Theories for mutagenesis

The problem here is to predict the mutagenicity of a set of 230 aromatic and heteroaromatic nitro compounds. Mutagenicity is measured using the Ames test using S. typhimurium TA98. This data is based on the results in [1]. The prediction of mutagenesis is important as it is relevant to the understanding and prediction of carcinogenesis. Not all compounds can be empirically tested for mutagenesis, e.g. antibiotics. The compounds here are more heterogeneous structurally than any of those in other ILP datasets concerning chemical structure activity (see Figure 1).

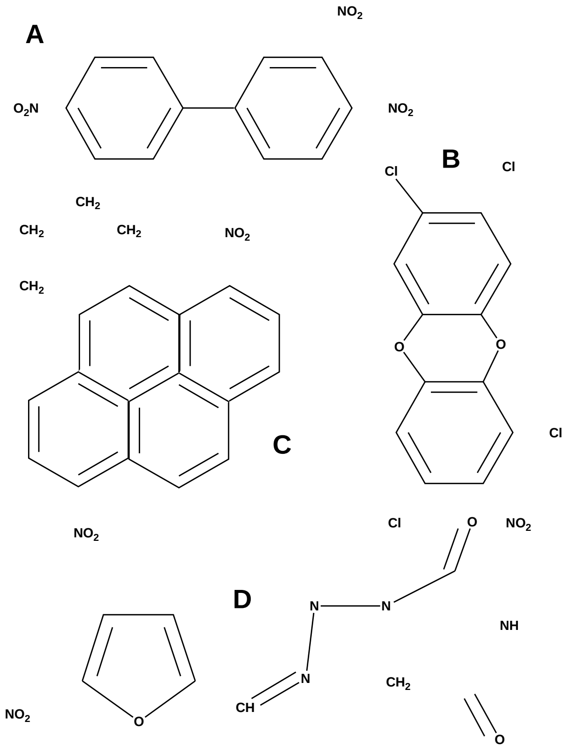

Figure 1: Examples of compounds used in the mutagenesis study. (A) 3,4,4'-trinitrobiphenyl (B) 2-nitro-1,3,7,8-tetrachlorodibenzo-1,4-dioxin (C) 1,6,-dinitro-9,10,11,12-tetrahydrobenzo[e]pyrene (D) nitrofurantoin

The data here comes from ILP experiments conducted with *Progol* [3]. Results of interest to the Machine Learning community are available in [4, 5, 6]. Relevant chemical results are in [2]. Of the 230 compounds, 138 have positive levels of log mutagenicity, these are labelled "active" and constitute the positive examples: the remaining 92 compounds are labelled "inactive" and constitute the negative examples. Of course, algorithms that are capable of full regression can attempt to predict the log mutagenicity values directly.

The original Debnath *et al* paper recognised two subsets of data: 188 compounds that could be fitted using linear regression, and 42 compounds that could not. For the *Progol* experiments, accuracies of theories constructed for the 188 compounds were estimated from a 10-fold cross-validation. The accuracy of theories for the 42 compounds were estimated by a leave-one-out procedure.

The ILP experiments used the obvious generic description of compounds consisting of atoms and their bond connectivities. The compounds were input into the molecular modelling program $QUANTA$ using its chemical editing facility. $QUANTA$ then automatically adds typing information and calculates approximate partial charges associated with each atom. The choice of QUANTA was arbitrary, any similar molecular modelling package would have been suitable. The result is that each compound is represented by a sets of facts of the form:

atm(127, 127_1, c, 22, 0.191)
bond(127, 127_1, 127_6, 7)

These two predicates give, in conjunction with ILP, the first completely generic method of describing molecular structure in drug design. The predicates also allow a straightforward definition of generic chemistry knowledge. This knowledge takes the form of a series of Prolog programs that define higher level chemical concepts (for example, ring structures).

In [1], four attributes are provided for analysis of the compounds. These can be used directly by both propositional and ILP learners. They are:

1. The hydrophobicity of the compound (termed logP);
2. The energy level of the lowest unoccupied moelcular orbital (termed LUMO);
3. A boolean attribute identifying compounds with 3 or more benzyl rings (termed indicator variable I1); and
4. A boolean attribute identifying a sub-class of compounds termed acenthryles (termed indicator variable Ia).

All data is as used in the *Progol* experiments, stored as a compressed TAR file. Within this, the Progol data is in files with a ".pl" suffix. Positive and negative examples for the subsets of 188 and 42 compounds are in the directories "188" and "42" respectively. All other information is in the directory "common". This includes the atom and bond information for each molecule, the values for the four attributes above, log mutagenicity, definition of ring concepts and a tabulation of these for ILP programs that require ground background knowledge. Also included are the language constraints used by *Progol*.

References

[1] Debnath, A.K. Lopez de Compadre, R.L., Debnath, G., Shusterman, A.J., and Hansch, C. (1991). Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. Correlation with molecular orbital energies and hydrophobicity. *J. Med. Chem.* 34:786-797.

[2] King, R.D., Muggleton, S.H., Srinivasan, A., and Sternberg, M.J.E. (1995) Representing molecular structure in structure activity relationships: The use of atoms and their bond connectivities to predict mutagenicity using inductive logic programming. Submitted to *J. Am. Chem. Soc.*

[3] Muggleton S.H. (1995) Inverse Entailment and Progol. *New Gen. Comput.* (to appear).

[4] Srinivasan, A., Muggleton, S., King, R.D., and Sternberg, M.J.E. (1994) Mutagenesis: ILP experiments in a non-determinate biological domain. *Proceedings of the Fourth Inductive Logic Programming Workshop.*

[5] Srinivasan, A., Muggleton, S.H., Sternberg, M.J.E., and King, R.D. (1995) Theories for mutagenicity: a study of first-order and feature based induction. *PRG-TR-8-95*, Oxford University Computing Laboratory.

[6] Srinivasan, A., Muggleton, S.H., Sternberg, M.J.E., and King, R.D. (1995) The effect of background knowledge in Inductive Logic Programing: a case study. *PRG-TR-9-95*, Oxford University Computing Laboratory.