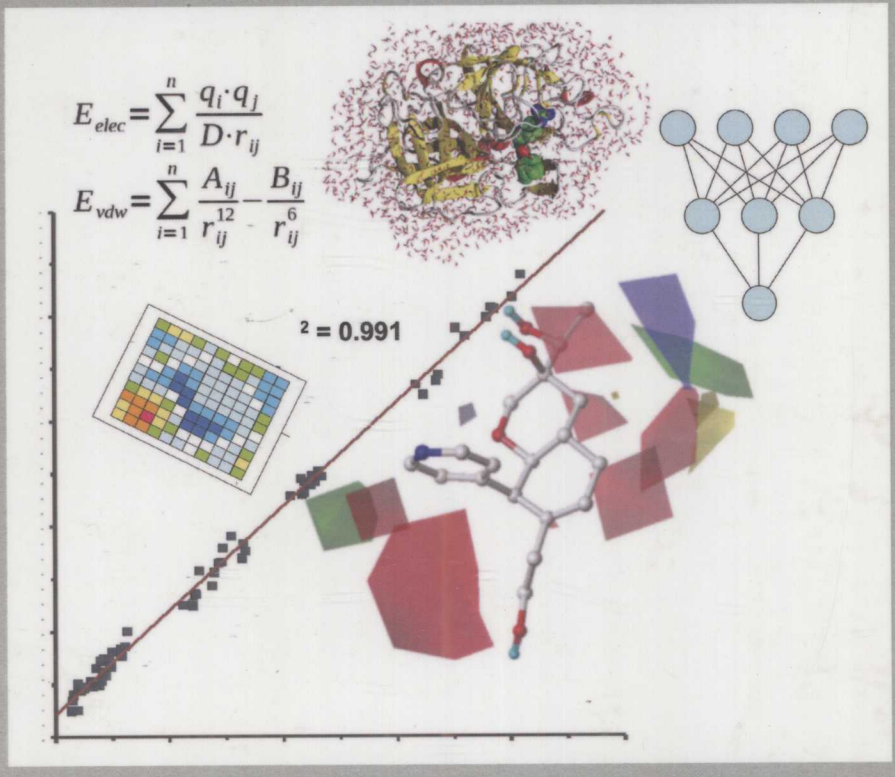
QSAR in Environmental and Health Sciences

THREE DIMENSIONAL QSAR

Applications in Pharmacology and Toxicology



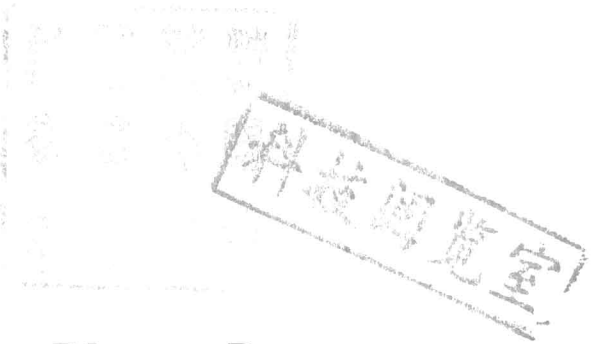
Jean Pierre Doucet
Annick Panaye

CRC Press
Taylor & Francis Group

QSAR in Environmental and Health Sciences

THREE DIMENSIONAL QSAR

Applications in Pharmacology and Toxicology



Jean Pierre Doucet
Annick Panaye

沈阳药科大学图书馆
Y2006917

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2010 by Taylor and Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number: 978-1-4200-9115-1 (Hardback)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (http://www.copyright.com/) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

¥ 1396.

THREE DIMENSIONAL QSAR

Applications in Pharmacology and Toxicology

QSAR in Environmental and Health Sciences

Series Editor

James Devillers
CTIS-Centre de Traitement de
l'Information Scientifique
Rillieux La Pape, France

Aims & Scope

The aim of the book series is to publish cutting-edge research and the latest developments in QSAR modeling applied to environmental and health issues. Its aim is also to publish routinely used QSAR methodologies to provide newcomers to the field with a basic grounding in the correct use of these computer tools. The series is of primary interest to those whose research or professional activity is directly concerned with the development and application of SAR and QSAR models in toxicology and ecotoxicology. It is also intended to provide the graduate and postgraduate students with clear and accessible books covering the different aspects of QSARs.

Published Titles

Endocrine Disruption Modeling, *James Devillers,* 2009

Three dimensional QSAR: Applications in Pharmacology and Toxicology, *Jean Pierre Doucet and Annick Panaye,* 2010

此为试读,需要完整PDF请访问: www.ertongbook.com

Series Introduction

The correlation between the toxicity of molecules and their physicochemical properties can be traced back to the nineteenth century. Indeed, in a French thesis entitled *Action de l'alcool amylique sur l'organisme* (Action of amyl alcohol on the body), which was presented in 1863 by A. Cros before the Faculty of Medicine at the University of Strasbourg, an empirical relationship was made between the toxicity of alcohols and their number of carbon atoms as well as their solubility. In 1875, Dujardin-Beaumetz and Audigé were the first to stress the mathematical character of the relationship between the toxicity of alcohols and their chain length and molecular weight. In 1899, Hans Horst Meyer and Fritz Baum, at the University of Marburg, showed that narcosis or hypnotic activity was in fact linked to the affinity of substances to water and lipid sites within the organism. At the same time, at the University of Zurich, Ernest Overton came to the same conclusion, providing the foundation of the lipoid theory of narcosis. The next important step was made in the 1930s in St. Petersburg by Lazarev, who first demonstrated that different physiological and toxicological effects of molecules were correlated with their oil–water partition coefficient through formal mathematical equations in the form $\log C = a \log P_{oil/water} + b$. Thus, the quantitative structure–activity relationship (QSAR) discipline was born. Its foundations were definitively fixed in the early 1960s by the seminal works of C. Hansch and T. Fujita. Since then, the discipline has gained tremendous interest, and QSAR models now represent key tools in the development of drugs as well as in the hazard assessment of chemicals. The new REACH (Registration, Evaluation, Authorization, and Restriction of Chemicals) legislation on substances, which recommends the use of QSARs and other alternative approaches instead of laboratory tests on vertebrates, clearly reveals that this discipline is now well established and is an accepted practice in regulatory systems.

In 1993, the journal *SAR and QSAR in Environmental Research* was launched by Gordon and Breach to focus on all the important works published in the field and to provide an international forum for the rapid publication of structure–activity relationship (SAR) and QSAR models in (eco)toxicology, agrochemistry, and pharmacology. Today, the journal, which is now owned by Taylor & Francis, publishes twice as many issues per year and continues to promote research in the field of QSAR by favoring the publication of new molecular descriptors, statistical techniques, and original SAR and QSAR models. This field continues to grow rapidly, and many subject areas that require larger development are unsuitable for publication in a journal due to space limitations.

This prompted us to develop a series of books entitled *QSAR in Environmental and Health Sciences* to act in synergy with the journal. I am extremely grateful to Colin Bulpitt and Fiona Macdonald for their enthusiasm and invaluable help in making the project a reality.

This book is the second of the series. Its purpose is at least twofold: On one hand, it introduces the theory and practical applications of 3D-QSAR approaches in

pharmacology and toxicology to both the neophytes and the experienced scientists; on the other, it provides a clear overview of the strengths and weaknesses of these methods.

At the time of going to press, two other books are in the pipeline. One deals with reproductive and developmental toxicology modeling and the other focuses on the topological description of molecules. I gratefully acknowledge Hilary Rowe for her willingness to assist me in the development of this series.

James Devillers

Preface

Computational chemistry is today playing a major role in the studies of complex processes involved in the design and development of new drugs. Based on structural similarity, data mining from huge chemical databases allows the selection of compounds having the pharmacophore likely to give them an adequate biological activity. The calculation of intermolecular interactions between a drug and its receptor specifies the mechanisms at the molecular scale and may suggest structural modifications that would be able to increase activity.

QSAR models establish relationships between a molecular structure and its activity. Their ability to predict the behavior of untested and even unsynthesized molecules is a valued asset in the quest for new drugs. QSAR models direct research toward the more promising structures from the initial stages of development, avoiding wrong tracks, reducing laboratory tests, and limiting animal experimentation.

Another important field of application is toxicology and ecotoxicology. With the widespread introduction of new chemicals in the market, sometimes with important tonnages, it is crucial to have quantitative models at one's disposal that are able to identify pollutants acting on human health or wildlife and prioritize chemicals to be submitted for in-depth experimental tests.

In the past, QSARs were generally received with skepticism, not always unjustified, on account of being rather crude models and sometimes providing inaccurate results with regard to robustness and applicability range of the models. QSARs today, however, have greatly improved and have made rapid strides in the various fields they are employed in as indicated in the following:

- Introduction of new statistical or mathematical tools for data analysis with nonlinear or nondeterministic approaches such as neural networks and genetic algorithms. These methods also benefit from the increasing power of computers for data processing.
- Improvements of the methods for structural representation and development of large chemical databases.
- Availability of graphical display and interactive visualization tools.

QSARs now constitute, in their own right, an important element of drug design approaches. The newly introducted 3D-QSAR models take into account the spatial characteristics of molecules (geometry, shape, and electron distribution), and even evaluate the fields they create in their surrounding or their interactions with neighboring structures (solvents, or receptors). This results in definite improvement in the field. A more detailed and accurate picture of the molecular behaviors is thus accessible.

Two elements explain the now well-established interest in QSAR models:

- On the one hand, international policies (such as the European REACH project) proposed the use of QSARs (if correctly designed) for hazard identification and risk assessment.
- On the other hand, there has been a progressive fusion of approaches that were previously limited to distinct areas of molecular modeling. 3D-QSARs routinely call for molecular mechanics, quantum mechanics, or molecular dynamics to define the privileged conformations of drugs and their possible interconversions. Docking is currently used to specify the best binding mode of drugs in their receptor pocket. Free energy calculations, reserved in the past to some specific or illuminating examples, can now be performed for a series of molecules and incorporated in 3D models. Conversely, QSARs become an efficient tool for screening chemical databases (possibly after a preliminary filtering process) in the search for new leads.

We would like to express our gratitude to our colleagues and coworkers for their wholehearted support and fruitful discussions and specially to Dr F. Barbault who also designed the cover artwork. At last, we emotionally remember our colleague and friend, Prof. B.T. Fan, who departed prematurely.

Introduction

The term "QSARs" (quantitative structure–activity relationships) encompasses a set of methodologies relating, for a specific process, the biological activity of molecules to some selected features of their physicochemical structure by means of a statistical or mathematical tool. The derived model is then used to analyze the results and to predict the activity of untested compounds.

Property prediction is of paramount importance in Chemistry. From a practical point of view, the interest is not so much on the molecular structure itself but rather on the properties the structure may have. It is therefore not surprising that the search for relationships (more specifically, *quantitative* relationships) between structures and properties or activities presented itself as a major concern several years ago. For example, in the Shanghai Museum it is reported that the "Kaogongji" (roughly, something like the book of the craftsman techniques) proposed in the fifth century BC a qualitative relationship between the composition of bronze and its properties (such as quality of the cutting edge, ease of polishing, and sparkling aspect).

With the boom of combinatorial chemistry, a large number of new chemicals can be readily obtained. It is imperative to have efficient methods for activity prediction not only because such methods save time and resources, but also because they avoid large-scale tests, orienting synthesis toward selected, potentially interesting compounds. Toxicology and ecotoxicology are now faced with the widespread diffusion of many long-life chemicals, for example, polychlorinated aromatics that are able to bind nuclear receptors and disrupt the normal hormonal processes of the endocrine system in humans and animals. The outburst in the number of xenobiotics present in the ecosystem makes such prediction tools a privileged way for prioritizing tests on chemicals the more suspect. Thus, they play an important role in environmental policy and adhere to international regulations for risk assessment and hazard identification. Furthermore, they are in line with policies that recommend a decrease in animal experimentation.

After initial skepticism, justified in part by several "meaningless" models (as quoted by Kubinyi [1]), QSARs are now regarded as valuable, scientifically credible tools in drug discovery and environmental toxicology programs such as the REACH (Registration, Evaluation, Authorization and restriction of CHemicals) policy for chemicals in the EU (http://europa.eu.int/comm/environment/chemicals/reach.html) and the Chemical Assessment and Control Program of EPA (U.S. Environment Protection Agency) (see, e.g., the reviews of Schultz et al. [2] and Schmeider et al. [3,4]).

Trying to relate the properties of a molecule with its structure is an old problem, but models widely evolved in parallel to the advances in chemical knowledge. QSARs also took advantage of the development of calculation methods and graphical display tools. How did we pass from the simple count of a sequence of atoms to the concept of the complementarity ligand–receptor?

BRIEF HISTORICAL EVOLUTION

The Beginnings

In the following, we will briefly mention some milestones in the development of QSAR models. For a detailed historical presentation, see Kubinyi [1], Rekker [5], Parascandola [6], and Güner [7]. The concept of structure–activity relationship may be traced back to Crum-Brown and Fraser [8,9], who indicated, in 1869, that "there can be no reasonable doubt but that a relation exists between the physiological action of a substance and its composition and constitution" (quoted from [10]). A few years before this, in his thesis at the Faculty of Medicine in Strasbourg (France) in 1863, Cros observed an increase in the toxicity of alcohols to mammals with decreasing water solubility up to a maximum potency (quoted from Kubinyi [1]).

An important notion arose from Langley's work [11] on the antagonism between pilocarpine and saliva. He suggested the formation of a complex between exogenic compounds introduced and a material present on the nervous terminations. This was the concept of the receptor, which later became very useful in choosing the active conformation of a drug and in constructing receptor-based models, as well as in the neighboring field of molecular modeling. The hypothesis of specific interactions was also formulated by Fisher in 1894 [12–14], with the image of "the key and the lock," and was later modified in 1966 by Koshland [15], who envisaged the possibility of receptor deformation on ligand binding. This was the notion of "induced fit."

At nearly the same time, Richet [16] correlated toxicities of narcotics with the inverse of their solubility in water in 1893, and in 1899–1901 Meyer and Overton [17,18] independently found linear relationships between the toxicity of organic compounds and their lipophilicity (ability to partition between water and a lipophilic biophase, the system olive oil–water being proposed as a reference medium).

Bases of Modern QSARs

The basic hypothesis is that the structure of a molecule contains features (geometric and/or electronic) responsible for its physical, chemical, or biological properties. Thus, for a given biological process from a set of active molecules assumed to have the same (or very similar) mode of action (MOA), it becomes possible to define a model relating structure and activity provided that the molecular structure can be represented by a set of structural descriptors (numerical values, fragments, etc). This corresponds to "ligand-based" models.

The parameters characterizing the molecular structure may be as follows:

- Descriptors calculated from the 2D molecular formula or the actual 3D geometry.
- Physicochemical quantities (measured or calculated) such as partition coefficients, vapor pressures, ionization constants, and orbital energies.

More precisely, QSARs generally relate, in a series of compounds, the *variations* of the activity to *variations* in the values of computed or experimental characteristics or properties of the structures.

此为试读,需要完整PDF请访问: www.ertongbook.com

The biological action of a chemical is generally associated from an early stage with (non-covalent) interactions with a specific "receptor" (protein, enzyme, etc.) in the living organism ("receptor-mediated" mechanism). Evaluation of these interactions leads to "receptor-based" models in contrast to the previously mentioned "ligand-based" models that consider only the active molecules and ignore their biological receptors. However, these approaches, which rely more closely on the actual mechanism of action, are, in most cases, more intricate, and to date less frequently developed. Nevertheless, with the increasing number of protein structures (free or bound to a ligand) now available (from X-ray crystallography, NMR, and homology modeling), the number of such applications is also rapidly increasing.

A traditional distinction remains between QSARs and QSPRs (quantitative structure property relationships), which deal with the prediction of physicochemical quantities. Although there are some differences in the nature of the data that are treated (biological data are often "softer" than physicochemical data), this distinction looks rather artificial since mathematical and statistical models are generally the same, and some physicochemical parameters (such as partition coefficient, pK values, orbital energy) may be introduced in the QSAR formulation.

2D MODELS

Linear free energy relationships: The birth of modern QSAR models is generally associated with the pioneering work of Hammett (1937) [19], who defined substituent constants for describing the electronic properties of aromatic compounds, and Taft (1952), who introduced steric substituent constants [20,21]. QSAR models have also benefited from the work of Ferguson [22], who proposed a thermodynamic interpretation of the relationship between nonspecific narcotic effect levels and lipophilicity.

Development was then stalled for some years until a new impetus was provided in the 1960s, when Hansch and Fujita (1964) [23] used the formalism of the linear free energy relationships (also known as "extrathermodynamic relationships") to correlate biological activities with physicochemical properties. We recall here that Hammett proposed to quantify electronic effects in substituted aromatic compounds by σ constants, measured in the dissociation of substituted benzoic acids:

$$\text{For the substituent X, } \sigma_X = \log (K_X/K_H)$$

where K_H and K_X are the dissociation constants of the benzoic acid (the reference) and the X-substituted benzoic acid, respectively. The advantage of this σ scale is that the variations in the equilibrium (K) or rate (k) constants of many reactions can be correlated by equations such as

$$\log (k_X/k_H) \quad \text{or} \quad \log (K_X/K_H) = \rho\sigma_X$$

where k_X (resp. K_X) are the rate (or equilibrium) constant of the X-substituted compound and k_H (resp. K_H) are the corresponding values for the reference (X = H). The reaction parameter ρ characterizes the sensitivity of the process to electronic

effects (and may, for instance for rate constants, be related to the partial charge developed in the transition state). This first scale was followed by many other substituent constant scales (σ^-, σ^+, σ_0, σ_R, F, R) for a more precise evaluation of electronic effects, and, later, for aliphatic systems by the σ^*, E_S scales for polar and steric effects, respectively [20,21,24,25]; however, a more precise representation of steric effects (taking into account the shape of the substituent group) was introduced by Verloop et al. [26].

After the definition of the π constants characterizing lipophilicity contribution, Hansch and Fujita proposed a σ-π-analysis on various processes such as activity of benzoic acids on mosquito larvae or of diethylaminoethyl benzoates on guinea pigs [23]. The general form of such a correlations is

$$A = a\sigma + bE_S + c\pi + d\pi^2$$

The biological activity A is expressed as the concentration of the chemical for a given end point (50% mortality or effect, log IC_{50}), the inhibitory power or the dissociation constant of the drug–receptor complex log K_i, etc. Parabolic terms (in π^2) were introduced in Hansch relationships [27] to express the fact that very polar drugs will not reach the receptor site due to their inability to cross lipid membranes, whereas very lipophilic drugs will just stay "trapped" in these membranes and will not pass through aqueous phases. Only compounds with intermediate lipophilicity have a good chance of arriving at the receptor site in a reasonable time frame and with sufficient concentration.

These pioneering studies were followed by several QSARs built with these substituent constants (of experimental origin) [28,29].

Incremental models: Free and Wilson [30] and Fujita and Ban [31] developed a different type of model using additive indicator variables (set to 0 or 1) to encode the presence (or the absence) of certain chemical groups. This method had also been explored earlier by Bruice et al. [32]. In the same vein, structural fragments were used in classification (separation of active, weakly active, or inactive compounds).

In the framework of molecular graphs, the DARC (Description, Acquisition, Restitution and Conception) system of Dubois [33,34] considered contributions of ordered atomic positions (sites) in a hierarchically ordered concentric description of the environment of a focus (the atomic position, or the bond, where the property was localized). For properties not localized on a given site or bond (e.g., ^{13}C shifts or IR frequencies), a "defocalized treatment" is possible. Numerous examples (related to QSPR models) are provided in Ref. [34]. In addition to QSAR/QSPR models, this type of description of the molecular structure has been applied to several other fields of cheminformatics: database management, structural elucidation, spectral simulation, and computer-aided organic synthesis.

Topological indices: Another effective approach took advantage of the similarity between chemical structural formulas and mathematical graphs (atoms corresponding to vertices, and bonds to edges). This led to the definition of topological indices, which started in 1947 with Wiener's work on the boiling point of paraffins,

where molecules were described by path counts determined on the molecular graph [35]. This application was followed by the introduction of a deluge of indices (more than 400) encoding, for example, ramification (Randić), shape (Kier and Hall), and cyclization (Balaban). Electronic aspects were also taken into consideration (E state indices of Kier and Hall, electrotopologic indices, or Galvez charge distribution indices). For more details, see Devillers and Balaban [36] or Todeschini and Consoni [37].

3D AND 2.5D MODELS

The preceding topological models are based on the structural formula of molecules (a 2D only representation), and, despite this, led to many satisfactory correlations (predicted vs. observed activities). However, it is clear that the properties of molecules actually depend on their 3D structures. Numerous examples can be found of the influence of geometrical isomerism or conformational preferences in spectroscopy or reactivity even in elementary organic chemistry; hence, the efforts to develop 3D models taking the actual molecular geometry into consideration. This was not without its share of problems. Considering 3D structures implies the choice of a "good conformation." The minimum energy conformation attainable by more or less lengthy geometry optimizations were frequently used but things worked better when the "active conformation" (that bound the receptor) was known (or might be reasonably inferred). Schematically, two avenues were explored:

3D models: A first type of method, exemplified by the well-known, and still widely used, CoMFA method [38] exploits structural information on discrete and individualized points in the neighborhood of the molecules under scrutiny. This corresponds to actual 3D methods.

With the pioneering work of Cramer et al., after the limited success of DYLOMMS (DYnamic lattice-oriented molecular modeling system) [39], the CoMFA methodology [38] asserted itself as the 3D method of reference, and is still widely used 20 years after its inception. The method relies on the calculation of steric and electrostatic potentials on nodes of a lattice surrounding the molecules, all aligned on a common reference in their supposed active conformation. Additional potentials have also been proposed. The steepness of the steric potential caused some problems and a few years later, the CoMSIA approach [40] replaced the evaluation of potentials by similarity calculations with smoother functions.

However, although several programs are now available, alignment (a critical step that must be addressed) remains a problem, especially when considering a series of non-congeneric compounds. But with the increasing number of ligand–receptor complexes solved by X-ray crystallography, NMR, or homology modeling studies (to infer the receptor-binding site structure), docking calculations (determination of the best ligand arrangement in the receptor site) now make the choice of the active conformation of the drug relatively easier.

In addition to CoMFA and CoMSIA, several other approaches were developed. Rather than nodes on a lattice external to the molecular shapes, they considered nodes in the occupied molecular volumes or points scattered on the surface as,

for example, molecular shape analysis (MSA) [41], hypothetical active site lattice (HASL) [42], and comparative molecular active site analysis (CoMASA) [43].

2.5D models: The other explored avenue corresponds to what may be called *2.5D approaches* [44]. They take into account descriptors or quantities that obviously depend on the molecular geometry, but the information is condensed in a single numerical value or a vector of a few components (without explicitly specifying the location of the point where information is collected). For example, the delicate alignment phase of CoMFA and CoMSIA is avoided in GRIND [45] by means of an *auto-* and *cross-correlation* transform.

The *similarity concept* (similar molecules have similar property), largely verified but with some exceptions [46], prompted the use of similarity indices (steric, electrostatic, or quantum) as structural descriptors. For a population of N compounds, each molecule is described by its similarity with the other molecules of the set in an N*N symmetrical matrix. One advantage is that only pairwise comparisons are carried out without alignment of the whole set on a common reference [47–50].

Relevant to the same concern, QSDARs (quantitative spectroscopic data activity relationships) took advantage of the sensitivity of spectroscopic data to structural or geometrical modifications to convey molecular information: The underlying idea was that spectra very sharply reflected the molecular structure and so were able to characterize it. Descriptors are here measured or calculated spectral parameters (^{1}H or ^{13}C NMR shifts, IR frequencies, orbital energies) rather than abstract calculated descriptors [51,52].

Alternately, other approaches may be viewed as a direct expansion of the "classical" 2D-QSARs. In addition to constitutional (molecular weight, number of atoms) and topological descriptors (calculated from the structural formula) or substituent constants (σ, π), other descriptors were introduced involving the molecular geometry. HOMO, LUMO energies, for example, have been largely used, as well as molecular surface area and similar quantities. Topographical indices and extrapolation of topological indices, where the actual 3D interatomic Euclidian distances replace the topological distances (number of bonds between atoms), were also introduced, but, seemingly, with a more limited diffusion.

Various packages (e.g., CODESSA [53], DRAGON [54], ALMOND [55], Cerius [56], ADAPT [57]) are now available for the calculation of a large number of such 2D and 2.5D descriptors (nearly 3000). Faced with this deluge, an important problem arose: the selection of the descriptors relevant to the specific biological application looked for (vide infra).

Receptor-Based Methods

All these approaches (ligand based) consider only the drug and provide no information on the receptor (there are, of course, exceptions as cited above). On the other hand, "receptor-based" methods try to evaluate interaction energies between the receptor and a potential ligand. The GRID approach developed by Goodford [58], who determined positions for favorable interactions on nodes in the vicinity of a protein, was the starting point. These methods required more sophisticated potential functions than those generally used in ligand-based models and often more refined thermodynamic

paths to provide reliable results [59–61]. However, with the ever-increasing sophistication of computers, such thermodynamic-type models, which were until some years limited to specific examples, can now be implemented in QSAR studies.

4D MODELS AND FURTHER

All these models implicitly assume that ligands act under a single conformation and bind in a unique or similar mode. The possibility of several simultaneous binding modes was investigated by Lukakova and Balaz [62] on rigid aromatic hydrocarbons. Another problem is that flexible compounds may exist as a mixture of several coexisting conformations. This problem was addressed in the 4D methodology proposed by Hopfinger [63–65] whereas Funatsu et al. [66,67] used *n-way* PLS analysis to choose the good alignment and the good conformation among various possibilities.

Until recently, the receptor was assumed to be rigid. However, several experimental observations showed that the receptor may undergo some deformation to accept a ligand. In other words, the old "key and lock" image must be replaced by something like "the hand and the glove," which corresponds to "induced fit." In the "fifth-dimension" models, different adaptation protocols, with possible dynamic interchange between them, are considered [68–71]. The possibility of different solvation models was even introduced in the 6D-QSAR approach.

2D- vs. 3D-QSAR Models: What Is the Best Choice?

In the 1990s, with the increasing number of topological indices, this question was the subject of heated debates in several QSAR meetings. However, it seems to us to be more a problem of resources and objective rather than a dilemma. As expressed aptly by a Chinese proverb "no matter a cat is white or black if it catches mice well."

From the various comparisons carried out so far on specific properties, with often (more or less) limited populations of compounds, no definite general conclusion can be drawn. However, on an extended data set, Sutherland et al. [72] indicated a better predictive capability of field-based 3D methods, and emphasized the importance of interpretability of the models. A naive empirical guideline might be as follows:

- If we need only a numerical equation or a mathematical model allowing for reasonably predicting activities, a 2D approach (duly validated) may be sufficient, and, in agreement with Occam's razor principle, it will then be useless to search for a complex 3D model since the calculation of 2D indices is very rapid and requires only knowledge of the structural formula.
- Conversely, if the problem is to get some insight into the main intervening interactions at a physicochemical level, and to access geometrical information as to the spatial areas involved, a 3D approach is far more efficient. It may readily suggest, for example, what structural modifications would be interesting for the synthesis of new active compounds. However, it is more time-consuming and resource-demanding as several processes, such as conformational analysis, choice of conformation, and (sometimes) alignment have to be covered.

QSARs and Related Fields: Database Mining, Molecular Modeling, and Post-3D-QSAR Models

We would like to conclude this short historical survey indicating the synergy arising from QSARs and molecular modeling not only in receptor-based approaches but also in ligand-directed studies. Molecular docking (determining the best conformation and orientation of a ligand in its receptor binding site) is of definite help for aligning compounds in CoMFA and CoMSIA treatments. At the same time, X-ray crystallography and homology modeling of proteins supply additional tools that are of great value. Molecular dynamics, quantifying the conformational flexibility of ligands, has now become an integrated part of numerous QSAR applications.

At the beginning, QSARs were restricted to linear relationships, on strictly congeneric compounds (a common core and some substituent groups), involving standard "substituent constants" in an LFER formalism. With 3D methods such as CoMFA and CoMSIA, the numerical prediction is refined, taking into account the various fields acting on the ligands, and even (with Quasar and Raptor [69,73]) the geometrical adaptation of the ligand, the variation in the dynamic processes, and the solvation modes.

Finally, QSARs, initially devoted to small populations of compounds with a common activity, now constitute new tools for mining chemical databases in view of possibly finding active compounds. Although this operation was mainly treated in the past as a classification problem relying on structure or substructure recognition, the development of a QSAR model, even a crude one, can be efficient to categorize compounds, provided the treatment is automated, for an acceptable speed. Various recent models have exemplified their efficiency [74–76].

BUILDING A QSAR MODEL

The Main Steps

The principal phases in building a QSAR model (Figure 1), that is, establishing a mathematical or statistical relationship that links the biological activity to a description of the molecular structure

$$\text{Biological activity } y = F \text{ (molecular descriptors)}$$

can be summarized as follows:

- *Constitution of the data set*, and splitting it into a training ("learning") set that will be used to adjust the model, and a test set, which the model has never seen and that will be used subsequently to check its predictive capacity. The training set must span, as widely as possible, the structural space with a rather limited number of compounds. The test set must correspond to a "reasonable" extrapolation.
- *Generation of the descriptors* characterizing the molecular structures under scrutiny. The nature and the number of the descriptors depend on the

此为试读,需要完整PDF请访问: www.ertongbook.com