2 Toxicology and prediction

2.1 Toxicology principles: modes of action

Toxicology (from the Greek words *toxicos* and *logos*) is the study of the adverse effects of chemicals on living organisms. It is the study of symptoms, mechanisms, treatments and detection of poisoning, especially the poisoning of people and the environment pollution.

Poisons are substances that can cause damage to organisms, leading to a deterioration of their main vital functions, illness, or death. Virtually, every chemical substance may be harmful or lethal when a sufficient concentration is absorbed by an organism. Paracelsus, sometimes called the father of toxicology, wrote: "All things are poison and nothing is without poison, only the dose permits something not to be poisonous." That is to say, substances often considered toxic can be benign or beneficial in small doses, and conversely an ordinarily benign substance can be deadly if over-consumed.

The toxic dose differs a lot according to the specific chemical that is considered. Some substances induce death by concentration of few micrograms per kilo, while others may be quite toxic even if their concentrations are much higher (some grams per kilo). At the same time it is not easy to quantify the toxicity of a substance because several factors need to be considered to understand and eventually predict the main phenomena in the field of toxicology. Thus several indexes have been proposed to estimate the toxicity of a substance and to make comparison among different chemicals. One of the most used indexes is LD_{50} (Lethal Dose), which is the dose (mg/kg body weight) that is responsible of the death of the 50% of the animals exposed to the different chemical agents.

Substance	Animal, Route	LD_{50}
Vitamin C (ascorbic acid)	rat, oral	11,900 mg/kg
Grain alcohol	young rat, oral	10,600 mg/kg
Table Salt	rat, oral	3,000 mg/kg
THC (main psychoactive substance in Cannabis)	rat, oral	1,270 mg/kg males; 730 mg/kg females
Caffeine	rat, oral	192 mg/kg
Nicotine	rat, oral	50 mg/kg
Strychnine	rat, oral	16 mg/kg
Aflatoxin B1 (from *Aspergillus flavus*)	rat, oral	0.048 mg/kg
Batrachotoxin (from poison dart frog)	human, sub-cutaneous injection	0.002-0.007 mg/kg (estimated)
Polonium 210	human, inhalation	0.00001 mg/kg (estimated)
Botulinum toxin (Botox)	human, oral, injection	0.000001 mg/kg (estimated)

Figure 3: LD_{50} values for different substances

When a chemical agent or one of its metabolites produces a toxic effect, it has to interact with specific sites of the organism and to be present in adequate concentration for a sufficient period of time. As a consequence it is important to known which effects a particular substance may cause, information about its chemical structure, the characteristics of the exposure (administration mode, exposure time and rate) and the features of the organism.

In particular, the toxic actions of all substances are carried out by the alteration of biochemical and physiologic processes of cells. The cellular death is the direct consequence of the damage induced by chemical agents and it may lead to serious effects on the tissue in exam. In fact a lot of toxic responses come from the cellular death and the loss of efficiency of crucial organs, that affect the functionality of the whole organism. However different responses exist, which are not due to the cellular death, but they depend on the unbalancing of the physiological and biochemical processes.

Chemicals influence such processes by different modes of action, that can be synthesized as follows:

- interference in normal ligand-receptor interactions;
- interference in membrane functions;
- interference in cellular energy production;
- binding/influence on biomolecules;
- alteration of calcium homeostasis;
- toxicity due to the death of specific cells;
- nonlethal genetic alteration of somatic cells.

Many substances have toxic effects because they interfere in the normal ligand-receptor interaction. Ligand-receptor interaction can be defined as the interaction between a molecule (usually of an extracellular origin) and a protein on or within a target cell. Receptors are macromolecular tissue components, that a drug or a chemical agent (ligand) interact with to produce a biological effect [12]. This particular kind of bond is reversible and highly selective, that is to say also little changes of the chemical structure may drastically reduce or cancel the binding effect. An example of this mode of action is given by neurotoxic substances, because the functions fulfilled by the central nervous system (CNS) are highly dependent on neurotransmitter-receptor interaction.

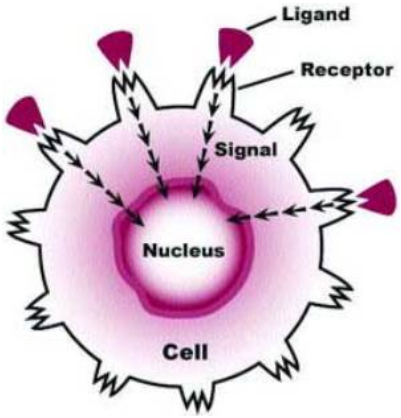

Figure 4: Functional scheme of the ligand-receptor interaction. This interaction is very selective and it influences the inner cell mechanisms.

Other substances may alter the functionality of excitable membranes, of membranes of cellular organelles, of lysosomal membranes and of mitochondrial membranes. The DDT insecticide, for example, interferes in the closing of sodium channel and then alters speed of repolarization, which is specific of excitable membranes.

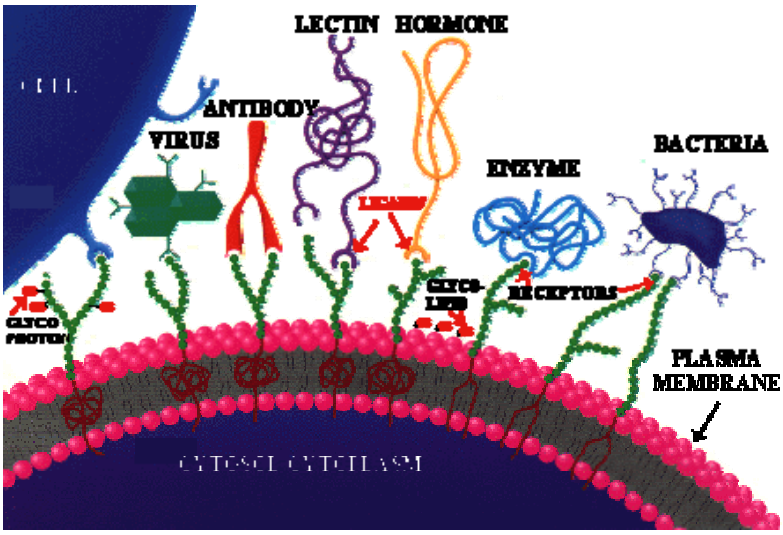

Figura 5: Different types of ligand-receptor interaction: different cell membrane receptors interact with their specific extracellular ligands.

In other cases chemicals become toxic interfering in the normal oxidation process of carbohydrates, that lead to the production of Adenosine Tri-Phosphate (ATP). Some of them interfere in the oxygen supply to tissues, others like cyanide stop the use of oxygen in tissues, because of the affinity to a specific enzyme. The subsequent reduction of ATP stores may cause a lot of consequences, such as the alteration of the functionality of the membranes and of the ion pumps, and the inhibition of protein synthesis, leading to the loss of the functionality and the death of cells.

Some toxicants often bind or at least affect the normal use of biomolecules (proteins, lipids and nucleic acids). One of the most famous example is the carbon monoxide, which binds iron in the hemoglobin with high affinity, causing the reduction of the oxygen supply to tissues. Other agents promote the production of intermediates, especially free radicals, that bind biomolecules (like lipids and nucleic acid), inducing the loss of the functionality of the cell membrane and the alteration of basic intracellular functions, such as protein synthesis.

The inference in the normal processes that are responsible of the intracellular calcium homeostasis seems to play a crucial role in the cellular damage or death, caused by chemical agents. Some membrane abnormalities develop when great amounts of calcium ions accumulate in isolated cells, as a consequence of the toxic effect of some chemicals. Furthermore, calcium is a second messenger in the regulation of numerous functions. For example, the normal organization of cellular cytoskeleton is altered when calcium concentration rises or some endonucleases, activated by calcium ions, may induce the DNA fragmentation and the chromatin condensation, which are important processes involved in cellular apoptosis.

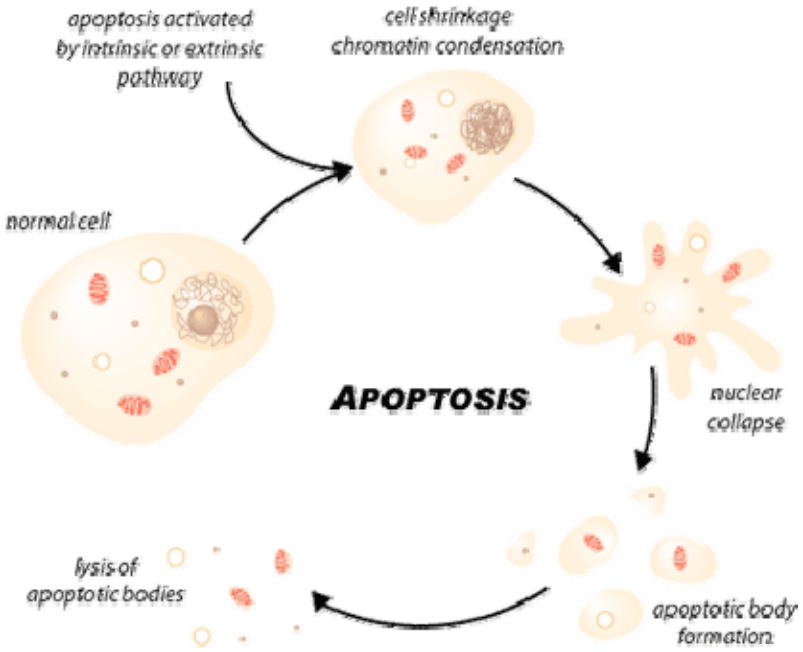

Figure 6: Description of the apoptosis process. Although many pathways and signals lead to apoptosis, there is only one mechanism that actually causes the death of the cell in this process; after the appropriate stimulus has been received by the cell, that cell will undergo the organized degradation of cellular organelles by activated proteolytic caspases, through different phases: cell shrinkage and rounding, chromatin condensation into compact patches against the nuclear envelope, DNA fragmentation, nucleus breaking into several discrete chromatin bodies or nucleosomal units due to the degradation of DNA, irregular buds formation on the cell membrane (know as blebs), cell breaking apart into several vesicles called apoptotic bodies, which are then phagocytosed.

Another kind of toxic effect is the selective death of cells within an organ or a tissue, that sometimes looks like particular pathologic processes. The human embryo, for example, is very sensitive to the action of many toxicants. The administration of a particular drug against nausea (Thalidomide) to pregnant women may cause the loss of some undifferentiated embryonic cells, leading to abortion or to some congenital malformations.

At last a particular group of chemical substances (xenobiotics) binds DNA molecules, inducing cellular death or promoting a complex series of events that may generate cancer. Such substances are called genotoxic carcinogens. Most of the lesions chemically induced to DNA are repaired, but some of them may be missed or incorrectly repaired, causing the introduction of an altered gene, which the new cellular generation will inherit. If the mutation affects a somatic cell, the genetic lesion will not be transmitted to the future generation, but it would generate a cancer. Genotoxic substances seem to be able to induce cancer, since they activate some proto-oncogenes, whose expression is strictly controlled in normal cells. However the induction of cancer is a process which depends on several factors, since also substances that are not genotoxic may increase the incidence of the pathology, perhaps by some mechanisms which are different from the DNA damage (cancer promoters).

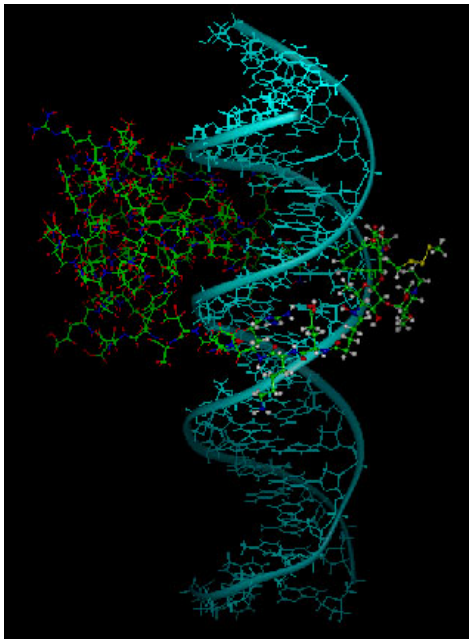

Figure 7: The DNA cleavage made by Calicheamicin. This is a highly toxic substance but has found pharmacological use as a conjugate in antibody cancer therapeutics (e.g. Gemtuzumab), because it binds DNA, chops it into pieces and thus kills the tumor cells..

Furthermore it is necessary to note that the classification of the mode of action referring to a specific process or site of action is difficult to be made and above all it is not exclusive. The cyanide binds a particular enzyme with a specific affinity, so this bond looks like the ligand-receptor interaction. On the other hand this interaction inhibits the enzymatic activity, causing the reduction of energy stores, which may influence biomolecular use and alter the calcium homeostasis. Another example is given by lead, whose several toxic effects may be partly imputed to the bond with specific proteins, while other ones are not easily attributable to a specific biochemical mode or a particular enzyme [12].

Finally it is important to emphasize that the mode and the site of action of the majority of substances is not well known yet. This knowledge gap is actually the main problem that has to be faced assessing the toxicity of a chemical by alternative methods, such as QSARs. Nowadays several steps have been made in the field of the development of structure-activity relationships, thanks to innovative algorithms and calculation tools, but great efforts should still be made to improve toxicological knowledge in order to promote a wider and more effective use of such methods.

2.2 Theory of QSAR models

2.2.1 What are molecular descriptors

Nowadays, the concept of molecular structure is one of the most important concepts in the scientific development. The reasoning based on the molecular structure has played a major role in the development of physical chemistry, molecular physics, organic chemistry, quantum chemistry, chemical synthesis, medicinal chemistry, etc.

A system is complex by definition when its behaviour as a whole can not be derived from the properties of its parts. It is evident that a molecule, with its embedded concept of molecular structure, exactly fulfils these conditions. The properties of a molecule do not depend only on the properties of the component atoms but also on their mutual connections. Therefore, they are inherent to the whole molecular organisation and stability and cannot be derived as the sum of the properties of the component atoms. In principle, we have a holistic system.

Due to its complexity, molecular structure cannot be represented by a unique formal model. Depending on the level of the underlying theoretical approach, several molecular representations can represent the same molecule. These representations however are often not derivable from each other.

Different molecular representations have been proposed, for example the 3-dimensional Euclidean representation, the 2-dimensional representation based on the graph theory, or the vector representations called fingerprints where the frequencies of several molecular fragments are stored. Each of these representations constitutes a different conceptual model of the molecule and by each model different sources of chemical information become available.

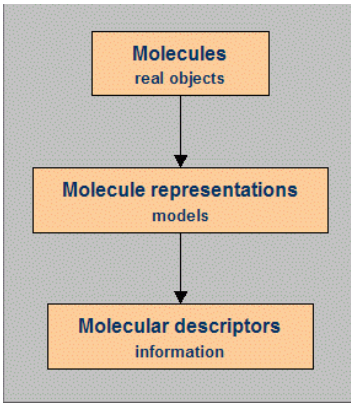

Figure 8: From molecules to molecular descriptors

The molecule implicitly contains all the chemical information, however, only a part of this information can be extracted by means of experimental measurements. This is where the concept of molecular descriptors comes into use: they are numbers able to extract small pieces of chemical information from the different molecular representations.

So what exactly is a molecular descriptor? It is any molecular property that characterizes the molecule, or more specifically: "The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment." The term “useful” in this previous definition has a double meaning: it means that the number can give more insight into the interpretation of the molecular properties and/or is able to take part in a model for the prediction of some interesting property of other molecules.

In the last decades, many researches have been focused on studying ways to capture and convert the information encoded in the molecular structure into one or more molecular descriptors. In such way, quantitative relationships can be established between structures and properties on the one hand, and biological activities and other experimental properties on the other.

Therefore, molecular descriptors are now playing a key role in scientific research. Evidence of the interest of the scientific community in the molecular descriptors is provided by the huge number of descriptors proposed until today: more than 3000 descriptors are actually defined and computable by using dedicated software tools. Each molecular descriptor represents a small part of the whole chemical information contained into the real molecule and, as a consequence, the large number of descriptors is continuously increasing with the increasing of the complexity of the investigated chemical systems.

The molecular descriptors range from simple counts of atoms, functional groups, or molecular weight, to properties based on 3D structure. They may require measurement of a physicochemical property, which is not recommended for a predictive technique, or may be calculated from knowledge of the contributions of functional groups, molecular topology or electron distribution. Despite their diversity, they are considered to describe only three aspects of a molecule – its hydrophobic, electronic and steric properties, which are held responsible for the biological activity in an organism.

Physicochemical properties tend to describe fundamental molecular effects, such as the partition coefficient being a measure of hydrophobicity. It is believed that these properties may be related to the mechanism of action of a molecule and are therefore less susceptible to spurious correlation. That is why many modellers claim that predictive techniques should be based on fundamental physicochemical properties, such as the partition coefficient, water solubility, ionization or dissociation constant, melting point, boiling point, etc. While the properties of a pure chemical substance may be measured, the computation of physicochemical properties has many advantages. These include the speed and low cost of calculation and, more importantly, the fact that calculation may be performed for chemicals that are not available.

A subgroup of descriptors, named topological indices (for example, Wiener Index, Harary Index, Zagreb Indices, etc.), has become a point of interest in recent years. The topological indices are derived from knowledge of the connections within a molecule and are formulated from graph theory. They account for many molecular properties, such as molecular size, branching and to a certain degree shape.

Molecular orbital theory has also been widely applied to the prediction of toxicity and fate, for example metabolism, persistence, or biochemical reactivity. The vast increase in the speed of computers has enabled the rapid calculation of numerous atomic and molecular orbital descriptors, such as charges, dipole moment, energy levels, etc.

The field of molecular descriptors is interdisciplinary and encompasses many different theories and disciplines. Knowledge of algebra, graph theory, information theory, computational chemistry, physical chemistry, quantum chemistry as well as theories of organic reactivity is needed for the definition of molecular descriptors. For their use, knowledge of statistics and chemometrics in addition to the specific knowledge of the problem is needed, as well as the principles of the QSAR approaches which will be explained in more detail later on. In addition to all this, sophisticated software and good programming skills are often a requirement. Molecular descriptors constitute a field where the most diverse strategies for scientific discovery can be found.

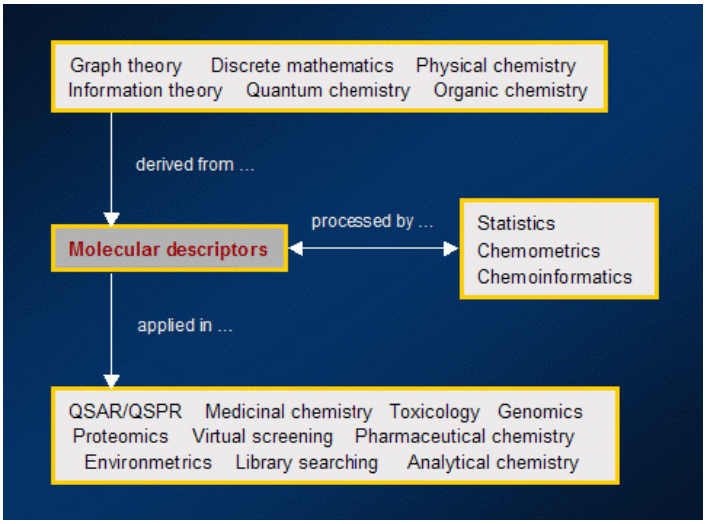

Figure 9: Theories and disciplines encompassed by the filed of molecular descriptors

2.2.2 Which are the endpoints of interest

Molecular descriptors have become part of the most important variables used in molecular modelling. Until about 30 years ago, molecular modelling was based on discovering mathematical relationships between experimentally measured quantities. Nowadays, it is performed modelling a measured property by the use of molecular descriptors which capture structural chemical information.

There is a remarkable number of diverse activities that have been successfully modelled. The activities of interest here are related to toxicology and drug design, for example the toxicity of a chemical to an environmental organism or human being, the fate of a pollutant in an ecosystem, the pharmacokinetic properties of a xenobiotic in humans, etc. The measures of these activities, which are quantities somehow related to the toxic activity of the molecules, are called endpoints and represent the output variables of the modelling.

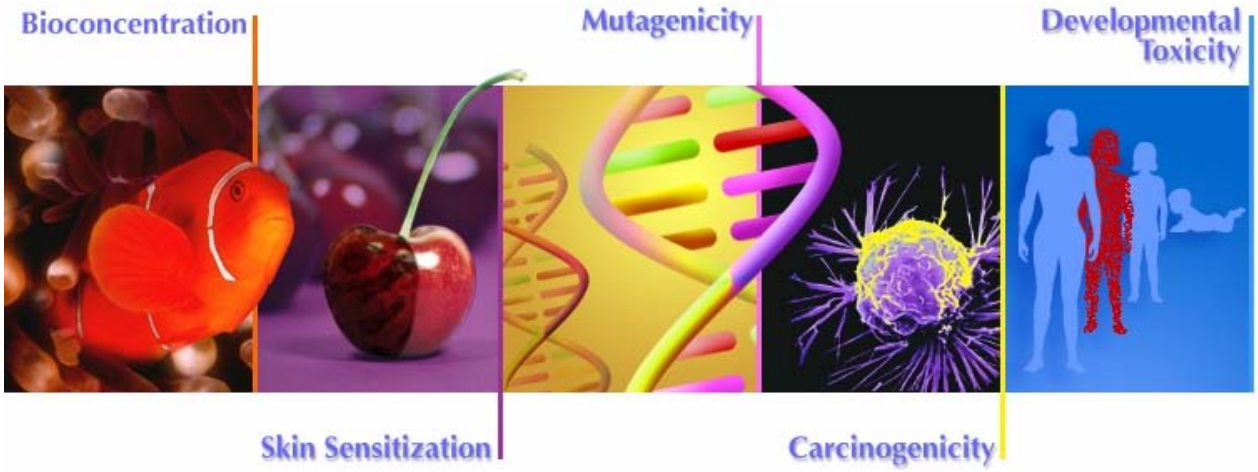

Figure 10: examples of endpoints. These are the endpoints of interest in CAESAR project.[1]

[1] For further information see: http://www.caesar-project.eu/

From the point of view of the type of biological activity that is being modelled, these toxicological endpoints can be divided in two major groups: human health endpoints and environmental toxicity endpoints.

Two very important human health endpoints are chemical carcinogenicity and mutagenicity. Carcinogens can be genotoxic, which interact directly with DNA and are thought to work by inducing mutations, or epigenetic, which act through mechanisms that do not involve direct DNA damage, however, no unifying theory exists for their mode of action. On the other hand, mutagens provoke heritable changes to the genetic material. The modelling of chemical carcinogenicity and mutagenicity is a very important goal in toxicology because of the huge impact they have on the quality of life and because of the enormous investment in time, money and animal lives needed to test chemicals adequately.

Other human health endpoints of great significance are chemical metabolism and biotransformation of chemicals within biological organisms, in order to determine their biological effects such as their effectiveness or toxicity as agents, for example pharmaceuticals or agrochemicals. Their modelling is motivated by the need for an efficient screening of large numbers of chemicals and the concern to reduce animal testing. Until recent years, a less intensely researched area has been the modelling of pharmacokinetic parameters. As the focus of the pharmaceutical or agrochemical industry has shifted towards faster and smarter screening in order to avoid failure of candidate chemicals, pharmacokinetic parameters such as absorption, distribution, metabolism and excretion (ADME) are becoming toxicological endpoints with increasing significance. When developing a new drug, it is important that the drug reaches its site of action in adequate concentration without accumulating in the body or producing side-effects. The pharmacokinetic properties must be optimized such that the drug will be readily absorbed, transported to the appropriate site and eliminated from the body in a timely manner. Approximately 40% of drug candidates fail during preclinical tests as a result of unacceptable pharmacokinetics. Due to the high economic cost of failure at this late stage, the benefits of the ability to screen out the likely failures at an earlier stage are evident.

In the area of environmental toxicity, some of the more important endpoints are persistence, bioaccumulation and soil sorption.

Persistence relates to the length of time that a predefined fraction of a substance remains in a particular environment before it is physically transported to another compartment and/or chemically or biologically transformed. Knowledge of persistence is crucial for evaluating the hazard and risk from chemicals released into the environment. In general, higher risk is assumed with increasing persistence of compounds. Persistence is not a simple endpoint, but a complex phenomenon comprising biodegradation/transformation, accumulation, distribution and transport processes.

Bioaccumulation is defined as uptake by an organism of a chemical from the environment by any possible pathway. On the other hand, soil sorption or sorption of a chemical (usually from water) by soil prevents the chemical from being transported further in the environment, although temporarily in most of the cases. In order to quantify the risks that environmental pollution poses, an understanding of the many factors that affect the distribution of a chemical in the environment is needed, two of which are precisely bioaccumulation and soil sorption.

From a statistical point of view two types of biological activity can be modelled: continuous and categorical.

Continuous data are numerical values that describe a concentration that provokes a particular effect. Usually this is a 50% effect lethal concentration, such as 96-h LC_{50}, defined as the concentration that kills half of an animal test population in 96 hours. A subgroup of continuous data for modelling is not necessarily biological in nature, and it includes physicochemical properties (partition

coefficient, melting point), fate descriptors (persistence, bioaccumulation), pharmacokinetic properties (bioavailability, metabolism) and many others.

Categorical data are typically yes/no data that classify a chemical as being active or inactive in a particular toxicological assay (for example, carcinogenic or non-carcinogenic), or may be descriptive such as high or low bioavailability. For some activities and endpoints a classification may be assigned on the basis of a particular number of criteria, originally developed from quantitative measures of toxicity.

2.2.3 Statistic theory of QSAR / How to construct models

Chemicals have various effects on organisms to which they are exposed, some of which are desirable and some are undesirable. Nowadays, the number of these chemicals is rapidly increasing, in pace with the fast industrial development. That is why it is extremely important to assess their potential toxicological effects on organisms and the environment.

Two main streams have been developed in order to explain the complex relationships between molecules and observed quantities, or endpoints. The first one is related to the search for relationships between molecular structures and physicochemical properties and is called QSPR (Quantitative Structure-Property Relationships). The second one, which is the focus of our work, is related to the search for relationships between molecular structures and biological activities and is called QSAR (Quantitative Structure-Activity Relationships).

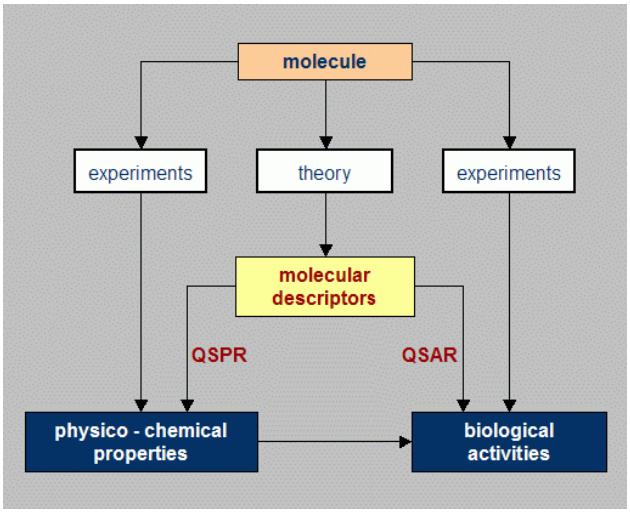

Figure 11: QSPRs/QSARs relating molecular structures with endpoints

Since chemical structure was elucidated, the relationship between chemical structure and biological activity has intrigued scientists. It has been recognized that the investigation of QSARs may provide useful tools for obtaining information regarding the effects of chemicals on man and the environment. Initially developed to assess the value of drugs, QSARs are now proposed as a method to assess general toxicity.

QSARs are based on the assumption that the structure of a molecule (its geometric, steric and electronic properties) contains the features responsible for its biological activity. For example, as already explained in the previous sections, biological activity can be expressed quantitatively as in

the concentration of a substance required to give a certain biological response. When the information encoded in the molecular structure is expressed by molecular descriptors in the form of numbers, one can form a quantitative structure-activity relationship between the two. By QSAR models, the biological activity of a new or untested chemical can be inferred from the molecular structure of similar compounds whose activities have already been assessed.

QSAR's most general mathematical form is:

Activity = f (physiochemical properties and/or structural properties)

It is therefore evident that the three key components required for the development of a QSAR model are:

- Some measure of the activity (in this case toxicity) for a group of chemicals in a biological or environmental system – toxicological endpoint
- A description of the physicochemical properties and/or structure for this group of chemicals – molecular descriptors
- A form of statistical relationship to link activity and descriptors

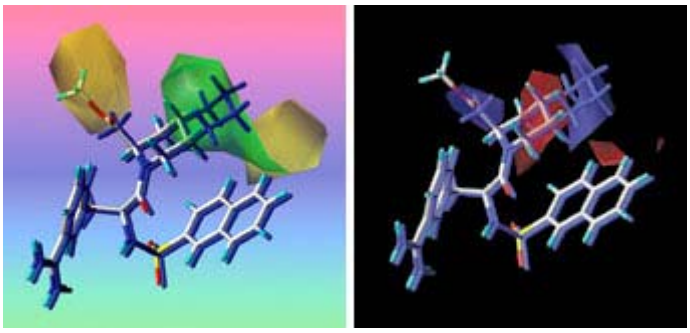

Figure 12: QSAR model visualization: building graphical models that relate biological activity of molecules to their structure

At first sight, the selection of compounds for developing QSAR models may appear to be self-evident. If we are interested in the biological effects of a certain group of chemicals we collect or measure all the compounds of that group that can be found. However, this strategy is not the best way to gather data and it may happen that too many results for the wrong compounds prevent the establishment of a good QSAR model. The successful construction of QSAR models requires experimental design, in which each compound included corresponds to a design point and the experimental factors that need to be varied in order to create the design are the physicochemical properties that characterize the compounds. The toxicological endpoint can also be an experimental factor and the goal is to develop a model that links the endpoint to the physicochemical descriptors. It is crucial that the design includes compounds that give both high and low values of the endpoint of interest, and if possible, a uniformly-spread range of intermediate values.

The response data, which are measures of the biological activity of compounds and represent the output variables in the QSAR models, can be measured directly by the investigators or collected from the literature. Knowledge of the precision and range of these data is of high importance. Some measurements have a natural range, but others may cover many orders of magnitude which may be

deceptive. It is therefore dangerous to take these data at face value. Examination of their distribution can be very useful because it can indicate where a certain type of processing is required. The precision is another property of interest because the model should have a standard error no better than the measurement errors. This is because of the fact that it should not be possible to calculate something more precisely that it can be measured. A standard error that is better (less) than the experimental one is a good indication that the model has been overfitted, which means that it fits the training data set well, but cannot generalize to other sets, which is the purpose for fitting a model.

The descriptor data, which capture information about the chemical structure of compounds and represent the input variables in the QSAR models, can be obtained from a variety of sources. In the early period of QSAR modelling, the choice of the descriptors was limited because they were generally tabulated physicochemical properties. Nowadays, there are over 3000 different molecular descriptors and it is common to use many more descriptor variables than there are compounds in the set when building a QSAR model. This leads to the need for dimension reduction, variable elimination and variable selection, which are different techniques for reducing the complexity of a problem in order to be able to recognize useful and informative patterns in the data. Dimension reduction is the process of reducing the number of random variables under consideration and is usually performed by a mathematical procedure called Principal Component Analysis (PCA) in which new variables called principal components are created from linear combinations of the original variables. Variable elimination is the process by which unhelpful or unnecessary variables are removed from a data set. Common procedures for variable elimination are Corchop and unsupervised forward selection. Even after eliminating unnecessary variables from a data set, there may still be many variables to choose from when building a model. In this case variable selection is used, whose aim is to choose descriptors that will be useful in some sort of mathematical model and will lead to a model that will generalize to other unseen compounds. There are many diverse procedures for variable selection and some are built in to the process of model building, such as the forward stepping multiple regression.

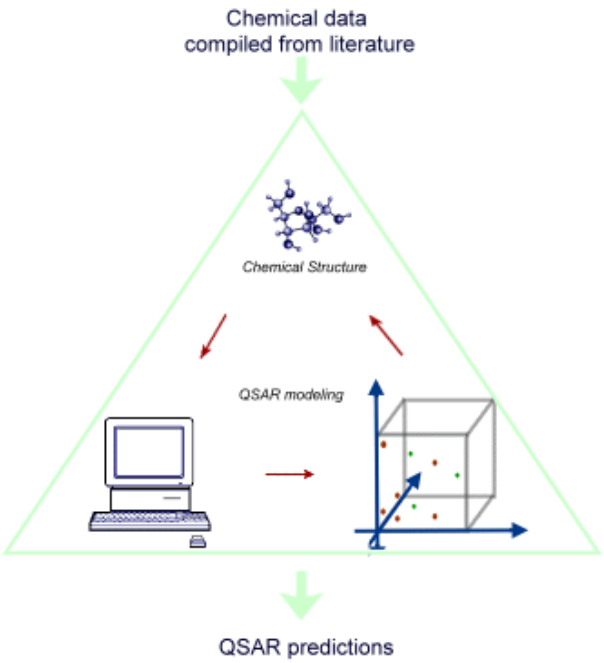

Figure 13: The process of QSAR modelling for predicting the biological activity of novel compounds

In this context, it must be mentioned that one of the major problems in QSAR modelling is the availability of high quality experimental data for building the models. The input data must be both accurate and precise in order to develop a meaningful model. Any developed QSAR model is statistically as valid as the data that led to its development.

In addition to this, a problem related to molecular descriptors is their reproducibility: experimental values can differ greatly even when referred to the same compound. As an illustration, several approaches have been developed for the theoretical calculation of the partition coefficient (log P), but in these calculations it is not uncommon to have differences of several orders of magnitude. In modern QSAR approaches, it is common to use a wide set of theoretical molecular descriptors of different kinds which take into account the various features of the chemical structure. There are many software packages that calculate wide sets of different theoretical descriptors. The greatest advantage of theoretical descriptors is the fact that they can be calculated homogeneously by defined software for all chemicals, including those not yet synthesized but represented by a hypothesized chemical structure, and therefore they are reproducible.

A variety of methods for building QSAR models exists. These methods are called pattern recognition methods because their aim is to devise algorithms that could learn to distinguish patterns in a data set. They can be classified as supervised (for example, Multiple Linear Regression, Discriminant Analysis, Partial Least Squares, Classification and Regression Trees, Neural Networks, etc.) or unsupervised (for example, Principal Component Analysis, Cluster Analysis, k-Nearest Neighbours, Nonlinear Mapping, etc.), where supervision refers to the use of the response data which are being modelled. Unsupervised learning makes no use of the response, meaning that the algorithms seek to recognize patterns in the descriptor data only. The advantage of unsupervised learning is the lower likelihood of chance effects, due to the fact that the algorithm is not trying to fit a model. On the other hand, supervised learning does use the response data and care needs to be taken to avoid chance effects. Another significant difference between supervised and unsupervised learning methods is the ratio of compounds (p) to variables (n) in a data set. When $n \geq p$, some supervised learning techniques may not work due to failure to invert a matrix, while others may give a false, apparently correct, classification. Even though this is not a problem for unsupervised methods, the presence of extra variables that have no useful information may obscure meaningful patterns.

The nature of the response data that they are capable of handling is another important feature of modelling methods. In this context, there are two types of methods: methods that deal with classified responses (for example, mutagen / not mutagen, toxic / slightly toxic / non toxic) and methods that handle continuous data (the response is a potency of an end-point). For the modelling of categories, a wide range of classification methods exists, including: Discriminant Analysis, k-Nearest Neighbours (KNN), Classification and Regression Trees (CART), Support Vector Machine, etc. For the modelling of continuous data, the most widely used method is Multiple Regression Analysis (MRA), a simple approach that leads to a result that is easy to understand. MRA is a powerful means for establishing a correlation between independent variables (molecular descriptors) and a dependent variable (biological activity). In addition, Artificial Neural Networks can be used for modelling both classified and continuous data.

After the model is developed, regardless of the type, it is of crucial importance to assess its performance by validating its predictive application. Most statistics packages generate a variety of statistical quantities for the common modelling approaches which will enable a judgement of significance and will give some guidance on whether the model may have arisen by chance. This is

based on the assumption that the data conform to some statistical distribution, usually multivariate normal. Unfortunately, this only indicates how well the model fits the data within the modelling assumptions and does not really give any information on how well the model might work. The best fit models are not the best ones for prediction. As a consequence, the only way to know how well a model may work is to try it out.

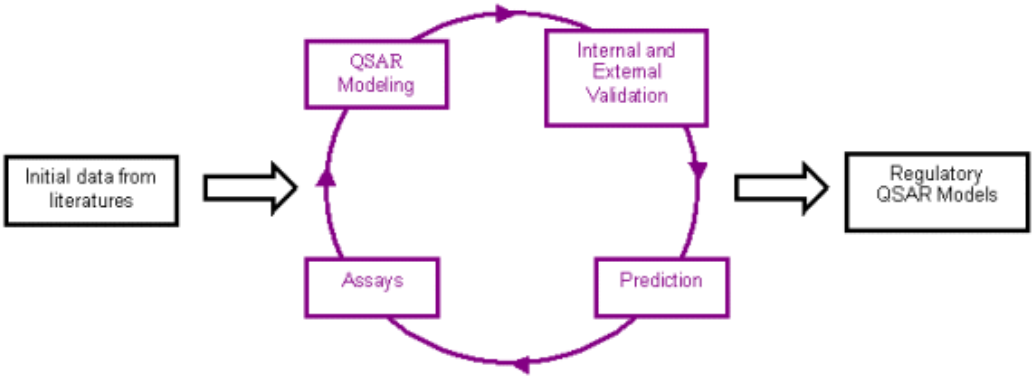

Figure 14: Depiction of the recursive process for developing QSAR models

One common approach is the Leave-One-Out Cross Validation (LOO or CV), which involves leaving out one compound, fitting the model to the remainder of the set, making a prediction for the left out compound and repeating the process for each of the compounds in the set. A variety of statistics can be generated using this procedure, for example LOO R_2 (called Q_2) and a predictive residual sum of squares (PRESS). The disadvantage of LOO is that only a small part of the data set is omitted and if outliers occur in pairs or groups they will not be identified. A better approach is to leave out some larger portion of the set (10 or 20%) and to repeat this a number of times. This allows the generation of a set of predicted values for the compounds so that estimates may be made of the likely errors in prediction. The disadvantage of this approach is that it is computationally intensive and suffers from a combinatorial explosion as the sample size is increased.

However, none of these procedures allows us to judge whether a relationship is real or it has happened by chance. One way to check for chance effects is to scramble the response values and then try to build models using the scrambled data. This can be repeated a number of times and some fit statistics, such as R_2, can be tabulated for the resulting models. If the R_2 value for the model of the unscrambled response is higher than the R_2 value for the scrambled sets, it is reasonable to assume that the model is not a chance fit.

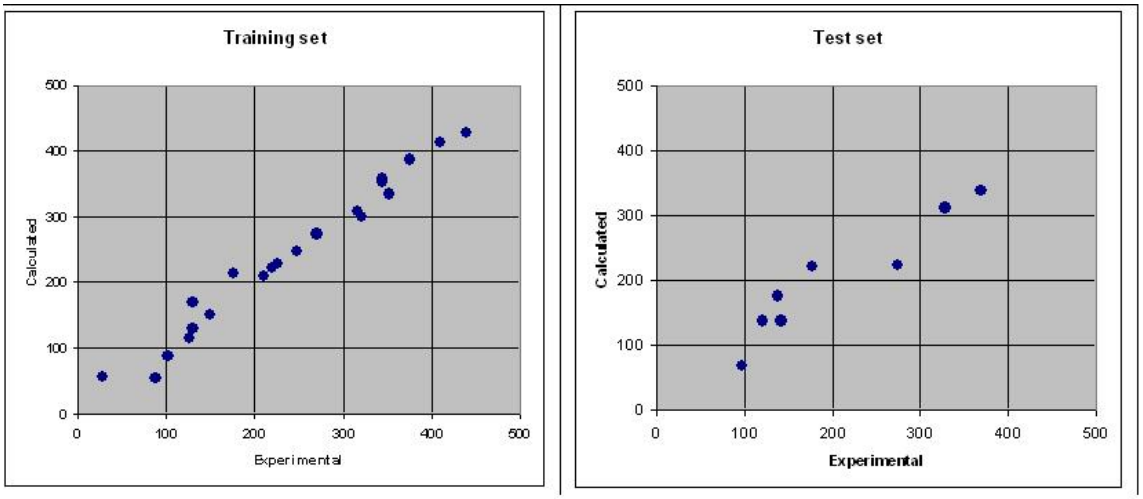

Figura 15: QSAR model validation by using a training set and a test set

Obviously, the best test of a model is to present it with unseen data, ether by holding back some of the original data to form a test set or by synthesizing or testing some more compounds once the model has been built. Only a stable and predictive model can be considered a reliable model and can be usefully interpreted for its mechanistic meaning.

There is an argument that, if the main aim of QSAR modelling is simply prediction, the attention should be focused on model quality and it is not necessary to try to interpret models. Another argument is that it is dangerous to attempt to interpret models, since correlation does not imply causality. Regarding the interpretability of QSAR models, Livingstone 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". On this basis, we can differentiate predictive QSARs, where the focus is best prediction quality, from descriptive QSARs, where the focus is descriptor interpretability.

To summarize all that was previously mentioned, what makes a successful QSAR model? The ideal QSAR model should: (1) consider an adequate number of molecules for sufficient statistical representation, (2) have a wide range of quantified end-point potency (for example, several orders of magnitude) for regression models or adequate distribution of molecules in each class (for example, active and inactive) for classification models, (3) be applicable for reliable predictions of new chemicals (validation and applicability domain) and (4) allow to obtain mechanistic information about the modelled end-point.

2.2.4 Ethic and economic impacts of QSAR

It has been more than 40 years since QSAR modelling was first used in the practice of agrochemistry, drug design, toxicology, industrial and environmental chemistry. Its growing power in the following years may be attributed to the rapid and extensive development in of methodologies and computational techniques that have allowed to delineate and refine many variables and approaches used in this modelling approach. Initially developed to assess the value of drugs, QSARs are now proposed as a method to assess general toxicity. They initiated a radical change in the way of thinking and leading toxicological studies. They aim at going beyond the limits of the traditional approach and facing the complexity of the biological world through a deeper analysis of the intrinsic toxicity mechanisms of actions and their driving forces. QSAR modelling is

a challenging approach and the whole scientific community agrees on the numerous potential advantages that could come from the application of QSARs.

There are many reasons why one may wish to predict the toxicity of chemicals. It is fundamental that computer models allow for the effects of chemicals to be predicted and these predictions may be obtained from knowledge of chemical structure alone. For most methods, provided that the chemical structure can be described in two or three dimensions, the effects may be predicted. Information regarding the chemicals may be gained without chemical testing, or even the need to synthesize the chemical. QSARs are therefore often employed to establish a correlation between structural features of potential drug candidates and their binding affinity towards a macromolecular target in order to identify chemical structures that could have good inhibitory effects on specific targets and have low toxicity. In addition to designing in attractive features of molecules in drug and pesticide design, it is now possible to design out toxic features.

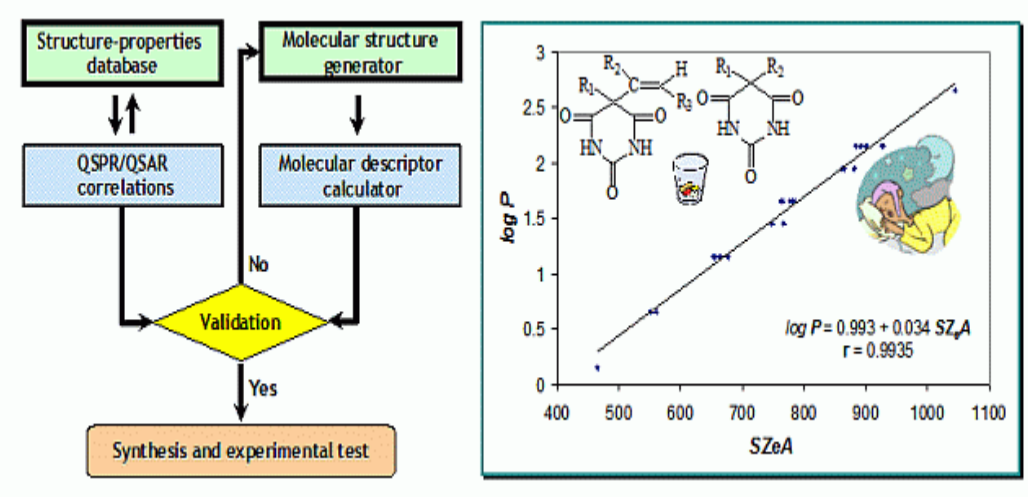

Figure 16: Predicting the toxicity of chemicals in drug design

Furthermore, approximately 100000 separate chemicals may be released into the environment annually and it is therefore frightening to consider that reliable toxicity data exist for only a tiny proportion of these chemicals, probably less than 5 percent. Computer-aided prediction of toxicity has the capability to assist in the prioritisation of chemicals for testing, and for predicting specific toxicities to allow for labelling.

For several decades there has been a growing public concern regarding the use of animals in testing, especially in toxicology and medical research. This has resulted in the boycotting of companies, organizations and individuals associated with animal testing. Campaigners for animal welfare cite a number of approaches to reduce and ultimately replace animal tests. There is clearly a role for predictive techniques in the replacement of animal tests, either as stand-alone methods, or more commonly as part of a tiered assessment strategy. The integration of computational methods in combination with the judicious use of physicochemical properties is a viable alternative to animal testing. Having in mind that in 2002 in Europe it is estimated that 10.7 million animals were used for experimental aims, QSARs would save a lot of animals and solve ethic problems about their use.

Moreover, animal testing takes about 1-2 years per compound, so companies sometimes prefer to continue using tried and tested substances rather than starting such a long testing procedure. A

single QSAR may also take 1-2 years to be developed, but once the model is ready to be applied it could drastically reduce the time requested for testing because a lot of compounds may be tested almost immediately (it depends on the type of descriptors and the complexity of the model, but these tests do not go over a pair of days).

In this context, it is also necessary to underline that recent studies have proved that animal testing is not as valid as it is believed to be. The main problem is in the difference between men and animals, thus often some results can not be directly applied to the latter in the same way as the former. QSARs, on the other hand, are based on a mechanistic interpretation of biological activities and so, if we are able to apply them with an adequate level of uncertainty, we could have a more general, and therefore applicable, analytical approach.

Unfortunately, the level of uncertainty that is associated to QSARs is still too high to proceed with the complete substitution of animal testing, even though the legislative framework has accepted their application (at least in theory). QSAR information will most often be used to supplement test data within chemical categories and endpoint-specific Integrated Testing Strategies (ITS). Uncertainty remains the only barrier for QSARs to fully replace animal testing.

Toxicological testing is costly financially as well as in terms of the animals used and the time taken. By using the methods to predict toxicity these costs are greatly reduced, which should allow for faster and less expensive product development, e.g. pharmaceuticals, as well as assessment of environmental effects. According to the study of Pedersen and colleagues, presented on the *Stakeholder Workshop on Impact Assessment of REACH* on 2 November 2003 in Brussels, the cost-saving potential of valid QSARs is estimated to be 700-940 million euros.

An often ignored spin-off from the development of QSARs is the increased understanding they can provide in both the biology and chemistry of active compounds. There are countless examples where knowledge of biology and chemistry has been advanced by modelling in the field of toxicological effects.

Thus, it is expected that huge efforts will be made in order to improve the current incomplete knowledge on toxicological mechanisms and to develop more reliable QSAR models with a lower degree of uncertainty.