Theories for chemical carcinogenesis

Prevention of environmentally-induced cancers is a health issue of unquestionable importance. Almost every sphere of human activity in an industrialised society faces potential chemical hazards of some form. It is estimated that nearly 100,000 chemicals are in use in large amounts every day. A further 500–1000 are added every year. Only a small fraction of these chemicals have been evaluated for toxic effects like carcinogenicity. The U.S. National Toxicology Program contributes to this enterprise by conducting standardised chemical bioassays – exposure of rodents to a range of chemicals – to help identify substances that may have carcinogenic effects on humans. However, obtaining empirical evidence from such bioassays is expensive and usually too slow to cope with the number of chemicals that can result in adverse effects on human exposure. This has resulted in an urgent need for models that propose molecular mechanisms for carcinogenesis.

The data here come from tests conducted by the NTP. These have so far resulted in a data base of more than 300 compounds that have been shown to be carcinigenic or otherwise in rodents. Amongst other criteria, the chemicals have been selected on the basis of their carcinogenic potential – for example, positive mutagenicity tests – and on evidence of substantial human exposure. Using rat and mouse strains (of both genders) as predictive surrogates for humans, levels of evidence of carcinogenicity are obtained from the incidence of tumors on long-term (two years) exposure to the chemicals. The NTP assigns the following levels of evidence: CE, clear evidence; SE, some evidence; E, equivocal evidence; and NE, no evidence. For the experiments here, we are concerned only with overall levels of activity: +, if CE or SE; or -, otherwise.

A complete listing of all chemicals tested is available at the NTP Home Page *http://ntp-server.niehs.nih.gov*. The diversity of these compounds present a general problem to many conventional SAR techniques. Most of these, such as the regression-based techniques under the broad category called Hansch Analysis can only be applied to model compounds that have similar mechanisms of action. This "congeneric" assumption does not hold for the chemicals in the NTP data base, thus limiting the applicability of such methods. The Predictive Toxicology Evaluation project undertaken by the NIEHS aims to obtain an unbiased comparison of prediction methods by specifying compounds for blind trials. One such trial, PTE-1, is now complete. Complete results of NTP tests for compounds in the second trial, PTE-2, will be available by mid 1998.

ILP experiments with this data use Progol to predict carcinogenic activity for compounds in PTE-1 and PTE-2. These experiments use the obvious generic description of compounds consisting of atoms and their bond connectivities. These two predicates follow the same representation as used by us earlier in predicting the mutagenic activity of nitro-aromatic compounds. The background knowledge thus available for the ILP program was as follows.

Atom-bond description. These are ground facts representing the atom and bond structures of the compounds. The representation first introduced in our mutagenesis experiments is retained. These are Prolog translations of the output of the molecular modelling package QUANTA. Bond information consist of facts of the form *bond(compound,atom1,atom2,bondtype)* stating that *compound* has a bond of *bondtype* between the atoms *atom1* and *atom2*. Atomic structure consists of facts of the form *atm(compound,atom,element,atomtype,charg* stating that in *compound*, *atom* has element *element* of *atomtype* and partial charge *charge*.

Generic structural groups. This represents generic structural groups (methyl groups, benzene rings etc.) that can be defined directly using the atom and bond description of the compounds. Here we use definitions for 29 different structural groups, which expands on the 12 definitions used in the mutagenesesis studies. We pre-compute these structural groups for efficiency. An example fact that results is in the form *methyl(compound,atom_list)*, which states that the list of atoms *atom_list* in *compound* form a methyl group. Connectivity amongst groups is defined using these lists of atoms.

Genotoxicity. These are results of short-term assays used to detect and characterize chemicals that may pose genetic risks. These assays include the *Salmonella* assay, in-vivo tests for the induction of micro-nuclei in rat and mouse bone marrow etc. A full report available at the NTP Home Page lists the results from such tests in one of 12 types. Results are usually + or − indicating positive or negative response. These results are encoded into Prolog facts of the form *has_property(compound,type,result)*, which states that the *compound* in genetic toxicology *type* returned *result*. Here *result* is one of p (positive) or n (negative). In cases where more than 1 set of results are available for a given type, we have adopted the position of returning the majority result. When positive and negative results are returned in equal numbers, then no result is recorded for that test.

Mutagenicity. Progol rules from the earlier experiments on obtaining structural rules for mutagenesis are included. Mutagenic chemicals have often been found to be carcinogenic and we use all the rules found with Progol.

Structural indicators. We have been able to encode some of the carcinogenic structural alerts from the literature. At the time of writing this paper, the NTP proposes to make available nearly 80 additional structural attributes for the chemicals. Unfortunately, this is not yet in place for reuse in experiments here.