Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 11322–11326, December 1992
Biophysics

Drug design by machine learning: The use of inductive logic programming to model the structure–activity relationships of trimethoprim analogues binding to dihydrofolate reductase

(artificial intelligence/enzyme activity/protein modeling/active sites)

ROSS D. KING*, STEPHEN MUGGLETON†, RICHARD A. LEWIS‡§, AND MICHAEL J. E. STERNBERG‡¶

*Department of Statistics, Strathclyde University, Glasgow, G1 1XH, United Kingdom; †Turing Institute, George House, 36 North Hanover Street, Glasgow, G1 2AD, United Kingdom; and ‡Biomolecular Modelling Laboratory, Imperial Cancer Research Fund, 44 Lincoln's Inn Fields, London WC2A 3PX, United Kingdom

Communicated by David Phillips, August 7, 1992 (received for review December 13, 1991)

ABSTRACT **The machine learning program GOLEM from the field of inductive logic programming was applied to the drug design problem of modeling structure–activity relationships. The training data for the program were 44 trimethoprim analogues and their observed inhibition of *Escherichia coli* dihydrofolate reductase. A further 11 compounds were used as unseen test data. GOLEM obtained rules that were statistically more accurate on the training data and also better on the test data than a Hansch linear regression model. Importantly machine learning yields understandable rules that characterized the chemistry of favored inhibitors in terms of polarity, flexibility, and hydrogen-bonding character. These rules agree with the stereochemistry of the interaction observed crystallographically.**

The design of a potent pharmaceutical agent from a lead compound is often based on an understanding of the quantitative structure–activity relationship (QSAR) in a related series of ligands (e.g., refs. 1–3).

A standard method for modeling a QSAR was proposed by Hansch and coworkers (4–6), in which the physicochemical properties of a series of similar compounds are linked by an empirical equation to their biological activity. However the equation gives little insight into the structure of the binding site. Recently, neural network models have been applied to QSAR with some success (7), but the design of the network is highly subjective and the numerical results are difficult to interpret.

Pattern recognition methods, such as principal component analysis (8), are widely used to identify the chemical properties that contribute most to the activity of a compound. An alternative pattern recognition method with potential advantages for QSAR is machine learning. Machine learning methods are nonparametric and nonlinear and work best when using human understandable symbols to represent a problem. Thus in drug design the concepts used, such as size, polarity, and flexibility, relate naturally to stereochemistry. The use of such symbols has two potential advantages over the numerical representation. First, the problem can be set up and changed more easily because the designer can work in comprehensible terms. Background knowledge, in particular the stereochemistry of the compounds, can be directly added, whereas in a statistical method it would typically be represented by some form of prior probabilities or in a neural net by connection weights and topology. Second and more importantly, the production of humanly comprehensible rules from a machine learning system allows the rules to be checked for consistency with existing knowledge and opens the possibility that the rules may provide fresh insight.

In this paper we apply the machine learning program GOLEM (9) from the newly developed field of inductive logic programming (ILP) (10) to QSAR. In the development of methodologies, it is advantageous to consider systems for which atomic structural information of the drug–receptor complex is available. An ideal system is the complex formed between analogues of the drug trimethoprim and the enzyme dihydrofolate reductase (DHFR) from *Escherichia coli*, which has been studied crystallographically (11, 12). Thus one can compare the predicted QSAR models with the x-ray stereochemistry of interaction. These compounds have been studied by Hansch *et al.* (6) and so provide an ideal system to compare the performance of machine learning with the Hansch method.

METHODS

Data. The study was performed with a training set of 44 trimethoprim analogues (Table 1 and Fig. 1*A*) from Hansch *et al.* (6) and a testing set of 11 further cogeners (Table 1) from Roth and coworkers (14, 15) (Table 1). Biological activities were measured as $\log(1/K_i)$, where K_i is the equilibrium constant for the association of the drug to DHFR.

GOLEM. GOLEM (9) is a program for machine learning by ILP. The ILP methodology (10) was chosen because it is designed specifically to learn relationships between objects (e.g., molecular structures). In ILP, each language is a subset of first-order predicate calculus, which is expressive enough to describe most mathematical concepts and, having a strong link with natural language, leads to ease of comprehension. GOLEM is written in the programming language C but implements predicate logic in the language Prolog.

GOLEM takes as input positive examples, negative examples, and background knowledge described as Prolog facts. It produces as output Prolog rules that are generalizations of the examples. Observations are collected from the outside world (the activities of trimethoprim analogues). These are then combined by an ILP program with background knowledge (the stereochemistry of the compounds) to form inductive hypotheses (rules relating the structure of an analogue with its activity). These rules are then experimentally tested on additional data. If experimentation leads to high confidence in the validity of the hypotheses, the rules are added to the

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: QSAR, quantitative structure–activity relationship; ILP, inductive logic programming; DHFR, dihydrofolate reductase.

§Present address: Rhone-Poulenc Rorer, Dagenham Research Centre, Rainham Road South, Dagenham Essex RU10 7XS, United Kingdom.
¶To whom reprint requests should be addressed.

Table 1. Predicted and observed activity of trimethoprim analogues

Substituent	log $(1/K_{i,app})$	Observed rank	Rank by Golem	Rank by Hansch	Substituent	log $(1/K_{i,app})$	Observed rank	Rank by Golem	Rank by Hansch
Training set									
3,5-$(OH)_2$	3.04	1	17	2	4-$NHCOCH_3$	6.89	30.5	25	12
4-$O(CH2)_6CH_3$	5.60	2	4.5	4	3-OSO_2CH_3	6.92	32	33	25
4-$O(CH_2)_5CH_3$	6.07	3	4.5	10	3-OCH_3	6.93	33	36	38
H	6.18	4	1	6.5	3-Br	6.96	34	37	35.5
4-NO_2	6.20	5	7.5	20.5	3-NO_2, 4-$NHCOCH_3$	6.97	35	34	37
3-F	6.23	6	6	6.5	3-$OCH_2C_6H_5$	6.99	36	35	31
3-$O(CH_2)_7CH_3$	6.25	7	15	6.5	3-CF_3	7.02	37	32	19
3-CH_2OH	6.28	8	2	16	3,4-$(OCH_2CH_2OCH_3)_2$	7.22	38	39	40
4-NH_2	6.30	9	7.5	3	3-I	7.23	39	38	33
3,5-$(CH_2OH)_2$	6.31	10	3	23	3-CF_3, 4-OCH_3	7.69	40	41.5	39
4-F	6.35	11	9.5	6.5	3,4-$(OCH_3)_2$	7.72	41	41.5	41
3-$O(CH_2)_6CH_3$	6.39	12	16	11	3,5-$(OCH_3)_2$, 4-$O(CH_2)_2OCH_3$	8.35	42	43	43
4-$HCH_2CH_2OCH_3$	6.40	13	20	20.5	3,5-$(OCH_3)_2$	8.38	43	40	42
4-Cl	6.45	14	12	17.5	3,4,5-$(OCH_3)_3$	8.87	44	44	44
3,4-$(OH)_2$	6.46	15	18	1	*Testing set*				
3-OH	6.47	16	13	9	3,5-$(CH_3)_2$, 4-OCH_3	7.56	40 (1)	52.5 (9)	45.5 (4.5)
4-CH_3	6.48	17	9.5	17.5	3-Cl, 4-NH_2, 5-CH_3	7.74	43 (2)	40 (2)	44 (3)
3-$OCH_2CH_2OCH_3$	6.53	18	21	34	3,5-$(CH_3)_2$, 4-OH	7.87	44.5 (3.5)	40 (2)	41 (1)
3-$CH_2O(CH_2)_3CH_3$	6.55	19	24	35.5	3,5-Cl_2, 4-NH_2	7.87	44.5 (3.5)	40 (2)	45.5 (4.5)
3-OCH_2CONH_2	6.57	20.5	14	13	3,5-Br_2, 4-NH_2	8.42	48 (5)	44 (4)	53 (10)
4-OCF_3	6.57	20.5	19	22	3,5-$(OCH_3)_2$, 4-$OCH_2C_6H_5$	8.57	49 (6)	52.5 (9)	52 (9)
3-CH_2OCH_3	6.59	22	28	29.5	3,5-$(OCH_3)_2$, 4-CH_3	8.82	50.5 (7.5)	47.5 (5.5)	54.5 (11)
3-Cl	6.65	23	30.5	29.5	3,5-$(OCH_3)_2$, 4-$O(CH_2)_7CH_3$	8.82	50.5 (7.5)	52.5 (9)	43 (2)
3-CH_3	6.70	24	30.5	27.5	3,5-$(OCH_3)_2$, 4-$O(CH_2)_5CH_3$	8.85	52 (9)	52.5 (9)	49 (7)
4-$N(CH_3)_2$	6.78	25	22.5	27.5	3,5-I_2, 4-OCH_3	8.87	54.5 (10.5)	52.5 (9)	50 (8)
4-Br	6.82	27	11	24	3,5-I_2,4-OH	8.87	54.5 (10.5)	47.5 (5.5)	47 (6)
4-OCH_3	6.82	27	22.5	26					
3-$O(CH_2)_3CH_3$	6.82	27	29	32					
3-$O(CH_2)_5CH_3$	6.86	29	26	14					
4-$O(CH_2)_3CH_3$	6.89	30.5	27	15					

The observed value of the affinity is expressed as $\log(1/K_{i,app})$. The first 44 drugs were used in the training set, and the observed rank ranged from 1 to 44. The final 11 drugs were the testing set; the observed rank is the rank in the 55 drugs, and the number in parentheses is the rank for the 11. The rank corresponding rank values obtained by machine learning (GOLEM) and by the application of the Hansch equation are in the subsequent columns. GOLEM uses paired comparisons, which are then converted to rankings by the algorithm of David (13).

background knowledge. The method of generalization is based on the logical idea of "relative least general generalization."

The basic algorithm used in GOLEM is as follows. First, it takes a random sample of pairs of positive examples. In this application, this will be a set of pairs of compared drugs chosen randomly from the set of all examples represented (see below). For each of these pairs, GOLEM computes the set of all properties that are common to both examples. These properties are then made into a rule that is true of both the examples in the pair under consideration. Having built such a rule for all chosen pairs of examples, GOLEM takes each rule and computes the number of examples that rule could be used to predict. GOLEM chooses the rule that predicts the most true examples while predicting less than a predefined threshold of false examples. Having found the rule for the best pair, GOLEM then takes a further sample of as yet unpredicted examples and forms rules that express the common properties of this pair together with each of the individual residues in the sample. Again the rule that gives best predictions on the training set is chosen. The process of sampling and rule building is continued until no improvement in prediction is produced.

GOLEM avoids over-fitting the data by using the "minimal description length" as implemented in the compression model of Muggleton *et al.* (16).

Application of GOLEM to the QSAR of the Trimethoprim Series. To apply GOLEM to the QSAR of the trimethoprim series, the data has to be converted into a form suitable for GOLEM. QSAR is generally a regression problem, in which a real number is predicted from the description of a compound. However, GOLEM is designed to carry out classification (discrimination) tasks in which a small number of discrete classes are predicted. This difficulty is bypassed by considering pairs of drugs and comparing their activities (pairs where the activities are equal or within the margin of experimental error are discarded). The output is a set of rules that decides which of a pair of drugs has higher activity. Paired comparisons are then converted to a ranking by the method of David (13).

The input to GOLEM is three types of facts: positive, negative, and background. The positive examples are the paired examples of greater activity—e.g.,

great(d20, d15).

which states that drug no. 20 has higher activity than drug no. 15. The negative examples are the paired examples of lower activity.

The background facts are the chemical structures of the drugs and the properties of the substituents. Chemical structure is represented in the form:

struc(d35, NO_2, $NHCOCH_3$, H).

which states that drug no. 35 has NO_2 substituted at position 3, $NHCOCH_3$ substituted at position 4, and no substitution at position 5. In addition, if either position 3 or 5 has no substitution, as in drug no. 35, the position with no substitution is assumed to be position 5 (this is used in ref. 6).

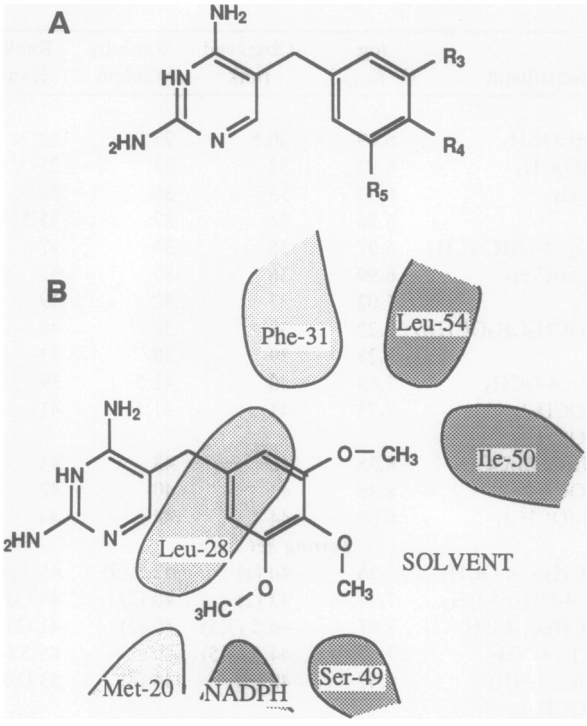

FIG. 1. (*A*) Structure of trimethoprim analogues. (*B*) Cartoon of the interaction of trimethoprim with DHFR from x-ray structures (11, 12). Faint stippling indicates that the residue lies below the plane of the phenyl ring; darker stippling indicates that the atoms are above.

Chemical properties were then assigned heuristically to substituents (Table 2). The properties, chosen to make the approach generally applicable to drug-design problems, are polarity, size, flexibility, hydrogen-bond donor, hydrogen-bond acceptor, π donor, π acceptor, polarizability, and σ effect. Each of the 24 nonhydrogen substituents was given an integer value for each of these properties. This was represented by using different predicates for each property and value; for example,

polar(Br, polar3).

states that Br has polarity of value 3.

Information was also given about the relative values of these properties for the substituent groups; for example,

polar(OCF_3, polar4).

Finally information was given about the relative values of the properties compared to fixed values; for example,

great0_polar(polar1).

states that a polarity of 1 is greater than a polarity of 0.

The input to GOLEM was 871 positive facts, 871 negative facts, and 2976 facts in the background information. The run time was about 30 central processing unit minutes on a SUN SparcStation 1 per rule.

RESULTS

GOLEM derived nine rules that predict the relative activities of two drugs. Table 3 lists the machine-learned rules in Prolog syntax together with an English interpretation. The coverage indicates the number of pairs of relationship that are correctly and incorrectly represented. Each rule relates the relative activities of two drugs (A and B) and identifies the chemical properties of substituents that yield a drug of higher activity.

Table 2. Chemical properties of substituents

Group	PL	SZ	FLX	H-D	H-A	P-D	P-A	POL	SIG
H	—	—	—	—	—	—	—	—	—
OH	3	1	0	2	2	2	0	1	2
$O(CH_2)_6CH_3$	2	5	7	0	1	1	0	1	1
$O(CH_2)_5CH_3$	2	5	6	0	1	1	0	1	1
NO_2	5	2	0	0	0	0	2	0	3
F	5	1	0	0	1	0	0	0	5
$O(CH_2)_7CH_3$	2	5	8	0	1	1	0	1	1
CH2OH	2	2	2	2	2	0	0	1	0
NH_2	3	1	0	2	0	2	0	0	1
$OCH_2CH_2OCH_3$	2	3	4	0	1	1	0	1	1
Cl	3	1	0	0	0	0	0	1	3
CH_3	0	1	0	0	0	0	0	1	0
$CH_2O(CH_2)_3CH_3$	0	4	6	0	0	0	0	1	0
OCH_2CONH_2	3	3	2	1	1	1	0	0	1
OCF_3	4	3	1	0	0	0	2	0	3
CH_2OCH_3	0	2	3	0	1	0	0	1	0
$N(CH_3)_2$	1	2	0	0	1	2	0	1	1
Br	3	1	0	0	0	1	0	2	3
$O(CH_2)_3CH_3$	2	3	4	0	1	1	0	1	1
$NHCOCH_3$	3	2	0	1	1	1	0	1	1
OSO_2CH_3	4	2	1	0	0	0	1	2	3
$OCH_2C_6H_5$	2	4	2	0	1	1	0	1	1
CF_3	3	1	0	0	0	0	0	0	3
I	3	1	0	0	0	1	0	3	3
OCH_3	2	2	1	0	1	1	0	1	1

PL, polarity indicates the amount of residual charge on the α and β atoms of the substituent; SZ, size is a measure of the extended volume of the group; FLX, flexibility is assigned from the number of rotatable bonds; H-D and H-A indicate the presence and strength of hydrogen-bonding acceptors and donors; P-D and P-A indicate the presence and strength of π-acceptors and π-donors. POL indicates polarizability of the molecular orbitals, and SIG is its σ-property.

Table 1 gives the performance on training data of 44 compounds. The prediction from machine learning gave a rank correlation with the observed order of 0.916 [using the Spearman method (17)] (Fig. 2*A*). As a benchmark, the Hansch equation had a rank correlation of 0.794 (Fig. 2*B*). The significance of the difference in these rankings was evaluated by Fisher's z transformation (17). The value $z = 2.18$ is significant at the 5% level and almost significant at the 1% level ($P = 0.985$).

A better test of a prediction method is its performance on data not used in developing the algorithm. The structure and activity of 11 trimethoprim analogues not used in the original paper (6) on which the Hansch equation was derived was used as a test set for the two approaches. A ranking of the 11 additional drugs relative to all 55 drugs was obtained by (*i*) forming all paired comparisons involving the 11 additional drugs, (*ii*) adding these to the predicted results of the paired comparisons of the original 44 drugs, and (*iii*) producing a ranking from all the paired comparisons. From this ranking of 55 drugs, a rank order for the 11 additional drugs was extracted (Table 1), and this was compared by a rank correlation coefficient to the observed order. The rank correlation for the 11 additional drugs by machine learning was 0.457 compared to 0.415 for the Hansch method (Fig. 2). The Fisher z value is 0.10, which is not significant ($P = 0.540$) and reflects the similar rank correlations obtained on a small test set. Thus on the test set, the machine learning is as accurate as the regression approach of Hansch. Both methods predict well that the tests drugs have high activity (Fig. 2).

A further test of the GOLEM approach was a cross-validation study in which 44 of the 55 drugs were chosen at random as a training set with the remaining 11 as the test data. The resultant Spearman rank correlation coefficients for the training sets are similar to those for the main trial.

Table 3. Rules for QSAR derived by machine learning

Rule 3.1 - (coverage 119/0 train: 105/0 test)

```
great(A,B) :- struc(A, D, E, F), struc(B, h, C, h), h_donor(D,
              h_don0), pi_donor(D, pi_don1), flex(D, G),
              less4_flex(G).
Drug A is better than drug B if
  drug B has no substitutions at positions 3 and 5 and
  drug A at position 3 has hydrogen donor = 0 and
  drug A at position 3 has π-donor = 1 and
  drug A at position 3 has flexibility <4.
```

Rule 3.2 - (coverage 244/71 train: 248/4 test)

```
great(A, B) :- struc(A, C, D, E), struc(B, F, h, G), not
               except3.2(A, B).
except3.2(A, B) :- struc(A, C, D, h), struc(B, E, h, F),
    h_donor(E, h_don0).
Drug A is better than drug B if
  drug B has no substitution at position 4 unless
    drug A has no substitution at position 5 and
    drug B at position 3 has hydrogen donor = 0.
```

Rule 3.3 - (coverage 102/13 train: 33/0 test)

```
great(A, B) :- struc(A, G, H, I), struc(B, C, h, D), pi_donor(C,
               pi_don0), polar(C, E), great0_polar(E), h_acceptor(C,
               F), great0_h_acc(F).
Drug A is better than drug B if
  drug B has no substitutions at position 4 and
  drug B at position 3 has π-donor = 0 and
  drug B at position 3 has polarity >0 and
  drug B at position 3 has hydrogen acceptor >0.
```

Rule 3.4 - (coverage 129/2 train: 126/0 test)

```
great(A, B) :- struc(A, C, D, E), struc(B, G, h, h), h_donor(C,
               h_don0), pi_donor(C, pi_don1), flex(C, F),
               less4_flex(F), polarisable(G, H), less3_polari(H).
Drug A is better than drug B if
  drug B has no substitutions at positions 4 and 5 and
  drug B at position 3 has polarizability <3 and
  drug A at position 3 has hydrogen donor = 0 and
  drug A at position 3 has π-donor = 1 and
  drug A at position 3 has flexibility <4.
```

Rule 3.5 - (coverage 84/0 train: 52/0 test)

```
great(A, B) :- struc(A, C, D, E), struc(B, F, h, h), size(C, size2),
               h_donor(C, h_don0), polarisable(F, polari1), polar(C,
               G), great0_polar(G).
Drug A is better than drug B if
  drug B has no substitutions at positions 4 and 5 and
  drug B at position 3 has polarizability = 1 and
  drug A at position 3 has size = 2 and
  drug A at position 3 has hydrogen donor = 0 and
  drug A at position 3 has polarity >0.
```

Rule 3.6 - (coverage 29/0 train: 16/0 test)

```
great(A, B) :- struc(A, C, D, E), struc(B, F, h, h),
               h_donor(C,h_don0), polarisable(C, polari1), flex(F,
               G), flex(C, H), great_flex(G, H), great6_flex(G).
Drug A is better than drug B if
  drug B has no substitutions at positions 4 and 5, and
  drug B at position 3 has flexibility >6 and
  drug A at position 3 has polarizability = 1 and
  drug A at position 3 has hydrogen donor = 0 and
  drug A at position 3 is less flexible than drug B at position 3.
```

Rule 4.1 - (coverage 289/72 train: 99/0 test)

```
great(A, B) :- struc(A, D, E, F), struc(B, h, C, h), not
               except4.1(A,B).
  except4.1(A,B) :- struc(B, h, C, h), size(C, size3).
  except4.1(A,B) :- struc(B, h, C, h), size(C, size2),
    h_acceptor(C,h_acc1).
Drug A is better than drug B if
  drug B has no substitutions at positions 3 and 5 unless
    drug B at position 4 has size = 3 or
    drug B at position 4 has size = 2 and hydrogen acceptor = 1.
```

Rule 4.2 - (coverage 187/2 train: 193/2 test)

```
great(A, B) :- struc(A, E, F, G), struc(B, C, D, h), not_h(E),
               polar(F, polar2).
Drug A is better than drug B if
  drug B has no substitution at position 5 and
  drug A has a substitution at position 3 and
  drug A at position 4 has polarity = 2.
```

Rule 3 and 4.1 - (coverage 85/0 train: 55/0 test)

```
great(A,B) :- struc(A,C,D,E), struc(B,H,I,h), h_donor(C,h_don0),
              polar(C,F), less5_polar(F), size(C,G), less5_size(G),
              polarisable(I,J), less2_polari(J), sigma(I,K),
              great1_sigma(K).
Drug A is better than drug B if
  drug B has no substitution at position 5 and
  drug B at position 4 has polarizability <2 and
  drug B at position 4 has σ >1 and
  drug A at position 3 has hydrogen donor = 0 and
  drug A at position 3 has polarity <5 and
  drug A at position 3 has size <5.
```

The rules are first given as Prolog clauses in which ":-" is a logical implication and a comma is a logical "and". Then an exact interpretation is given. The rules have been classified into those primarily relating to substituent 3 (rules 3.1–3.6), to substituent 4 (rules 4.1 and 4.2), and to both positions (rules 3 and 4.1).

DISCUSSION AND CONCLUSIONS

The x-ray crystallographic structures of the trimethoprim–*E. coli* DHFR complex (11) and of the ternary complex (12) with NADPH have been solved (see Fig. 1*B*). In the ternary complex, the phenyl ring of trimethoprim is sandwiched in a hydrophobic cleft between Phe-31, Leu-28, and Met-20 on one side and Leu-54, Ile-54, Ser-49 and with the NADPH cofactor on the other. The aromatic ring is thus effectively buried while the environments of the 3, 4, and 5 substituents vary. The 4 (i.e., para) position is the most exposed to solvent while the meta positions (i.e., the 3 and 5 substituents) are restricted in size by the surrounding protein and cofactor atoms.

A main aim in using machine learning was to obtain rules that could provide insight into stereochemistry of drug–DHFR interactions. We examined the features that favor the better drugs (i.e., the properties of drug A in the rules). For

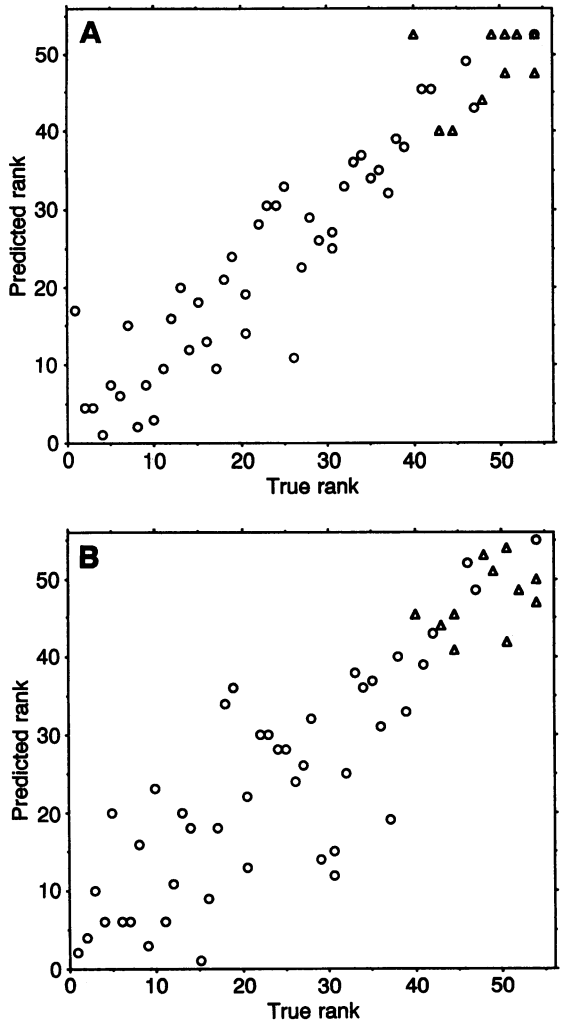

FIG. 2. Scattergram of the observed rank versus that predicted by Golem (*A*) and by Hansch (*B*). ○, Train; △, test.

3 and/or 5 positions a favorable substituent D is defined as

h_donor(D,h_don0),
pi_donor(D,pi_don1),
flex(D,G),less4_flex(G),
size(D,size2),
polar(D,V),great0_polar(V),
polarisable(D,polar1).

The properties are, therefore, not a hydrogen-bond donor, a π-donor of 1, flexibility <4, a size of 2, a polarity greater than zero, and a polarizability of 1. Only the methoxy (OCH_3) substituent satisfies these conditions. These principles are in keeping with the location of both meta sites (i.e., both 3 and 5 positions) in the crystal structures. Both meta sites are partially buried in a hydrophobic environment and hence have a restriction on size and flexibility. The absence of solvent at these sites might explain the requirement that the group should not be a hydrogen-bond donor. Substituents that are π-donors will force this group to lie in the plane of the aromatic ring, and this has been suggested as a requirement for a favorable drug (15). Finally, because both meta positions have similar chemical locations in DHFR, one cannot decide whether the rules in Table 3 for a 3 position on the chemical compound (Fig. 1*A*) relate exclusively to the upper meta position or exclusively to the lower meta position or to both locations in the three-dimensional location (Fig. 1*B*).

The only positive feature for the 4 position is that it should have a polarity of 2. This property is consistent with a site that is exposed to solvent and should be polar. Matthews *et al.* (11) proposed that the oxygen of the methoxy group might form a hydrogen bond with a neighboring water molecule. In addition, the rules suggest that each of the 3, 4, and 5 positions should not be hydrogen. This is in keeping with the suggestion (14) that an important role of the 4 position is to force the meta substituents away from the 4 position toward the 2 and 6 positions.

For drug design, we have shown that machine learning can yield rules that model a QSAR of a series of DHFR inhibitors better than one of the standard methods widely used. In addition one automatically derives a stereochemical description of the drug–receptor interaction. In another recent study, GOLEM (18) has produced predictions of the secondary structure of α/α proteins of 80% accuracy. We consider that this demonstrates the wide-ranging potential of ILP in the domain of modeling biological information.

We thank Drs. B. Roth and C. Beddell for helpful comments.

1. Goodford, P. J. (1984) *J. Med. Chem.* **27**, 557–564.
2. Dean, P. M. (1987) *Molecular Foundations of Drug-Receptor Interactions* (Cambridge Univ. Press, Cambridge, U.K.).
3. Marshall, G. R. & Cramer, R. D. (1988) *Trends Pharmacol. Sci.* **9**, 285–289.
4. Hansch, C., Maloney, P. P., Fujita, T. & Muir, R. M. (1962) *Nature (London)* **194**, 178–180.
5. Hansch, C. (1969) *Acc. Chem. Res.* **2**, 232–239.
6. Hansch, C., Li, R.-I., Blaney, J. M. & Langridge, R. (1982) *J. Med. Chem.* **25**, 777–784.
7. Andrea, T. A. & Kalayeh, H. (1991) *J. Med. Chem.* **34**, 2824–2836.
8. Wold, S., Dunn, W. J. & Hellberg, S. (1984) in *Drug Design: Fact or Fantasy?*, eds. Jolles, G. & Wooldridge, K. R. H. (Academic, London), pp. 95–115.
9. Muggleton, S. & Feng, C. (1990) in *Proceedings of the First Conference on Algorithmic Learning Theory*, eds. Arikawa, S., Goto, S., Ohsuga, S., and Yokomori, T. (Jpn. Soc. Artificial Intelligence, Tokyo), pp. 368–381.
10. Muggleton, S. (1991) *New Gener. Comp.* **8**, 295–318.
11. Matthews, D. A., Bolin, J. T., Burridge, J. M., Filman, D. J., Volz, K. W., Kaufman, B. T., Beddell, C. R., Champness, J. N., Stammers, D. K. & Kraut, J. (1985) *J. Biol. Chem.* **260**, 381–391.
12. Champness, J. N., Stammers, D. K. & Beddell, C. R. (1986) *FEBS Lett.* **199**, 61–67.
13. David, H. A. (1987) *Biometrika* **74**, 432–436.
14. Roth, B., Aig, E., Rauckman, B. S., Strelitz, J. Z., Phillips, A. P., Ferone, R., Bushby, S. R. M. & Sigel, C. W. (1981) *J. Med. Chem.* **24**, 933–941.
15. Roth, B., Rauckman, B. S., Ferone, R., Baccanari, D. P., Champness, J. N. & Hyde, R. M. (1987) *J. Med. Chem.* **30**, 348–356.
16. Muggleton, S., Srinivasan, A. & Bain, M. (1992) in *Proceedings of 9th International Conference on Machine Learning* (Morgan–Kaufman, San Diego).
17. Kendall, M. & Stuart, A. (1977) *The Advanced Theory of Statistics* (Griffen, London).
18. Muggleton, S., King, R. D. & Sternberg, M. J. E. (1992) *Protein Eng.*, in press.