Recent Advances in QSAR Studies

CHALLENGES AND ADVANCES IN COMPUTATIONAL CHEMISTRY AND PHYSICS

Volume 8

Series Editor:

JERZY LESZCZYNSKI
Department of Chemistry, Jackson State University, U.S.A.

For further volumes:
http://www.springer.com/series/6918

Recent Advances in QSAR Studies

Methods and Applications

Edited by

Tomasz Puzyn
University of Gdańsk, Gdańsk, Poland

Jerzy Leszczynski
Jackson State University, Jackson, MS, USA

Mark T.D. Cronin
Liverpool John Moores University, Liverpool, UK

Springer

Editors

Dr. Tomasz Puzyn
Laboratory of Environmental
Chemometrics
Faculty of Chemistry
University of Gdańsk
ul. Sobieskiego 18/19
80-952 Gdańsk
Poland
puzi@qsar.eu.org

Prof. Jerzy Leszczynski
Interdisciplinary Nanotoxicity
Center
Department of Chemistry
Jackson State University
1325 Lynch St
Jackson, MS 39217-0510
USA
jerzy@icnanotox.org

Dr. Mark T. D. Cronin
School of Pharmacy and Chemistry
Liverpool John Moores University
Byrom Street
Liverpool L3 3AF
England
m.t.cronin@ljmu.ac.uk

ISBN 978-1-4020-9782-9 e-ISBN 978-1-4020-9783-6
DOI 10.1007/978-1-4020-9783-6
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2009930952

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

PREFACE

Since the inception of this volume, the world's financial climate has radically changed. The emphasis has shifted from booming economies and economic growth to the reality of recession and diminishing outlook. With economic downturn comes opportunity, in all areas of chemistry from research and development through to product registration and risk assessment, replacements are being sought for costly time-consuming processes. Leading amongst the replacements are models with true predictive capability. Of these computational models are preferred.

This volume addresses a broad need within various areas of the "chemical industries", from pharmaceuticals and pesticides to personal products to provide computational methods to predict the effects, activities and properties of molecules. It addresses the use of models to design new molecules and assess their fate and effects both to the environment and to human health. There is an emphasis running throughout this volume to produce robust models suitable for purpose. The volume aims to allow the reader to find data and descriptors and develop, discover and utilise valid models.

Gdańsk, Poland
Jackson, MS, USA
Liverpool, UK
May 2009

Tomasz Puzyn
Jerzy Leszczynski
Mark T.D. Cronin

CONTENTS

Part I Theory of QSAR

Part II Practical Application

Part I
Theory of QSAR

CHAPTER 1

QUANTITATIVE STRUCTURE–ACTIVITY RELATIONSHIPS (QSARs) – APPLICATIONS AND METHODOLOGY

MARK T. D. CRONIN

School of Pharmacy and Chemistry, Liverpool John Moores University, Liverpool L3 3AF, England, e-mail: m.t.cronin@ljmu.ac.uk

Abstract: The aim of this introduction is to describe briefly the applications and methodologies involved in (Q)SAR and relate these to the various chapters in this volume. This chapter gives the reader an overview of how, why and where in silico methods, including (Q)SAR, have been utilized to predict endpoints as diverse as those from pharmacology and toxicology. It provides an illustration of how all the various topics in this book interweave to form a single coherent area of science.

Keywords: QSAR, In silico methods, Resources for QSAR

1.1. INTRODUCTION

If we can understand how a molecular structure brings about a particular effect in a biological system, we have a key to unlocking the relationship and using that information to our advantage. Formal development of these relationships on this premise has proved to be the foundation for the development of predictive models. If we take a series of chemicals and attempt to form a *quantitative relationship* between the biological effects (i.e. the *activity*) and the chemistry (i.e. the *structure*) of each of the chemicals, then we are able to form a *quantitative structure–activity relationship* or QSAR.

Less complex, or quantitative, understanding of the role of structure to govern effects, i.e. that a fragment or sub-structure could result in a certain activity, is often simply termed a *structure–activity relationship* or SAR. Together SARs and QSARs can be referred to as (Q)SARs and fall within a range of techniques known as in silico approaches. Generally, although there is no formal definition, in silico includes SARs and QSARs, as well as the use of existing data (e.g. searching within databases), category formation and read-across. It also borders into various other areas of chemoinformatics and bioinformatics.

T. Puzyn et al. (eds.), Recent Advances in QSAR Studies, 3–11.
DOI 10.1007/978-1-4020-9783-6_1, © Springer Science+Business Media B.V. 2010

A (Q)SAR comprises three parts: the (activity) data to be modelled and hence predicted, data with which to model and a method to formulate the model. These three components are described below and in greater detail in subsequent chapters.

1.2. PURPOSE OF QSAR

QSAR should not be seen as an academic tool to allow for the post-rationalization of data. We wish to derive the relationships between molecular structure, chemistry and biology for good reason. From these relationships we can develop models, and with luck, good judgment and expertise these will be predictive. There are many practical purposes of a QSAR and these techniques are utilized widely in many situations. The purpose of in silico studies, therefore, includes the following:

- To predict biological activity and physico-chemical properties by rational means.
- To comprehend and rationalize the mechanisms of action within a series of chemicals.

Underlying these aims, the reasons for wishing to develop these models include

- Savings in the cost of product development (e.g. in the pharmaceutical, pesticide, personal products, etc. areas).
- Predictions could reduce the requirement for lengthy and expensive animal tests.
- Reduction (and even, in some cases, replacement) of animal tests, thus reducing animal use and obviously pain and discomfort to animals.
- Other areas of promoting green and greener chemistry to increase efficiency and eliminate waste by not following leads unlikely to be successful.

1.3. APPLICATIONS OF QSAR

The ability to predict a biological activity is valuable in any number of industries. Whilst some QSARs appear to be little more than academic studies, there are a large number of applications of these models within industry, academia and governmental (regulatory) agencies. A small number of potential uses are listed below:

- The rational identification of new leads with pharmacological, biocidal or pesticidal activity.
- The optimization of pharmacological, biocidal or pesticidal activity.
- The rational design of numerous other products such as surface-active agents, perfumes, dyes, and fine chemicals.
- The identification of hazardous compounds at early stages of product development or the screening of inventories of existing compounds.
- The designing out of toxicity and side-effects in new compounds.
- The prediction of toxicity to humans through deliberate, occasional and occupational exposure.
- The prediction of toxicity to environmental species.
- The selection of compounds with optimal pharmacokinetic properties, whether it be stability or availability in biological systems.

- The prediction of a variety of physico-chemical properties of molecules (whether they be pharmaceuticals, pesticides, personal products, fine chemicals, etc.).
- The prediction of the fate of molecules which are released into the environment.
- The rationalization and prediction of the combined effects of molecules, whether it be in mixtures or formulations.

The key feature of the role of in silico technologies in all of these areas is that predictions can be made from molecular structure alone.

1.4. METHODS

Predictive models of all types are reliant on the data on which they are based, the technique to develop the model and the overall quality of the information including the item to be modelled. In silico models for the prediction of the properties and effects of molecules are no different. In almost all cases two types of information are required for a model (the effect to be modelled and descriptors on the chemicals) and a technique(s) to formulate the relationship(s). These are denoted in Figure 1-1 in a typical spreadsheet for organizing the data. The data to be modelled are denoted as the X-matrix, the descriptors as the Y-matrix. From such a matrix various types of relationship may be obtained by statistical, or other, means. For instance, a structure–activity relationship will be formed for a categorical endpoint, e.g. active/non-active or toxic/non-toxic. In this case a molecular fragment or substructure is associated with an effect. A quantitative structure–activity relationship is based on a continuous endpoint, e.g. potency where activity (X) is a function of one or more descriptors (Y).

To develop a SAR, as few as a single compound might be required – should there be a very firm basis (such as a well-established mechanism of action) for developing the relationship. For instance, if a compound is known to elicit a particular effect, and the structural determinant is recognized, that structural fragment can be extracted. This may be in the form of a “structural alert” which can be coded easily into software. Obviously, the greater the number of compounds with the same structural determinant demonstrating the same effect, the greater the confidence that

Chemical Identifier	Activity (to be modelled)	Property/ Descriptor/ Fragment 1	Property/ Descriptor/ Fragment 2	Property/ Descriptor/ Fragment 3	…	Property/ Descriptor/ Fragment *n*
Molecule *i*	X_i	Y_{1i}	Y_{2i}	Y_{3i}	…	Y_{ni}
Molecule *ii*	X_{ii}	Y_{1ii}	Y_{2ii}	Y_{3ii}	…	Y_{nii}
Molecule *iii*	X_{iii}	Y_{1iii}	Y_{2iii}	Y_{3iii}	…	Y_{niii}
…	…	…	…	…	…	
Molecule *n*	X_n	Y_{1n}	Y_{2n}	Y_{3n}	…	Y_{nn}

Figure 1-1. Typical data matrix for a (Q)SAR study

can be demonstrated in the alert. The formation of SARs is usually appropriate for a qualitative (i.e. yes/no; active/inactive; presence of toxicity/absence of toxicity, etc.) endpoint.

To develop a QSAR, a more significant number of compounds is required to develop a meaningful relationship. An often asked question is "how many compounds are required to develop a QSAR?" There is no direct and simple response to this question – other than "as many as possible!" To provide some guide, it is widely accepted that between five and ten compounds are required for every descriptor in a QSAR [1, 2]. This does suggest that a one descriptor regression-based QSAR could be developed on five compounds. This is possible, but is very reliant on issues such as data distribution and range. Ideally "many more" compounds are required to obtain statistically robust QSARs, with some modelling techniques being considerably more data hungry than regression analysis.

In the history of developing in silico models, there have been many types of information integrated into (Q)SARs. These are summarized in Table 1-1. The biological effects are normally (though not exclusively) the property to be modelled; some aspect from the physical or structural chemistry of the molecules is related to the effects. Readers are welcome to extend this list according to their experience and requirements!

There has been a wide range of modelling approaches. A brief overview of these is given in Table 1-2. These can be very simplistic to extremely complex.

Table 1-1. Types of information included in in silico modelling approaches and reference to chapters for further reading

- Data to be modelled
  - Pharmacological effects (Chapter 9)
  - Toxicological effects (Chapters 7, 11, 12 and 14)
  - Physico-chemical properties (Chapters 12 and 14)
  - Pharmacokinetic properties governing bioavailability (Chapters 9 and 10)
  - Environmental fate (Chapter 12)
- Chemistry
  - Physico-chemical properties (Chapters 12 and 14)
  - Structural properties – 2-D and 3-D (Chapters 4, 5, 8 and 14)
  - Presence, absence and counts of atoms, fragments, sub-structures (Chapters 3 and7)
  - Quantum and computational chemistry (Chapters 2 and 14)
- Modelling
  - Formation of categories of "similar molecules" (Chapters 7, 13 and 14)
  - Statistical (Chapters 5, 6 and 12)
  - 3D/4D QSAR (Chapters 2, 4, 5, 9 and 14)
- Other issues
  - Data quality and reliability (Chapter 11)
  - Model and prediction reporting formats (Chapter 13)
  - Applicability domain (Chapters 12 and 13)
  - Robustness of model and validity of a prediction (Chapters 6 and 12)

Table 1-2. Summary of the main modelling approaches for the development of (Q)SARs and in silico techniques and where further details are available in this volume

(Q)SAR method	Chapters
Hansch analysis	9
Free-Wilson	9
Structural fragments and alerts	7, 12
Category formation and read-across	7
Linear regression analysis	5, 6, 12
Partial least squares	5, 6
Pattern recognition	6
Robust methods, outliers	6
Pharmacophores	4, 5, 9
3-D models	2, 4, 14
CoMFA	4

1.5. THE CORNERSTONES OF SUCCESSFUL PREDICTIVE MODELS

Predictive and intuitive models are widely used in all aspects of society and science. The user of a model accepts that it is a model and the results, or information it provides, should be used with circumspection. This is true whether one is accepting an actuarial prediction for one's pension planning or a weather forecast to determine whether to wear a raincoat. The same is true for a prediction, or any information (e.g. mechanism of action), that may be determined from a (Q)SAR. Therefore, the user must put the model in the context in which it exists and be aware of a number of possible problems and pitfalls.

Much has been written and said about the reality of using (Q)SARs. The concern is that scientists who are introduced to the field can place too much confidence in either a model or predictive system, only to see their expectations dashed. Alternatively, there is a point of view that (Q)SARs will not work and models are not to be trusted. A healthy dose of scepticism is important, but some form of balance is required to meet the hopes of the optimists and criticisms of the sceptics. In order to do that, some comment is required on the "successful use of (Q)SAR".

The requirements for a good model are quite straightforward. Some of the fundamentals are noted below and expanded upon in more detail in various chapters of this book.

(i) *The data to model.* The modeller, and user of a model, must consider the data to model. Data should, ideally, be of high quality, meaning they are reliable and consistent across the data set to be modelled. The definition of data quality is, at best, subjective and is likely to be different for any effect, endpoint or property. Therefore, the modeller or user should determine whether the data are performed in a standard manner, to a recognized protocol, and if they are taken from a single or multiple laboratories.

This author is of the belief that, within reason, poor-quality data can be used in models, but their limitations must be clearly understood, and the implications for the model appreciated. Therefore, to use a (Q)SAR successfully there should be complete access to the data used and a complete description of those data. Even then, producing public models from confidential business information may make this restrictive. To provide the source data is the responsibility of the modeller, and to assess the source data in terms of the model is the responsibility of the model user.

(ii) *Reasonable and honest use of statistics to describe a (Q)SAR*. Many (Q)SARs are accompanied by performance statistics of some kind. These statistics may assess statistical fit and predictivity for a QSAR or the predictive capability of a SAR. Generally statistics are helpful to the interpretation of a model. One would prefer to use a model with a good statistical fit between the effect to be modelled and the descriptors of the chemicals. However, it is important to ensure that the statistical fit of a model does not go beyond the experimental error of the data being modelled – should that happen it would suggest an overfit model. To develop a significant quantitative model, a significant range of effect values are ideally required. Also, one must be cautious of comparing the ubiquitous correlation coefficient between different data sets.

In the opinion of the author, whilst neither the model developer nor the user needs be a statistician, it is of great help to discuss the issues with a competent statistician. In addition, the developer or user must have confidence in the statistics they are applying and interpreting.

(iii) *The molecules for which predictions are being made must be within the applicability domain of the model*. The applicability domain of the model is the chemical, structural, molecular, biological and/or mechanistical space of the data set of the model. The definition of the applicability domain will vary for different types of model (e.g. SAR vs. QSAR), endpoints and effects. There are also a wide variety of methods to define it. The important fact is that the user of a model must assess whether a molecule is within the domain of a model, and thus how much confidence they can place in a predicted value.

(iv) *Ideally a (Q)SAR should be simple, transparent, interpretable and mechanistically relevant*. A simple model will have only one or a very small number of descriptors to form the relationship with the effect data. Transparency is usually dependent on the modelling approach itself; thus linear regression analysis can be thought of as being highly transparent, i.e. the algorithm is available, and predictions can be made easily. For the more multivariate and non-linear modelling techniques (e.g. a neural network), it is generally accepted that there is lower transparency.

The mechanistic relevance of a model is more difficult to define. Some data sets are based around a single, well-defined and understood mechanism of action. Other models comprise data where the mechanism may not be known or where there are many mechanisms. There is also a difference between biological mechanisms (e.g. receptor binding, concentration at an active site, accumulation in a membrane) and physico-chemical effects (e.g. the properties affecting solubility, ionization), which may be general across the chemical universe. There is no reason to exclude a model where the mechanism is not known or if there are multiple mechanisms. However, the advantages of a strong mechanistic basis to model are that it provides a clear capability to understand the model and should the descriptors be relevant to that mechanism it provides the user with extra confidence to use the model. Another advantage is that it can aid a priori descriptors selection.

In reality, (Q)SARs span the range from simple models to highly complex multivariate. It is important to remember that whilst a simple model is preferable in many circumstances if it provides comparable performance to something more complex, many multivariate models are routinely and successfully used. The requirements for model simplicity are highly dependent on the context and application of the model.

1.6. A VALIDATED (Q)SAR OR A VALID PREDICTION?

Historically, much effort has been placed into performing some form of validation on a (Q)SAR. Often this has been in terms of a model's statistical fit; more recently the focus has turned to using an external test set, i.e. group of molecules not in the original data set on which the model has been developed. Confusion has arisen in some areas, due to the term "validated" which has a specific regulatory, and hence legal, connotation in replacing animal tests in toxicology.

As a result of the efforts to use (Q)SARs correctly, for the statistical validation of models it is more usual to refer to those algorithms that may be applied in drug discovery and lead optimization. Whilst statistical approaches may be applied to toxicological endpoints of regulatory significance, for the validation of a toxicological (Q)SAR to be used to assess hazard, for example for the purposes of registration of a product, a more formal validation process may be required.

In terms of toxicological predictive models, "Principles for the Validation of (Q)SARs" have been proposed by the Organization for Economic Co-operation and Development (OECD) and promoted widely. These principles are described in more detail in Chapter 13, and whilst they were originally derived with toxicity and fate endpoints in mind, they are generally applicable across all models to determine whether a (Q)SAR may be valid. The use of the OECD principles has brought to the forefront of whether a (Q)SAR can be "validated" in terms of being an acceptable alternative method. Probably of more importance is using these principles to evaluate and characterize a (Q)SAR and hence determine whether an individual prediction is valid.

1.7. USING IN SILICO TECHNIQUES

This book will make it apparent that there are many models available for use in QSAR. Publication on paper is, of course, essential, but to make these models usable they must be presented in a user-friendly format. Thus, there have been many attempts to computerize these models. As computational power has increased, and hardware platforms became more sophisticated, the possibilities to produce useable algorithms have improved. Accessibility to software has also, of course, been made so much more convenient through the use of the Internet. As a result of the progress in these areas, many algorithms are now freely available. Sources of some, as well as other essential resources for (Q)SAR, are noted in Table 1-3.

Table 1-3. Invaluable resources for QSAR

Internet

There are obviously many Internet sites, wikis and blogs devoted to (Q)SAR, molecular modelling, drug design and predictive ADMET. Two of the most well established are

- The homepage of the International Chemoinformatics and QSAR Society: www.qsar.org – this is a good starting place for those in the field of QSAR; it also contains excellent listings of upcoming meetings and resources.
- The homepage of the Computational Chemistry List: www.ccl.net – this also contains excellent listings resources and freely downloadable software.

Journals

Papers relating to (Q)SAR are published in a very wide variety of journals from those in pure and applied chemistry to pharmacology, toxicology and risk assessment and as far as chemoinformatics and statistics. The following is a small number that is commonly used by the author; whilst the reader will hopefully find these suggestions useful, they are, by no means, an exhaustive list (see the resources section of www.qsar.org which lists over 250 journal titles).

- *Chemical Research in Toxicology*
- *Chemical Reviews*
- *Journal of Chemical Information and Modeling*
- *Journal of Enzyme Inhibition and Medicinal Chemistry*
- *Journal of Medicinal Chemistry*
- *Journal of Molecular Modelling*
- *"Molecular Informatics (formerly QSAR and Combinatorial Science)"*
- *SAR and QSAR in Environmental Research*

Books

There are many hundreds of books available in areas related to (Q)SAR. Again, the reader is referred to the resource section of www.qsar.org. A very short list is given below, clearly biased by the author's own interests and experience. Apologies are given for omission of other "favourite" or "essential" books that have not been listed.

- Cronin MTD, Livingstone DJ (eds) (2004) *Predicting Chemical Toxicity and Fate*, CRC Press, Boca Raton, FL.
- Helma C (ed) (2005) *Predictive Toxicology*, CRC Press, Boca Raton, FL.
- Livingstone DJ (1995) *Data Analysis for Chemists – Application to QSAR and Chemical Product Design*, Oxford University Press, Oxford.
- Todeschini R, Consonni V (2001) *Handbook of Molecular Descriptor*. Wiley, New York.
- Triggle DJ, Taylor JB (series eds) (2006) *Comprehensive Medicinal Chemistry II – Volumes 1–8*. Elsevier, Oxford.

Software

It is well beyond the scope or possibility of this section to note individual software for use in (Q)SAR. Experienced QSAR practitioners will no doubt be familiar with many of the freely available and commercial packages available. For the novice, in addition to the resources listed on www.qsar.org and www.ccl.net, there is information in the following chapters of this book in the three key areas to formulate a (Q)SAR:

- Activity to be modelled: Pharmacology (Chapters 4, 5, 9 and 10), ADMET (Chapters 4, 7, 10, 11, 12 and 14), physico-chemical properties (Chapters 8, 12 and 14)
- Descriptor calculation (Chapters 2, 3, 4, 5 and 14)
- Statistical analysis (Chapters 5, 6 and 12)

1.8. NEW AREAS FOR IN SILICO MODELS

Understanding and forming the relationships between the effect of a molecule and its structure has a long history [3] – its nearly 50 years since Hansch, Fujita and co-workers first published in this area [4], over 150 years since the foundations of modern chemistry and millennia since man first determined the beneficial and harmful effects of plants. It is surprising therefore that there continues to be such continued interest in developing technologies for in silico models.

There are many reasons for the growth of in silico techniques. In particular, these can be in response to new problems. Areas where in silico approaches can play a particular role include

- integrating and harnessing new computational technologies and increasing speed and power of processing;
- ability to react to new disease states (e.g. HIV);
- ability to react to new toxicological problems (e.g. cardio-toxicity);
- modelling the new problems with regard to the impact of chemicals on the environment;
- new and emerging issues, problems and opportunities, e.g. nano-technology, properties of crystals, extension into other areas of chemistry, e.g. design of formulations;
- integration with the -omics technologies to improve all areas of molecular design.

1.9. CONCLUSIONS

QSAR is a broadly used tool for developing relationships between the effects (e.g. activities and properties of interest) of a series of molecules with their structural properties. It is used in many areas of science. It is a dynamic area that integrates new technologies at a staggering rate. There have been many recent advances in the applications and methodologies of QSAR, which are summarized partially in Table 1-3 and more thoroughly described in this book.

REFERENCES

1. Topliss JG, Costello RJ (1972) Chance correlations in structure-activity studies using multiple regression analysis. J Med Chem 15:1066–1068.
2. Schultz TW, Netzeva TI, Cronin MTD (2003) Selection of data sets for QSARs: analyses of *Tetrahymena* toxicity from aromatic compounds. SAR QSAR Environ Res 14:59–81.
3. Selassie CD (2003) History of quantitative structure-activity relationships. In: Abraham DJ (ed) Burger’s Medicinal Chemistry and Drug Discovery, 6th edn., Volume 1: Drug Discovery. John Wiley and Sons, Inc., New York.
4. Hansch C, Maloney PP, Fujita T et al. (1962) Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. Nature 194:178–180.

CHAPTER 2

THE USE OF QUANTUM MECHANICS DERIVED DESCRIPTORS IN COMPUTATIONAL TOXICOLOGY

STEVEN J. ENOCH

School of Pharmacy and Chemistry, Liverpool John Moores University, Liverpool L3 3AF, England, e-mail: s.j.enoch@ljmu.ac.uk

Abstract: The aim of this chapter is to outline the theoretical background and application of quantum mechanics (QM) derived descriptors in computational toxicology, specifically in (quantitative) structure–activity relationship models ((Q)SARs). The chapter includes a discussion of the mechanistic rationale for the need for such descriptors in terms of the underlying chemistry. Having established the mechanistic rationale for quantum mechanical descriptors, a brief discussion of the underlying mathematical theory to quantum mechanical methodologies is presented, the aim being to help the reader understand (in simple terms) the differences between the commonly used levels of theory that one finds when surveying the computational toxicological literature. Finally, the chapter highlights a number of (Q)SAR models in which QM descriptors have been utilised to model a range of toxicological effects

Keywords: Geometry optimisation, Semi-empirical methods, Density functional theory, Quantum mechanical descriptors

2.1. INTRODUCTION

Computational toxicology is concerned with rationalising the toxic effects of chemicals, with the hypothesis being that if the factors that are responsible for a given chemical's toxicity can be understood, then the toxicity of related chemicals can be predicted without the need for animal experiments. Unfortunately, there are many factors, some of them extremely complex, that govern whether even the simplest industrial chemical will be toxic. The majority of these factors (e.g. metabolism, bioavailability) are outside of the scope of this chapter. Instead the focus of this chapter is to highlight the importance of assessing the electronic state of a potentially toxic chemical, and how this information enables one to begin to rationalise and subsequently predict certain aspects of human health and environmental toxicology.

Knowledge of a chemical's mechanism of action is important if a chemical's potential toxic effects are to be understood. Broadly speaking, potential non-

T. Puzyn et al. (eds.), Recent Advances in QSAR Studies, 13–28.
DOI 10.1007/978-1-4020-9783-6_2, © Springer Science+Business Media B.V. 2010

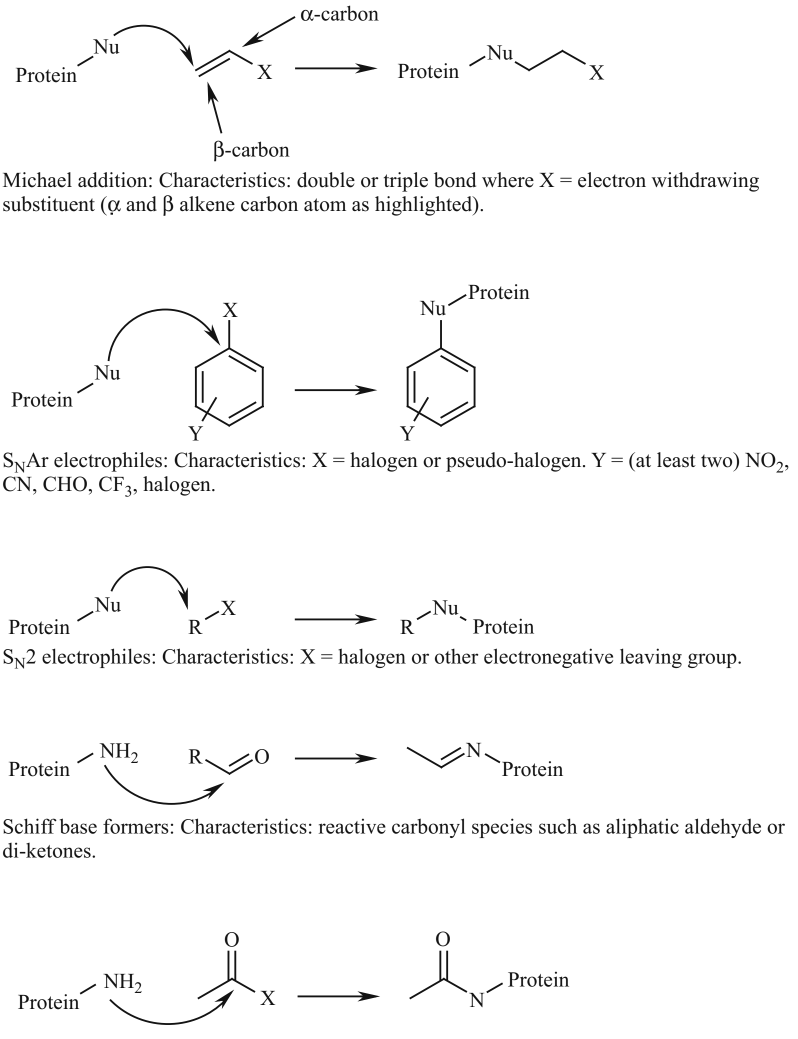
receptor-mediated mechanisms of toxic action can be divided into non-covalent and covalent categories. One of the most important non-covalent mechanisms in aquatic systems involves the accumulation of a chemical within the cell membrane resulting in narcosis. Chemicals able to cause narcosis can be split into a number of mechanisms, the two most frequent being non-polar and polar narcosis. Non-polar narcotics are well modelled using hydrophobicity alone, whilst the modelling of the polar chemicals may require the inclusion of a parameter to account for the polarisation effect of an electronegative centre in the molecule. Such effects are well modelled using quantum mechanics derived descriptors such as E_{LUMO} and A_{max} (see Table 2-1 for definitions).

In contrast, covalent mechanisms of toxicity involve the formation of a chemical bond between proteins (or DNA) and the toxic chemical. Such mechanisms are irreversible and have little or no correlation with the chemical's hydrophobicity (assuming the hydrophobicity of the chemical is within a range that allows it to get to the reactive site). In order for a chemical to be toxic via a covalent mechanism, it must be electrophilic, that is to say some portion of it must be susceptible to attack (either directly or after either oxidative or metabolic conversion) from electron-rich amino acid (or nucleic acid) side chains. These covalent mechanisms have recently been rationalised in terms of simple electrophilic–nucleophilic organic chemistry reactions [1] (Figure 2-1).

The chemical reactions between toxicant and biomolecule can be rationalised in terms of hard–soft acid–base theory which states that for a chemical reaction to occur like should react with like, i.e. a soft electrophile (where an electrophile can be considered as an acid) prefers to react with a soft nucleophile (where the nucleophile can be considered as a base), whilst a hard electrophile preferentially reacts with a hard nucleophile [2]. This is related directly to the energies of the frontier molecular orbitals as a soft electrophile has a low E_{LUMO} which can readily interact with the energetically close high E_{HOMO} of the soft nucleophile. In contrast, a hard electrophile has a high E_{LUMO} that can readily interact with the energetically close low E_{HOMO} of a hard nucleophile.

In the simplest terms it is the relative differences between the nucleophile and electrophile orbitals that govern how reactive a given nucleophile–electrophile interaction will be (assuming factors such as entropy and steric hindrance at the reaction centre are equal). Clearly, in terms of covalent toxicity mechanisms, the more reactive a nucleophile–electrophile interaction is (in which the nucleophile is a protein or DNA and the electrophile is a chemical) the more toxic the chemical is likely to be. However, it is important to remember that toxicokinetics and toxicodynamics play an important role in a chemical's ability to produce a toxic effect, with the relative importance (compared to intrinsic reactivity) of such effects being mechanism dependent. Given the importance of the frontier molecular orbitals in hard–soft acid–base theory, it is clear that quantum mechanics methods that enable the molecular orbitals to be calculated play an important role in the rationalising and the subsequent modelling of such reactions.

Table 2-1 highlights some common descriptors used to model both covalent and non-covalent mechanisms. In addition, Schüürmann provides an excellent recent review of the theoretical background of such descriptors in more detail [3].



Michael addition: Characteristics: double or triple bond where X = electron withdrawing substituent (α and β alkene carbon atom as highlighted).

S_NAr electrophiles: Characteristics: X = halogen or pseudo-halogen. Y = (at least two) NO_2, CN, CHO, CF_3, halogen.

S_N2 electrophiles: Characteristics: X = halogen or other electronegative leaving group.

Schiff base formers: Characteristics: reactive carbonyl species such as aliphatic aldehyde or di-ketones.

Acylating agents: Characteristics: X = halogen or electronegative leaving group

Figure 2-1. Electrophilic–nucleophilic reactions responsible for covalent mechanisms of toxic action

2.2. THE SCHRÖDINGER EQUATION

Given the importance of the ability to calculate the electronic structure of a molecule in computational toxicology, it is important to outline, albeit briefly, the underlying theory that both the commonly used semi-empirical and density functional methods attempt to solve. The mathematics is complex and will be kept to an absolute minimum, the aim being to set the scene concerning the various components that must be dealt with if quantum mechanics is to be utilised to help understand the electronic structure of chemicals. The subsequent sections dealing with the commonly

Table 2-1. Common quantum mechanics derived molecular and atom-based descriptors

Name	Definition
E_{LUMO}	Energy of the lowest unoccupied molecular orbital
E_{HOMO}	Energy of the highest occupied molecular orbital
μ	Chemical potential (negative of electronegativity) $\mu = (E_{LUMO} + E_{HOMO})/2$
η	Chemical hardness $\eta = (E_{LUMO} - E_{HOMO})/2$
σ	Chemical softness $\sigma = 1 - \eta$
ω	Electrophilicity $\omega = \mu^2/2\eta$
AEI	Activation energy index $AEI = \Delta E_{HOMO-1} + \Delta E_{HOMO}$ ΔE_{HOMO} and ΔE_{HOMO-1} are the changes in energy of the highest occupied molecular orbital and second highest occupied molecular orbital on going from the ground state to transition state in an S_NAr reaction
A_{max}	Maximum atomic acceptor superdelocalisability within a molecule, where acceptor superdelocalisability is a measure of an atom's ability to accept electron density
D_{max}	Maximum atomic donor superdelocalisability within a molecule, where donor superdelocalisability is a measure of an atom's ability to donate electron density
A_N	Atomic acceptor superdelocalisability for atom N
D_N	Atomic donor superdelocalisability for atom N
$\omega_m{}^+/\omega_m{}^-$	Atomic local philicity. Derived from Fukui functions [4], electrophilicity index ω and then applied to individual atoms using a charge scheme
Q_N	Atomic charge on atom N
B_{a-b}	Bond order between atom a and b

used semi-empirical and density functional approaches will highlight how each of these methods approximates these important mathematical components. The starting point of any discussion into quantum mechanics is always the time-independent Schrödinger equation (2-1):

$$\mathbf{H}\Psi = E\Psi \tag{2-1}$$

where **H** is the *Hamiltonian operator*, E is the energy of the molecule and ψ is the wavefunction which is a function of the position of the electrons and nuclei within the molecule.

A number of solutions exist for Eq. (2-1), with each one representing a different electronic state of the molecule. Importantly the lowest energy solution represents the ground state. It is worth stating that the Schrödinger equation is an eigenvalue equation, which in mathematical terms means that the equation contains an operator acting upon a function that produces a multiple of the function itself as the result [Eq. (2-2)]:

$$\mathbf{Operator}^{*}\text{function} = \text{constant}^{*}\text{function} \tag{2-2}$$

In Eq. (2-1) the wavefunction (ψ) can be approximated to the electronic state, this being the configuration of the electrons in a series of molecular orbitals. It is then possible to evaluate differing electronic configurations of the wavefunction in terms of their energies, with the lowest energy configuration being the ground state. It is the ground state energy that corresponds to the ground state geometry of a given molecule. For a given wavefunction the associated *Hamiltonian* is calculated according to Eq. (2-3):

$$\mathbf{H}= \mathrm{KE}_{\mathrm{total}}+\mathrm{PE}_{\mathrm{total}} \tag{2-3}$$

where

$\mathrm{KE}_{\mathrm{total}}$ = total kinetic energy = $\sum$ (coulomb repulsion between each pair of charged entities)

$\mathrm{PE}_{\mathrm{total}}$ = total potential energy = $\sum$ (electron–nuclei attraction) + $\sum$ (electron– electron repulsion) + $\sum$ (nuclei–nuclei repulsion).

In order to evaluate the components of Eq. (2-3), a number of approximations are required that are complex and out of the scope of this chapter. A number of excellent texts exist that discuss these approximations in great detail [5, 6].

2.3. HARTREE–FOCK THEORY

Having established the importance of the electronic wavefunction (ψ) in Eq. (2-1), it is now necessary to discuss the methods that enable the derivation of the electronic states for which Eq. (2-1) holds true. The following discussion is an outline of the fundamentals of Hartree–Fock theory from which both semi-empirical and density functional methods have been developed.

The first step towards obtaining an optimised electronic structure (i.e. the ground state) for a molecule is to consider the wavefunction as a series of molecular orbitals with differing electronic occupations. One of these sets of molecular orbitals will correspond to the ground state and hence have the lowest energy. Approximating the wavefunction to a series of molecular orbitals allows the substitution of the wavefunction in Eq. (2-1) with Eq. (2-4) resulting in Eq. (2-5) (both Eqs. (2-4) and (2-5) are simplified to illustrate the important conceptual idea that in Hartree–Fock theory the wavefunction is represented by a series of molecular orbitals).

$$\psi = \phi_1\phi_2\phi_3 \ldots \phi_n \tag{2-4}$$

$$\mathrm{H}(\phi_1\phi_2\phi_3 \ldots \phi_n) = \mathrm{E}(\phi_1\phi_2\phi_3 \ldots \phi_n) \tag{2-5}$$

where ϕ_i is the ith molecular orbital.

Having broken down the electronic wavefunction into a series of molecular orbitals, Hartree–Fock theory then makes use of so-called “basis functions”. These functions are a series of one-electron mathematical representations that are localised on individual atoms, which can be thought of as representing the atomic orbitals.

The more basis functions included in the molecular orbital calculation, the more accurate the final representation. However as might be expected, this results in an increase in computational time. Both semi-empirical and density functional methods make use of basis functions to represent atomic orbitals (so-called basis sets). It is then possible to calculate the ground state electronic structure by making use of a mathematical procedure known as the variational principle.

The significant drawback within the Hartree–Fock formalisation is the incomplete treatment of so-called exchange–correlation effects when evaluating the energy of the wavefunction. These effects relate to the interactions between pairs of electrons with the same spin (exchange) and pairs of electrons with opposing spins (correlation). Thus, when evaluating the energy of the wavefunction within Hartree–Fock theory correlation effects are completely neglected, leading to an underestimation of the true energy of a given electronic state.

2.4. SEMI-EMPIRICAL METHODS: AM1 AND RM1

Initial usage of Hartree–Fock theory was limited to very small systems for which the iterative process of locating the lowest energy wavefunction was amenable to early computers. Such limitations led to the development of so-called semi-empirical quantum mechanics methods, with the aim of allowing chemically meaningful systems to be investigated. As would be expected, one of the most time-consuming steps in the Hartree–Fock optimisation procedure is the manipulation of the mathematical representations of the molecular orbitals. In contrast, the semi-empirical Austin Method 1 (AM1) deals only with the valence electrons, thus significantly reducing the complexity and hence time of one of the most computationally expensive steps [7]. Additional computational savings are made in the use of parameterised functions for some of the terms in the *Hamiltonian*. These functions are developed using experimental data such as heats of formation, the aim being that the functions are optimised (often manually) until the resulting calculations can reproduce a series of experimental molecular properties. Such approximations obviously reduce the accuracy of the AM1 method (and semi-empirical methods in general), this being the major limitation. Semi-empirical methods generally perform well for calculations upon molecular systems for which the basis functions were optimised (for example, heats of formations are frequently well reproduced). However (and as might be expected) calculations into systems for which no experimental data existed (or was used) in the parameterisation procedure often perform poorly. The significant advantage of the computational efficiency resulting from the various approximations in the AM1 methodology is that it allows for a high number of chemicals to be investigated in a reasonable timeframe, and for calculations upon large molecular systems.

A recent re-parameterisation of the AM1 model has led to the development of the Recife Model 1 (RM1) semi-empirical method [8]. This methodology has been suggested to be a significant improvement over the original AM1 model as additional parameterisation data were included in its development. These data came from high-level density functional calculations allowing for a better definition of common geometrical variables poorly defined by existing experimental data. In addition, the description of the electron repulsion portion of the wavefunction was also improved.

2.5. AB INITIO: DENSITY FUNCTIONAL THEORY

Density functional theory (DFT) is a closely related methodology to Hartree–Fock theory in that it attempts to provide a solution to the electronic state of a molecule directly from the electron density. One can view the methodologies as essentially analogous, for the purpose of this discussion, in terms of using basis functions for orbitals and in the use of the variational principle to locate the lowest energy wavefunction. However, the major difference is the inclusion of terms to account for both exchange and correlation when evaluating the energy of the wavefunction, resulting in a significantly improved description of the electronic structure. Differing functionals (for example, B3LYP) use differing mathematical approximations to describe the *Hamiltonian* and thus evaluate the energy of a given wavefunction. The discussion of how such functionals are calculated and thus their relative strengths and weaknesses is well outside the scope of this chapter. It is important only to realise that DFT (whatever the chosen functional) is a more complete description of the electronic structure than that offered from Hartree–Fock theory and is significantly more complete than semi-empirical methods. However, as would be expected by the inclusion of more complex mathematics, it is also the most time consuming. A more complete discussion of DFT and functionals can be found in several texts [6, 9].

2.6. QSAR FOR NON-REACTIVE MECHANISMS OF ACUTE (AQUATIC) TOXICITY

The importance of quantum mechanics electronic parameters in toxicology becomes apparent when one examines the descriptors required to model the polar narcotic chemicals. Such relationships frequently involve the use of the energy of the lowest unoccupied molecular orbital (E_{LUMO}) to account for the increased electronegativity of these chemicals (compared to those that cause baseline narcosis). The most commonly used level of theory is the AM1 Hamiltonian. The descriptor E_{LUMO} in combination with the logarithm of the octanol–water partition coefficient (log P) (or other descriptor describing hydrophobicity) leads to excellent statistical relationships. Such two-parameter QSARs are commonly referred to as response-surface models [10]. Cronin and Schultz investigated the acute toxicity of 166 phenols to the ciliated protozoan *Tetrahymena pyriformis*, for which potential toxic mechanisms of action had previously been assigned [11]. Of the 166 chemicals in the training data, 120 were assigned as acting via polar narcosis, and response-surface analysis of the toxicity for these chemicals (IGC_{50}) produced the following relationship:

$$\begin{aligned}&\text{Log }(IGC_{50})^{-1} = 0.67(0.02)\log P - 0.67(0.06)E_{LUMO} - 1.123(0.13)\\ &n = 120,\ r^2 = 0.90,\ r^2_{cv} = 0.89,\ s = 0.26,\ F = 523\end{aligned} \tag{2-6}$$