Journal of Computer-Aided Molecular Design, 8 (1994) 405–420
ESCOM

J-CAMD 250

Quantitative structure–activity relationships by neural networks and inductive logic programming. I. The inhibition of dihydrofolate reductase by pyrimidines

Jonathan D. Hirst*, Ross D. King and Michael J.E. Sternberg**

Biomolecular Modelling Laboratory, Imperial Cancer Research Fund, 44 Lincoln's Inn Fields, P.O. Box 123, London WC2A 3PX, U.K.

Received 4 January 1994
Accepted 5 March 1994

Key words: QSAR; Artificial intelligence; Neural networks; DHFR inhibitors

SUMMARY

Neural networks and inductive logic programming (ILP) have been compared to linear regression for modelling the QSAR of the inhibition of *E. coli* dihydrofolate reductase (DHFR) by 2,4-diamino-5-(substituted benzyl)pyrimidines, and, in the subsequent paper [Hirst, J.D., King, R.D. and Sternberg, M.J.E., J. Comput.-Aided Mol. Design, 8 (1994) 421], the inhibition of rodent DHFR by 2,4-diamino-6,6-dimethyl-5-phenyl-dihydrotriazines. Cross-validation trials provide a statistically rigorous assessment of the predictive capabilities of the methods, with training and testing data selected randomly and all the methods developed using identical training data. For the ILP analysis, molecules are represented by attributes other than Hansch parameters. Neural networks and ILP perform better than linear regression using the attribute representation, but the difference is not statistically significant. The major benefit from the ILP analysis is the formulation of understandable rules relating the activity of the inhibitors to their chemical structure.

INTRODUCTION

Recently, backpropagating neural networks have been applied to QSAR, using the Hansch description of molecules [1,2]. Analyses of the QSARs of 2,4-diamino-5-(substituted benzyl)pyrimidines [3] and of 2,4-diamino-6-dimethyl-5-phenyldihydrotriazines [4] as dihydrofolate reductase (DHFR) inhibitors and other work [5–7], have suggested that neural networks can perform better than traditional regression methods.

*Present address: Box 77, Department of Chemistry, Mellon Institute, 4400 Fifth Avenue, Pittsburgh, PA 15213, U.S.A.
**To whom correspondence should be addressed.
The ILP program GOLEM is freely available to academic users. The data in these studies has been deposited at the UCI Machine Learning Repository (anonymous ftp: ics.uci.edu) and at the GMD Machine Learning Archive (anonymous ftp: ftp.gmd.de).

Another computer learning method, inductive logic programming (ILP), has been used to model the QSAR of trimethoprim analogues binding to DHFR from *E. coli* [8]. Physicochemical attributes were assigned heuristically to substituents, and were chosen to make the approach generally applicable to drug design problems. While not significantly better than the traditional QSAR, this method produced rules that could provide insight into the stereochemistry of drug–DHFR interactions.

More evidence is required to assess these new methods properly. All comparative QSAR trials should:

- use randomly selected data for training and testing
- use cross-validation to avoid biases in the data
- compare the different methods on identical data
- test the significance of differences between methods
- examine more than one system
- use large data sets, for statistical sensitivity

This paper and the subsequent one [9] use these techniques to provide a thorough evaluation of the predictive capabilities of neural networks and ILP.

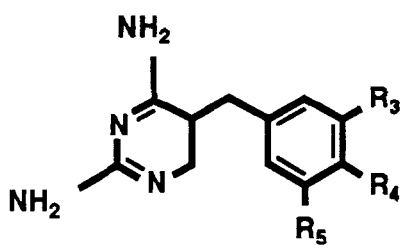
The insight into the drug–receptor interaction that the QSAR provides is another important criterion for the assessment. ILP is designed to produce understandable rules. The formulation of potentially insightful rules is facilitated by representing molecules by a set of physicochemical attributes instead of Hansch parameters. We have investigated whether similar information can be provided by linear regression and neural networks, using the attribute representation.

METHODS

Data

The data studied were 74 2,4-diamino-5-(substituted benzyl)pyrimidines (**I**) (Table 1). Biological activities have been measured by the association constant (log K_i) to DHFR from MB1428 *E. coli* [10]. These data have been extensively studied by QSAR methods, and there are also crystallographic studies of the complex formed between trimethoprim (2,4-diamino-5-(3,4,5-trimethoxybenzyl)pyrimidine) and DHFR from *E. coli* [11,12]. It is possible therefore, in this test case, to compare the QSAR models with the X-ray stereochemistry of interaction, although it must be stressed that the QSAR methods examined here are for application to drug design problems where there is no receptor structure.

The QSAR methods studied in this paper have been assessed using fivefold cross-validation, in which the 55 molecules (numbers 1–55) from the machine learning study [8] were randomly



I

TABLE 1
PYRIMIDINES USED IN THIS STUDY[a]

Index no.	Activity (log K_i)	Substituent		
		3-position	4-position	5-position
1	3.04	OH	H	OH
2	5.60	H	$O(CH_2)_6CH_3$	H
3	6.07	H	$O(CH_2)_5CH_3$	H
4	6.18	H	H	H
5	6.20	H	NO_2	H
6	6.23	F	H	H
7	6.25	$O(CH_2)_7CH_3$	H	H
8	6.28	CH_2OH	H	H
9	6.30	H	NH_2	H
10	6.31	CH_2OH	H	CH_2OH
11	6.35	H	F	H
12	6.39	$O(CH_2)_6CH_3$	H	H
13	6.40	H	$O(CH_2)_2OCH_3$	H
14	6.45	H	Cl	H
15	6.46	OH	OH	H
16	6.47	OH	H	H
17	6.48	H	CH_3	H
18	6.53	$O(CH_2)_2OCH_3$	H	H
19	6.55	$CH_2O(CH_2)_3CH_3$	H	H
20	6.57	OCH_2CONH_2	H	H
21	6.57	H	OCF_3	H
22	6.59	CH_2OCH_3	H	H
23	6.65	Cl	H	H
24	6.70	CH_3	H	H
25	6.78	H	$N(CH_3)_2$	H
26	6.82	H	Br	H
27	6.82	H	OCH_3	H
28	6.82	$O(CH_2)_3CH_3$	H	H
29	6.86	$O(CH_2)_5CH_3$	H	H
30	6.89	H	$O(CH_2)_3CH_3$	H
31	6.89	H	$NHCOCH_3$	H
32	6.92	OSO_2CH_3	H	H
33	6.93	OCH_3	H	H
34	6.96	Br	H	H
35	6.97	NO_2	$NHCOCH_3$	H
36	6.99	$OCH_2C_6H_5$	H	H
37	7.02	CF_3	H	H
38	7.22	$O(CH_2)_2OCH_3$	$O(CH_2)_2OCH_3$	H
39	7.23	I	H	H
40	7.69	CF_3	OCH_3	H
41	7.72	OCH_3	OCH_3	H
42	8.35	OCH_3	$O(CH_2)_2OCH_3$	OCH_3
43	8.38	OCH_3	H	OCH_3
44	8.87	OCH_3	OCH_3	OCH_3
45	7.56	CH_3	OH	CH_3
46	7.74	CH_3	OCH_3	CH_3
47	7.87	OCH_3	$O(CH_2)_5CH_3$	OCH_3
48	7.87	OCH_3	$O(CH_2)_7CH_3$	OCH_3
49	8.42	OCH_3	$OCH_2C_6H_5$	OCH_3
50	8.57	OCH_3	CH_3	OCH_3

TABLE 1 (continued)

Index no.	Activity (log K_i)	Substituent		
		3-position	4-position	5-position
51	8.82	I	OCH_3	I
52	8.82	I	OH	I
53	8.85	Br	NH_2	Br
54	8.87	Cl	NH_2	Cl
55	8.87	Cl	NH_2	CH_3
56	6.45	H	OH	H
57	6.60	H	OSO_2CH_3	H
58	6.84	OH	OCH_3	H
59	6.89	H	$OCH_2C_6H_5$	H
60	6.93	H	C_6H_5	H
61	7.04	CH_3	H	CH_3
62	7.13	OCH_2O	H	OCH_2O
63	7.16	$O(CH_2)_7CH_3$	OCH_3	H
64	7.20	OCH_3	$O(CH_2)_7CH_3$	OCH_3
65	7.41	OC_3H_7	H	OCH_3H_7
66	7.53	OCH_3	$OCH_2C_6H_5$	H
67	7.54	OCH_3	OH	H
68	7.66	$OCH_2C_6H_5$	OCH_3	H
69	7.71	OCH_3	$N(CH_3)_2$	OCH_3
70	7.77	OCH_3	$O(CH_2)_2OCH_3$	H
71	7.80	OSO_2CH_3	OCH_3	H
72	7.82	CH_2CH_3	CH_2CH_3	CH_2CH_3
73	7.94	OCH_3	OSO_2CH_3	H
74	8.18	OCH_3	Br	OCH_3

[a] Numbers 1–44 have been analysed by linear regression [14]; numbers 45–55 are from Roth and co-workers [17,18]; numbers 2–44 and 56–74 were used in a more recent linear regression study [13] and in a neural network analysis [3]. Six of the 25 have not been included, because the complete set of Hansch parameters for the substituents was not available at the time of this study.

divided into five equal sets, each containing 11 molecules (Table 2). Each of these sets was used as a test set, with the other four sets forming corresponding training sets of 44 molecules. Each of the 55 pyrimidines in the cross-validation study appears once only in only one of the test sets. The 19 more recently characterised derivatives [13] (numbers 56–74) were used as an additional test set. This division of the data maintains consistency with the earlier QSAR analyses [8,14].

Hansch parameters

Linear regression studies [13–15] have correlated the activity of the pyrimidines to the chemical properties of the 3-, 4- and 5-substituents of the phenyl ring. The activity was measured by log $1/K_i$, where K_i is the inhibition constant as experimentally assayed [10,16–18]. The chemical properties of the substituents were represented by the hydrophobic parameter π and the molar refractivity MR, where π is derived from partition coefficients between 1-octanol and water [19], and MR is related to the size of the substituent. π and MR values were taken from the literature. In previous analyses [3,8,13,14], MR values were scaled by 0.1 to make them equiscalar with the π values; furthermore, the MR values were truncated with an upper limit of 0.79. The truncated MR values have been used here to maintain consistency with previous studies.

TABLE 2
TRAINING AND TESTING SETS

Split number	Index numbers[a]
Set 1	11, 31, 34, 42, 20, 24, 30, 23, 37, 08, 39
Set 2	16, 54, 43, 19, 22, 10, 41, 33, 35, 01, 09
Set 3	14, 03, 55, 06, 04, 47, 28, 25, 27, 50, 07
Set 4	51, 49, 53, 46, 12, 21, 44, 52, 38, 15, 18
Set 5	36, 45, 26, 32, 02, 17, 40, 13, 48, 29, 05
Set 6	56–74
Training set 1	sets 2, 3, 4, 5
Training set 2	sets 1, 3, 4, 5
Training set 3	sets 1, 2, 4, 5
Training set 4	sets 1, 2, 3, 5
Training set 5	sets 1, 2, 3, 4

[a] The numbers in this column correspond to those in column 1 of Table 1.

Physicochemical attributes

The attribute representation, developed for the ILP approach [8], describes explicitly volume and electrostatic interactions encoded in the Hansch parameters. The substituents of the drugs were heuristically assigned attributes: polarity, size, flexibility, number of hydrogen-bond donors and acceptors, presence and strength of π-acceptors and π-donors, polarisability of the molecular orbitals, the σ-effect and branching. The attributes were assigned to reproduce general trends rather than precise values. The attributes, based on previous work [8], are unrefined, and an automatic and more rigorous assignment would need to be developed for general use of ILP. The main focus of our papers, however, is the evaluation of methods and not the exact nature of the representation of molecules in QSAR. Both are of importance, and by using the same representation for all the methods, we have sought to minimise the influence of deficiencies in the representation on the evaluation of the methods.

The size (SZ) of a substituent was based on the number of carbon, nitrogen and oxygen atoms it contains, with SZ = 0 for hydrogen, SZ = 1 for single atoms and SZ = 2 for substituents containing two to four C, N or O atoms. For larger substituents, the size assigned depended on the compactness as well as the number of C, N or O atoms, with the ranges as follows: SZ = 3 (3–6 C, N or O atoms); SZ = 4 (4–8); SZ = 5 (7–10); SZ = 6 (9–12); SZ = 7 (13); SZ = 8 (13–15).

The flexibility (FL) was equated with the number of freely rotatable carbon-carbon, carbon-nitrogen or carbon-oxygen bonds. The number of branch points provided the branching (BR) attribute. Substituents had between zero and three hydrogen-bond donors or acceptors (the strength of the possible hydrogen bond was also taken into account).

The influence of the substituent on the availability of electrons has been modelled using several attributes. The electron-releasing or -withdrawing effect of the substituent may be transmitted through either the σ-bonds or the π-bonds. High and low σ-effects imply high and low electronegativity, respectively. The σ-effect was assigned as follows: σ = 0 for CH_2X and C_6H_5; σ = 1 for I, NR_2, NHCOR and OR; σ = 2 for Br, OH and SO_2R; σ = 3 for CF_3, Cl, OCF_3 and NO_2; and σ = 5 for F. The attributes πA (π-acceptor) and πD (π-donor) describe the nature of the π-orbitals. Most substituents were π-acceptors (πA) of zero; the exceptions with πA = 1 were CF_3 and $CONR_2$, and those with πA = 2 were CN, NO_2 and OCF_3. π-Donors (πD) were assigned

as follows: $\pi D = 0$ for OCF_3, NO_2 and alkyl groups; $\pi D = 1$ for Br, Cl, I, NHCOR, OR and SO_2R; $\pi D = 2$ for C_6H_5, NR_2 and OH. Polarisability (PO) depends on how loosely or tightly electrons are held by the substituent. In the halogen family, for example, polarisability increases in the order F < Cl < Br < I. Polarisability was assigned as follows: PO = 0 for CF_3, F, H, NH_2, NO_2, OCF_3 and OCH_2OX; PO = 1 for CH_2X, Cl, $CONR_2$, NR_2, NHCOX and OR; PO = 2 for Br, CN, O and SO_2X; PO = 3 for C_6H_5 and I. Polarity (PL) was based on the amount of residual charge on the α and β atoms of the substituent; PL = 0 for C_6H_5; PL = 1 for alkyl groups; PL = 2 for CH_2CONR, CH_2O and NR_2; PL = 3 for Br, CF_3, $CONR_2$, OR and Cl; PL = 4 for OCF_3, OH and SO_2X; PL = 5 for F and NO_2.

Each property of each substituent was represented by an integer value, as required by the inductive logic program GOLEM [20]. Neural network algorithms do not cope with inputs that are exactly zero, so all data have been rescaled to lie between 0.1 and 0.9 for the neural network and, for consistency, the linear regression analyses. The pyrimidines are substituted at up to three positions, and so each molecule is represented by 27 attributes (three positions and nine attributes per position). The attributes assigned to 114 different fragments, used in this paper and the subsequent one [9], are given in Table 3.

Linear regression

The Hansch parameters π_3, π_4, π_5, MR_3, MR_4 and MR_5 were assigned to the molecules in the five cross-validation training sets. To provide a benchmark on the data in this study, a stepwise linear regression was performed on these variables and their squares; the regression equation was derived fully automatically, using the STEP command in Minitab [21]. A predictor variable was included in the regression according to the maximum F-statistic criterion, with $F \geq 4$. Stepwise linear regression was also performed using the 27 attributes and their squares, analogous to the approach using the Hansch parameters.

The squared terms were included to allow some basic nonlinear behaviour. Hansch suggested that the parabolic dependence of activity on the hydrophobicity of a drug could be due to the phenomenon of very highly hydrophobic drugs not entering the cell, but remaining in the lipophilic membrane. Nonlinear dependences on MR values have also been reported [14]. More complex relationships have been derived from the consideration of thousands of possible correlation equations [13–15], some of which include logarithms and terms based on the sum of Hansch parameters. It was infeasible to determine such equations automatically for each split of the data here. One advantage of neural networks and ILP is that they provide a mechanism for exploring nonlinear terms without considering each different functional form, e.g., reciprocals, logs, cubes, etc., explicitly.

Neural networks

In outline (see Refs. 3–6), a neural network consists of a large number of simple computational units that have many connections to one another. Each unit performs a weighted sum X of incoming signals I, and sends out to other units a signal O, which is a function of this weighted sum. The connections are the weights w of the weighted sum and are adjustable real numbers. For the *i*th unit, with connected units indexed by j,

$$X_i = \sum w_{ij} I_j$$

TABLE 3
PHYSICOCHEMICAL ATTRIBUTES OF FRAGMENTS

Group	PL	SZ	FL	HD	HA	πD	πA	PO	σ	BR[a]
$(CH_2)_2$	1	2	2	0	0	0	0	1	0	0
$(CH_2)_2COCH_2Cl$	1	3	3	0	1	0	0	1	0	1
$(CH_2)_2CON(CH_2CH_2)_2O$	1	6	2	0	1	0	0	1	0	1
$(CH_2)_2CONEt_2$	1	5	2	0	1	0	0	1	0	2
$(CH_2)_2CONMe_2$	1	4	2	0	1	0	0	1	0	2
$(CH_2)_3CH(CH_2NHCOCH_2Br)$	1	6	6	0	1	0	0	1	0	0
$(CH_2)_3O$	1	4	4	0	1	0	0	1	0	0
$(CH_2)_4$	1	4	4	0	0	0	0	1	0	0
$(CH_2)_4COCH_2Cl$	1	5	5	0	1	0	0	1	0	1
$(CH_2)_6$	1	4	6	0	0	0	0	1	0	0
$(CH_2)_2CON(C_3H_7)_2$	2	5	3	0	1	0	0	1	0	2
Br	3	1	0	0	0	1	0	2	2	0
C_6H_5	0	3	0	0	0	2	0	3	0	1
CF_3	3	1	0	0	0	0	1	0	3	1
$CH(CH_2NHCOCH_2Br)CH_2$	2	5	2	0	1	0	0	1	0	2
CH_2	1	1	0	0	0	0	0	1	0	0
CH_2CH_3	1	2	2	0	0	0	0	1	0	0
CH_2CN	2	2	1	0	1	0	0	1	0	0
$CH_2CON(CH_2CH_2)_2O$	2	5	1	0	2	0	0	1	0	2
CH_2CONEt_2	2	4	3	0	1	0	0	1	0	2
CH_2CONMe_2	2	3	2	0	1	0	0	1	0	2
$CH_2N(Me)CO(CH_2)_2$	2	6	4	0	1	0	0	1	0	2
$CH_2N(Me)COCH_2$	2	5	3	0	1	0	0	1	0	2
$CH_2NHCOCH_2Br$	2	4	3	1	1	0	0	1	0	1
$CH_2NHCONEt_2$	2	5	3	1	2	0	0	1	0	2
$CH_2NHCON(CH_2CH_2)_2O$	2	5	1	1	3	0	0	1	0	2
CH_2O	2	2	2	0	1	0	0	1	0	0
$CH_2O(CH_2)_3CH_3$	1	4	6	0	1	0	0	1	0	0
CH_2OCH_3	1	2	3	0	1	0	0	1	0	0
CH_2OH	2	2	2	2	2	0	0	1	0	0
CH_3	1	1	0	0	0	0	0	1	0	0
Cl	3	1	0	0	0	1	0	1	3	0
CN	4	1	0	0	1	0	2	2	0	0
$COCH_2Cl$	3	2	1	0	1	0	1	1	0	1
$CON(CH_2)_4$	3	4	0	0	1	0	1	1	0	2
$CON(CH_2)_5$	3	5	0	0	1	0	1	1	0	2
$CON(CH_2CH_2)_2O$	3	5	0	0	2	0	1	1	0	2
$CONEt_2$	3	4	0	0	1	0	1	1	0	2
$CONH(CH_2)_2$	3	3	2	1	1	0	1	1	0	1
$CONH(CH_2)_4O$	3	5	4	1	2	0	1	1	0	1
$CONHCH_2$	3	2	1	1	1	0	1	1	0	1
$CONMe_2$	3	3	0	0	1	0	1	1	0	2
F	5	1	0	0	1	0	0	0	5	0
H	1	0	0	0	0	0	0	0	0	0
I	2	1	0	0	0	1	0	3	1	0
$N(CH_3)_2$	1	2	0	0	1	2	0	1	1	1
$N(CH_3)COCH_2O$	2	4	2	0	2	2	0	1	1	2
$N(Me)CO(CH_2)_2$	2	4	2	0	1	2	0	1	1	2
$N(Me)COCH_2$	2	3	1	0	1	2	0	1	1	2
NH_2	2	1	0	2	0	2	0	0	1	0
NHCO	2	2	0	1	1	2	0	1	1	2
$NHCO(CH_2)_2$	2	3	2	1	1	1	0	1	1	1

TABLE 3 (continued)

Group	PL	SZ	FL	HD	HA	πD	πA	PO	σ	BR[a]
$NHCO(CH_2)_2S$	2	4	3	1	1	1	0	1	1	1
$NHCO(CH_2)_3$	2	4	3	1	1	1	0	1	1	1
$NHCO(CH_2)_3O$	2	5	4	1	2	1	0	1	1	1
$NHCO(CH_2)_4O$	2	5	5	1	1	1	0	1	1	1
$NHCOCH(\alpha\text{-}C_{10}H_7)CH_2$	2	8	1	1	1	1	0	1	1	1
$NHCOCH(CH_2CH_2Ph)CH_2$	2	8	1	1	1	1	0	1	1	3
$NHCOCH(CH_3)O$	2	4	2	1	2	1	0	1	1	2
$NHCOCH(Me)CH_2$	2	4	2	1	1	1	0	1	1	2
$NHCOCH(Ph)CH_2$	2	6	2	1	1	1	0	1	1	3
$NHCOCH(Ph\text{-}2''\text{-}CH_3)CH_2$	2	6	1	1	1	1	0	1	1	4
$NHCOCH(Ph\text{-}2''\text{-}OCH_3)CH_2$	2	7	1	1	2	1	0	1	1	4
$NHCOCH(Ph\text{-}3''\text{-}CH_3)CH_2$	2	6	1	1	1	1	0	1	1	4
$NHCOCH(Ph\text{-}3''\text{-}OCH_3)CH_2$	2	7	1	1	2	1	0	1	1	4
$NHCOCH(Ph\text{-}4''\text{-}CH_3)CH_2$	2	6	1	1	1	1	0	1	1	4
$NHCOCH_2$	2	3	1	1	1	1	0	1	1	1
$NHCOCH_2Br$	2	3	1	1	1	1	0	1	1	1
$NHCOCH_2CH(CH_3)$	2	3	2	1	1	1	0	1	1	2
$NHCOCH_2CH(Ph)$	2	6	2	1	1	1	0	1	1	3
$NHCOCH_2CH(Ph)CH_2$	2	6	3	1	1	1	0	1	1	3
$NHCOCH_2O$	2	3	2	1	2	1	0	1	1	1
$NHCOCH_2S$	2	3	2	1	1	1	0	2	1	1
$NHCOCH_3$	2	2	0	1	1	1	0	1	1	0
NHCOCHCH	2	3	0	1	1	1	0	1	1	1
$NHCONH(CH_2)_2O$	2	5	3	1	3	1	0	1	1	1
$NHCONH(CH_2)_3O$	2	6	4	2	3	1	0	1	1	1
$NHCONH(CH_2)_4O$	2	6	5	2	3	1	0	1	1	1
$NHCONHCH_2$	2	4	1	2	2	1	0	1	1	1
NO_2	5	2	0	0	0	0	2	0	3	1
O	2	1	0	0	1	1	0	2	1	0
$O(CH_2)_2$	2	3	2	0	1	1	0	1	1	0
$O(CH_2)_2O$	2	3	3	0	2	1	0	1	1	0
$O(CH_2)_2O(CH_2)_2O$	2	4	6	0	3	1	0	1	1	0
$O(CH_2)_3CH_3$	2	3	4	0	1	1	0	1	1	0
$O(CH_2)_3O$	2	3	4	0	2	1	0	1	1	0
$O(CH_2)_4$	2	4	4	0	1	1	0	1	1	0
$O(CH_2)_4O$	2	4	5	0	2	1	0	1	1	0
$O(CH_2)_5CH_3$	2	5	6	0	1	1	0	1	1	0
$O(CH_2)_5O$	2	5	6	0	2	1	0	1	1	0
$O(CH_2)_6O$	2	5	7	0	2	1	0	1	1	0
$O(CH_2)_6CH_3$	2	5	7	0	1	1	0	1	1	0
$O(CH_2)_7CH_3$	2	5	8	0	1	1	0	1	1	0
OC_3H_7	2	4	3	0	1	1	0	1	1	0
OCF_3	4	2	0	0	0	0	2	0	3	1
$OCH_2C_6H_{10}CH_2O$	2	6	2	0	1	1	0	1	1	1
$OCH_2C_6H_5$	2	4	2	0	1	1	0	1	1	1
$O(CH_2)_2OCH_3$	2	3	4	0	2	1	0	1	1	0
OCH_2CH_3	2	3	2	0	1	1	0	1	1	0
$OCH_2CON(CH_2)_4$	2	5	1	0	2	1	0	0	1	2
$OCH_2CON(CH_2)_5$	2	5	1	0	2	1	0	0	1	2
$OCH_2CON(CH_2CH_2)_2O$	2	5	2	0	3	1	0	0	1	2
OCH_2CONMe_2	2	4	1	0	2	1	0	0	1	1
$OCH_2CONMePh$	2	6	1	0	2	1	0	0	1	1
OCH_2CONEt_2	2	5	2	0	2	1	0	0	1	2

TABLE 3 (continued)

Group	PL	SZ	FL	HD	HA	πD	πA	PO	σ	BR[a]
OCH_2CONH_2	2	3	1	1	2	1	0	0	1	1
OCH_3	2	2	0	0	1	1	0	1	1	0
OCH_2O	2	3	3	0	2	1	0	1	1	0
OH	3	1	0	2	2	2	0	1	2	0
$ON(CH_3)COCH_2O$	3	4	2	0	3	1	0	1	2	2
OSO_2CH_3	4	3	1	0	0	0	1	2	2	2
SO_2F	4	2	0	0	0	0	1	2	2	2
$SO_2NH(CH_2)_2$	4	4	2	0	1	0	1	2	2	2
SO_2NMe_2	4	4	0	0	1	0	1	2	2	3

PL: polarity; SZ: size; FL: flexibility; HD: number of hydrogen-bond donors; HA: number of hydrogen-bond acceptors; πD: strength and presence of π-donors; πA: strength and presence of π-acceptors; PO: polarisability; σ: σ-effect; and BR: branching.
[a] The attribute representing branching was not used in this study.

$$O_i = \frac{1}{1 + e^{-X_i}}$$

The units may be organised into layers: an input layer, hidden layers, and an output layer, with signals being propagated forward from the input layer to the output layer via any hidden layers. The neural network is trained to learn a mapping between input and output signals, by presenting the network with training data (input: output pairs) and altering the weights in a well-defined manner using a learning rule, such that the sum of the squared error between the desired output signals and the actual output signals is minimised. A learning rule, the backpropagation of errors [22], has been implemented here, as in the subsequent paper [4,9], using an approach to increase the speed of learning [23] where the changes to the weights are calculated by solving a set of stiff differential equations [24]. Code was written in FORTRAN to simulate the neural network, incorporating the original code of Gear [24], and was implemented on a VAX 4600 and on a SUN Sparc workstation.

The neural network methodology has several empirically determined parameters. These include:

- when to stop training (i.e., the number of epochs; here called the convergence criterion)
- the number of hidden units
- the learning rate
- a momentum term

The Gear algorithm obviates the need for a learning rate or a momentum term. Memorisation [3,25,26] occurs when the neural network is overtrained. If the convergence criterion is too stringent, i.e., training is continued until the change in the error on the training set is too small, then the performance on the test set will be impaired, although the training set performance may increase. The method used here determines the convergence criterion and the number of hidden units, without examination of the test data, by the performance of the neural network on a third set of data.

The data were divided into three sets: training, monitoring and testing. The neural network was trained directly on the training data, and its performance was monitored using the monitor set. For each test molecule, the weights from the training cycle that gave the maximum performance on the appropriate monitor set were used to evaluate the test molecule.

Two studies with neural networks were performed to consider both representations of the data. In an approach similar to that which has been previously used [3], a neural network with six input units, zero to five hidden units and one output unit was trained to predict the activity of drugs, given π_3, π_4, π_5, MR_3, MR_4 and MR_5. A neural network with 27 input units, zero to five hidden units and one output unit (Fig. 1) was trained using the attribute representation, with the same monitoring procedure.

For each test set, one molecule was used as the completely unseen test set, and the others were used to monitor the neural network. This was repeated for each molecule in the test set in turn. The neural network was only trained once per test set (but simultaneously monitored using the 11 different monitor sets), rather than once per molecule, i.e., 11 different sets of weights for testing (the 11 different molecules in a test set) were selected at the appropriate iteration from the training procedure, without having to retrain the neural network for every test molecule. This use of some of the test set for monitoring is legitimate, because the activity of the test molecule is not used in training. This procedure is a combination of cross-validation and leave-one-out, and the performance of the neural network may be slightly over-estimated compared to the other methods, because information from both the training and monitoring sets is used.

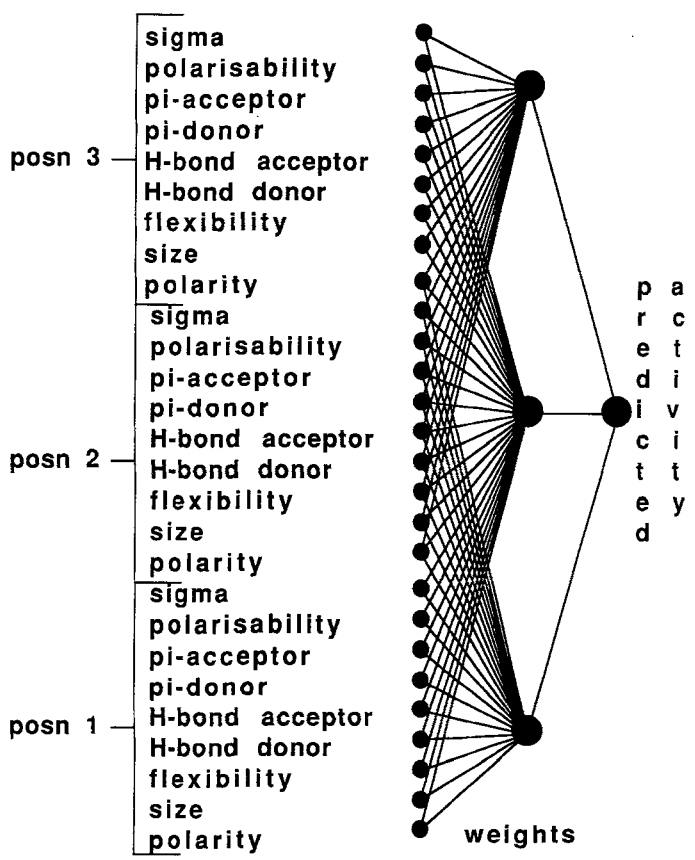


Fig. 1. Schematic representation of a neural network, trained using the attribute representation. Each input unit is connected to each of the three hidden units by a weight (not shown).

Inductive logic programming (ILP)

GOLEM is a machine learning program that uses ILP. The ILP methodology [20] is, in theory, well suited for drug design problems, as it is designed specifically to learn relationships between objects, e.g. molecular structures. In ILP, logical relationships, expressed as a subset of predicate calculus [27], are used to represent learned rules. Predicate calculus is expressive enough to describe most mathematical concepts and has a strong link with natural language. GOLEM is written in C, but implements predicate logic in Prolog. The representation used by GOLEM is similar to that described by King et al. [8] for the modelling of the QSAR of trimethoprim analogues binding to DHFR.

QSAR is often treated as 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 difference is reconciled by considering pairs of drugs and comparing their activities. Paired comparisons are then converted to a ranking by the method of David [28].

GOLEM takes three types of facts as input: positive, negative and background. The positive facts are the paired examples of greater activity, e.g. great (d50, d9), which states that molecule number 50 has higher activity than molecule number 9. The paired examples of lower activity, e.g. great (d9, d50), are negative facts (or false statements). GOLEM requires both positive and negative facts to give balanced generalisation.

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). This Prolog representation of molecule number 35 states that the molecule has NO_2 substituted at position 3, $NHCOCH_3$ substituted at position 4, and no substituent at position 5. By convention, if only one of positions 3 and 5 is substituted, as in molecule number 35, the position with no substitution is assumed to be position 5.

In the training phase, only facts relating to the training set were used; for testing, additional new facts about the test set were required. For the fivefold cross-validation study, involving 55 molecules, there were 2198 background facts, 1394 positive facts and 1394 negative facts. In the separate test set of 19 molecules, there were 2388 background facts (a superset of the previous 2198, with the addition of extra background facts to describe the new groups), 965 positive facts and 965 negative facts. Rules were selected sequentially, using the 'minimal description length'

TABLE 4
CROSS-VALIDATION TRAINING SET PERFORMANCES AS MEASURED BY THE SPEARMAN RANK CORRELATION COEFFICIENTS

Method	Set 1	Set 2	Set 3	Set 4	Set 5	Mean[a] (σ)
LR on Hansch + squares[b]	0.883	0.832	0.846	0.830	0.773	0.833 (0.035)
LR on attributes + squares	0.941	0.922	0.899	0.881	0.796	0.888 (0.050)
Neural network on Hansch parameters	0.821	0.565	0.903	0.746	0.764	0.760 (0.112)
Neural network on attributes	0.719	0.529	0.895	0.917	0.943	0.801 (0.157)
GOLEM[c]	0.966	0.929	0.952	0.952	0.943	0.948 (0.012)

[a] Each method was trained on the five cross-validation training sets. The mean and the standard deviation (σ) of the five performances are given.

[b] Linear regression (LR) on the Hansch parameters and their squares.

[c] The parameter settings for GOLEM were: depth: i = 5; clause parameter: j = 5; error level: *noise* = 2; sample size: *rlggsample* = 8 (as defined in the original GOLEM work [20]).

TABLE 5
CROSS-VALIDATION TEST SET PERFORMANCES AS MEASURED BY THE SPEARMAN RANK CORRELATION COEFFICIENTS

Method	Set 1	Set 2	Set 3	Set 4	Set 5	Mean[a] (σ)
LR on Hansch + squares[b]	0.717	0.478	0.521	0.845	0.905	0.693 (0.170)
LR on attributes + squares	0.654	0.506	0.694	0.819	0.596	0.654 (0.104)
Neural network on Hansch parameters	0.661	0.482	0.683	0.852	0.709	0.677 (0.118)
Neural network on attributes	0.788	0.228	0.702	0.838	0.753	0.660 (0.225)
GOLEM	0.751	0.574	0.753	0.757	0.627	0.692 (0.077)

[a] Each method was trained on the five cross-validation training sets. The mean and the standard deviation (σ) of the five performances are given.
[b] Linear regression (LR) on the Hansch parameters and their squares.

principle to avoid overfitting the data (as implemented in the compression model [29]). Simplistically, this means that if the data can be described by a rule that has less information than the data, then the rule is significant.

RESULTS

The performances of the methods on the cross-validation training and testing data, as measured by the Spearman rank correlation coefficient [30], are given in Tables 4 and 5, respectively. Each of these cross-validation training sets has also been tested with the independent test set of 19 molecules (Table 6). The Spearman rank correlation coefficient is based on nonparametric correlation, which is more robust than linear correlation [31]. The Pearson r, based on linear correlation, is not a good statistic for deciding whether an observed correlation is statistically significant, because the method is ignorant of the individual distributions of the two variables, and so there is no universal way of computing the distribution of the Pearson r in the case of the null hypothesis. GOLEM predicts the rank order of the molecules, and not their absolute activity, so its performance is not measurable by the Pearson r.

GOLEM performed better than linear regression using the attribute representation. However, the improvement was not statistically significant, as determined by a Fisher z test (two-tailed); for a data set of this size, a difference in Spearman rank correlation coefficients of about 0.2 would have been required.

TABLE 6
THE MEAN SPEARMAN RANK CORRELATION COEFFICIENTS ON THE INDEPENDENT TEST SET OF 19 MOLECULES

Method	Mean performance[a] (σ)
Linear regression on Hansch parameters + squares	0.657 (0.036)
Linear regression on attributes + squares	0.509 (0.152)
Neural network on Hansch parameters	0.627 (0.179)
Neural network on attributes	0.510 (0.115)
GOLEM	0.738 (0.095)

[a] Each method was trained on the five cross-validation training sets. For each training set, a corresponding performance on the independent test set of 19 drugs was calculated. The mean and the standard deviation (σ) of the five performances are given here.

Linear regression

The regressions from the cross-validation trial consistently found a negative dependence on MR_5 and π_4^2, a positive dependence on MR_3^2 and MR_4^2, and a parabolic dependence on π_5. Linear regression on the attributes alone gave a cross-validated Spearman rank correlation coefficient of 0.505 on the test set. This was improved by the inclusion of the squares of the attributes. The regression equations from the five cross-validation trials were similar, with many of the same variables appearing in all the equations with similar coefficients. These regression equations were combined by taking the mean of the coefficients of the variables that appeared in at least three of the five equations. This indicated the following relationship:

$$\log(1/K_i) = 0.44 + 0.66(\pm 0.21)PO_3 - 0.82(\pm 0.23)PO_3^2 + 0.11(\pm 0.05)\pi A_4 + 0.79(\pm 0.15)SZ_5 - 0.94(\pm 0.10)FL_5 + 1.10(\pm 0.18)\pi A_5 - 0.18(\pm 0.14)\pi A_5^2$$

$$\bar{n} = 44;\ \bar{\sigma} = 0.04;\ \bar{r}^2 = 0.93$$

where the attributes are represented by their two-letter abbreviations, with the subscripts denoting the position of substitution. The statistics associated with the equation, $\bar{n}$, $\bar{\sigma}$ and $\bar{r}^2$, are the mean values calculated from the five equations from the cross-validation trial. The 95% confidence intervals of the coefficients in the regression equation are given in parentheses.

Neural networks

When the data were represented using the Hansch parameters, a neural network without hidden units performed as well as those with hidden units. This permits a more straightforward analysis of the weights of the neural network than when there are hidden units [3], because without hidden units, the neural network is essentially just performing a weighted sum. Although nonlinear, the sigmoid transfer function is monotonically increasing, so that without hidden layers, large positive weights indicate that the respective input should be high for high activity and large negative weights indicate that the respective input should be low for high activity. A separate trial was performed for the analysis of the weights, where the weights giving optimal performance on the test set were averaged for the five cross-validation data sets. This would not be legitimate for evaluating the predictive performance, but it provides a well-defined set of weights for the purposes of analysis. The following proportionality was indicated:

$$\log(1/K_i) \propto 0.40\pi_4 + 0.29\pi_5 + 0.51MR_3 - 0.48MR_5$$

The proportionality indicates which variables are important in determining activity, and their relative contributions. Statistics analogous to those of the regression equation cannot be calculated, and the significance of the relationship is measured by the performance on the test set. The result is broadly in accord with a previous neural network analysis using hidden units [3], which found that higher MR_3 values improved activity, as well as a more complex dependence on MR_5; negative π_3 values decreased activity, and π_4 and π_5 had not been used as inputs.

The neural network trained on the attributes also showed no improvement in test set performance with the addition of hidden units. A similar analysis of the weights using the attribute representation suggested the following function:

$$\log(1/K_i) \circ -0.22HA_3 + 0.16\pi A_3 + 0.25\sigma_3 + 0.20\pi D_4 + 0.19\pi A_4 + 0.13\sigma_4 + 0.24SZ_5 + 0.28FL_5$$

TABLE 7
EXAMPLE OF A GOLEM RULE

PROLOG format	English translation
great(A,B) :-	Drug A is better than drug B if:-
struc(A,Pos_a3,Pos_a4,_), h_donor(Pos_a3,h_don0), pi_acceptor(Pos_a3,pi_acc0), polar(Pos_a3,Pol_a3), great0_polar(Pol_a3), size(Pos_a3,Siz_a3), less3_size(Siz_a3), polar(Pos_a4,Pol_a4), struc(B,_,_,h).	drug A has a substituent at position 3 with hydrogen-bond donor = 0 and π-acceptor = 0 and polarity > 0 and size < 3 and drug A has a substituent at position 4 and drug B has no substituent at position 5.

The test accuracy, i.e., the number of cases for which the rule is correct, divided by the coverage, was 0.918. The coverage, i.e., the number of pair comparisons that the rule covers, was 440. Both are based on the 19 new molecules.

Inductive logic programming

Although it is possible to use the Hansch parameters as input to GOLEM, this is not a useful comparison, as the major motivation for using GOLEM is to generate easily understandable rules, which the Hansch parameters do not facilitate. The rules have the form of Prolog clauses, and are readily translated into English (an example is given in Table 7). In total, 59 rules were found for the five cross-validation runs. From these rules, nine consensus rules were formed manually (Table 8), by selecting the most commonly found features. The consensus rules, which are consistent with previous work [8], have a simpler form than the automatically generated GOLEM rules, making them easier to understand. In forming these consensus rules, the substitutions at each position were assumed to be independent. The consensus rules gave an average Spearman rank correlation coefficient of 0.845 on the cross-validation test data, and 0.793 on the separate test set of 19 molecules. The rank correlations were similar for training and test sets, which indicates that the rules are not overfitting the data.

The best predicted drug or drugs can be generated from the consensus rules. Consider positions 3 and 5. The only possible substituents with the conjunction of polarity < 5 (Table 8: rule b); size < 3 (rule c); hydrogen-bond donor = 0 (rule d); π-donor = 1 (rule e); σ < 5 (rule f); and flexibility < 3 (rule g), are OCH_3, I, Cl and Br (O is excluded, because it is a linking moiety). These are, therefore, the substituents recommended for positions 3 and 5; the rules do not distin-

TABLE 8
ENGLISH TRANSLATIONS OF THE CONSENSUS RULES

A is more active than B if:-

Rule a	A has a substituent at position 3 and B does not.
Rule b	A has a substituent at position 3 with polarity < 5 and B does not.
Rule c	A has a substituent at position 3 with size < 3 and B does not.
Rule d	A has a substituent at position 3 with hydrogen-bond donor = 0 and B does not.
Rule e	A has a substituent at position 3 with π-donor = 1 and B does not.
Rule f	A has a substituent at position 3 with σ < 5 and B does not.
Rule g	A has a substituent at position 3 with flexibility < 3 and B does not.
Rule h	A has a substituent at position 4 with polarity < 5 and B does not.
Rule i	A has a substituent at position 4 with hydrogen-bond donor = 0 and B does not.

guish between these groups. There are fewer constraints at position 4, only the conjunction of polarity < 5 (rule h) and hydrogen-bond donor = 0 (rule i). There are 68 substituents with these attributes (the substituents suggested for positions 3 and 5 are a subset of this set).

DISCUSSION AND CONCLUSIONS

The distribution of training and test sets in independent variable space is crucial to the predictive capabilities of the methods. The selection of representative training sets can lead to over-estimation of the predictive capability [9]. This is why cross-validation trials, with random partitioning of data, are important. For example, all the methods performed below average on the second split of the data, which suggests that the second test set was not representative of the whole data set. It contains one data point, 3,5-$(OH)_2$, which Hansch et al. [14] found to be 6000 times less active than expected. Excluding this point improved their regression from a Pearson r^2 of 0.650 to one of 0.815. Omission of this data point gave similar improvements in the performances of the methods in this study, but such an omission would preclude an unbiased assessment of predictive capability. In the independent test set of 19 molecules, the 3,5-dimethoxy, 4-(dimethylamino) derivative (molecule number 69) is overpredicted by all the methods. For this derivative, it has been suggested that the amino substituent would be forced to lie almost perpendicular to the phenyl ring, and the projection above and below the ring plane reduces the expected activity of this drug [13].

Both the linear regression and the neural network methods suggest that the 5-substituent should have a low MR value, that the 3-substituent should have a high MR value, and that π_3 does not have a large effect on activity. However, the analyses disagree on the importance of MR_4, π_4 and π_5. The methods using the attribute representation generate a variety of possible influences of structure on activity. GOLEM suggests that the 3-substituent should not be a hydrogen-bond donor. Linear regression and GOLEM both indicate that the 3-substituent should have a small size and low flexibility, and that the 5-substituent should have low flexibility. Linear regression and neural networks suggest that the 4-substituent should be involved in π-bonding. The σ-effect of the 3-substituent is identified as an influencing factor by neural networks and GOLEM. Analysis of the crystal structure of the complex formed between trimethoprim and DHFR [12] indicates that both meta sites, i.e., the 3- and 5-substituents, are buried in a hydrophobic environment, and restrictions on size and flexibility are consistent with this. It has also been proposed that the 4-substituent should lie in the plane of the aromatic ring [18], and a substituent involved in π-bonding would have such a constraint.

The addition of hidden units to the neural network did not give an improvement, even though the linear regression analysis suggested that squared attribute terms were important. The neural network and regression analyses also differed in their identification of important attributes. This indicates that the power of neural networks may not be fully exploited using the attribute representation, especially for small data sets.

The main conclusions from this study are:

- neural networks and ILP do not predict test data significantly better than linear regression
- ILP produces understandable rules using the attribute representation
- neural networks tend to overfit using the attribute representation
- hidden units do not significantly improve the neural network predictions for this QSAR.

In the following paper [9], a larger data set of DHFR inhibitors provides a statistically more sensitive test, and demonstrates the general applicability of attributes with an extension to many more substituents.

REFERENCES

1 Hansch, C., Maloney, P.P., Fujita, T. and Muir, R.M., Nature, 194 (1962) 178.
2 Hansch, C., Acc. Chem. Res., 2 (1969) 232.
3 So, S.-S. and Richards, W.G., J. Med. Chem., 35 (1992) 3201.
4 Andrea, T.A. and Kalayeh, H., J. Med. Chem., 34 (1991) 2824.
5 Aoyama, T., Suzuki, Y. and Ichikawa, H., J. Med. Chem., 33 (1990) 905.
6 Aoyama, T. and Ichikawa, H., J. Chem. Inf. Comput. Sci., 32 (1992) 492.
7 Tetko, I.V., Luik, A.I. and Poda, G.I., J. Med. Chem., 36 (1993) 811.
8 King, R.D., Muggleton, S., Lewis, R.A. and Sternberg, M.J.E., Proc. Natl. Acad. Sci. USA, 89 (1992) 11322.
9 Hirst, J.D., King, R.D. and Sternberg, M.J.E., J. Comput.-Aided Mol. Design, 8 (1994) 421.
10 Li, R.L., Hansch, C. and Kaufman, B.T., J. Med. Chem., 25 (1982) 435.
11 Champness, J.N., Stammers, D.K. and Beddell, C.R., FEBS Lett., 199 (1986) 61.
12 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. and Kraut, J., J. Biol. Chem., 260 (1985) 381.
13 Selassie, C.D., Li, R.-L., Poe, M. and Hansch, C., J. Med. Chem., 34 (1991) 46.
14 Hansch, C., Li, R.-I., Blaney, J.M. and Langridge, R., J. Med. Chem., 25 (1982) 777.
15 Li, R.-L. and Poe, M., J. Med. Chem., 31 (1988) 366.
16 Dietrich, S.W., Blaney, J.M., Reynolds, M.A., Jow, P.Y.C. and Hansch, C., J. Med. Chem., 23 (1980) 1205.
17 Roth, B., Aig, E., Rauckman, B.S., Srelitz, J.Z., Phillips, A.P., Ferone, R., Bushby, S.R.M. and Siegel, C.W., J. Med. Chem., 24 (1981) 933.
18 Roth, B., Rauckman, B.S., Ferone, R., Baccanari, D.P., Champness, J.N. and Hyde, R.M., J. Med. Chem., 30 (1987) 348.
19 Leo, A., Hansch, C. and Elkins, D., Chem. Rev., 71 (1971) 525.
20 Muggleton, S. and Feng, C., In Arikawa, S., Goto, S., Ohsuga, S. and Yokomori, T. (Eds.) Proceedings of the First Conference on Algorithmic Learning Theory, Japanese Society of Artificial Intelligence, Ohmsha Press, Tokyo, 1990, pp. 368–381.
21 Minitab, release 7.2, VAX/VMS version, Minitab, Inc., Pennsylvania State University, Philadelphia, PA, 1989.
22 Rumelhart, D.E., Hinton, G.E. and Williams, R.J., Nature, 323 (1986) 533.
23 Owens, A.J. and Filkin, D.L., In IEEE/INNS International Joint Conference of Neural Networks, Washington, DC, 1989, pp. 381–386.
24 Gear, C.W., Numerical Initial Value Problems in Ordinary Differential Equations, Prentice Hall, Englewood Cliffs, NJ, 1971.
25 Livingstone, D.J. and Salt, D.W., Bioorg. Med. Chem. Lett., 2 (1992) 213.
26 Livingstone, D.J. and Mallanack, D.T., J. Med. Chem., 36 (1993) 1295.
27 DeLong, H., A Profile of Mathematical Logic, Addison-Wesley, Reading, MA, 1970.
28 David, H.A., Biometrika, 74 (1987) 432.
29 Muggleton, S., Srinivasan, A. and Bain, M., In Sleeman, D. and Edwards, P. (Eds.) Proceedings of the 9th International Conference on Machine Learning, Morgan-Kaufman, San Mateo, CA, 1992, pp. 338–347.
30 Kendall, M. and Stuart, A., The Advanced Theory of Statistics, Griffen, London, 1977.
31 Press, W.H., Teukolsky, S.A., Vettering, W.T. and Flannery, B.P., Numerical Recipes, Cambridge University Press, Cambridge, 1992.