4 Users' Requirements

4.1 Research Institutions' requirements

QSAR is a very attractive field that involves large part of research activity in chemistry and biology. Both academic and private research institution are interested in this area. The reasons are about the potentiality of QSAR models: to find a QSAR model means to acquire a deeper understanding of all the processes that are related with chemical compounds and of course get the access to an enormous quantity of information about properties and effects even for unknown molecules.

Unfortunately, first studies suggest that a real model that connects each molecular compound with its activity and with its physical and chemical properties is just a dream. For the enormous number of variables a complete QSAR requires a so complicated mathematical model that at the moment is out of the capacity of human knowledge.

Besides this strong statement, the practical use of the QSAR analysis is restricted to only some characteristics to predict and only for a specific class of molecules. From this point of view the problem seems much more affordable. In fact the research field is full of QSAR models that predict only some particular characteristics.

To build a QSAR model the starting point is to select a list of descriptors that describe the important properties of class of chemical compounds. Then several machine learning techniques or statistical tools can be applied to extract information and a deductive model that also work for unseen molecules. The first problem of research is at the beginning of this process: which descriptors to use?

In the literature there are thousands of descriptors that have been used during all the years of research in this field. But it is impossible to use all the descriptors available in literature to build a QSAR model. There are two principal reasons. The first one is that it is computational heavy to calculate all the possible descriptors for all the chemical compounds taken in exam for a particular model. Second, and most important, the machine learning techniques and the statistical tool have difficulties that increases exponentially with the dimensionality of the data and so with the number of descriptors. These difficulties are caused from the presence of redundant descriptors or in particular from those descriptors that do not give useful information to predict a specific activity.

So the first need of the research community is the reduction of the number of descriptors to use in the developing of QSAR models. This task resulted to be particularly difficult because a specific descriptor can be much useful to predict a certain activity and useless in predicting another. Other descriptors can be useful for a particular set of molecules while useless for another set.

The evidence suggests that such minimization of the number of descriptors is possible only after that the activity to predict and class of compounds have been chosen. This means that the first part of the process devoted to build a QSAR model must deal with the minimization of descriptors: this is not an easy process because requires very specific knowledge. And even when specific knowledge is available there can be some useful descriptors that can be cut of from the analysis. This can be the case of topological descriptors that can appear a priori uncorrelated with a certain activity but can also give useful information.

Another big problem of research in QSAR field is about sharing knowledge. In this field there is a large part of knowledge that is private. This is due to the economical importance that some QSAR models have for pharmacological companies, and in general for those companies that lead a private research looking for chemical compound with innovative properties.

The problem of sharing knowledge is a big one: it brakes the research. Many researches institutes can work on the same model but they can not benefit of the results of the others.

This lack of sharing knowledge can involve both the entire QSAR model or only the calculation of descriptors: the situation is complicated in both cases. Many papers show that good results have been obtained with QSAR approach but they do not reveal the model used: so both the descriptors used and machine learning techniques are unknowns. This kind of results has no utility for another research institute that is interested to performs a similar study.

Much often scientific papers describe accurately the QSAR model and the descriptors used but they do not publish the source code of descriptors. This is another big problem in research community because descriptors are not standardized. For each descriptor there are tens of different variations and each variation can be implemented with tens of small variations. Also less important variations such as changing the type of a variable can modify the final value of the descriptor. So, until descriptors source code is not published it is impossible to compare the results obtained between two different research institutions. For example some researchers can get bad result from the same model presented with good results from other researchers: the difference can be caused by different versions of the chosen descriptors.

Some researchers can not publish the source code because they used for the calculation some professional non-free programs like Codessa and Dragon. This kind of software can be very expensive and this means that institutions with less found are cut off from research in some areas of QSAR.

To summarize this part, the principal requirements of research institutions are two: minimization of the number of descriptors to use and the sharing of knowledge.

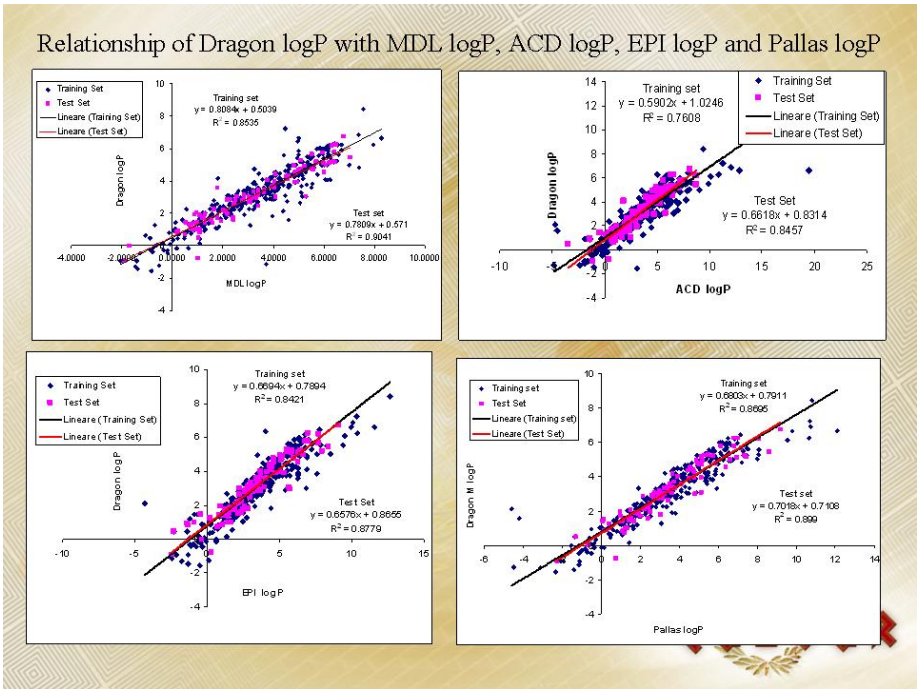

Figure19: how different software (Dragon, MDL, ACD, EPI, Pallas, see section 4.2) gets different results when calculating the same descriptor, in this case logP. From CAESAR project[2]

4.2 Firms' requirements

Since the REACH entered into force, manufacturers and importers of the chemical substances that are produced or imported in quantities of one tonne or above per year have to submit a registration dossier to the ECHA. Chemical firms have to demonstrate that risk coming from the substances they deal with are limited and can be controlled, because the registration is the necessary condition for the commercialization of products.

The innovative aspect of REACH, that is changing firms' approach to the risk assessment of their chemicals, is the acceptance of alternative methods (such as QSARs, in-vitro studies and chemical grouping) for regulatory purposes. In addition, reducing animal testing is one of the main objectives of REACH and this new regulation foresees the use of QSARs when testing does not appear necessary because the same information can be obtained by other means. As a consequence, interest in QSAR models is increasing more and more also among firms, so it is necessary to consider them as one of the most important stakeholder of Vichem project. Even if all chemical firms need to gain complete and clear information about REACH provisions, this necessity is more problematic for Small and Medium Enterprises (SMEs), which do not have the adequate know-how and economical resources to face the new regulatory framework.

The needs of these stakeholders may be ascribable to two different aspects. The first one is the need of obtaining knowledge about REACH provisions that is structured, comprehensive but not wasteful, and easy to be applied in practice. Currently a lot of information about REACH is obtainable by surfing the net, but it does not fit the features just mentioned and it may create confusion and misunderstandings. The second need is still to gain complete and well organized

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

information, but the focus is on those aspects present in REACH regulation, that refer to the use of QSARs for regulatory purposes. In this case there is the opposite problem, because this kind of information is poor and difficult to find out.

Both of these needs can be translated into similar firms' requirements, that are technical. The final requirements may be exhaustive guidance tools, which do not directly contain the overall knowledge related to REACH in general, on the one hand, and the regulatory provisions about QSARs, on the other hand, but which provide a simple and brief way to find and manage useful information.

In order to assess the users' requirements a questionnaire was chosen as the most effective method to be used. Its structure was divided into three main sections and each of them was dedicated to a specific topic: REACH, REACH focusing on QSARs, and QSARs models. All the questions had the same purpose to understand what kind of information firms would be interested in to satisfy their needs. In the REACH section, the questions mainly concerned with the usefulness of a guidance on the websites dealing with REACH, of an interactive tool to manage information about workshops and conferences, and of a glossary of recurring terms. The second section aimed at finding out if consulting services or specific summaries of the articles about QSAR regulatory uses may be interesting. The last section investigated the interest of firms on the availability of open source codes for descriptors calculation, of a guidance on free computational tools and specific websites, and of article repository. This questionnaire was send to Federchimica to be widely spread among Italian chemical firms, but unfortunately no replies were obtained.

Thus, the users' requirements were obtained joining the information coming from the analysis of the state of art about the different relevant topics and from the experience of Mario Negri institute's team.