albeit tragic, example of this is thalidomide. One stereoisomer exhibited a beneficial sedative/hypnotic effect and was not teratogenic, while the mirror image enantiomer was a less effective sedative and highly teratogenic.

The often-complex relationship between molecules and their properties occurs not only in biomolecular systems, but in physicochemical properties (eg water solubility), materials properties (eg glass transition temperature) and many other processes as well. These structure–property relationships can often be studied by very computationally intensive methods such as molecular dynamics, or *ab initio* molecular orbital methods. In biochemical systems, methods such as molecular docking and transition state analogue design can be used to understand relationships between molecular structure and biological properties in considerable detail, providing molecular information (eg a protein crystal structure) about the target, or the biochemical reaction catalysed, eg by an enzyme, is available. Even when this detailed information is available, the

David A. Winkler,
Senior Principal Research Scientist,
CSIRO Molecular Science,
Private Bag 10,
Clayton South MDC,
Clayton 3169,
Australia

Tel: +61 3 9545 2222
Fax: +61 3 9545 2446
E-mail:
Dave.winkler@molsci.csiro.au

High level methods often intractable

details of the relationship between structure and activity are so complex that many assumptions and simplifications must be used in order to make the calculations tractable.

Empirical QSAR methods deal with complexity

An alternative approach is a phenomenological one – studying a series of molecules of differing structure and different observed properties, and attempting to find empirical relationships between structure and property. This method is useful even when considerable information is available about biological mechanisms, but is essential when this information is not available – the majority of situations. There are families of methods that use this approach that differ mainly in the types of properties they model. Quantitative structure-activity relationships (QSAR) is the name usually applied to methods of this type which correlate molecular structure to some kind of *in vitro* or *in vivo* biological property. When this approach is applied to modelling of toxicological data, it is termed quantitative structure–toxicity relationships (QSTR). When applied to modelling of physicochemical properties it is called quantitative structure–property relationships (QSPR). QSAR in particular, first developed by Hansch and Fujita 40 years ago, has been invaluable for understanding drug structure–activity relationships for lead discovery and optimisation.

Many biological, toxicological and physical properties can be predicted

The ability to generate predictive models for toxicological, physicochemical and absorption, disposition, metabolism and excretion (ADME) processes has gained new momentum in the past decade. This is because of the recognition of the importance of these models in building 'developability' into drug leads, resulting in fewer expensive downstream failures in the drug development process. This paper introduces QSAR methods, and additionally discusses several important developments in the field that have taken place in the last decade. For more comprehensive information on QSAR and related methods, the reader should consult some of the many reviews of the role of QSAR in drug design and development,[1–10] and papers that summarise the successes of the QSAR method.[11,12] Table 1 lists a number of sources of data, descriptors, software and information on QSAR.

QSAR methods that deal with classification rather than quantitative model building (so-called qSAR methods) have been dealt with only briefly in this review. Readers are referred to recent papers on classification methods for more in-depth discussions of qSAR.[13,14]

QSAR METHOD

The QSAR method involves recognition that a molecule (organic, peptide, protein, etc.) is really a three-dimensional distribution of properties. The most important of these properties are steric (eg shape and volume), electronic (eg electric charge and electrostatic potential), and what are termed 'lipophilic' properties (how polar or non-polar the sections of the molecular are, usually exemplified by the log of the octanol–water partition coefficient, log *P*). Scientists are used to visualising mainly steric properties of

Predicting 'developability' is important to save costs

Table 1: Some sources of software, data and information relevant to QSAR

Information	Source
Topological index software (MolconnZ)	http://www.edusoft-lc.com
CODESSA descriptor software	http://www.semichem.com/codessa.html
Dragon descriptor software	http://www.disat.unimib.it/chm/Dragon.htm
Pomona College QSAR site	http://clogp.pomona.edu/medchem/chem/qsar-db
CoMFA 3-D QSAR method	http://www.tripos.com/software/qsar.html
QSAR and Modelling Society	http://www.ndsu.nodak.edu/qsar_soc/index.htm
Summaries of QSAR studies in literature	Journal *Quantitative Structure–Activity Relationships*

molecules. However, molecules 'look' different when viewed in electrostatic, or lipophilic 'space' (Figure 1).

QSAR involves four main steps

The QSAR method (and analogously QSTR and QSPR) involves a number of key steps:

- Converting molecular structures into mathematical descriptors that encapsulate the key properties of the molecules relevant to the activity or property being modelled.
- Selecting the best descriptors from a larger set of accessible, relevant descriptors.
- Mapping the molecular descriptors into the properties, preferably using a 'model-free' mapping system in which no assumptions are needed as to the functional form of the structure–activity relationship. These relationships are often complex, unknown and non-linear.
- Validating the model to determine how predictive it is, and how well it will generalise to new molecules not in the data set used to generate the model (the training set).

This review will be concerned mainly with the application of QSAR to the quantitative modelling of biological properties. The relationship between molecular structure and some biological response, BR (eg IC_{50}, LD_{50}, ED_{90}) can be expressed as:[1]

$$\log(\mathrm{BR}) = f(x_1, x_2, \ldots, x_N)$$

where f is usually an unknown, complex, non-linear function, and $x_1, \ldots, x_N$ are molecular descriptors. Building of a QSAR model via the four steps outlined above involves finding the best form of function f.

GENERATION OF DESCRIPTORS

Many types of descriptors are available

There are a myriad methods for generating molecular descriptors. Packages such as Dragon are able to generate over a thousand descriptors, while methods such as CoMFA[15] (discussed below) generate many thousands. Molecular descriptors can be of diverse types. We have chosen to categorise them into fragment descriptors, involving properties of sections of molecules, and whole molecule descriptors, based on the properties of the intact molecule.

Fragment descriptors

The very earliest descriptors used in QSAR were of this type.[1,10] QSAR was performed using 'substituent constants'

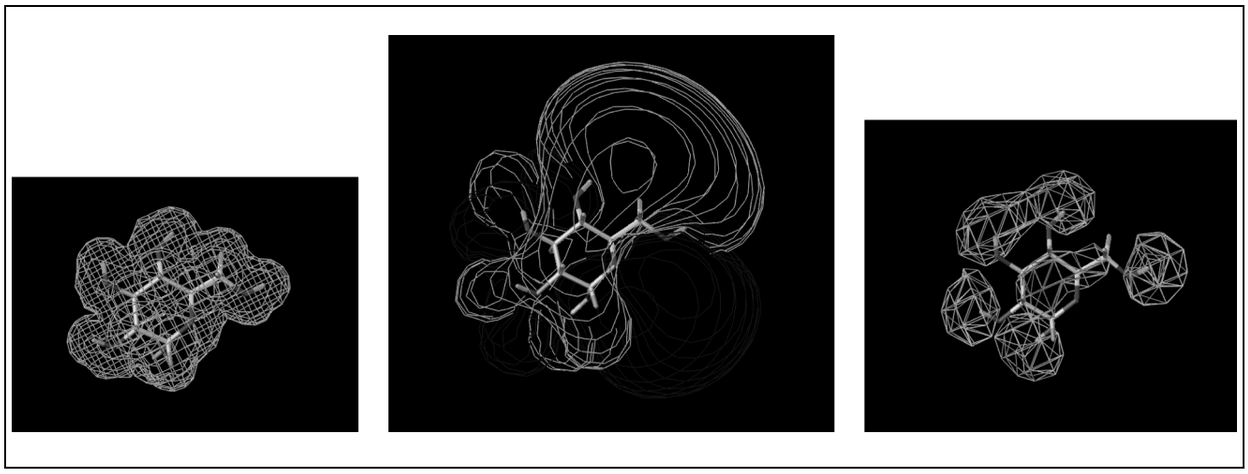

Figure 1: A small organic molecule (glucopyranose) viewed in steric (left), electrostatic (centre) and lipophilic (right) space

such as hydrophobic constants π, molar refractivity MR, Hammett constants σ and several other, less well-known constants. The hydrophobic constant π is related to the difference between the log of the partition coefficient of an unsubstituted molecule and that of a molecule with given substituent. It measures the polar/non-polar character of the substituent. MR is a measure of size and polarisability of substituents and is derived from the Lorentz–Lorenz equation and is a function of refractive index, density and molecular weight. The Hammett constant is a measure of electron-withdrawing or electron-donating properties of substituents and was derived originally from the differences between the log of the ionisation constant of benzoic acid and that of substituted benzoic acids. An alternative approach is the Free Wilson method where indicator variables are used to show whether a particular chemical group is present (1) or absent (0) at a given position on a chemical structure.[16] Both of these methods work well for fairly homologous series of molecules where there was a common molecular scaffold, but are unsuitable for data sets in which the core structure was different. Compilations of these and related substituents constants are available in the literature.[10]

Descriptors need to be informative

The recent explosion in the number of molecular descriptors is partly due to the ease by which they may be generated by computational methods, such as molecular orbital calculations[17–19] (recent examples where this has been done on a grand scale may be found in the papers by Clark and coworkers).[20] There has also been a focus on developing fragment descriptors that are very computationally efficient. The reason is that rapid searching for leads in large chemical libraries (databases of real chemical compounds) or virtual libraries (databases of chemically reasonable molecules that have not yet been synthesised) require efficient, information-rich descriptors. Surprisingly simple descriptors can yield useful models. For example, molecules may be represented simply by counting the numbers of atoms of specific elemental type, with specific numbers of connections (a measure of atomic hybridisation).[21,22] Although simple, this representation is adequate to encode not only physicochemical parameters, such as lipophilicity and molar refractivity, but also biological activity (eg $GABA_A$ receptor activity of benzodiazepines). Subtle, higher-level information is captured by these simple descriptors.[23–25]

Fragment descriptors are computationally efficient and do not consider conformation

A current trend is to employ fragment descriptors based on important molecular properties such as hydrophobes (eg aromatic rings), hydrogen bond donors (eg amines), hydrogen bond acceptors (eg carbonyls), positive charges (eg NH_4^+), and negative charges (eg PO_3^-). The rationale for this approach was first described by Andrews and coworkers.[26] Other fingerprint and general fragment based methods such as molecular holograms[27,28] generalise this approach of breaking molecules into fragments. Another important class of fragment-based descriptors, the van der Waals surface area descriptors (VSA), have been reported by Labute to have attributes that make them a widely applicable QSAR descriptors.[29] VSA descriptors are derived by adding together the van der Waals surface area contributions of atoms exhibiting a given property (chosen from steric, electrostatic and lipophilic properties) within a given binned property range. Linear combinations of VSA descriptors correlate well with most other commonly used descriptors. Fragment-based descriptors have advantages of being computationally efficient and independent of molecular conformation or 3D structure. However, they are usually not as informative, being relatively poor at accounting for effects of stereochemistry in activity.

Whole molecule descriptors

Some of the earliest molecular descriptors were of this type.[1,10] They typically capture information on molecular size and lipophilicity through properties such as

Octanol-water partition is an important property

the molecular weight or molecular volume and log of the octanol–water partition coefficient (log *P*). Log *P* in particular was found to be important in many QSAR models. The relationship between log *P* and some biological responses was often inverse parabolic, in which a maximum in the biological response occurred at some optimum log *P* value. The explanation for this relationship was that it described the partitioning of drug molecules into biological membranes. Hansch *et al.* recently reviewed the subclass of QSAR problems for which hydrophobicity was not an important factor.[30]

Topological descriptors are easy to calculate and often useful

An important class of whole molecule descriptors are the topological descriptors.[31–36] These involve treating molecules as topological objects where atoms become the vertices, and bonds the edges, of a molecular graph. Figure 2 shows the conversion of a molecular structure into a molecular graph. It is possible to characterise molecular graphs using a number of indices. The most well known of these are the Randic indices, and the Kier and Hall electrotopological indices.[37–39] In essence, these indices describe molecules in terms of connected paths through the hydrogen-suppressed molecular graph. For instance, the zeroth and first order Randic indices are:

Eigenvalue descriptors have unique properties

$$^0\chi = \sum_{i=1}^{n} \delta_i^{-1/2}$$

$$^1\chi = \sum_{s=1}^{N_e} (\delta_i \delta_j)^{-1/2}$$

where the Γ_i represent the hydrogen-suppressed valence of vertex (atom) *i*, *n* is the number of vertices (atoms) and N_e the number of edges (bonds). Higher-order indices are calculated from progressively longer paths in the molecule. The Kier and Hall *E*-state indices extended this idea to include the effects of bond orders and electronegativity of atoms, and the influence of more distant atoms in a molecule (the 'field').[40]

Recently, descriptors derived from eigenvalues of molecular matrices derived from graphs have shown promise in generating descriptors useful for QSAR[41–43] and for molecular diversity purposes (eg characterisation of chemical libraries and databases, and for design of optimally diverse combinatorial libraries).[44] Modified adjacency matrices describe how atoms in a molecule are connected. They provide a means of combining the molecular properties with topological information encoding the way a molecule is connected. Figure 3 shows an example of a modified adjacency matrix. Diagonalisation of these matrices provide eigenvalue descriptors. A modification of this eigenvalue approach

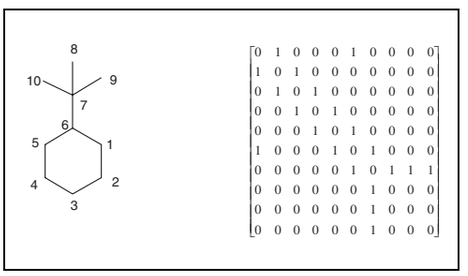

Figure 3. Conversion of molecule into an adjacency matrix. Off-diagonal elements are 1 if the two atoms are bonded, 0 if not

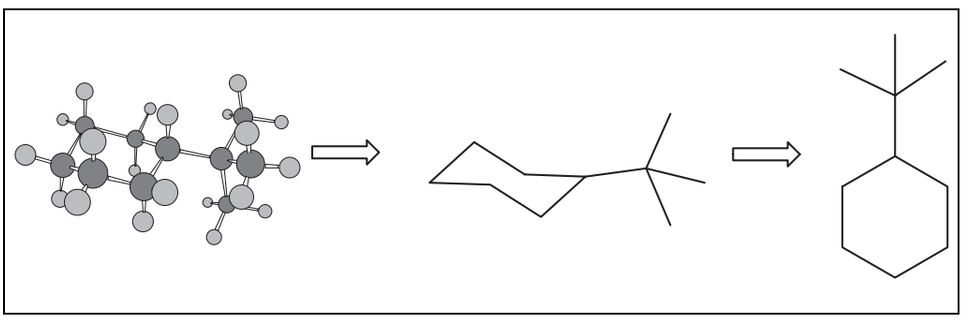

Figure 2: Relationship between a molecule and its molecular graph. Three-dimensional structure (left), two-dimensional, hydrogen-suppressed structure (centre) and hydrogen-suppressed molecular graph (right)

has been particularly useful in the description of molecular diversity (dissimilarity between molecules). Pearlman's BCUT (Burden, CAS, University of Texas) descriptors have been widely used to quantify molecular library and database diversity.[44] Although Pearlman warns against using BCUT descriptors for QSAR, several recent studies (eg Stanton[42]) have shown that they are effective QSAR descriptors that encode molecular information not provided by other descriptors.

The descriptors discussed so far encode 2D representations of molecules (eg molecular graphs). Several widely used molecular descriptors types are calculated from 3D properties of molecules. The CoMFA[15] (comparative molecular field analysis) method surrounds 3D structures of molecules by an array of grid points. At each grid point outside of the molecule a probe atom or functional groups is used to calculate steric, electrostatic and (optionally) lipophilic fields at that point.

Molecular field methods are very useful for 3D QSAR

This generates a molecular field representation of molecules in a training data set. The large number of descriptors this method generates must be dealt with by special regression techniques such as partial least squares (PLS) analysis. This and other field-based methods have been very useful in finding QSAR models. However, all 3D QSAR methods suffer from the problems of requiring a valid 3D structure (assumptions must be made concerning the biologically active conformer or shape), and an alignment rule to superimpose molecules in the training set. An example of a molecular alignment and CoMFA steric map is provided in Figure 4.

Silverman attempted to overcome this problem by expanding molecular properties as a series of molecular multiple moments.[45,46] There are many other types of whole molecule descriptors which have been devised ranging from those describing molecular shape, through molecular similarity, to molecular

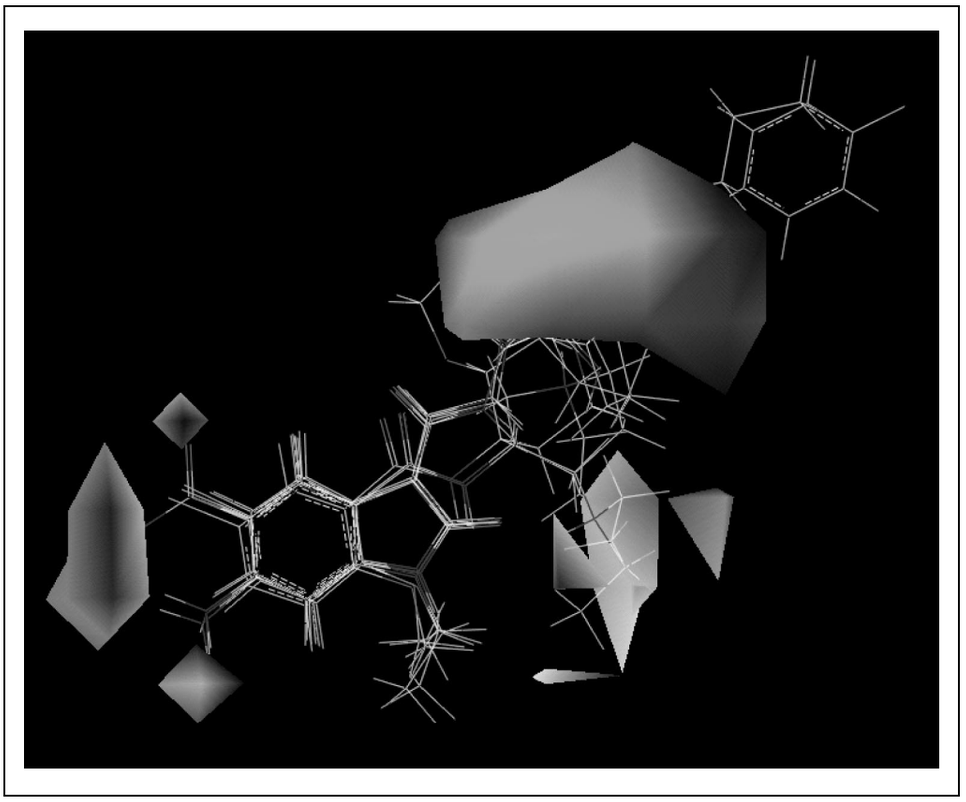

Figure 4. CoMFA map of a series of $5HT_4$ receptor ligands showing molecular alignments and molecular field contributions favouring (large region near top) and disfavouring (small regions near bottom) increased steric bulk in the molecules.

autocorrelation functions. Several recent papers have compared the efficacy of many types of molecular descriptors.[47,48]

Whole molecule descriptors are the most information dense and usually the most informative. QSAR models derived from 3D molecular descriptors are often more easily interpreted than are those from fragment descriptors or 2D whole molecule descriptors (eg topological indices). They are able to account for effects of conformation and isomerism on activity, but this flexibility comes at a computational cost.

Development of better descriptors is an active research area

The development of new descriptors is an active research area for the relatively few major groups who work on discovery of new QSAR and chemometrics methods. There is still considerable scope to discover more information rich, more generally applicable molecular descriptors than the myriad of those now available.

DESCRIPTOR SELECTION

To build a good QSAR model, a minimal set of information-rich descriptors is required. The large number of possible indices creates several problems for the modeller:[49,50]

Descriptor selection must be done very carefully

- Many descriptors do not contain molecular information relevant to the problem.
- Many descriptors are linearly dependent (contain essentially the same information).
- Use of poor descriptors in QSAR yields poor and misleading models.

PCA is a widely-used linear method

- Including too many descriptors in the model, even if they contain relevant information, can result in overfitting of the model, and loss of ability of the model to generalise to unseen molecules.
- Many methods of screening this large pool of potential descriptors for relevant ones can lead to chance correlations (correlations that arise by chance because so many descriptors have been tried in models). In other words, if a large number of random numbers are generated as potential descriptors (which clearly do not contain any useful molecular information), and various subsets of these are used to build models, apparently significant models can arise by chance.

These factors have all led to some poor models being published in the literature. It is important to consider these points when building a QSAR model, and there are methods that we discuss in more depth that avoid these pitfalls. Variable selection involves choosing descriptors containing relevant information, either by experience, variable reduction methods or intelligent selection methods.

Knowledge of the biological process being modelled can provide insight into the types of descriptor that are likely to be important. For instance, in modelling toxicity due to uncoupling of mitochondrial oxidative phosphorylation, the mechanism of uncoupling suggests that lipophilic, weakly acidic molecules are likely to be most effective. Consequently, an experienced modeller would choose descriptors correlating with these molecular properties.

The earliest method of variable selection used stepwise regression. This was integrated with the model-building process and involved stepwise addition (or backwards elimination) of descriptors according to a statistical test, to find the best model. Another widely used variable reduction method is principal components analysis (PCA). This involves creating a smaller set of new orthogonal descriptors from linear combinations of the original descriptors and using these to generate QSAR models. Many of the papers, reviews and textbooks on QSAR referred to in this paper contain descriptions of PCA.

The most active research into variable selection methods is in the area of genetic algorithms.[51–54] These methods start with a pool of possible descriptors (often

Genetic methods of variable selection are currently popular

having undergone some rational preprocessing) and taking various combinations of these, chosen according to a selection operator. These combinations constitute a population of possible descriptor sets that are evolved under genetic selection rules to find the set that generates the best QSAR model. Evolutionary pressure is applied using a fitness function (often a measure of the validity of the QSAR model), and population members who are less fit (produce bad models) are eliminated.

Bayesian methods offer significant advantages

An alternative approach is to use Bayesian statistics to rank input descriptors according to relevance to the model, effectively removing uninformative descriptors. This method, known as automatic relevance determination (ARD),[55] has the advantage that it is a non-linear process, which has a sound statistical basis. Choosing descriptors sets for a non-linear QSAR model using a linear method such as PCA is not optimal. The reader is referred to recent reviews that address the variable selection methods and issues in greater detail.[56,57]

STRUCTURE–ACTIVITY MAPPING

Many QSARs are non linear

Many methods have been used to map molecular descriptors to properties. The majority are regression methods, of which multiple linear regression was the first used.[1] Regression methods attempt to fit a specific function with free parameters to a set of data. They usually do this using some gradient descent method such as least squares, which finds the best set of free parameters that minimise the sum of the squares of the errors between the measured values of the dependent variables, and those calculated by the fitted function. Some QSAR problems have relatively linear response surfaces that can be modelled successfully by linear regression methods. However, most QSAR problems involve at least some degree of non-linearity. Initially this was tackled by using bilinear, exponential, power law or polynomial regression, together with cross-terms, to find the relationship. However, these required the researcher to subjectively choose the functional relationships to create the model.[58] This created problems because the true, complex nature of the relationship was often not found. Frequently many models needed to be created until the 'best' one emerged, owing to the subjective nature of the choice of functional relationship.

Neural networks are very useful model free, non-linear mapping methods

In the last decade, neural networks have emerged as the most useful way to overcome many of these shortcomings.[59–67] Neural networks are regression methods, which automatically learn the functional relationships between molecular structure and properties without any subjective input from the researcher. The class of neural net most often used for empirical stucture–property modelling, the back-propagation neural net – learns in a similar way to the brain — by example. The network is initialised then shown a training set of molecular representations together with the property the molecules exhibit. The neural net learns the associations. A trained neural net can then be used in 'readback' mode to predict the properties of other molecules not in the training set. Back-propagation neural networks, in particular, have been very successful in modelling complex structure–activity relationships as they solved many of the outstanding problems with stucture–property mapping.

However, neural networks left some structure–activity mapping problems unsolved and introduced a few extra problems.[68,69] Like most other regression methods, neural networks can overfit the data. They can also be overtrained, where they get progressively better at predicting the behaviour of the training set, but worse at predicting the behaviour of other test molecules not used in training. Neural networks can be constructed in a number of ways with one or more 'hidden' layers, and varying numbers of neurodes in the hidden layers. Finding the optimum architecture for the neural net is a subjective, time-consuming problem.

Neural networks can be overtrained

When neural nets are trained several times using the same training data, and random initial weights in the neural nets, they do not always train to the same model. Even very similar models can have very different sets of neural network weights associated with them.

One solution is to use a special kind of back-propagation neural net, the Bayesian regularised neural net,[70–72] to build structure–property models.[73] This is a very robust, generally applicable regression method. It has a number of advantages in that it is resistant to overtraining, automatically optimises the neural net architecture, and can use all of the available data as cross-validation is not necessary. It produces a single optimum model. It can be shown mathematically that such a model is optimum, extracting the most information possible from the data. Any other method not approximating it will not do as well on average. Because it uses neural networks the method is a very good generalised mapping method that can automatically learn quite complex structure–property relationships. It is also capable of learning a number of models simultaneously. We have illustrated this by creating a single toxicity model for four distinct classes of aromatics with at least four different mechanisms of toxicity.[74] Such mapping modalities are also quite tolerant of noisy or missing data in a data set.

Bayesian neural nets solve most problems

There are many alternative mapping methods

Alternative methods of mapping structure to properties include recursive partitioning,[75] Gaussian processes,[76] radial basis functions (RBF)[77] and other types of neural networks.[78,79] The Support Vectors Machine (SVM), a classification algorithm widely used in pattern recognition, is showing considerable promise in finding good qSAR models in a very computationally efficient manner.[80] The SVM chooses from increasingly complex hypothesis spaces those that can simultaneously minimise training and generalisation errors. Burbidge and coworkers compared SVM with back-propagation neural networks, RBF and recursive partitioning and found they were superior to the latter two methods and as effective as the neural net (although quicker to train).

A number of variable selection and SAR mapping concepts have been combined. For example genetic variable selection methods have been incorporated into PLS[81] and neural net mapping[82] methods.

VALIDATION AND TESTING

It is important to know how predictive a model is, once derived, to show whether a structure–property mapping method has overfitted the data, a neural net has overtrained or that chance correlations are present. Several methods have been developed to estimate the validity or predictivity of the derived structure–property model. The most common method is 'leave-one-out' cross-validation.[83] This involves leaving each molecule out of the training set in turn, then creating a model using the remainder of the training set. The property of the omitted molecule is predicted using the model derived from all of the other molecules. This method is not a very rigorous test of the predictivity of the model and suffers from two other major deficiencies: the time to carry out the cross-validation increases as the square of the size of the training set; the method produces *n* final models (each corresponding to one of the training set molecules being left out) and it is not clear which is the 'best' model.

Validation is essential and may be time consuming

A better method is to remove a percentage of the training set into a test set.[84] The structure–property model is derived using the reduced training set, and the properties of the test set predicted using this model. This is a more rigorous test of the quality of the structure–property model but again suffers from problems: not all of the available data can be used to make the model as some must be held back for the test set; it is not clear how the test set is best selected from the training set, eg randomly or using cluster analysis.

Bayesian methods can eliminate validation bottlenecks

As Bayesian regularised neural nets do not strictly require a validation set, they overcome the validation bottleneck inherent in other regression methods.[85] Leave-one-out cross-validation methods are not difficult when data sets of 100 compounds are involved. However, with combinatorial data sets of 10,000 or more, the cross-validation effort would increase by $100^2 = 10{,}000$-fold. Bayesian neural nets can eliminate this overhead. Being able to use all of the data to generate the model is an advantage, especially where data sets are small and generation of additional data impossible or expensive. Several seminal papers and reviews provide clear descriptions of the methods and pitfalls of validating QSAR models.[69,86]

APPLICATIONS AND CONCLUSIONS

QSAR methods have very wide applicability

QSAR has been applied extensively and successfully over several decades to find predictive models for activity of bioactive agents. It has also been applied to areas related to discovery and subsequent development of bioactive agents: distinguishing drug-like from non-drug-like molecules,[87] drug resistance,[88] toxicity prediction,[89–94] physicochemical properties prediction (eg water solubility, lipophilicity),[95] gastrointestinal absorption,[96] activity of peptides,[97] data mining,[98] drug metabolism,[99] and prediction of other pharmacokinetic and ADME properties.[100,101] Recent reviews[102–112] summarise work in a number of these areas and a book[78] has summarised the application of neural networks to combinatorial discovery. The journal *Quantitative Structure–Activity Relationships* contains abstracts of QSAR studies in other journals in each issue.

It is clear that the number of potential applications for structure–property modelling, in the most general case, is extensive and growing daily. Improved molecular descriptors, based on a better understanding of which molecular attributes are most important for a given property being modelled, and increasing use of genetic and artificial intelligence methods will raise QSAR to even greater levels of usefulness than the current high level. A basic understanding of QSAR concepts is essential for most people, across a diverse range of skills, who design molecules.

References

1. Hansch, C. and Fujita, T. (1964), 'p-σ-π Analysis. A method for the correlation of biological activity and chemical structure', *J. Amer. Chem. Soc.*, Vol. 86, p. 1616.
2. Grover, M., Singh, B., Bakshi, M. and Singh S. (2000), 'Quantitative stucture–property relationships in pharmaceutical research – Part 1', *Pharm. Sci. Technol. Today*, Vol. 3(1), pp. 28–35.
3. Trinajstic, N., Randi, M. and Klein, D. J. (1986), 'On the quantitative structure–activity relationship in drug research', *Acta Pharm. Jugosl.*, Vol. 36(2), pp. 267–279.
4. Mager, P. P. (1984), 'Biometrics in medicinal chemistry: A difficult road ahead', *QSAR Des. Bioact. Compd.*, pp. 433–442.
5. Martin, Y. C. (1981), 'A practitioner's perspective of the role of quantitative structure–activity analysis in medicinal chemistry', *J. Med. Chem.*, Vol. 24(3), pp. 229–237.
6. Testa, B. (2000), 'Structure–activity-relationships – Challenges and context', *Pharm. News*, Vol. 7(1), pp. 13–22.
7. Kaiser, K. L. E. (1999), 'Quantitative structure–activity relationships in chemistry', *Can. Chem. News*, Vol. 51(1), pp. 23–24.
8. Guo, Z. (1995), 'Structure–activity relationships in medicinal chemistry: Development of drug candidates from lead compounds', *Pharmacochem. Libr.*, Vol. 23, pp. 299–320.
9. Chu, K. C. (1980), 'The quantitative analysis of structure–activity relationships', in 'Burger's Med. Chem.' (4th Edn) Wiley, New York, Vol. 1, pp. 393–418.
10. Hansch, C., Leo, L. and Hoekman, D. (1995), Monograph: 'Exploring the QSAR. Hydrophobic, Electronic, and Steric Constants', Heller S. R., Ed., ACS, Washington, DC.
11. Fujita, T. (1997), 'Recent success stories leading to commercialisable bioactive compounds with the aid of traditional QSAR procedures', *Quant. Struc.-Activ. Relat.*, Vol. 16, pp. 107–112.
12. Boyd, D. B. (1990), 'Successes of computer-

assisted molecular design', in Lipkowitz, K. and Boyd, D. B., Eds, 'Reviews in Computational Chemistry', Vol. 1, VCH, New York, pp. 355–371.

13. Jürgen Bajorath, J. (2001), 'Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening', *J. Chem. Inf. Comput. Sci.*, Vol. 41(2), pp. 233–245.

14. Brown, R. D. and Martin, Y. C. (1996), 'Use of structure–activity data to compare structure-based clustering methods and descriptors for use in compound selection', *J. Chem. Inf. Comput. Sci.*, Vol. 36, pp. 572–584.

15. Cramer, R. D., Patterson, D. E. and Bunce, J. D. (1988), 'Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins', *J. Amer. Chem. Soc.*, Vol. 110, pp. 5959–5967.

16. Free, S. M. and Wilson, J. W. (1964), 'A mathematical contribution to structure–activity studies', *J. Med. Chem.*, Vol. 7(4), pp. 395–399.

17. Warne, M. A. and Nicholson, J. K. (2000), 'Quantitative structure–activity relationships (QSARs) in environmental research. Part II. Molecular orbital approaches to property calculation', *Prog. Environ. Sci.*, Vol. 2(1), pp. 31–52.

18. Karelson, M., Lobanov, V. S. and Katritzky, A. R. (1996), 'Quantum-chemical descriptors in QSAR/QSPR studies', *Chem. Rev.*, Vol. 96(3), pp. 1027–1043.

19. Carbo-Dorca, R., Amat, L., Besalu, E. *et al.* (2000), 'Quantum mechanical origin of QSAR: theory and applications', *Theochem*, Vol. 504, pp. 181–228.

20. Beck, B., Horn, A., Carpenter, J. E. and Clark, T. (1998), 'Enhanced 3D-databases: A fully electrostatic database of AM1-optimized structures', *J. Chem. Inf. Comput. Sci.*, Vol. 38 (6), pp. 1214–1217.

21. Burden, F. R. (1996), 'Using artificial neural networks to predict biological activity from simple molecular structural considerations', *Quant. Struct.–Activ. Relat.*, Vol. 15, pp. 7–11.

22. Winkler, D. A., Burden, F. R. and Watkins, A. J. R. (1998), 'Atomistic topological indices applied to benzodiazepines using various regression methods', *Quant. Struct.-Activ. Relat.*, Vol. 17, pp. 14–19.

23. Brown, R. D. and Martin, Y. C. (1997), 'The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding', *J. Chem. Inf. Comput. Sci.*, Vol. 37, pp. 1–9.

24. Burden, F. R. and Winkler, D. A. (1999), 'New QSAR methods applied to structure–activity mapping and combinatorial chemistry', *J. Chem. Inf. Comput. Sci.*, Vol. 39(2), pp. 236–242.

25. Wildman, S. A. and Crippen, G. M. (1999), 'Prediction of physicochemical parameters by atomic contributions', *J. Chem. Inf. Comput. Sci.*, Vol. 39(5), pp. 868–873.

26. Andrews, P. R., Craik, D. J. and Martin, J. L. (1984), 'Functional group contributions to drug–receptor interactions', *J. Med. Chem*, Vol. 27, pp. 1648–1657.

27. Tong, W., Lowis, D. R., Perkins, R. *et al.* (1998), 'Evaluation of quantitative structure–activity relationship methods for large-scale prediction of chemicals binding to the estrogen receptor', *J. Chem. Inf. Comput. Sci.*, Vol. 38, pp. 669–677.

28. Winkler, D. A. (1998), 'Holographic QSAR of benzodiazepines', *Quant. Struct.-Activ. Relat.*, Vol. 17, p. 224.

29. Labute, P. (2000), 'A widely applicable set of molecular descriptors', *J. Mol. Graph. Mod.*, Vol. 18, pp. 464–477.

30. Hansch, C., Kurup, A., Garg, R. and Gao, H. (2001), 'Chem-bioinformatics and QSAR: A review of QSAR lacking positive hydrophobic terms', *Chem. Rev.*, Vol. 101(3), pp. 619–672.

31. Randíc, M. (1975), 'On characterization of molecular branching', *J. Amer. Chem. Soc.*, Vol. 97, pp. 6609–6615.

32. Randíc, M. (1991), 'On computation of optimum parameters for multivariate analysis of stucture–property relationship', *J. Comp. Chem.*, Vol. 12(8), pp. 970–980.

33. Balaban, A. T. (2001), 'A personal view about topological indices for QSAR/QSPR', *QSPR/QSAR Stud. Mol. Descriptors*, pp. 1–30.

34. Devillers, J. (2000), 'New trends in (Q)SAR modeling with topological indices', *Curr. Opin. Drug Discovery Dev.*, Vol. 3(3), pp. 275–279.

35. Estrada, E. (1999), 'Novel strategies in the search of topological indices', *Topol. Indices Relat. Descriptors QSAR QSPR*, pp. 403–453.

36. Bonchev, D. (1999), 'Overall connectivity and topological complexity. A new tool for QSPR/QSAR', *Topol. Indices Relat. Descriptors QSAR QSPR*, pp. 361–401.

37. Hall, L. H. and Kier, L. B. (1995), 'Electrotopological state indices for atom types: A novel combination of electronic, topological and valence state information', *J. Chem. Inf. Comput. Sci.*, Vol. 35, pp. 1039–1045.

38. Hall, L. H., Mohoney, B. and Kier, L. B. (1991), 'The electrotopological state: An atom index for QSAR', *Quant. Struct.-Activ. Relat.*, Vol. 10, pp. 43–51.

39. Kier, L. B. and Hall, L. H. (1995), 'The molecular connectivity chi indexes and kappa shape indexes in stucture–property modelling', in Lipkowitz, K. B. and Boyd, D. B., Eds, 'Reviews in Computational Chemistry'. Vol. 2, Verlag Chemie, New York, pp. 367–422.

40. Kier, L. B. and Hall, L. H. (1999), 'Molecular Structure Descriptions: The Electrotopological State', Academic Press, San Diego, CA.

41. Burden, F. R. (1997), 'A chemically intuitive molecular index based on eigenvalues of a modified adjacency matrix', *J. Chem. Inf. Comput. Sci.*, Vol. 16, pp. 309–314.

42. Stanton, D. T. (1999), 'Evaluation and use of BCUT descriptors in QSAR and QSPR studies', *J. Chem. Inf. Comput. Sci.*, Vol. 39, pp. 11–20.

43. Randíc, M., Vracko, M. and Novic, M. (2001), 'Eigenvalues as molecular descriptors', *QSPR/QSAR Stud. Mol. Descriptors*, pp. 147–211.

44. Pearlman, R. S. and Smith, K. M. (1998), 'Novel software tools for chemical diversity', in Kubinyi, H., Folkers, G. and Martin, Y. C., Eds, '3D QSAR in Drug Design', Vol. 2, Kluwer/ESCOM, London, pp. 339–353.

45. Silverman, B. D. and Platt, D. E. (1996), 'Comparative molecular moment analysis (CoMMA): 3D-QSAR without molecular superimposition', *J. Med. Chem.*, Vol. 39, pp. 2129–2140.

46. Platt, D. E. and Silverman, B. D. (1996), 'Registration, orientation, and similarity of molecular electrostatic potentials through multipole matching', *J. Comput. Chem.*, Vol. 17(3), pp. 358–366.

47. Dearden, J. C. and Ghafourian, T. (1999), 'Hydrogen bonding parameters for QSAR: Comparison of indicator variables, hydrogen bond counts, molecular orbital and other parameters', *J. Chem. Inf. Comput. Sci.*, Vol. 39(2), pp. 231–235.

48. Estrada, E. and Molina, E. (2001), 'QSPR/QSAR by graph theoretical descriptors beyond the frontiers', *QSPR/QSAR Stud. Mol. Descriptors*, pp. 83–107.

49. Topliss, J. G. and Edwards, R. P. (1979), 'Chance factors in studies of quantitative structure-activity relationships', *J. Med. Chem.*, Vol. 22(10), pp. 1238–1244.

50. Manallack, D. T. and Livingstone, D. J. (1992), 'Artificial neural networks: Application and chance effects for QSAR data analysis', *Med. Chem. Res.*, Vol. 2, pp. 181–190.

51. Yasri, A. and Hartsough, D. (2001), 'Toward an optimal procedure for variable selection and QSAR model building', *J. Chem. Inf. Comput. Sci.*, Vol. 41(5), pp. 1218–1227.

52. Hou, T. J., Wang, J. M., Liao, N. and Xu, X. J. (1999), 'Applications of genetic algorithms on the structure–activity relationship analysis of some cinnamamides', *J. Chem. Inf. Comput. Sci.*, Vol. 39(5), pp. 775–781.

53. Waller, C. L. and Bradley, M. P. (1999), 'Development and validation of a novel variable selection technique with application to multidimensional quantitative structure–activity relationship studies', *J. Chem. Inf. Comput. Sci.*, Vol. 39(2), pp. 345–355.

54. Kimura, T. and Funatsu, K. (1999), 'GA strategy for variable selection in QSAR studies: Application of GA–based region selection to a 3D-QSAR study of acetylcholinesterase inhibitors', *J. Chem. Inf. Comput. Sci.*, Vol. 39(1), pp. 112–120.

55. Burden, F. R., Ford, M., Whitley, D. and Winkler, D. A. (2000), 'The use of automatic relevance determination in QSAR studies using Bayesian neural nets', *J. Chem. Inf. Comput. Sci.*, Vol. 40(6), pp. 1423–1430.

56. Zheng, W. and Tropsha, A. (2000), 'Novel variable selection quantitative stucture–property relationship approach based on the *k*-nearest-neighbor principle', *J. Chem. Inf. Comput. Sci.*, Vol. 40(1), pp. 185–194.

57. Bajorath, J. (2001), 'Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening', *J. Chem. Inf. Comput. Sci.*, Vol. 41(2), pp. 233–245.

58. Constans, P. and Hirst, J. D., (2000), 'Nonparametric regression applied to quantitative structure–activity relationships'. *J. Chem. Inf. Comput. Sci.*, Vol. 40(2), pp. 452–459.

59. Salt, D. W., Yildiz, N., Livingston, D. J. and Tinsley, C. J. (1992), 'The use of artificial neural networks in QSAR', *Pestic. Sci.*, Vol. 36, pp. 161–170.

60. Tetko, I. V., Luik, A. I. and Poda, G. I. (1993), 'Applications of neural networks in structure–activity relationships of a small number of molecules', *J. Med. Chem.*, Vol. 36, pp. 811–814.

61. Andrea, T. A., Kalayeh, H. (1991), 'Applications of neural networks in quantitative structure–activity relationships of dihydrofolate reductase inhibitors', *J. Med. Chem.*, Vol. 34, pp. 2824–2836.

62. Burns, J. A. and Whitesides, G. M. (1993), 'Feed-forward neural networks in chemistry: Mathematical systems for classification and pattern recognition', *Chem. Rev.*, Vol. 93(8), pp. 2583–2601.

63. Gasteiger, J. and Zupan, J. (1993), 'Neural networks in chemistry', *Angew. Chem. Int. Ed. Engl.*, Vol. 32, pp. 503–527.

64. Weinstein, J. N., Kohn, K. W., Grever, M. R. *et al.* (1992), 'Neural computing in

cancer drug development: Predicting mechanisms of action', *Science*, Vol. 258, pp. 447–451.

65. Aoyama, T., Suzuki, Y. and Ichikawa, H. (1990), 'Neural networks applied to quantitative structure–activity relationship', *J. Med. Chem.*, Vol. 33, pp. 2583–2590.

66. Maggiora, G. M., Elrod, D. W. and Trenary, R. G. (1992), 'Computational neural nets as model-free mapping devices', *J. Chem. Inf. Comput. Sci.*, Vol. 32, pp. 732–741.

67. Ivanciuc, O. (2001), 'New neural networks for structure–property models', *QSPR/QSAR Stud. Mol. Descriptors*, pp. 213–231.

68. Manallack, D. T. and Livingston, D. J. (1999), 'Neural networks in drug discovery: Have they lived up to their promise?', *Eur. J. Med. Chem.*, Vol. 34, pp. 195–208.

69. Manallack, D. T. and Livingston, D. J. (1992), 'Artificial neural networks: Applications and chance effects for QSAR data analysis', *Med. Chem. Res.*, Vol. 2, pp. 181–190.

70. MacKay, D. J. C. (1992), 'A practical Bayesian framework for backprop networks', *Neural Computation*, Vol. 4, pp. 415–447.

71. Buntine, W. L. and Weigend, A. S. (1991), 'Bayesian back-propagation', *Complex Sys.*, Vol. 5, pp. 603–643.

72. Mackay, D. J. C. (1995), 'Probable networks and plausible predictions – a review of practical Bayesian methods for supervised neural networks', *Comput. Neural Sys.*, Vol. 6, pp. 469–505.

73. Burden, F. R. and Winkler, D. A. (1999), 'Robust QSAR models using Bayesian regularized neural networks', *J. Med. Chem.*, Vol. 42(16), pp. 3183–3187.

74. Burden, F. R. and Winkler, D. A. (2000), 'A QSAR model for the acute toxicity of substituted benzenes towards *Tetrahymena pyriformis* using Bayesian regularized neural networks', *Chem. Res. Toxicol.*, Vol. 13(6), pp. 436–440.

75. Rusinko, A., Farmen, M. W., Lambert, C. G. *et al.* (1999), 'Analysis of a large structure/biological activity data set using recursive partitioning', *J. Chem. Inf. Comput. Sci.*, Vol. 39(6), pp. 1017–1026.

76. Frank, R. and Burden, F. R. (2001), 'Quantitative structure–activity relationship studies using Gaussian processes', *J. Chem. Inf. Comput. Sci.*, Vol. 41(3), pp. 830–835.

77. Mayer-Bäse, A. and Watzel, R. (1998), 'Transformation radial basis neural network for relevant feature selection', *Patt. Recog. Lett.*, Vol. 19, pp. 1301–1306.

78. Winkler, D. A. and Burden, F. R. (2002), 'Application of neural networks to large dataset QSAR, virtual screening and library design', in Bellavance, L., Ed., 'Combinatorial Chemistry Methods and Protocols', Humana Press, Totowa, New Jersey.

79. Burden, F. R. (1998), 'Holographic neural networks as non-linear discriminants for chemical applications', *J. Chem. Inf. Comput. Sci.*, Vol. 38(1), pp. 47–53.

80. Burbidge, R., Trotter, M., Buxton, B., and Holden, S. (2001), 'Drug design by machine learning: support vector machines for pharmaceutical data analysis', *Comput. Chem.*, Vol. 26, pp. 5–14.

81. Hasegawa, K. and Funatsu, K. (2000), 'Partial least squares modeling and genetic algorithm optimization in quantitative structure–activity relationships', *SAR QSAR Environ. Res.*, Vol. 11(3–4), pp. 189–209.

82. So, S.-S. and Karplus, M. (1996), 'Genetic neural networks for quantitative structure–activity relationships: Improvements and application of benzodiazepine affinity for benzodiazepine/$GABA_A$ receptors', *J. Med. Chem.*, Vol. 39, pp. 5246–5256.

83. Wold, S. (1991), 'Validation of QSAR's', *Quant. Struct.-Activ. Relat.*, Vol. 10, pp. 191–193.

84. Tetko, I. V., Livingstone, D. J. and Luik, A. I. (1995), 'Neural network studies: 1. Comparison of overfitting and overtraining'. *J. Chem. Inf. Comput. Sci.*, Vol. 35(5), pp. 826–833.

85. Husmeier, D., Penny, W. D. and Roberts, S. J. (1999), 'An empirical evaluation of Bayesian sampling with hybrid Monte Carlo for training neural network classifiers', *Neural Networks*, Vol. 12, pp. 667–705.

86. Livingstone, D. J., Manallack, D. T. and Tetko, I. V. (1997), 'Data modelling with neural networks: Advantages and limitations'. *J. Comput.-Aided Mol. Des.*, Vol. 11, pp. 135–142.

87. Ajay, Walters, W. P. and Murcko, M. (1998), 'Can we learn to distinguish between ''drug-like'' and ''nondrug-like'' molecules?', *J. Med. Chem.*, Vol. 41, pp. 3314–3324.

88. Wiese, M. and Pajeva, I. K. (2001), 'Structure–activity relationships of multidrug resistance reversers', *Curr. Med. Chem.*, Vol. 8(6), pp. 685–713.

89. Schultz, T. W. and Seward, J. R. (2000), 'Health effects related stucture–toxicity relationships – a paradigm for the first decade of the new millennium', *Sci. Total Environ.*, Vol. 249(1–3), pp. 73–84.

90. Benigni, R., Giuliani, A., Franke, R. and Gruska, A. (2000), 'Quantitative structure–activity relationships of mutagenic and carcinogenic aromatic amines', *Chem. Rev.*, Vol. 100(10), pp. 3697–3714.

91. Garg, R., Kurup, A. and Hansch, C. (2001), 'Comparative QSAR: On the toxicology of the phenolic OH moiety', *Crit. Rev. Toxicol.*, Vol. 31(2), pp. 223–245.

92. Bashir, S. J. and Maibach, H. I. (2000), 'Quantitative structure analysis relationships in the prediction of skin sensitization potential', *Biochem. Modulation Skin React.*, pp. 61–64.

93. Cronin, M. T. D. (2000), 'Computational methods for the prediction of drug toxicity', *Curr. Opin. Drug Discovery Dev.*, Vol. 3(3), pp. 292–297.

94. Freidig, A. P. and Hermens, J. L. M. (2001), 'Narcosis and chemical reactivity QSARs for acute fish toxicity', *Quant. Struct.-Activ. Relat.*, Vol. 19(6), pp. 547–553.

95. Gombar, V. K. and Enslein, K. (1996), 'Assessment of *n* -octanol/water partition coefficient: When is the assessment reliable?' *J. Chem. Inf. Comput. Sci.*, Vol. 36(6), pp. 1127–1134.

96. Agatonovic-Kustrin, S., Beresford, R., Yusof, A. and Pauzi, M. (2001), 'Theoretically-derived molecular descriptors important in human intestinal absorption', *J. Pharm. Biomed. Anal.*, Vol. 25(2), pp. 227–237.

97. Brusic, V., Bucci, K., Schönbach, C. *et al.* (2001), 'Efficient discovery of immune response targets by cyclical refinement of QSAR models of peptide binding', *J. Mol. Graph. Modell.*, Vol. 19(5), pp. 405–411.

98. Burden, F. R. and Winkler, D. A. (2000), 'The computer simulation of high throughput screening of bioactive molecules', *Mol. Model. Predict. Bioact*, [Proceedings of the 12th European Symposium on Quantitative Structure – Activity Relationships], pp. 175–180.

99. Lewis, D. F. V. (2000), 'Structural characteristics of human P450s involved in drug metabolism: QSARs and lipophilicity profiles', *Toxicology*, Vol. 144(1–3), pp. 197–203.

100. Vedani, A. and Dobler, M. (2000), 'Multi-dimensional QSAR in drug research: predicting binding affinities, toxicity and pharmacokinetic parameters', *Prog. Drug Res.*, Vol. 55, pp. 105–135.

101. Guba, W. and Cruciani, G. (2000), 'Molecular field-derived descriptors for the multivariate modeling of pharmacokinetic data', *Mol. Model. Predict. Bioact.*, [Proceedings of the 12th European Symposium on Quantitative Structure-Activity Relationships], pp. 89–94.

102. Katritzky, A. R., Petrukhin, R., Tatham, D. *et al.* (2001), 'Interpretation of quantitative structure–property and –activity relationships', *J. Chem. Inf. Comput. Sci.*, Vol. 41(3), pp. 679–685.

103. Gupta, S. P. (2000), 'Quantitative structure–activity relationships of cardiotonic agents', *Prog. Drug Res.*, Vol. 55, pp. 235–282.

104. Klebe, G. (2000), 'Recent developments in structure-based drug design', *J. Mol. Med.*, Vol. 78(5), pp. 269–281.

105. Podlogar, B. L. and Ferguson, D. M. (2000), 'QSAR and CoMFA: a perspective on the practical application to drug discovery', *Drug Des. Discovery*, Vol. 17(1), pp. 4–12.

106. Lien, E. J. and Ren, S. (2000), 'QSAR and molecular modeling of bioactive phyto-phenolics', *Phytochem. Bioact. Agents*, pp. 21–41.

107. Mickle, T. and Nair, V. (2000), 'Predictive QSAR analysis of anti-HIV agents', *Drugs Future*, Vol. 25(4), pp. 393–400.

108. Warne, M. A. and Nicholson, J. K. (1999), 'Quantitative structure–activity relationships (QSARs) in environmental research. Part I. An overview of techniques and physicochemical properties', *Prog. Environ. Sci.*, Vol. 1(4), pp. 327–344.

109. Hadjipavlou-Litina, D. (2000), 'Quantitative structure–activity relationship (QSAR) studies on non steroidal anti-inflammatory drugs (NSAIDs)', *Curr. Med. Chem.*, Vol. 7(4), pp. 375–388.

110. Kurup, A., Garg, R. and Hansch, C. (2000), 'Comparative QSAR analysis of 5 α-reductase inhibitors', *Chem. Rev.*, Vol. 100(3), pp. 909–924.

111. Leo, A. J. and Hansch, C. (1999), 'Role of hydrophobic effects in mechanistic QSAR', *Perspect. Drug Discovery Des.*, Vol. 17, pp. 1–25.

112. Livingstone, D. J. (2000), 'The characterization of chemical structures using molecular properties. A survey', *J. Chem. Inf. Comput. Sci.*, Vol. 40(2), pp. 195–209.