Chapter 2
How Far Chemistry and Toxicology are Computational Sciences?

Giuseppina Gini

Abstract In this chapter we describe the basis of computational chemistry and discuss how computational methods have been extended to biology, and toxicology in particular. Since about 20 years, chemical experimentation is more and more replaced by modelling and virtual experimentation. Computer modelling of biological properties is still a debated topic. However, the need of safety assessment of chemicals is pushing toxicology towards computer modelling. The term *in silico* discovery is now applied to chemical design, to computational toxicology, and to drug discovery. We discuss how the experimental practice in biological science is moving more and more towards computer modelling and simulation. Such virtual experiments confirm hypotheses, provide data for regulation, and help in designing new chemicals.

Keywords *In silico* experiments · Chemoinformatics · QSAR · Toxicology

2.1 Introduction

"All science is computer science". When a New York Times article in 2001 used this title, the general public was aware that the introduction of computers has changed the way that experimental sciences develop. A first example of the historical connection between chemistry and computer science is the development of fragment codes, usually called fingerprints, used to filter large data sets of molecules for the presence or absence of a particular sub-structure. For that M. Lynch, J. Ziv, and A. Lempel produced the Ziv-Lempel algorithms, which are the basis of the wide used algorithms for data compression [1].

Giuseppina Gini (✉)
Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy
e-mail: giuseppina.gini@polimi.it

F. Amigoni and V. Schiaffonati (eds.), *Methods and Experimental Techniques in Computer Engineering*, PoliMI SpringerBriefs,
DOI: 10.1007/978-3-319-00272-9_2,

Chemistry and physics are among the best examples of such a new way of making science. A new discipline, *chemoinformatics* has been in existence for the past two decades [2, 3]. Many of the activities performed in chemoinformatics are information retrieval [4], aimed at searching for new molecules of interest when a single molecule has been identified as being relevant. However chemoinformatics is more than "chemical information"; it requires strong algorithmic development. Simulations now are widely used in chemistry, material sciences, engineering, and process control, to name a few fields. What about life sciences?

Starting from chemistry and reviewing the role of mathematics and then algorithms in chemical research, in this chapter we will move to some part of biological experimentation. In particular we will see how animal experiments, aimed at providing a standardized result about a biological property, as bioavailability or even death, can be mimicked by new *in silico* methods. Our emphasis is on toxicology and QSAR (Quantitative Structure Activity Relationships) methods [5–7].

The aim of this chapter is to briefly review how and why life sciences are moving more and more towards modelling and simulation. Here computing is a tool for scientific disciplines, with the interesting consideration that many basic computing tools (as graph representation, simulators, efficient data hashing) had their origin in the arising needs of reasoning with atoms and molecules.

In Sect. 2.2 we introduce the role of computer-based models and algorithms in chemistry. In Sects. 2.3 and 2.4 we see how biological modelling and toxicology has evolved from the first animal models to the *in silico* models. Section 2.5 presents the problems related to human toxicology and drug development. Section 2.6 discusses environmental toxicology and risk assessment. Section 2.7 points out the pitfalls and new trends of *in silico* methods, while Sect. 2.8 presents some conclusions.

2.2 Chemistry

The word 'chemistry' is from the Greek khymeia ($\chi\upsilon\mu\epsilon\iota\alpha$) "to fuse together". For centuries it has been seen as a kind of magic, a way to transform elements. All the activities connected to chemistry, from cooking to metallurgy, to medicinal remedia have been dominated by empirical rules. In his 1830 book *Course of Positive Philosophy* Auguste Compte wrote:

> Any attempt to use mathematical methods in the study of chemical problems should be considered not rationale and contrary to the spirit of chemistry. In case mathematical analysis would assume a prominent role in chemistry—an aberration that fortunately is almost impossible—that would bring to a rapid degeneration of this science.

2.2.1 Atoms and Elements

The basic components of matter have been investigated in the Greek philosophy. Democritos postulated the existence of atoms. "There are atoms and space.

Everything else is opinion". Aristotle declared the existence of four elements: fire, air, water, and earth, stating that all matter is made up of these four elements.

In much more recent times, atoms and elements have been scientifically defined. In 1789 Antoine Lavoisier published a list of 33 chemical elements grouped into gases, metals, non metals, earths. In 1803 John Dalton in his *Atomic Theory* stated that all matter is composed of atoms, which are small and indivisible.

Theoretically-based models of atoms were defined much later. Rutherford in 1909 adapted the solar system model: the atom is mostly empty space with a dense positively charged nucleus surrounded by negative electrons. In 1913 Bohr proposed that electrons travelled in circular orbits. Finally in the 1920's the electron cloud model was defined by Schrödinger; in it an atom consists of a dense nucleus composed of protons and neutrons surrounded by electrons. The most probable locations of the electron predicted by Schrödinger's equation coincide with the locations specified in Bohr's model.

About the elements, their accepted definition was the periodic table, published by Mendeleev in 1869. It organized the elements into a table, listing them in order of atomic weight, and starting a new row when the characteristics of the elements began to repeat [8]. It happened much before the development of theories of atomic structure; after that, it became apparent that Mendeleev had listed the elements in order of increasing atomic number.

So in the first 30 years of the XX century chemistry (and physics) has been refunded. Quantum theory and material physics have changed the way that materials are studied. Dirac, just one century after Compte, wrote:

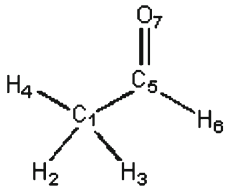
> The physical rules necessary for a mathematical theory of the whole chemistry and of part of physics are known, and the only difficulty is that the application of those rules generates equations too complicated to be solved.

2.2.2 Computer-Based Representation for Molecules

A molecule is an electrically neutral group of atoms held together by covalent bonds. A molecule is represented as atoms joined by semi-rigid bonds.

The graph theory, established back in XVIII century, initially evolved through chemistry; the name 'graph' indeed derives from its use in drawing molecules. The valence model naturally transforms a molecule into a graph, where the atoms are represented as vertices and the bonds as edges. The edges are assigned weights according to the kind of bond. Today hydrogens are not represented in the graph since they are assumed to fill the unused valences [9]. This representation is called 2D chemical structure.

A common representation of the graph is the adjacency matrix, a square matrix with dimension N equal to the number of atoms. Each position (i, j) specifies the absence (0 value) or the presence of a bond connecting the atoms i and j, filled with 1, 2, 3 to indicate simple, double, or triple bond, 4 for amide bond, 5 for aromatic bond. An example of a matrix representation is in Fig. 2.1.



	C1	*H2*	*H3*	*H4*	*C5*	*H6*	*O7*
C1	0	1	1	1	1	0	0
H2	1	0	0	0	0	0	0
H3	1	0	0	0	0	0	0
H4	1	0	0	0	0	0	0
C5	1	0	0	0	0	1	2
H6	0	0	0	0	1	0	0
O7	0	0	0	0	2	0	0

Fig. 2.1 Graph representation and adjacency matrix of methil-aldehaide

To write the molecule as a text string, Simplified Molecular Input Line Entry Specification (SMILES) [10] is popular. SMILES is a context free language expressing the graph visit in a depth first style:

- Atoms are represented by their atomic symbols in upper-case.
- Bonds are Single, implicit; Double, "="; or Triple, "#".
- Branches are placed between round parentheses.
- Cycles are represented breaking one bond in each ring.

The SMILES notation suffers the lack of a unique representation, since a molecule can be encoded beginning anywhere. Therefore canonical SMILES [11] was proposed.

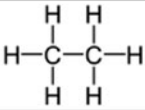
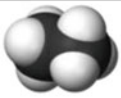
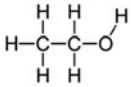
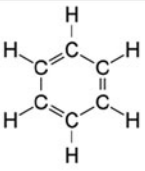
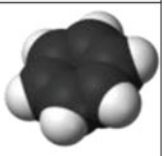
SMILES	*NAME*	*FORMULA*	*GRAPH*	*3D*
CC	Ethane	CH3CH3		
C=O O=C	Formaldehyde	CH2O		
CCO OCC C(C)O C(O)C	Ethanol	CH3CH2OH		
c1ccccc1	Benzene	C6H6		

Fig. 2.2 Some 2D and 3D representations of molecules

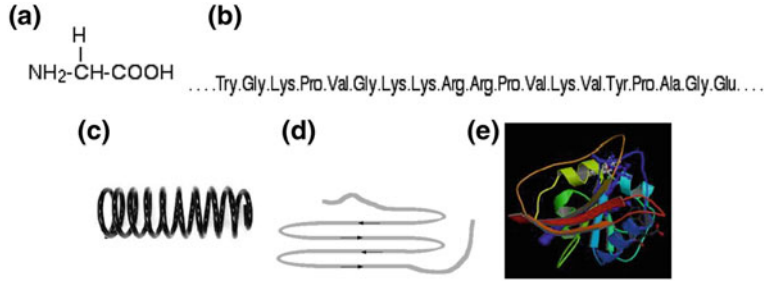


Fig. 2.3 **a** Glycine, an aminoacid. **b** A part of the primary structure of a protein. **c** The secondary structure alpha helixes. **d** The secondary structure beta sheets. **e** A tertiary structure from http://www.rcsb.org/pdb/

What about the real shape of molecules? They are 3D objects, and as such they should be represented. Figure 2.2 shows some examples of different representations. For biological molecules the representations should consider the real 3D shape in space. Figure 2.3 shows the aminoacid glycine, its primary structure (the sequence of aminoacids indicated by the first 3 letters of the name), the secondary structure, that is the organization of regions in alpha helixes and beta sheets, and 3D folding.

The point about defining the 3D shape of a molecule will take us to the basic methods of computational chemistry.

2.2.3 Computational Chemistry

Computational chemistry is a branch of chemistry that uses computers to assist in solving chemical problems, studying the electronic structure of matter. It uses the results of theoretical chemistry, incorporated into computer programs, to calculate chemical structures and properties. The methods cover both static and dynamic situations through accurate methods, *ab initio*, and less accurate methods, *semi-empirical*. *Ab initio* methods are based entirely on theory from first principles of quantum chemistry. Semi-empirical methods employ also experimental results from related molecules to approximate some parameters. *Ab initio* are computationally expensive, so the size of the molecules that can be modelled is limited to a few hundred atoms. For big molecules it is necessary to introduce empirical parameters. When the molecular system is even bigger, the simulation is statistically based.

Computational chemistry is a way to move away from the traditional approach of solving scientific problems using only direct experimentation, but it does not remove experimentation. Experiments produce new data; the role of theory is to situate all the new data into a framework, based on mathematical rules.

It is now possible to simulate with the computer an experiment before running it. It all happened in about 50 years, from the first *ab initio* calculations done in 1956

Table 2.1 The number of stars and molecules in real and virtual spaces

	Stars	Small molecules
Existent	10^{22}	10^{7} currently in CAS registry
Virtual	0	10^{80}

at MIT to the Nobel prize for Chemistry, assigned in 1998 to J. Pople and W. Kohn for computational chemistry.

Through computer simulations a very large number of different but structurally related molecules are created using the methods of combinatorial chemistry, able to create a "library" of thousand of different but related compounds [12].

Today in the CAS (Chemical Abstracts Service) registry there are more than 71 millions of unique numerical identifiers assigned to every chemical described in the open scientific literature; the expected dimension of the chemical space is tens orders of magnitude bigger. Table 2.1 shows a comparison between the number of potential chemical compounds and the number of stars.

Advances in robotics have led to an industrial approach to combinatorial synthesis, enabling companies to routinely produce over 100,000 new and unique compounds per year. Today new products are designed and checked on the computer before they are synthesized in the laboratory. Those methods are part of the *in silico* methods that will be described in the following.

2.3 Biological Models and Toxicology

In antiquity, the physiology research was carried out on animals. Although these observations and their interpretations were frequently erroneous, they established a discipline. The explosion of molecular biology in the second half of the XX century increased the importance of in vivo models [13].

In biomedical research the investigations may be classified as observational or experimental.

- Observational studies are carried out when the variables influencing the outcomes of the phenomena under study cannot be controlled directly. These variables are observed and an attempt is made to determine the correlations between them.
- Experimental studies require to directly control selected variables and to measure the effects of these variables on some outcome. The results of experimental studies tend to be more robust compared to observational studies.

Experimental studies may be carried out on in vitro biological systems such as cells, microorganisms, tissue slice preparations. Experiments using in vitro systems are useful where the screening of large number of potential therapeutic candidates may be necessary, or in making fast tests for possible pollutants. In vitro systems are, however, non-physiological and have important limitations since their biological complexity is much lower than that of most of the animal species including

humans. While data from experiments carried out in vitro can establish mechanisms, in vivo biological systems, using whole organisms, are required to study how such mechanisms behave in real conditions.

Models are meant to mimic the subject under study. Biomedical research models can also be either analogues or homologues.

- Analogous models relate one structure or process to another. An analogue animal model can explain some of the mechanisms of humans.
- Homologous models reflect the genetic sequence of the organism under study.

Often animal models are both analogues and homologues. Of course the ideal model for a human is a human, but animal models are so far the best approximation.

All models have their limitations; their prediction can be poor, and their transferability to the real phenomena they model can be unsatisfactory. So extrapolating data from animal models to the environment or to human health depends on the degree to which the animal model is an appropriate reflection of the condition under investigation. These limitations are, however, an intrinsic part of all modelling approaches. Most of the questions about animal models are ethical more than scientific. In public health the use of animal models is imposed by different regulations, and it is unlikely that any health authority will allow the adoption of novel drugs without supporting animal data.

2.3.1 Bioassays for Toxicity

Toxicity is the degree to which a substance can damage an organism. Paracelsus (1493–1541) wrote: "All things are poison and nothing is without poison; only the dose makes a thing not a poison." The relationship between dose and its effects on the exposed organism is of high significance in toxicology.

Animals have been used for assessing toxicity in pioneering experiments since more than one century. In more recent times the process of using animal testing to asses toxicity has been defined in the following way:

- Toxicity can be measured by its effects on the target.
- Because individuals have different levels of response to the same dose of a toxin, a population-level measure of toxicity is often used which relates the probabilities of an outcome for a given individual in a population. Example is median lethal dose LD_{50}: the dose that causes the death of 50% of the population after a specified test duration. Examples of some doses are in Table 2.2.
- When the dose is individuated, "safety factors" are defined. For example, if a dose is safe for a rat, one might assume that one tenth that dose would be safe for a human, allowing a safety factor of 10.

This process is based on assumptions that usually are very crude. It presents many open issues. For instance it is more difficult to determine the toxicity of chemical

Table 2.2 Examples of LD_{50} for common chemicals

Chemical	Target	LD_{50}
Water	Rat, oral	90000 mg/kg
Sucrose	Rat, oral	29700 mg/kg
Table salt $NaCl$	Rat, oral	3000 mg/kg
Paracetamol	Rat, oral	1944 mg/kg
Caffeine	Rat, oral	192 mg/kg
Nicotine	Rat, oral	50 mg/kg
Dioxin	Rat, oral	20 μg/kg

mixtures (gasoline, cigarette smoke, waste) since the percentages of the chemicals can vary, and the combination of the effects is not exactly a summation.

Perhaps the most common continuous measure of biological activity is the IC_{50} (inhibitory concentration), which measures the concentration of a compound necessary to induce a 50% inhibition of the biological activity under investigation.

The dose-response curve describes the change in effect on an organism caused by different levels of doses after a certain exposure time. A dose-response curve is a x-y graph relating the magnitude of a stressor to the response of the organism.

- The measured dose (usually milligrams per kilogram of body-weight for oral exposures) is plotted on the x axis and the response is plotted on the y axis.
- The response may be a physiological or biochemical response.
- LD_{50} is used in human toxicology; IC_{50}, inhibition concentration, and its dual EC_{50}, effect concentration, are fundamental to pharmacology.

Usually the logarithm of the dose is plotted on the x axis, so the curve is typically sigmoidal, with the steepest portion in the middle. In Fig. 2.4 we see an example of the dose response curve for LD_{50}.

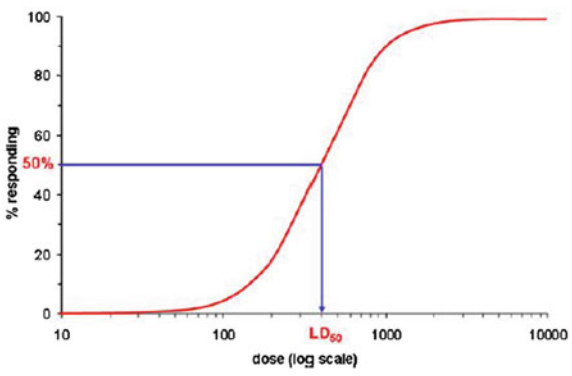


Fig. 2.4 A curve for $\log(LD_{50})$

2.3.2 Animal Testing Versus in Vitro and in Silico Testing

Worldwide it is estimated that the number of vertebrate animals used annually in testing ranges from the tens of millions to more than 100 million. Animal testing was introduced for drugs, in particular in response to many deaths from the Sulfanilamide in 1937. In the 1960s, in reaction to the Thalidomide tragedy, further laws were passed requiring safety testing also on pregnant animals.

Those in vivo models give doses for some species, and are used to extrapolate data to human health or to the environment. As we said above, the extrapolation of data from species to species is not obvious. For instance, the lethal doses for rats and for mice are sometimes very different. In vitro toxicity testing is the scientific analysis of the effects of a chemical on cultured bacteria or mammalian cells. It is known that their results poorly correlate with the results of in vivo.

How to construct a model that relates a chemical structure to its effect was investigated even before computers were available. The term *in silico* covers the current methods devoted to this end.

2.4 *In Silico* Methods

The term '*in silico*' refers to the fact that computers are used, and computers have silicon in their hardware. The most known *in silico* methods are the QSAR (Quantitative Structure Activity Relationships) methods, derived from the suggestion made in 1868 by A. Crum Brown and T. Fraser that a mathematical relationship can be defined between the physiological action of a molecule and its chemical constitution [14].

2.4.1 QSAR

Given quantitative data, we can build a QSAR model that seeks to correlate our particular response with molecular descriptors that have been computed or even measured from the molecules themselves [15]. QSAR methods were first pioneered by Corwin Hansch in the 1940s, who analyzed congeneric series of compound and formulated the QSAR equation:

$$\log(1/C) = a \cdot \log P + b \cdot Hs + c \cdot Es + \text{const}$$

where C is effect concentration, $\log P$ is octanol-water partition coefficient, Hs is Hammett substituent constant (electronic), Es is Taft's substituent constant, and a, b, and c are parameters. The octanol-water partition coefficient $\log P$ is the ratio of concentrations of a compound in the two phases of a mixture of two immiscible solvents at equilibrium. It is a measure of the difference in solubility of the compound in these

two solvents. With high octanol-water partition coefficient the chemical substance is hydrophobic and preferentially distributed to hydrophobic compartments such as cell membrane, while hydrophilic substances are found in hydrophilic compartments such as blood serum [16].

Sometimes the QSAR methods take more specific names as: QSPR (Quantitative Structure Property Relationship) or QSTR (Quantitative Structure Toxicity Relationship). They all correlate a dependent variable (the effect) with a set of independent variables (usually calculated properties, or descriptors).

2.4.1.1 Molecular Descriptors

The generation of informative data from molecular structures is of high importance in chemoinformatics. There are many possible approaches to calculating molecular descriptors [17], that represent local or global salient characteristics of the molecule. Different classes are:

- Constitutional descriptors, depending on the number and type of atoms, bonds, and functional groups.
- Geometrical descriptors, which give molecular surface area and volume, moments of inertia, shadow area projections, and gravitational indices.
- Topological indices, based on the topology of molecular graph [9]. Examples are the Wiener index (the sum of the number of bonds between all nodes) and the Randic index (the branching of a molecule).
- Physicochemical properties attempt to estimate the physical properties of molecules. Example are molecular weight, hydrogen bond acceptors, hydrogen bond donors, and partition coefficients, as log P.
- Electrostatic descriptors, such as partial atomic charges, depending on the possibility to form hydrogen bonds.
- Quantum chemical descriptors, related to the molecular orbitals.
- Fingerprints. Since subgraph isomorphism (substructure searching) in large molecular databases is time consuming, substructure screening was developed as a rapid method of filtering out those molecules that definitely do not contain the substructure of interest. The method uses fingerprints, binary strings encoding a molecule, where the 1 or 0 in a position means whether the substructure of this position in the dictionary is present or not.

2.4.1.2 Model Construction

After selecting the relevant descriptors, whatever method is chosen to develop predictive models, it is important to take heed of the model quality statistics and ensure a correct modelling methodology is used, such as testing the model against an external and unseen test set to ensure it is not overfitting to the training set. Model extrapolation is another concern that frequently occurs when models are applied outside

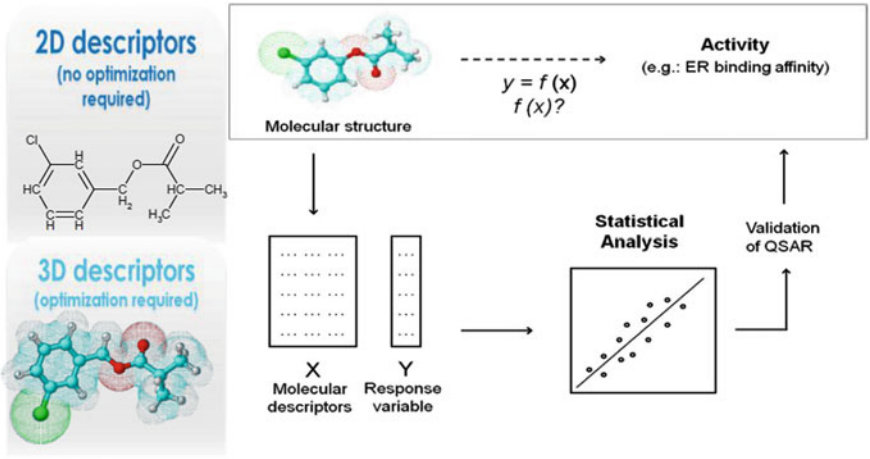


Fig. 2.5 The steps in constructing a QSAR model

the space from which the models were generated. Numerous model statistics are available that can indicate if new data points can be predicted by the model [18].

Two types of supervised learning methods are applied widely: classification and regression. Classification methods assign new objects, in our case molecules, to two or more classes—either biologically active or inactive. Regression methods attempt to use continuous data, such as a measured biological response variable, to correlate molecules with that data so as to predict a continuous numeric value for unseen molecules using the generated model [6]. Figure 2.5 illustrates the flow chart of the activities in the QSAR construction.

There is generally a trade-off between prediction quality and interpretation quality. Interpretable models are generally desired in situations where the model is expected to provide information about the problem domain. However, these models tend to suffer in terms of prediction quality as they become more interpretable. The reverse is true with predictive models, in that their interpretation suffers as they become more predictive. Models that are highly predictive tend to use molecular descriptors that are not readily interpretable by the chemist. However, predictive models are generally not intended to provide transparency, but predictions that are more reliable and can therefore be used as high-throughput models. If interpretability is of concern, other methods are available, more or less as a kind of expert systems, or SAR [19, 20].

2.4.2 SAR

SAR (Structure-Activity Relationships) typically makes use of rules created by experts to produce models that relate subgroups of the molecule atoms to a biological

property. The SAR approach consists in detecting particular structural fragments of molecule already known to be responsible for the toxic property under investigation. Structural rules usually can be explained in terms of reactivity or activation of biological pathways.

In the mutagenicity/carcinogenicity domain, the key contribution in the definition of such toxicophores came from [21], who compiled a list of 19 Structural Alerts (SAs) for DNA reactivity. Practically SAs are rules which state the condition of mutagenicity by the presence of peculiar chemical substructures. SAs have a sort of mechanistic interpretation; however, their presence alone is not a definitive method to prove the property under investigation, since the substituents present could change the classification.

To conclude, if the main aim of (Q)SAR is simply prediction, the attention should be focused on the quality of the model, and not on its interpretation. Regarding the interpretability of QSAR models [14] states:

> The need for interpretability depends on the application, since a validated mathematical model relating a target property to chemical features may, in some cases, be all that is necessary, though it is obviously desirable to attempt some explanation of the mechanism in chemical terms, but it is often not necessary, *per se*.

It is worth mentioning that the modern QSAR paradigm extends the initial one proposed by Hansch in many directions: from congeneric to heterogeneous compounds, from single to multiple modes of actions, from linear regression to non-linear models, from simple to complex endpoints.

2.5 Human Toxicology and Drug Design

The discovery of new medical treatments is time consuming, and incredibly expensive. Drug discovery is the area in which chemoinformatics is routinely used. Drug discovery starts from the identification of a biological target that is screened against many thousands of molecules to identify the hits (molecules that are active). A number of those hits will produce a lead, a fragment that appears responsible for the wanted effects.

A lead has some desirable biological activity [22]: it is not extremely polar, does not contain toxic or reactive functional groups, has a small molecular weight and a low log P, has a series of congeners to allow structural modification. The leads are then combined with other elements to obtain the candidate drugs, in a process that requires multiple optimizations: reduced size, reduced toxicity, bioavailability.

Chemical space is the term given to the space that contains all of the theoretically possible molecules. However, when considering drug-like chemicals, the space becomes bounded according to known conditions such as the Lipinski rule-of-five [23] where a set of empirically derived rules is used to define molecules that are more likely to be orally available as drugs. This drug-like chemistry space is estimated to contain at least 10^{12} molecules [16], a very huge number.

To be able to explore this vast chemical space, it is necessary to deploy computer systems. There are two general approaches to drug design: one optimizes the molecule directly so to satisfy the binding for a particular target and the other one optimizes the molecule for a desired biological activity. The former is 3D based, while the latter uses only topological information. Often the two methods are integrated: using combinatorial chemistry and then QSAR, a small set of molecules with a high desired activity are selected. Then their shape is studied to see how they can fit the constraints of the binding site. In this last part, a new approach that simulates the binding using methods derived from robotics has also been developed [24].

Toxicity testing typically involves studying adverse health outcomes in animals administered with doses of drugs or toxicants, with subsequent extrapolation to expected human responses. The system is expensive, time-consuming, low-throughput, and often provides results of limited predictive value for human health. The toxicity testing methods are largely the same for industrial chemicals, pesticides and drugs, and have led to a backlog of more than 80,000 chemicals to which humans are exposed but whose potential toxicity remains largely unknown. This potential risk has urged national and international organizations in making a plan for assessing the toxicity of those chemicals.

In the USA, EPA (Environmental Protection Agency) routinely uses predictive QSAR based on existent animal testing to authorize new chemicals. Recently in the USA, a new toxicity testing plan, "Human Toxome Project", has been launched which will make extensive experimentation using predictive, high-throughput cell-based assays of human organs to evaluate perturbations in key pathways of toxicity. There is no consensus about this concept of "toxicity pathway" that in the opinion of many should be instead "disruption of biological pathways". The target of the project is to gain more information directly from human data, so to check in a future, with specific experiments, the most important pathways.

In the European Union, the REACH legislation for industrial chemicals has been introduced together with specific regulations for cosmetics, pesticide, food additives.

2.6 Environmental Toxicology

One of the most known episodes that draw the attention to the environmental pollution happened in the 1950s in Japan. Outbreaks of methylmercury poisoning occurred in several places in Japan due to industrial discharges of mercury into rivers and coastal waters. In Minamata bay alone, more than 600 people died. After that, mercury has been recognized as a pollutant and its presence in food monitored. In December 1970 a chemistry professor at New York University bought canned tuna and found a mercury dose 20 times higher than the limits of the FDA (Food and Drug Administration). It confirmed that the mercury poisoning was much more diffused, mainly in fish. Fishes have a natural tendency to concentrate mercury, especially the ones that are high on the food chain.

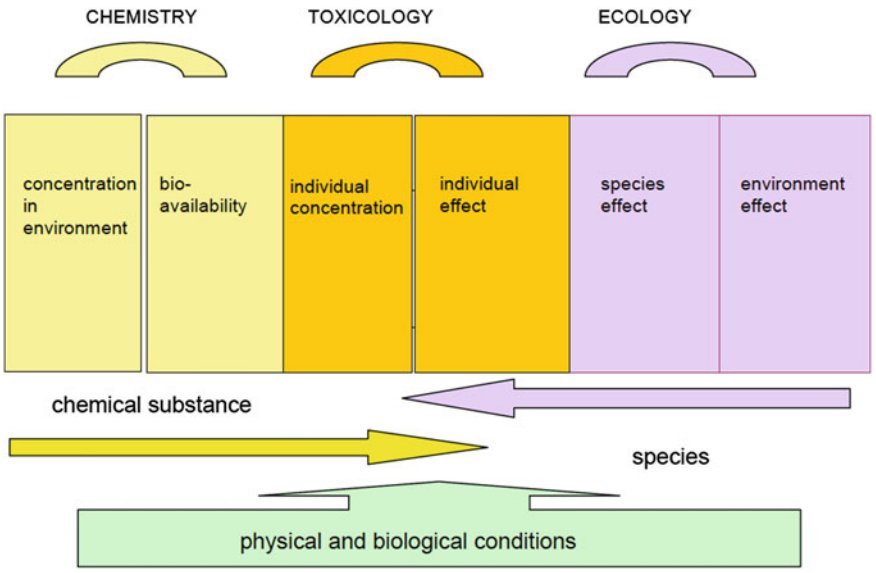


Fig. 2.6 Chemistry, toxicology, and ecology

The recognition that environment pollutants can harm humans in various ways, brought in the 1960s to the development of *environmental toxicology*. Environmental toxicology is concerned with the study of the harmful effects of various chemical, biological, and physical agents on living organisms. The steps of ecotoxicity studies are:

- entry, distribution, and fate of pollutants within the environment;
- entry and fate of pollutants in living organisms of the ecosystem;
- harmful effects on the constituents of ecosystems (which include humans).

Figure 2.6 indicates the main relationships between chemistry, toxicology, and ecology.

To address environmental toxicology some tests are used. Ideally a pollutant has to be tested on invertebrates and vertebrates, in air and in water species. Acute toxicity is usually the property studied on all those species, and *in silico* tools are applicable.

The regulations for human and environmental protection are out of the scope of this chapter. We only indicate that different regulations apply for air pollutants, industrial products (e.g., REACH), food, drinking water, cosmetics and detergents, pesticides, and drugs. There is only limited international agreement on the regulations and doses.

2.7 Problems

Of the many open problems in assessing toxicology using *in silico* models we discuss about a few points. The first one is the causal or mechanistic value of the QSAR equation. The QSAR for LD_{50}, for instance, does not have a simple interpretation in term of logic sentences. This will lead to the problem of mechanistic interpretation. Another point is about ethical issues. Is it really needed to make experiments on animals?

2.7.1 Logical and Probabilistic Knowledge

At the origin of any model there is a core hypothesis. In the case of QSAR for toxicology, we assume that the toxicity is related to the chemical structure:

$$Tox = f(Chem)$$

where Tox is toxicity, $f(\cdot)$ is a mathematical function and $Chem$ represents the chemical compound.

However, we have to better understand the implications and limits of this equation.

- From the classical work of toxicologists we know that the dose makes a compound toxic. Toxicologists have defined a kind of standardized effect, such as the dose which produces a given effect (e.g., death in 50% of the cells). For instance, chemical A will give the same toxic effect of chemical B using a dose double of that of chemical B; what changes is the dose, not the effect. Thus we can compare different chemicals only on the basis of their chemical nature, because we have defined a standard effect.
- We understand that different chemicals require different toxic doses to produce the same effect.
- The toxic effect refers to a cell or organism. Does this have an influence? If we consider LD_{50}, immediately we see that the same dose on 50% animals produces an effect which is opposite to that on the other 50%, because half die and half stay alive. The toxic effect is also dependent on the organism.
- From previous point we see that the basic QSAR equation, which appeared as a deterministic one, can be better considered from a stochastic point of view.
- It is well known that the chemical effect is mediated by processes, which can be, in many cases, unknown. Thus, Tox can be better described as

$$Tox = tox_1 + tox_2 + \cdots + tox_n$$

each of these factors tox_i describes a different process related to toxicity.

- The chemical part is actually much more complicated than a chemical formula; for instance, biochemical processes can transform the original compound in new compounds, more or less toxic than the original one.

The debate between models and their use can be taken at large, using the recent Internet posts of P. Norvig about N. Chomsky. Norvig [25] recalls a paper by L. Breiman [26] that presents two cultures: the data modelling culture, saying that nature can be described as a black box that has a relatively simple underlying model which maps input variables to output variables, and the algorithmic modelling culture, considering that nature cannot be described by a simple model, but proposing complex statistical and probabilistic models. People working in the algorithmic approach, as people using Artificial Intelligence (AI) techniques, aim at finding functions that map from input to output variables, but without expectation that the form of the function reflects the true underlying nature. None of the approaches can describe why something happens. The first since causality cannot be expressed in a purely statistical approach, the second since the underlining hypothesis of (the weak approach to) AI implies that the model simply emulates the effects of reality.

Breiman explains his objections to the data modelling culture. Data modelling makes conclusions about the model, not about reality. There is no way to uniquely model the true form of nature's function from pairs of inputs/outputs. What this model can do is to generalize to new data, not to give us the true form of a function. Whether this true form exist or not, it is not the task of a modelling and simulation method, but a matter of the right generalization process, something where humans are still superior to machines.

2.7.2 Mechanism or Causality

Hume argued that causality cannot be perceived and instead we can only perceive correlation. And indeed the basic biological experiments aim at finding a correlation (positive or negative) between some features and effects.

Biologists then want to understand why the effect can be explained in terms of metabolism, transformation substances, and so on. This is often called with the vague terms of "mode of action", or "mechanistic interpretation". Vagueness derives from the fact that there is no unique definition of mode of action: in some cases this is an observed behaviour as narcosis, in other cases it is a supposed chemical transformation. This is more complex than considering the organic chemical transformations since they happen in an organism where different biological pathways are usually supposed.

Inferring causality from data through Bayesian networks is today an active area of research and hopefully some answers could be found using those tools [27].

2.7.3 Ethical Issues

Toxicity testing typically involves studying adverse health outcomes in animals subjected to high doses of toxicants with subsequent extrapolation to expected human responses at lower doses. The system is expensive, time-consuming, low-throughput and often provides results of limited predictive value for human health. There are more than 80,000 chemicals to which humans are potentially exposed but whose potential toxicity remains largely unknown. Each year a few hundred new substances are registered. Is it really necessary to test all of them on animals?

The Declaration of Bologna, 1999, called the 3 R (Reduce, Refine, Replace), proposed a manifesto to develop alternative methods that could save millions of animals. In this scenario, the ethical issues are advocated also by authorities that have to protect humans, and that see use of animals as ethical than that of humans.

Generally, the stakeholders, often with competing needs, in the toxicity assessment are:

- Scientists and producers: they want modelling and discovery of properties. In other words, they want to build knowledge and translate it in products.
- Regulators and standardization organizations: they want be convinced by some general rule (mechanism of action). In other words, they want to reduce the risk of erroneous evaluations and be fast in decisions.
- Public, media, and opinion makers; they want to be fully protected against risk.

2.8 Conclusions

Since about 20 years chemical experimentation is more and more replaced by modelling and virtual experimentation. It has even been speculated that the vast majority of the discovery process for novel chemical entities will one day be performed *in silico* rather than in vitro or in vivo.

However *in silico* modelling of biological properties is a debated topic. Alongside classical methods as in vivo and in vitro experiments, the use of computational tools is gaining more and more interest. The usage of predictive QSARs is growing, since they provide fast, reliable, and quite accurate responses. They are candidates as accompaniment or replacement of existing techniques.

Finally, as shown in this chapter, the use of computers in chemistry and life sciences brings better tools to science and an open question: is computing (i.e., algorithms) able to capture and express knowledge about physical systems, and biological phenomena in particular?

References

1. Lynch M (2004) Introduction of computers in chemical structure information systems, or what is not recorded in the annals. In: Proceedings of 2002 conference on the history and heritage of scientific and technological, information systems, pp 137–148
2. Brown N (2009) Chemoinformatics—An introduction for computer scientists. ACM Comput Surv 41(2):8:1–8:38
3. Gasteiger J, Engel T (2003) Chemoinformatics: a textbook. Wiley-VCH, Weinheim
4. Willett P, Barnard J, Downs G (1998) Chemical similarity searching. J Chem Inf Comput Sci 38:983–996
5. Benfenati E, Gini G (1997) Computational predictive programs (expert systems) in toxicology. Toxicology 119:213–225
6. Gini G, Katritzky A (Eds) (1999) Predictive toxicology of chemicals: experiences and impact of Artificial Intelligence tools. In: Proceedings of AAAI spring symposium on predictive, toxicology, SS-99-01
7. Hartung T (2009) Toxicology for the twenty-first century. Nature 460(9):208–212
8. Scerri E (2006) The periodic table: its story and its significance. Oxford University Press, New York
9. Balaban A (1985) Applications of graph theory in chemistry. J Chem Inf Comput Sci 25:334–343
10. Weininger D (1988) SMILES, a chemical language and information system. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 28:31–36
11. Morgan HL (1965) The generation of a unique machine description for chemical structures—A technique developed at chemical abstracts service. J Chem Docum 5:107–113
12. Gillet V, Willet P, Bradshaw J, Green D (1999) Selecting combinatorial libraries to optimize diversity and physical properties. J Chem Inf Comput Sci 39:169–177
13. Chow P, Ng R, Ogden B (eds) (2008) Using animal model in biomedical research. World Scientific Publishing Co, Singapore
14. Livingston D (2000) The characterization of chemical structures using molecular properties. A survey. J Chem Inf Comput Sci 40:195–209
15. Hansch C, Malony P, Fujita T, Muir R (1962) Correlation of biological activity of phenoxyacetic acids with hammett substituent constants with partition coefficents. Nature 194: 178–180
16. Ghose A, Crippen G (1986) Atomic physicochemical parameters for three-dimensional structure directed quantitative structure-activity relationships I. Partition coefficients as a measure of hydrophobicity. J Comp Chem 7:565–577
17. Karelson M (2000) Molecular descriptors in QSAR/QSPR. Wiley-VCH, Weinheim
18. Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning: data mining, inference, and prediction. Springer, New York
19. Ferrari T, Gini G, Golbamaki Bakhtyari N, Benfenati E (2011) Mining structural alerts from SMILES: a new way to derive structure-activity relationships. In: Proceedings of 2011 IEEE CIDM, pp 120–127
20. Gini G, Benfenati E (2007) E-modelling: foundations and cases for applying AI to life sciences. Int J Artif Intell T 16(2):243–268
21. Ashby J (1985) Fundamental SAs to potential carcinogenicity or noncarcinogenicity. Environ Mutagen 7:919–921
22. Jorgensen W (2004) The many roles of computation in drug discovery. Science 303:1813–1818
23. Lipinski C, Lombardo F, Dominy B, Feeney P (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 46:3–26
24. Cortes J, Jaillet L, Simeon T (2007) Molecular disassembly with RRT-like algorithms. In: Proceedings of 2007 IEEE ICRA, pp 3301–3306

25. Norvig P (2012) http://norvig.com/chomsky.html, accessed July 2013
26. Breiman L (2001) Statistical modelling: the two cultures. Stat Sci 16(3):199–231
27. Kalisch M, Mächler M, Colombo D, Maathuis M, Bühlmann P (2012) Causal inference using graphical models with the R package pcalg. J Stat Softw 47(11):1–26