1 Summary

1.1 Outline

In our everyday life we have to deal with an exponentially increasing number of different chemical compounds: the number of registered chemicals is about 28 millions, including food colouring and preservatives, drugs, varnishes and paints for wearing and objects, pesticides and many others. It is well recognised that chemicals may pose high risk to environment and to humans, and hence their toxic activity has to be assessed.

Biological active substances interact with bio-molecules, triggering specific mechanisms like activation of an enzyme cascade or opening of an ion channel, which finally leads to a biological response. These mechanisms, determined by the chemical composition of the considered substances, are unfortunately largely unknown, and thus one has to study toxicity experimentally.

It is possible to conduct at least three different kinds of experiments to assess the biological activity of a molecule: "in vivo" experiments, that is animal testing, "in vitro" experiments, that is using tissue culture cells, and "in silico" experiments, which refers to computer simulations.

Figure 1: the three experimental ways

Both animal testing and “in vitro” experiments are of course time consuming and expensive, and especially animal testing is considered ethically unacceptable by a growing majority of people. For these reasons, and also thanks to the increased power of computers, the scientific community and the industrial world focused on “in silico” experiments, which represent a more mathematical and computational approach, developing a number of models and strategies to predict physical and chemical properties of compounds.

Computational chemistry has changed the classical way to make experimental science. We are more and more moving from experiments to simulations. We are able to model molecules according to different views, from the basic valence model to graph representation, from electronic clouds to 3D structure. Algorithms are available to compute molecular descriptors, ranging from simple properties to complex molecular fingerprints. Descriptors can help in transforming the study of interactions of molecules with living organisms in a kind of data mining problem. Data mining could find relevant correlations between descriptors and the response of interest. It is not enough; if we want to make predictive models we need to assess also the predictive power of our relations.

In this project we will focus on a particular kind of data mining applications, called QSARs (Quantitative Structure Activity Relation). These algorithms look for correlations between the properties of the chemical structure of a compound and a measure of its activity/toxicity in a specific area, such as mutagenicity, cancinogenicity, skin sensitization, that is called “endpoint of interest”.

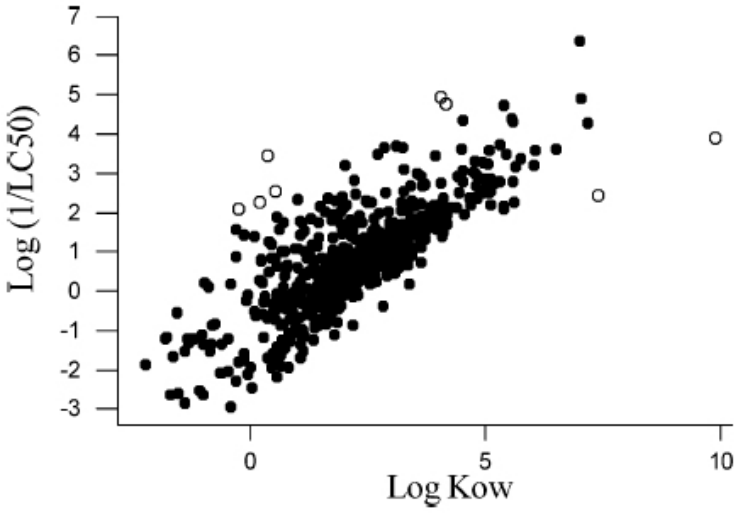

Figure 2: an example of QSAR analysis. The descriptor is Log Kow; the endpoint is death and it is measured as a lethal concentration (LC 50)

In other words, once the properties of a compound are quantified in a set of variables, called “descriptors”, these algorithms model the toxicity of a compound with respect to a given endpoint as a function of some descriptors, and the coefficients of the function are estimated through statistical analyses applied on a sufficiently large database.

The underlying idea of these models is that chemicals with similar structures, i.e. with similar values for the considered descriptors, must behave in a similar way; thus, once the model is built, it can be used as a forecasting tool in drug design, environment protection, hazard analysis, for all those compounds whose structure is similar to the structure of the ones used to tune the model.

The usage of these models is growing, since they aim to provide fast, reliable and quite accurate estimates of the chemicals’ activity; these features also make them suitable for legislative purposes, and that is why they have been included as an alternative tool for risk assessment in the new European legislation on chemical production, called R.E.A.C.H. (Registration, Evaluation, Authorisation and Restriction of Chemicals). This legislation fixes the rules for chemical production in E.U., and one of its key points is that it requires a risk analysis for each chemical placed in European market in amount greater than 1 ton/year. To further underline the broadness of this law, suffice it to say that it is 849 pages long, and international mass-media defined it as “*the most important legislation in European Union in 20 years*” (BBC News, 28 November 2005) and “*the strictest law to date regulating chemical substances*” (San Francisco Chronicle, 14 December 2006).

1.2 How this report is organized

In Section 2 we illustrate the three basics areas that constitute the necessary technical background needed to develop our work: toxicity mechanisms, QSAR building and calculation of descriptors. Section 3 is dedicated to a short presentation of the REACH legislation.

In Section 4 we analyze more in details the needs of stakeholders involved in the problem. For what concerns the researchers we will explore the issues caused by tight copyright policies on software, focussing especially on the difficulties in getting reliable values for descriptors due to the calculations performed with non Open Source code; we will also consider the lack of standardization in descriptors, and how to deal with unknown toxicity mechanisms. From the industrial point of view, we will discuss the problems due to REACH legislation and in particular to the QSAR usability for risk assessment.

In Section 5 we give a summary of the State of The Art: examples of effective QSAR models, existing software and other help desks for companies.