Data Quality Issues in Toxicological Knowledge Discovery

Christoph Helma
Institute for Tumor Biology–Cancer Research
Borschkegasse 8a, A-1090 Vienna
Christoph.Helma@univie.ac.at

Eva Gottmann
Institute for Environmental Hygiene
Kinderspitalgasse 15, A-1095 Vienna

Stefan Kramer and Bernhard Pfahringer
Austrian Research Institute for Artificial Intelligence
Schottengasse 3, A-1010 Vienna

Abstract
Every SAR technique for toxicity prediction relies on the exact estimation and representation of chemical and toxicological properties. We will present potential sources of errors associated with the utilization of large, noncongeneric datasets and complex toxicological endpoints (e.g., carcinogenicity). According to our experience we have identified the major problems in the areas of compound identification, descriptor calculation and toxicity data. Generally, we consider the chemical data as more reliable than the results from toxicity experiments. As it is impossible to tackle the data quality problem on a case by case basis for a large number of compounds, we will propose some possibilities for routine quality control of large datasets.

Introduction
The development of Structure Activity Relationships (SARs) relies on the comparison of chemical structures or their properties (descriptors) with their toxicological effects. Although it is generally accepted that the exact estimation and representation of chemical and toxicological properties is a prerequisite for a good SAR model, this topic has been rarely addressed in a systematic manner.

In this paper we will present our experience resulting from the application of Machine Learning techniques to large, noncongeneric datasets and complex toxicological endpoints (e.g., carcinogenicity). Our source of toxicological information is the Carcinogenic Potency Database (CPDB) (Gold & Zeiger 1997). It contains very detailed information from long-term in vivo carcinogenicity experiments and consists of two major parts. One dataset contains the results of carcinogenicity experiments performed within the National Toxicology Program (NTP). These studies were conducted in compliance with FDA Good Laboratory Practice Regulations. The second dataset contains data from the general literature which meet a set of standard criteria (details in (Gold & Zeiger 1997)).

Due to budget restrictions, we obtained structural data almost exclusively from free sources: the NCI Database1, Chemical Structures from NTP Technical Reports2 and ChemFinder from CambridgeSoft3 (Table 3).

<table>
<thead>
<tr>
<th></th>
<th>NTP</th>
<th>Literature</th>
<th>whole CPDB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compounds</td>
<td>393</td>
<td>1028</td>
<td>1299</td>
</tr>
</tbody>
</table>

Table 1: Summary of the Carcinogenic Potency Database (CPDB) (Gold & Zeiger 1997).

Identification of Compounds
The correct assignment of chemical structures and their toxicity data is crucial for the development of a valid SAR model, because faulty structures in the training set will prohibit the correct detection of features responsible for toxic action.

As we use data from different sources, we had to rely on Chemical Abstracts Registry (CAS) Numbers for their identification. We were able to find CAS Registry numbers for 1257 of 1299 CPDB compounds, 11 CAS numbers were associated with more than one compound and 14 compounds were marked as mixtures in the CPDB. Although the CAS was designed as an unique identification of chemicals, it is sometimes not a good identifier for toxicological purposes. Toxicologically irrelevant differences (e.g., crystal water) lead for example to different CAS numbers for similar structures. Typos may cause a wrong assignment of structures. On the other hand it is sometimes difficult to assign a CAS if only nonstandard nomenclatures are available. For the CPDB we were unable to find a CAS for 27 compounds. Table 2 lists some possible errors associated with CAS numbers and possibilities to check them.

Another source of variability is associated with finding the “correct” structures for a given CAS Number.

Another source of variability is arising from the possible presence of impurities or the incomplete knowledge about the structures (e.g. isomers) in the chemical tested. We tried to eliminate all compounds containing more than one structure from the CPDB (54 mixtures in 1169 identified compounds). This procedure relies to a large extent on correct information in the underlying database and the correct assignment of CAS Registry Numbers. In many cases it is virtually impossible to check the correctness of this data.

**Calculation of Descriptors**

In the development and particularly in the application of SARs it is essential to identify the structural or chemical properties that are predictive to the endpoint of interest. Presently the choice of structural and property descriptors for complex toxicological effects strongly relies on the intuition of the individual researcher, especially if no detailed knowledge of the underlying molecular mechanisms is available.

**Measured properties** Descriptors based on measured properties (e.g. lipophilicity, electrophilicity) have been historically the most favored approach when generating SARs. Their determination is expensive and time-consuming and they are therefore not very suitable for large datasets.

**Presence of substructures** Every chemist is used to think in terms of functional groups which compose a molecule. SAR models based on substructures are therefore very intuitive. Apart from predicting untested compounds they can be used to gain a deeper understanding of molecular mechanisms of toxic effects and they can provide good guidelines for the construction of new compounds. There are basically two approaches in using substructures as molecular descriptors.

The classical method checks whether predefined substructures are present in a molecule. Their presence or absence may be represented in tabular form. The disadvantage of this procedure is that substructures have to be defined in advance, and fragments not defined are consequently not considered.

The other possibility is to generate structural fragments automatically, and search for those which occur frequently in toxic compounds (Klopman & Rosenkranz 1994). This procedure is computationally very expensive, but it has the advantage that
fragments can be detected in an unbiased way. As this procedure results in different numbers of descriptors for every compound it is advisable to use a relational representation of the data.

Connectivity indices Molecular connectivity indices are a compact representation of the topological information of a molecular graph. In our opinion their application is becoming more and more obsolete, because more intuitive connectivity and topological information can be derived from higher level structural representations (Pfahringer et al. 1999) and 3-dimensional models (e.g. with NACCESS or VolSurf). The use of molecular connectivity indices was extensively discussed by Kier and Hall (Kier & Hall 1986).

Calculated structural and electronic descriptors Chemicals are toxic because they interact with biological macromolecules and reactivity is determined by electronic and steric properties. With the increasing speed of computers it is possible to calculate the “electronic nature” and three-dimensional structure of chemicals within a reasonable amount of time. As these properties can be calculated for almost every molecule, it should, at least theoretically, be possible to make predictions for compounds with novel substructures.

Programs for the calculation of 3D-structures and electronic descriptors were not designed for batch processing. It is often necessary to try different settings, if a calculation does not converge. As they use an iterative process to calculate the final structure, the results may depend on the initial structure, end in a local minimum or the solution may oscillate between different states (Clark 1985). In light of these results it may be advantageous to use rule based systems (e.g. CORINA, PETRA) for the calculation of 3-dimensional structures and electronic properties of large and diverse datasets. Once more it seems to be more important to obtain consistent instead of “correct” descriptors.

Calculated chemical properties Recent developments of Quantitative Structure–Property Relationships (QSPRs) have enabled the calculation of physico-chemical properties (e.g. logP (Meylan & Howard 1995)) which can be utilized in SAR models. They are usually very predictive and can be readily interpreted in terms of chemical and toxicological knowledge. Problems may arise from error propagation, if the results are not accurate enough or calculations fail for certain types of compounds. We use in our work the logP as a lipophilicity indicator. A review of algorithms for the calculation of lipophilicity parameters can be found in (Mannhold & Dross 1996).

In practice, the choice of descriptors will vary strongly with the scope of the desired model and the capabilities of the learning algorithm. Within a drug design process it is sensible to work with substructures, SARs for the elucidation of molecular mechanisms will contain structural, electronic or physico–chemical descriptors and high–predictivity models will use descriptors even if they are not instantly intuitive. Generally, it is advantageous to use different types of descriptors to enhance the predictivity of a SAR model (Helma, Kramer, & Pfahringer in press).

Despite the problems associated with descriptor selection and calculation, we consider chemical data as generally more reliable than toxicological data.

Toxicological Data

Most biological effects are a complex expression of several mechanisms happening in sequential and/or parallel order leading to highly variable endpoints.

Heterogeneous datasets are caused by the limited availability of good, validated data. Large standardized testing programs (e.g. NCI/NTP) were not designed for the development of SAR models, therefore they cover only a limited set of possible structures.

For rodent carcinogenicity assays the reproducibility was reported to be approximately 80% (Gold & Zeiger 1997) under standardized conditions. A comparison of the compounds present in both (Literature and NTP) parts of the CPDB shows, that 71% are classified similarly (Table 4). Figure 1 depicts a comparison of the tumorigenic doses TD50 for 43 compounds classified as carcinogens in both CPDB parts.

This comparison may underestimate the reproducibility of carcinogenicity experiments, because only NTP experiments were conducted with a standardized protocol. But it resembles closely the real world situation of many SAR modelers who have to aggregate data from different sources to obtain enough data for their investigations.

<table>
<thead>
<tr>
<th>concordant classifications</th>
<th>discordant classifications</th>
<th>Nr. compounds present in both parts</th>
</tr>
</thead>
<tbody>
<tr>
<td>73</td>
<td>30</td>
<td>103</td>
</tr>
</tbody>
</table>

Table 4: A comparison of the classifications in the Literature and NTP parts of the CPDB

It is beyond the scope of the present article to present a detailed discussion about the sources of variability in toxicity experiments, and in many cases it is impossible for the SAR modeler to check the validity of the data. Nevertheless, we want to stress the importance of using preferably results of standardized experiments. The definition of classifications derived from original data (e.g. rodent carcinogen/noncarcinogen from a variety of sex, strain, organ specific effects) should be documented to enable a comparison with other research groups. When adding data from non-standardized sources, it should be carefully considered, if the increased amount of data outweighs the additional variability. Including

