Knowledge Acquisition in Chemistry

by Joachim Johann Gasteiger

Computer-Chemie-Centrum, Universität Erlangen-Nürnberg, Erlangen (Germany)

Most chemical phenomena are too complex to be put into mathematical equations that can be solved. Chemical knowledge is gained to a large extent from individual observations and inductive learning by deriving general rules from a series of data. Data analysis is therefore of paramount importance in chemistry.

Statistical and pattern recognition methods are playing an increasing role in the analysis of chemical information. Flexible and robust methods are very much desired. This is underscored by the rapid acceptance of neural networks for chemical applications [1,2]. Quite a few problems in chemistry have to deal with the relationships between the structure of a compound and its physical, chemical or biological properties. The approach taken to find such relationships is outlined in Figure 1.

![Figure 1](image)

**Figure 1.** The building of models to derive relationships between chemical structure and data.

The endeavor to derive chemical knowledge from individual data is supported by the increasing availability of chemical information in databases. More than 13 million compounds have been stored, databases of several millions of chemical reactions are available, hundreds of thousands of spectra of chemical compounds...
have been stored in computer readable form together with information on the structure of those compounds.

Chemical Structures

A major problem in the analysis of chemical information is the representation of chemical structures [3]. Chemists have created a rich and internationally accepted language for representing the structure of molecules: the structural formula. A structural formula is essentially an undirected, weighted graph. Graph theory therefore plays a major role in the computer representation and manipulation of chemical structure information (Figure 2).

Figure 2. Structural formula of muscarine, the poison in the mushroom fly agaric. The wedge and dots indicate stereochemical information.

Graph theoretical methods are used to derive a unique representation of a chemical compound, an indispensable requirement for storing informations on compounds in databases. Furthermore, the perception of rings, of molecular symmetry, of aromaticity, and the coding of stereochemical information rest on graph theory.

However, with the increasing sophistication of our insight into chemical structures, the limitations of a structural formula are increasingly felt. Many applications in chemistry and biology have to deal with the three–dimensional aspects of chemical structures (Figure 3).

In fact, the representation of chemical structures has developed to a point that details of the three–dimensional electron distribution in molecules have to be taken into consideration. The graphical representation of such three–dimensional structures has matured to a high degree of sophistication (Figure 4). However, methods for extracting the important features of the three–dimensional electron distribution, for comparing molecules and for defining similarity at such a high level of structure representation are just being explored. Figure 5 shows a two–
Knowledge Acquisition in Chemistry

Figure 3. Ball and stick model (a) and space-filling-model (b)

dimensional map obtained by a Kohonen neural network [4] from the molecular electrostatic potential on the van der Waals surface of muscarine (cf. Figure 4) [5].

In essence, the Kohonen neural network allows a non-linear projection of the three-dimensional formation of a property on a molecular surface onto the surface of a torus that can be converted into a two-dimensional plane [5].

Chemical Reactions

Chemical reactions are usually represented by the structural formulae of the starting materials and products of a reaction (Figure 6). The extraction of knowledge from databases of chemical reactions requires finding: (a) the reaction center, i.e., the atoms and bonds directly involved in the reorganization of electrons during a reaction, (b) the required context, i.e., groups of atoms and bonds not directly
involved in a reaction but required to make a reaction go, and (c) the variable portion, i.e., those parts of the molecules that may vary in the individual instances
of a given reaction type.

Methods for the automatic assignment of reaction centers have been developed. Machine learning techniques for the classification and generalization of chemical reactions based on conceptual clustering methods are being developed [6].

**Figure 6.** Chemical reaction equation.

**Figure 7.** Reaction center (in bold) and required context (encircled) of a chemical reaction.

### Spectral Information

Spectra of compounds are rich in information that helps in the elucidation of the structure of compounds. However, the relationships between the structure and spectral data are again highly complex. For example, a mass spectrum records the fragmentation of a molecule after ionization by high energy electron impact.
Figure 8. Experimental and reconstructed mass spectrum of CH$_2$ClBr.

Figure 9. Fragmentation scheme used for explaining the mass spectrum in Figure 8.
The fragmentations and rearrangements occurring in the mass spectrometer can be elucidated by an assembly of conceptual and mathematical models [7]. Figure 8 (preceding page) shows the experimental mass spectrum of an organic compound and that part of the mass spectrum that can be explained by the fragmentation scheme shown in Figure 9.

Clearly, the structures analysed in Figures 8 and 9 are quite simple. More complex organic molecules lead to much more extensive fragmentation schemes that, nevertheless, can still be analysed.

References


Joachim Johann Gasteiger
Computer-Chemie-Centrum
Universität Erlangen-Nürnberg
Nagelbachstraße 25
D-91052 Erlangen, Germany