Discovering drug structure-activity rules

A central concern of chemistry is understanding the relationships between chemical structure and activity. In most cases, these relationships cannot be derived solely from physical theory and experimental evidence is essential. Such empirically derived relationships are called Structure Activity Relationships (SARs). In a typical SAR problem, a set of chemicals of known structure and activity are given, and the problem is to construct a predictive theory relating the structure of a compound to its activity. This relationship can them be used to select for structures with high or low activity. Typically, knowledge of such relationships form the basis for devising clinically effective, non-toxic drugs.

The data here concerns the classic drug design problem of inhibition of E. Coli Dihydrofolate Reductase by pyrimidines and triazines. Pyrimadine compounds are antibiotics. They act by inhibiting Dihydrolate Reductase, an enzyme on the pathway to forming DNA. They inhibit bacterial forms of the enzyme preferentially to the human form and therefore kill bacteria. Further details on using ILP for this problem are available in [1, 2].

The pyrimidine problem is as follows. The chemicals structures of all compounds used to induce the SAR can be considered to have a common template (see Figure 1). To this template, chemical groups can be added at three possible substitution positions, 3, 4, and 5. A chemical group is an atom or set of structurally connected atoms that can be substituted together as a unit and have well defined chemical properties.

The existence of a template with only three possible substitution positions gives the pyrimidine problem a relatively small structural component. The chemical structure of the example compound in Figure 1 is adequately represented by a Prolog fact of the form:

struc(d55, cl, nh2 ch3)

which is intended to represent that drug 55 has a chlorine atom substituted at position 3, a amino group (NH_2) group substituted at position 4, and a methyl (CH_3) group at position 5.

In [1], the authors use 9 integer-valued attributes to represent chemical properties of the substituents. These were encoded as predicates of the form:

polar(br, polar3)

which states that a bromine atom has a polarity of value 3

Positive examples are pairs of drugs, the activity of one being known to be higher than the other. For example:

great(d1,d2)

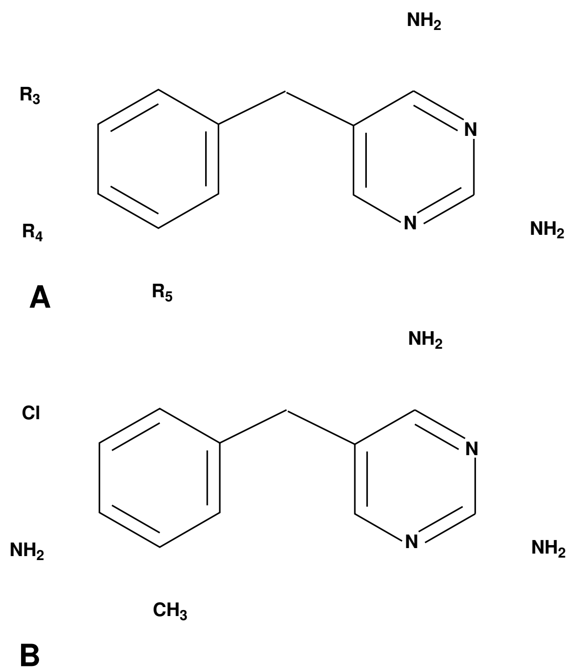

Figure 1: Chemical structure of pyramidine compounds. (A) Generic template for all compounds in the study; and (B) An example compound: $3-Cl, 4-NH_2$, $5-CH_3$

states that the E. Coli Dihydrofolate Reductase inhibition by drug $d1$ is higher than that by $d2$.

A more difficult problem concerns the inhibition of E. Coli Dihydrofolate reductase by triazines. Triazines act as anti-cancer agents by inhibiting the enzyme Dihydrolate Reductase. They act by preferentially inhibiting reproducing cells. Like the pyrimidines, the triazines can also be considered to have a common template structure (see Figure 2). However, the chemical groups substituted onto the template are much more complicated than the pyrimidine drugs. Further, many of the substituting groups can more naturally be considered as sub- templates with substitutions. There are seven (7) regions where a substituent might be present: the 2, 3, and 4 positions of the phenyl ring as shown in Figure 2. Each substituent can in turn, itself contain a ring structure. In this case, further substitutions are possible into positions 3 and 4 of these rings.

The first-order representation of the triazines is best explained using an example. The example compound in Figure 2 is represented by the following Prolog facts:

```
struc3(d217, cl, absent).
struc4(d217, (ch2)24, subst14).
subst(subst14, so2f, cl).
```

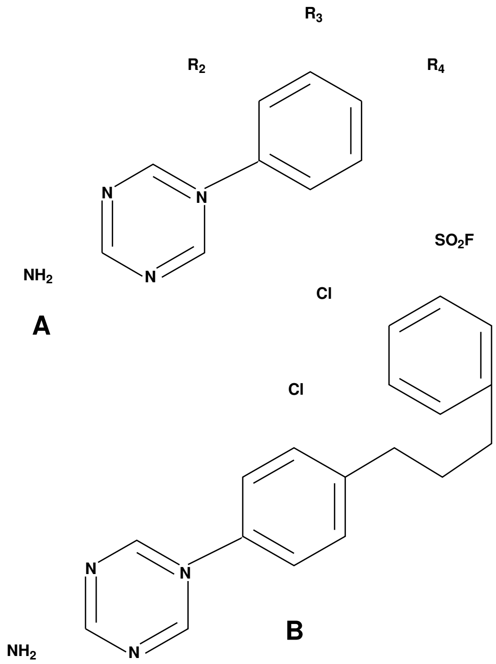

Figure 2: Chemical structure of triazine compounds. (A) Generic template for all compounds in the study; and (B) An example compound: $3 - Cl$, $4 - (CH_2)_2C_6H_3 - 4 - Cl$, $5 - CH_3$

The first clause represents substitutions at position 3 on the basic template: a Cl is present and there is an absence of a further phenyl ring. The second clause represents substitutions at position 4 on the basic template: there is a $(CH_2)4$ bridge to a second phenyl ring (implicit in the representation). This second phenyl ring has a $S0_2F$ group substituted at position 3 and a Cl group substituted at position 4. This is represented using the linker constant subst14 to the third clause. There is no substitution at position 2 on the basic template. Each of the chemical groups had 10 attributes, 9 of which were the same as used in the study of pyrimidines. One further attribute was added to capture flexibility of a substituent. The degree of flexibility is represented by one of 9 values.

The data files are stored in compressed TAR format. within that, they are as used in the original *Golem* [3] experiments. That is, background knowledge files have a ".b" suffix, positive example files have a ".f" suffix, and negative example files have a ".n" suffix.

References

[1] King R.D, Muggleton S., and Sternberg M.J.E. (1992). Drug design by machine learning: The use of inductive logic programming to model the structure-

activity relationships of trimethoprim analogues binding to dihydrofolate reductase. *Proc. of the National Academy of Sciences*, 89(23):11322–11326,

[2] King, R.D., Srinivasan, A. and Sternberg, M.J.E. (1995). Relating chemical activity to structure: an examination of ILP successes. *New Gen. Comput.* (to appear).

[3] Muggleton S.H. and Feng C. (1990). Efficient induction of logic programs. *Proceedings of the First Conference on Algorithmic Learning Theory*, Tokyo.