Journal of Computer-Aided Molecular Design, 8 (1994) 421–432
ESCOM

J-CAMD 251

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

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

One of the largest available data sets for developing a quantitative structure–activity relationship (QSAR) – the inhibition of dihydrofolate reductase (DHFR) by 2,4-diamino-6,6-dimethyl-5-phenyl-dihydrotriazine derivatives – has been used for a sixfold cross-validation trial of neural networks, inductive logic programming (ILP) and linear regression. No statistically significant difference was found between the predictive capabilities of the methods. However, the representation of molecules by attributes, which is integral to the ILP approach, provides understandable rules about drug–receptor interactions.

INTRODUCTION

The examination of several QSAR systems is essential to assess the performance of new methods. In this paper, neural networks, inductive logic programming (ILP) and linear regression, as applied in the preceding study [1], are tested on one of the largest available QSAR data sets. The same methods have been implemented with only minor alterations. The data come from work of Baker and co-workers in the 1960s, who synthesised many triazines and studied their inhibition of dihydrofolate reductase (DHFR). Silipo and Hansch [2] concluded that the data were 'an excellent testing ground for further new approaches to structure–activity analysis'.

This trial follows the evaluation procedures outlined in the previous paper [1]. The insight provided into the drug–receptor interaction is an important aspect in the assessment. A new method is presented for interpreting the QSAR modelled by a neural network with hidden units.

*Present address: Box 77, Department of Chemistry, Mellon Institute, 4400 Fifth Avenue, Pittsburgh, PA 15213, U.S.A.
**To whom correspondence should be addressed.

METHODS

Data

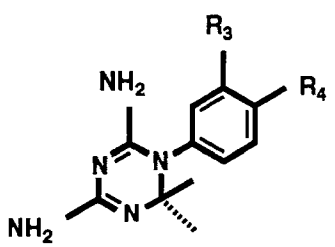
The data in this study, 186 2,4-diamino-6,6-dimethyl-5-phenyl-dihydrotriazines (**I**) (Table 1) are a subset of those used to derive a QSAR by multilinear regression [2] and by neural networks [3]. The 186 triazines have two or three rings, with the phenyl ring (the second ring) substituted at either the 3- or 4-position; ortho-substituents were not considered as there were only 11 examples, which is too small a set for generalisations [4]. The data were divided into six sets, each of 31 molecules; each of the 186 molecules appears once only in only one of the sets (Table 2). Each split of the data was used as a testing set, and for each testing set the other five sets were combined to give a training set of 155 molecules. All data were rescaled to lie between 0.1 and 0.9 for the neural networks, to avoid zero-valued inputs, and also for consistency with regard to the regression analyses.

Hansch parameters

Silipo and Hansch [2] correlated the activity of the triazines with the hydrophobic parameters π and the molar refractivities MR of the 3- and 4-substituents of the phenyl ring. The activity was measured by log 1/C, where C is the molar concentration that produces 50% reversible inhibition of dihydrofolate reductase, obtained from L1210 mouse leukaemia cells and Walker 256 rat tumours. A multilinear regression was performed using π_3, π_4, MR_3, MR_4 and also π_3^2, π_4^2, MR_3^2 and MR_4^2, to allow some nonlinear dependence. Six discrete variables (indicator variables: I-1, I-2, I-3, I-4, I-5 and I-6) were used to improve the regression. These variables take the value of 1 or 0 for structural features that could not be parameterised by the hydrophobic constants and the molar refractivities. Using the indicator variable I-1, with a value of 1 for Walker enzyme data and 0 for L1210 enzyme data, allowed the merging of the data from the two test systems. Ortho-substitution was flagged using I-2. Rigid groups directly attached to the 3- or 4-positions were marked by I-3. I-4 was 1 for all compounds containing 4-$OCH_2C_6H_4SO_2OC_6H_4$-X. I-5 was 1 for flexible bridges between the *N*-phenyl moiety and a second phenyl ring. I-6 was used for some other bridging groups. The hydrophobic constants, molar refractivities and indicator variables for all the molecules in this study have been listed in previous papers [2,3].

The attributes

The substituents were assigned the following attributes, using the heuristics given in the previous paper [1]: polarity, size, flexibility, number of hydrogen-bond donors and acceptors, presence and strength of π-acceptors and π-donors, polarisability, σ-effect and branching. There were six



I

TABLE 1
TRIAZINES USED IN THIS STUDY

Drug no.[a]	Activity (log 1/C)	Substituents	Drug no.[a]	Activity (log 1/C)	Substituents
9	4.68	4-$CONHC_6H_4$-4'-SO_2F	60	7.05	4-$(CH_2)_2CONMe_2$
10	4.68	4-$CONHC_6H_4$-3'-SO_2F	62	7.07	3-Cl-4-$O(CH_2)_3OC_6H_4$-4'-SO_2F
11	4.70	4-C_6H_5	63	7.07	3-NO_2
13	4.85	3-$OCH_2CON(CH_2CH_2)_2O$	64	7.10	3-$(CH_2)_2COCH_2Cl$
14	5.14	4-CN	65	7.10	3-$(CH_2)_4COCH_2Cl$
15	5.19	4-$CHCHCONHC_6H_4$-4'-SO_2F	66	7.12	4-$OCH_2CON(CH_2)_5$
16	5.44	3-OCH_2CONMe_2	67	7.12	4-$CH_2CON(CH_2CH_2)_2O$
17	5.74	4-$CH(Ph)NHCOCH_2C_6H_4$-4'-SO_2F	68	7.12	4-$(CH_2)_6C_6H_4$-4'-SO_2F
18	5.82	4-Cl-3-$(CH_2)_2C_6H_4$-4'-SO_2F	69	7.13	3-Cl-4-$OCH(CH_3)CONHC_6H_4$-4'-SO_2F
19	5.89	4-$CHCHCONHC_6H_4$-3'-SO_2F			
20	5.96	3-$CONHC_6H_4$-4'-SO_2F	70	7.13	4-$CH_2CH(Ph)CONHC_6H_4$-4'-SO_2F
21	6.11	3-$NHCOCH_2Br$-4-$O(CH_2)_3C_6H_5$	71	7.14	3-Cl-4-$O(CH_2)_2O(CH_2)_2OC_6H_4$-4'-$SO_2F$
22	6.11	3-$CH_2NHCONEt_2$			
23	6.17	3-OCH_3	72	7.15	3-Cl-4-$O(CH_2)_3CONHC_6H_4$-4'-SO_2F
24	6.17	4-$OCH_2CON(CH_3)C_6H_5$			
25	6.20	4-$CH_2CH(CH_2CH_2Ph)$ $CONHC_6H_4$-4'-SO_2F	73	7.16	3-Cl-4-OCH_2CONMe_2
			74	7.17	3-Cl-4-$O(CH_2)_3CONHC_6H_4$-3'-SO_2F
26	6.21	3-$COCH_2Cl$			
27	6.24	4-$CH_2CH(\alpha$-$C_{10}H_7)CONHC_6H_4$-4'-SO_2F	75	7.17	4-Cl-3-$O(CH_2)_4OC_6H_4$-4'-SO_2F
			76	7.17	4-$CH_2CH(Ph$-3"-$CH_3)CONHC_6H_4$-4'-SO_2F
28	6.26	4-OCH_2CONMe_2			
29	6.33	4-$CH_2CH(Ph$-2"-$OCH_3)$ $CONHC_6H_4$-4'-SO_2F	77	7.19	3-$(CH_2)_2CONHC_6H_4$-4'-SO_2F
			78	7.24	4-$CH_2CH(Ph$-4"-$CH_3)CONHC_6H_4$-4'-SO_2F
30	6.37	3-Cl-4-$OCH_2C_6H_{10}CH_2OC_6H_4$-4'-$SO_2F$	79	7.24	4-$CH_2CH(Ph$-2"-$CH_3)CONHC_6H_4$-4'-SO_2F
31	6.37	3-$CH(CH_2NHCOCH_2Br)$ $(CH_2)_3C_6H_5$	82	7.24	3-Cl-4-$O(CH_2)_4CONHC_6H_4$-4'-SO_2F
32	6.43	3-$CH_2NHCON(CH_2CH_2)_2O$			
33	6.45	4-$COCH_2Cl$	84	7.27	4-Cl-3-$O(CH_2)_2OC_6H_4$-4'-SO_2F
34	6.46	4-$CH_2CH(Ph$-3"-$OCH_3)$ $CONHC_6H_4$-4'-SO_2F	85	7.27	3-SO_2F
			86	7.28	3-Cl-4-$O(CH_2)_3NHCONHC_6H_4$-3'-SO_2F
35	6.52	4-$CH(CH_2NHCOCH_2Br)$ $(CH_2)_3C_6H_5$	87	7.28	4-$(CH_2)_2CONEt_2$
39	6.58	3-$CH_2NHCOCH_2Br$	88	7.29	3-Cl-4-$OCH_2CON(CH_2)_4$
40	6.60	3-$CONHC_6H_4$-3'-SO_2F	89	7.29	4-$OCH_2CON(CH_2CH_2)_2O$
41	6.63	4-CH_2CONMe_2	90	7.29	4-$CH(CH_3)CH_2CONHC_6H_4$-4'-SO_2F
42	6.66	4-$OCH_2CON(CH_2)_4$			
43	6.68	3-$OCH_2CONMePh$	91	7.30	4-$CH_2CON(Me)CH_2C_6H_5$
44	6.72	4-OCH_2CONEt_2	92	7.31	4-$(CH_2)_2CON(Me)CH_2C_6H_5$
45	6.72	3-$CH_2CH(CH_2NHCOCH_2Br)C_6H_5$	93	7.32	4-$(CH_2)_2CON(CH_2CH_2)_2O$
46	6.72	4-Cl-3-$O(CH_2)_5OC_6H_4$-4'-SO_2F	94	7.32	4-$O(CH_2)_3NHCONHC_6H_4$-3'-SO_2F
47	6.77	4-CH_2CONEt_2	95	7.34	3-Cl-4-$O(CH_2)_3NHCOC_6H_4$-4'-SO_2F
48	6.77	4-Cl-3-$(CH_2)_4C_6H_4$-4'-SO_2F			
50	6.85	3-$CH_2OCONHC_6H_5$	96	7.34	3-$CH_2CONHC_6H_4$-4'-SO_2F
51	6.85	3-C_6H_5	97	7.35	4-$CH_2NHCONHC_6H_4$-4'-SO_2F
52	6.89	4-$CH_2CH(Ph)CONHC_6H_4$-3'-SO_2F	98	7.35	4-$(CH_3)_2CON(C_3H_7)_2$
55	6.92	3-$OCH_2CONHC_6H_4$-4'-SO_2F	101	7.39	3-Cl-4-$S(CH_2)_2CONHC_6H_4$-4'-SO_2F
56	6.92	4-CH_2CN			
57	6.92	H	102	7.41	4-$(CH_2)_2C_6H_4$-4'-SO_2F
58	6.92	3-$OCH_2C_6H_4$-3'-$NHCOCH_2Br$	104	7.41	4-$(CH_2)_2NHSO_2C_6H_4$-4'-SO_2F
59	7.00	4-$CH_2CON(Me)C_6H_5$	105	7.42	3-Cl-4-$SCH_2CONHC_6H_4$-4'-SO_2F

TABLE 1 (continued)

Drug no.[a]	Activity (log 1/C)	Substituents	Drug no.[a]	Activity (log 1/C)	Substituents
107	7.43	3-Cl-4-$OCH_2CONHC_6H_4$-4'-SO_2F	162	7.77	3-$CH_2NHCOC_6H_4$-3'-$CONMe_2$
109	7.43	3-Cl-4-$OCH_2C_6H_3$-3'-Cl-4'-SO_2F	165	7.80	3-Cl-4-$CH_2NHCONHC_6H_3$-3'-CH_3-4'-SO_2F
111	7.44	3-Cl-4-$O(CH_2)_2OC_6H_4$-4'-SO_2F			
113	7.46	3-Cl-4-$O(CH_2)_6OC_6H_4$-4'-SO_2F	166	7.80	4-$O(CH_2)_2OC_6H_4$-4'-SO_2F
114	7.46	4-$(CH_2)_2CONHC_6H_3$-3'-OCH_3-4'-SO_2F	168	7.82	3-Cl-4-$O(CH_2)_2NHCONHC_6H_3$-3'-CH_3-4'-SO_2F
115	7.47	3-Cl-4-$OCH_2CON(CH_3)C_6H_4$-4'-SO_2F	169	7.82	3-$O(CH_2)_2OC_6H_4$-4'-SO_2F
			171	7.85	3-Cl-4-$(CH_2)_2C_6H_4$-4'-SO_2F
116	7.47	3-Cl-4-$OCH_2CON(CH_2)_5$	173	7.85	3-Cl-4-$(CH_2)_2C_6H_3$-3'-Cl-4'-SO_2F
117	7.48	3-Cl-4-$CH_2OC_6H_4$-4'-SO_2NMe_2	174	7.85	3-Cl-4-$OCH_2CON(CH_2CH_2)_2O$
119	7.49	3-$O(CH_2)_4OC_6H_4$-4'-SO_2F	175	7.85	3-Cl-4-$OCH_2C_6H_4$-3'-$CON(CH_2CH_2)_2O$
120	7.51	4-Cl-3-$O(CH_2)_3OC_6H_4$-4'-SO_2F			
121	7.51	3-Cl-4-$OCH_2C_6H_4$-3'-CN	176	7.85	3-Cl-4-$OCH_2C_6H_4$-3'-$CON(CH_2)_4$
122	7.52	3-Cl-4-$OCH_2C_6H_5$	177	7.89	3-Cl-4-$OCH_2CON(CH_3)C_6H_5$
123	7.52	4-$SCH_2CONHC_6H_4$-4'-SO_2F	178	7.89	4-$OCH_2CONHC_6H_5$
125	7.52	3-$CH_2NHCONHC_6H_5$	179	7.89	4-$(CH_2)_2C_6H_5$
126	7.55	4-$CH_2CH(Me)CONHC_6H_4$-4'-SO_2F	180	7.89	4-$(CH_2)_2CONHC_6H_3$-3'-CH_3-4'-SO_2F
127	7.55	3-$O(CH_2)_3OC_6H_4$-4'-$NHCOCH_2Br$	181	7.92	3-Cl-4-$CH_2NHCONHC_6H_4$-4'-SO_2F
128	7.56	4-$(CH_2)_2CON(Me)C_6H_5$			
129	7.57	3-Cl-4-$O(CH_2)_4OC_6H_4$-4'-SO_2F	182	7.92	3-Cl-4-$O(CH_2)_2NHCONHC_6H_4$-4'-SO_2F
130	7.57	3-Cl-4-$O(CH_2)_5OC_6H_4$-4'-SO_2F			
131	7.58	3-Cl-4-$OCH_2C_6H_4$-4'-SO_2F	183	7.92	4-$(CH_2)_3CONHC_6H_4$-3'-SO_2F
132	7.60	4-$(CH_2)_2CONHC_6H_4$-4'-SO_2F	184	7.92	4-$(CH_2)_2COCH_2Cl$
133	7.62	3-Cl-4-$(CH_2)_2CONHC_6H_4$-4'-SO_2F	185	7.92	3-OC_6H_4-4'-$NHCOCH_2Br$
134	7.62	3-$CH_2NHCONHC_6H_4$-3'-SO_2F	186	7.92	3-Cl-4-$(CH_2)_4C_6H_5$
135	7.64	4-$(CH_2)_2NHSO_2C_6H_4$-3'-SO_2F	189	7.96	3-$O(CH_2)_3OC_6H_4$-4'-SO_2F
136	7.64	3-Cl-4-OCH_2CONEt_2	192	8.00	3-Cl-4-$OCH_2C_6H_3$-3'-SO_2F-4'-Cl
137	7.64	3-$O(CH_2)_3OC_6H_4$-3'-$NHCOCH_2Br$	194	8.00	4-$OCH_2CONHC_6H_4$-3'-SO_2F
139	7.66	3-$O(CH_2)_2OC_6H_4$-3'-$NHCOCH_2Br$	196	8.00	3-$CH_2C_6H_5$
140	7.66	3-Cl-4-$SCH_2CONHC_6H_4$-3'-SO_2F	197	8.00	4-$(CH_2)_4C_6H_5$
141	7.66	3-Cl-4-$O(CH_2)_4NHCOC_6H_4$-4'-SO_2F	198	8.02	3-Cl-4-$OCH_2C_6H_4$-3'-$CON(CH_2)_5$
			199	8.02	3-$CH_2NHCONHC_6H_4$-3'-OCH_3
142	7.66	4-$(CH_2)_3CONHC_6H_4$-4'-SO_2F	200	8.02	4-$(CH_2)_2CONHC_6H_3$-3'-CH_3-4'-SO_2F
143	7.68	3-Cl-4-$O(CH_2)_3NHCONHC_6H_4$-4'-SO_2F			
			201	8.03	3-Cl-4-$(CH_2)_4C_6H_4$-3'-SO_2F
144	7.70	3-Cl-4-$O(CH_2)_4NHCONHC_6H_4$-3'-SO_2F	203	8.04	4-$CH_2NHCONHC_6H_4$-3'-SO_2F
			204	8.04	4-$(CH_2)_2CON(Me)C_6H_4$-4'-SO_2F
145	7.70	3-$(CH_2)_4C_6H_3$-3'-Cl-4'-SO_2F	206	8.05	4-$CH_2C_6H_5$
146	7.70	3-Cl-4-$(CH_2)_4C_6H_3$-3'-Cl-4'-SO_2F	207	8.05	3-$CH_2NHCONHC_6H_4$-3'-Cl
147	7.70	4-$(CH_2)_4C_6H_4$-4'-SO_2F	208	8.06	3-Cl-4-$O(CH_2)_3NHCONHC_6H_3$-3'-SO_2F-4'-CH_3
148	7.70	4-$CH_2CONHC_6H_4$-4'-SO_2F			
149	7.70	4-$O(CH_2)_2OC_6H_4$-4'-$NHCOCH_2Br$	209	8.06	4-$CH_2CONHC_6H_4$-3'-SO_2F
150	7.72	3-Cl-4-$OCH_2C_6H_4$-3'-$CONMe_2$	212	8.09	3-$CH_2NHCONHC_6H_4$-3'-NO_2
152	7.72	4-$OCH_2CONHC_6H_4$-4'-SO_2F	213	8.10	3-$(CH_2)_4C_6H_4$-4'-SO_2F
153	7.72	3-Cl-4-$OCH_2CONHC_6H_4$-3'-SO_2F	214	8.10	3-$(CH_2)_4C_6H_4$-3'-SO_2F
154	7.72	3-Cl-4-$OCH_2C_6H_4$-3'-SO_2F	215	8.10	3-$(CH_2)_2C_6H_4$-4'-SO_2F
156	7.72	4-$CH_2NHCONHC_6H_3$-3'-CH_3-4'-SO_2F	216	8.11	4-$(CH_2)_2NHCOC_6H_4$-4'-SO_2F
			217	8.11	3-Cl-4-$(CH_2)_4C_6H_3$-3'-SO_2F-4'-Cl
157	7.74	4-$(CH_2)_2CONHC_6H_4$-3'-SO_2F	219	8.13	3-$O(CH_2)_2OC_6H_4$-4'-$NHCOCH_2Br$
159	7.76	3-Cl	220	8.14	3-Cl-4-$OCH_2C_6H_4$-3'-$CONEt_2$
160	7.76	3-CF_3	221	8.14	3-Cl-4-$(CH_2)_4C_6H_4$-4'-SO_2F

TABLE 1 (continued)

Drug no.[a]	Activity (log 1/C)	Substituents	Drug no.[a]	Activity (log 1/C)	Substituents
222	8.14	3-Br-4-$OCH_2CONHC_6H_4$-4'-SO_2F	234	8.27	3-Cl-4-$(CH_2)_2C_6H_3$-3'-SO_2F-4'-Cl
223	8.14	4-$(CH_2)_4OC_6H_4$-4'-SO_2F	238	8.35	3-$(CH_2)_4OC_6H_5$
224	8.19	3-$(CH_2)_2C_6H_5$	239	8.35	3-$(CH_2)_4C_6H_5$
225	8.19	3-$CH_2NHCONHC_6H_4$-3'-CN	240	8.37	3-$(CH_2)_4C_6H_3$-3'-SO_2F-4'-Cl
229	8.24	4-$(CH_2)_2CONHC_6H_3$-3'-SO_2F-4'-OCH_3	241	8.38	3-$(CH_2)_4C_6H_4$-4'-$NHCOCH_2Br$
			246	8.41	3-$(CH_2)_4C_6H_4$-3'-$NHCOCH_2Br$
231	8.24	4-$O(CH_2)_4C_6H_5$	250	8.54	3,4-Cl_2
233	8.26	3-$(CH_2)_2C_6H_4$-4'-$NHCOCH_2Br$			

[a] The numbers in this column correspond to those in column 1 of Table 1 of Ref. 4 and also to those in column 1 of Table 1 of Ref. 3.

regions where a substituent might be present: the 3- and 4-positions of the phenyl ring (regions 1 and 4, respectively); if the substituent at the 3-position contained a ring itself, then the 3- and 4-positions of this third ring (regions 2 and 3, respectively; the attributes for these regions were set to zero if there was no third ring there); if the substituent at the 4-position of the phenyl ring contained a ring itself, then the 3- and 4-positions of this third ring (regions 5 and 6, respectively; the attributes for these regions were set to zero if there was no third ring there). With six possible regions of chemical variability and 10 attributes per region, each molecule is described by 60 parameters. The attributes assigned to the different fragments, some of which link two rings, are given in Table 3 of the preceding paper [1].

Linear regression

The six training sets were assigned the Hansch parameters π_3, π_4, MR_3, MR_4 and $\Sigma\sigma_{34}$ (the sum of the σ values of substituents at the 3- and 4-positions of the phenyl ring [3,5]) and the indicator variables I-1, I-3, I-4 and I-5 (I-2 indicates ortho-substitution and I-6 characterises a fourth ring; neither feature was present in this data). The primary comparison was provided by a stepwise linear regression on the Hansch parameters and their squares, in a procedure similar to that of

TABLE 2
SPLITS OF THE DATA USED FOR COMPARATIVE STUDY

Split no.	Data points[a]
1	184, 234, 24, 48, 107, 22, 85, 140, 171, 11, 31, 73, 63, 68, 241, 206, 55, 21, 30, 168, 40, 27, 199, 88, 84, 219, 117, 246, 239, 166, 240
2	9, 33, 23, 177, 114, 132, 50, 135, 123, 69, 120, 71, 58, 102, 156, 196, 15, 119, 45, 129, 98, 130, 143, 213, 225, 96, 165, 224, 176, 43, 233
3	62, 74, 220, 26, 186, 115, 150, 153, 72, 173, 18, 222, 194, 111, 44, 185, 207, 82, 169, 94, 104, 137, 97, 113, 41, 93, 147, 159, 67, 39, 178
4	105, 201, 86, 229, 35, 231, 152, 70, 76, 221, 78, 13, 125, 28, 180, 174, 16, 59, 20, 19, 214, 91, 212, 65, 17, 160, 126, 101, 148, 122, 34
5	146, 57, 157, 189, 116, 183, 87, 223, 154, 200, 162, 25, 95, 144, 136, 145, 46, 51, 181, 90, 89, 10, 121, 75, 42, 197, 109, 77, 192, 204, 216
6	215, 64, 29, 142, 128, 175, 250, 56, 182, 208, 131, 149, 60, 238, 179, 198, 52, 66, 141, 139, 209, 92, 14, 127, 47, 134, 217, 133, 32, 203, 79

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

TABLE 3
SUMMARY OF ALL METHODS – SPEARMAN RANK CORRELATION COEFFICIENT ON TRAINING SETS

Method	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Mean[a] (σ)
LR on Hansch + squares	0.271	0.379	0.295	0.321	0.296	0.238	0.300 (0.048)
LR on 60 attributes + squares[b]	0.488	0.565	0.553	0.495	0.610	0.530	0.540 (0.046)
Neural network on Hansch + I-1	0.482	0.693	0.618	0.730	0.686	0.688	0.650 (0.090)
Neural network on 60 attributes	0.862	0.936	0.803	0.781	0.641	0.773	0.799 (0.099)
GOLEM[c]	0.723	0.707	0.672	0.650	0.682	0.666	0.683 (0.027)

[a] Each method was trained on six cross-validation training sets. The mean and standard deviation (σ) of the six 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 = 3; error level: *noise* = 50; sample size: *rlggsample* = 20 (as defined in the original GOLEM work [7]); no examples were covered.

Silipo and Hansch [2], but with full automation, using the STEP command in Minitab [6], with a maximum F-statistic criterion of $F \geq 4$. In a separate trial, indicator variables were included in the regression. However, indicator variables are specific to individual systems, used when the limit has been reached for defining important structural features with known physicochemical constants [4], so they have not been used for the general evaluation of the QSAR methods. Stepwise linear regression was performed using the 60 attributes and their squares.

Neural networks

A benchmark was provided by a neural network similar to one used in a previous study on these data [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, MR_3, MR_4, $\Sigma\sigma_{34}$ and the indicator variable I-1. The neural network was trained using the backpropagation-of-errors learning rule and the Gear algorithm, as detailed in the previous paper [1]. For each of the 31 molecules in a test set, one was used as the unseen test set, and 30 were used to monitor the neural network. This was repeated for each drug in the test set in turn, so that the neural network was only trained once per test set (but simultaneously monitored using the 31 different monitor sets), rather than being trained 31 times. The same monitoring procedure was used for a neural network with 60 input units, zero to five hidden units and one output unit trained on the same data, represented by attributes.

Inductive logic programming

GOLEM [7], an inductive logic program, was implemented using the attribute representation. As before [1], the positive facts were the paired examples of greater activity and the negative facts were the paired examples of lower activity. The background facts were the chemical structures of the drugs and the properties of the substituents. Chemical structure was represented in the form:

struc3(d217, Cl, absent)
struc4(d217, $(CH_2)_4$, subst14)
subst(subst14, SO_2F, Cl)

This is the Prolog representation of the molecule number 217: 3-Cl, 4-$(CH_2)_4C_6H_3$-4'-Cl, 3'-SO_2F. The first clause represents substituents at position 3 on the *N*-phenyl moiety; a chlorine is present and there is no further phenyl ring. The second clause represents substituents at position 4 on the *N*-phenyl moiety; there is a $(CH_2)_4$ bridge to a second phenyl ring, implicit in the representation. This second phenyl ring has an SO_2F substituent at position 3 and a chlorine at position 4. This is represented using the linker constant subst14 in the third clause. This structural representation could be easily extended to include more substitution positions and more rings, e.g., the molecules included in previous studies [2,3], but not in this study.

For each of the six splits, the input to GOLEM was 2933 facts as background information and 1000 positive and negative facts. The positive and negative facts were the equivalent random sample of all possible pairs; not all the pairs could be used, because of computational complexity. One hundred rules were found for each split.

The method for selecting the best rules from the starting set of 100 rules was improved by recognising that all correct classifications do not have the same utility. For example, consider three molecules A, B and C, where B is slightly more active than C, and A is far more active than B. It is better to predict that A is more active than C, than to predict that B is more active than C. The measure used by GOLEM, utility, was the squared difference in rank, with correct predictions having a positive utility and incorrect predictions having a negative utility. The best rules were selected as follows:

Repeat
  Select the rule with the greatest utility that covers more than 50 examples;
  Remove examples covered by the most accurate rule;
Until the utility of the rule with greatest utility is less than 1 000 000.

RESULTS

Spearman rank correlation coefficients for the performances of the methods on the training data and test data are given in Tables 3 and 4, respectively. Comparing the same methods on the different data representations and different methods on the same data representations, the only significant difference was for the linear regression analysis, where the attribute representation gave a test set Spearman rank correlation coefficient of 0.446 and the Hansch parameters gave one of 0.272; the probability p of obtaining this difference by chance is 0.0414 (as determined by a two-tailed Fisher z test).

TABLE 4
SUMMARY OF ALL METHODS – SPEARMAN RANK CORRELATION COEFFICIENT ON TEST SETS

Method	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Mean[a] (σ)
LR on Hansch + squares[b]	0.463	–0.124	0.314	0.197	0.480	0.304	0.272 (0.220)
LR on 60 attributes + squares	0.425	0.216	0.724	0.291	0.491	0.529	0.446 (0.181)
Neural network on Hansch + I-1	0.588	0.103	0.534	0.400	0.438	0.197	0.377 (0.190)
Neural network on 60 attributes	0.329	0.301	0.676	0.399	0.559	0.620	0.481 (0.145)
GOLEM	0.583	0.536	0.321	0.610	0.212	0.325	0.431 (0.166)

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

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

Linear regression

The stepwise linear regression on only the Hansch parameters and their squares found a negative dependence on $\Sigma\sigma_{34}$ and a parabolic dependence on π_3, and gave a cross-validated Spearman rank correlation coefficient of 0.272 on the test set. Inclusion of indicator variables gave a cross-validated Spearman rank correlation coefficient of 0.635 on the test set, indicating that the Hansch parameters do not model all the relevant features. Combining the regression equations from the six cross-validation trials on the 60 attributes and their squares, by taking the mean of the coefficients of variables appearing in more than three of the regression equations, indicated the following relationship:

$$\log(1/C) = 0.52 - 0.17(\pm 0.11)HA_1 + 0.19(\pm 0.09)PO_1 + 0.36(\pm 0.11)PL_2 - 0.30(\pm 0.11)PL_2^2 + 0.31(\pm 0.16)SZ_2^2 + 1.31(\pm 0.38)FL_4 - 0.26(\pm 0.13)FL_4^2 - 1.29(\pm 0.45)HD_4^2 - 0.17(\pm 0.10)\pi A_4^2 - 0.69(\pm 0.41)PO_4^2$$

$$\bar{n} = 155;\ \bar{\sigma} = 0.13;\ \bar{r}^2 = 0.39$$

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

Neural networks

As expected, the neural network trained using Hansch parameters gave better predictions when hidden units were used, indicating that the activity of the drugs is nonlinearly dependent on the Hansch parameters. The capacity of neural networks to model nonlinear features and cross terms is the major motivation for their application to QSAR analysis. In contrast to traditional QSAR analyses, which generate regression equations, it is more difficult to interpret the relationship between the activity of a drug and its properties modelled by a neural network with hidden units, because nonlinear terms are considered implicitly in a multidimensional function.

The neural network model of the QSAR was interpreted by testing different combinations of inputs and analysing the common features of the combinations predicted by the neural network to have high activity. The five inputs π_3, π_4, MR_3, MR_4 and $\Sigma\sigma_{34}$ were varied between 0.1 and 0.9 in steps of 0.1, and I-1 was either 0.1 or 0.9. About 3.9 million (2×5^9) combinations of input values were evaluated. For each cross-validation trial the weights that gave the best test set performance were used, so that the generalised features would be optimised. These weights were generated in a separate trial; optimisation on the test set performance would not be legitimate for assessing the predictive ability, but is useful for investigating the features that are generalised.

The final weights from each cross-validation run were different. Exhaustive evaluation of the possible combinations of inputs, using the five different sets of weights, found different combinations of inputs to be highly active. The range of predicted activity also varied, with the weights from the third and sixth trials predicting no combinations to be more active than 0.99, whereas the fourth trial predicted more than 6000 combinations to have an activity greater than 0.99. A small number of the predicted high-activity combinations were clustered by eye (Table 5), to extract the general features. For each set of weights, the combinations predicted to be most active

TABLE 5
CLASSIFICATION OF COMBINATIONS OF HANSCH PARAMETERS PREDICTED BY A NEURAL NETWORK TO BE MORE ACTIVE THAN A GIVEN THRESHOLD

Hansch parameter	π_3[a]	π_4	MR_3	MR_4	$\Sigma\sigma_{34}$	Activity threshold[b]	No. above threshold[c]
Set 1a	–	+	+	–	+	0.99	18
Set 1b[d]	+	–	–	+	+		
Set 2	+	+	–	+	–	0.99	66
Set 3	+	–	–	–	–	0.98	10
Set 4	+	+	–	+/–	–	0.9994	50
Set 5	+	+/–	–	+	–	0.985	60
Set 6	+	+/–	–	–	+	0.96	84

[a] A '–' indicates that the value was consistently less than 0.5; a '+' indicates that the value was consistently greater than 0.5; a '+/–' indicates that the value was either greater or less than 0.5.
[b] The limit of predicted activity is unity. The threshold value was arbitrarily chosen so that the number of combinations with a higher predicted activity was reasonably small, for inspection by eye.
[c] A combination is a set of the five input values, ranging from 0.1 to 0.9 in discrete steps of 0.1. The number of such combinations with a predicted activity higher than the threshold is given here.
[d] Two classes of combination were found for set 1, denoted 1a and 1b.

were similar to each other, e.g., the weights optimised using the third cross-validation set of data predicted 10 combinations to have an activity greater than 0.98, and all 10 combinations have a high π_3, low π_4, low MR_3, low MR_4 and low $\Sigma\sigma_{34}$. Each trial generated one class of combinations of predicted high activity, except for the first trial, which generated two classes. All but one of these classes had a high π_3 value and a low MR_3 value; more variation occurred in π_4, MR_4 and $\Sigma\sigma_{34}$.

The neural network trained on the attribute representation gave the most accurate test set predictions without hidden units. As in the previous paper [1], this suggests that activity is essentially linearly dependent on the attributes, but it contradicts the linear regression analysis, which indicated that there was some nonlinear dependence. If some nonlinear dependence occurs, then the impairment of performance on addition of hidden units is probably due to overfitting because of the increase in the number of weights. An analysis of the neural network weights suggested the following function:

$$\log(1/C) \propto -0.28\pi D_1 - 0.58\pi A_1 + 0.37\sigma_1 + 0.37BR_1 - 0.60PL_2 + 0.64SZ_2 + 0.37PO_2 + 0.24BR_2 + 0.16\pi A_3 - 0.25PL_4 + 0.73HD_4 - 0.46\pi A_4 - 0.55PO_4 + 0.21\pi A_5 + 0.37\sigma_5.$$

Many of the weights are zero, with no attribute in region 6 having a large weight, but of the 10 attributes, only flexibility does not appear at all.

Inductive logic programming

For the six cross-validation runs, 45 rules were found. From these rules, seven simple consensus rules were generated manually (Table 6) from the most common features, assuming that substituents at each position are independent. The consensus rules gave average cross-validated Spearman rank correlations on the training and test sets of 0.498 and 0.457, respectively.

The consensus rules can be interpreted to generate the best predicted drug, or drugs. At position 4, rule 5 (Table 6) states that there should be a substituent with polarity = 1; and rule 7

TABLE 6
ENGLISH TRANSLATION OF THE CONSENSUS RULES FOR DRUGS A AND B

A is better than B if:–

Rule 1	A has a substituent on the first phenyl ring at position 3 with polarisability = 1 and B does not.
Rule 2	A has a substituent on the first phenyl ring at position 3 with π-donor = 1 and B does not.
Rule 3	A has a substituent on the first phenyl ring at position 3 with branching = 0 and B does not.
Rule 4	A has a substituent on the first phenyl ring at position 3 with hydrogen-bond acceptor = 0 and B does not.
Rule 5	A has a substituent on the first phenyl ring at position 4 with polarity = 1 and B does not.
Rule 6	A has a substituent that includes a phenyl ring at position 3 on the first phenyl ring and B does not.
Rule 7	A has a substituent that includes a phenyl ring at position 4 on the first phenyl ring and B does not.

states that this group should be connected to a phenyl ring. The rules do not specify what type of substituent should be on the second phenyl ring. Although the best drug in the present study only has a chlorine at position 4 (polarity = 3), most of the highly active drugs in the data (molecule numbers 246, 241, 240, 239, 234 and 233) comply with rules 4 and 7. Four attributes are specified: polarisability = 1 (rule 1), π-donor = 1 (rule 2), hydrogen-bond acceptor = 0 (rule 4), and branching = 0 (rule 3); the only substituent with all four specified attributes is Cl (see Fig. 1). This is the substituent found in the best drugs in this study and in other work [3,8]. However, rule 6 conflicts with this conclusion and suggests that a second phenyl ring be added. This indicates that it may be possible to achieve better results than with chlorine by relaxing the condition

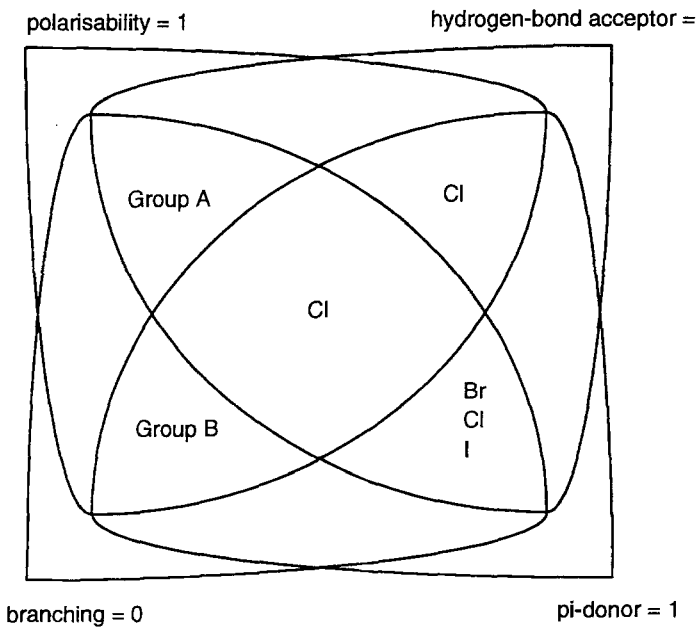


Fig. 1. Venn diagram of favoured properties for the first substitution at position 3. Chlorine is the only substituent that has: polarisability = 1, π-donor = 1, hydrogen-bond acceptor = 0 and branching = 0. Group A is the intersection of polarisability = 1, hydrogen-bond acceptor = 0 and branching = 0 and contains the fragments: Cl, $(CH_2)_2$, $(CH_2)_4$, $(CH_2)_6$, CH_2, CH_3. Group B is the intersection of polarisability = 1, π-donor = 1 and branching = 0 and contains: Cl, $NHCOCH_3$, $O(CH_2)_2$, $O(CH_2)_2O$, $O(CH_2)_2O(CH_2)_2O$, $O(CH_2)_3CH_3$, $O(CH_2)_3O$, $O(CH_2)_4$, $O(CH_2)_4O$, $O(CH_2)_5CH_3$, $O(CH_2)_5O$, $O(CH_2)_6O$, $O(CH_2)_6CH_3$, $O(CH_2)_7CH_3$, $OCH_2OCH_2OCH_3$, OCH_3.

π-donor = 1 and using a substituent from group A (Fig. 1) and a second phenyl ring, or relaxing the condition hydrogen-bond acceptor = 1 and using a substituent from group B and a second phenyl ring.

DISCUSSION

In this study, the methods performed surprisingly poorly compared to what might have been expected from the literature [3,4]. Estimates of test set performances are critically susceptible to biases in the data. This is particularly relevant in this study, where many pairs of the molecules are identically described by the same Hansch parameters and also, in many cases, have very similar activities. For these data, a leave-one-out cross-validation procedure will overestimate the predictive performance on new data, because much of the supposed test data is, in fact, the same as the training data. This is especially so for methods which fit the training data well.

In this study, a neural network, trained using the Hansch parameters, had a test set performance, as measured by the Pearson r^2, of 0.252. In a previous study comparing neural networks and linear regression on this data set [3], four estimates of the test set performance of neural networks were:

$r^2 = 0.804$ (100/32 drugs in the training/testing sets);
$r^2 = 0.672$ (66/66 drugs in the training/testing sets);
$r^2 = 0.511$ (57/56 drugs in the training/testing sets);
$r^2 = 0.787$ (leave-one-out cross-validation on 132 drugs).

These particular splits (100/32, 66/66 and 57/56) were deliberately selected, using a cluster analysis of the data prior to the assessment of the neural network to ensure that every point in the test set had points in the training set in its vicinity. While this ensures that the training and testing sets are evenly distributed in the space of independent variables, this is not a random split of the data and does not give an unbiased estimate of the test set performance.

For the linear regression analysis, the attribute representation gave significantly better results than the Hansch parameters. This may be because each of the molecules has a unique attribute representation, whereas there are over 50 pairs of molecules whose Hansch parameters have identical values. This is less likely to be a problem in a smaller data set.

CONCLUSIONS

The use of ILP and neural networks in QSAR analysis has been assessed in terms of their predictive capabilities and the insight into the QSAR that they provide. The predictive capability of ILP is not significantly better than those of linear regression or neural networks. In contrast to other work [3,9], our studies found no statistically significant difference between the predictive capabilities of neural networks and linear regression for the QSARs of pyrimidines and triazines as DHFR inhibitors. This demonstrates that comparative trials with random splits of the data and tests of statistical significance are essential for the proper evaluation of new QSAR methods.

The predictions using attributes are as accurate as the Hansch parameters, and the representation allows the formulation of readily understandable rules. These rules may be easily translated from the Prolog clauses generated by GOLEM. Extracting information from a neural network

with hidden units is more involved, but has been shown to be possible using the strategy of evaluating combinations of inputs, and clustering those predicted to have high activity to find the common features. Both neural networks and GOLEM have the capacity to model complex relationships, and advances in QSAR analysis may come from the full exploitation of this power.

REFERENCES

1 Hirst, J.D., King, R.D. and Sternberg, M.J.E., J. Comput.-Aided Mol. Design, 8 (1994) 405.
2 Silipo, C. and Hansch, C., J. Am. Chem. Soc., 97 (1975) 6849.
3 Andrea, T.A. and Kalayeh, H., J. Med. Chem., 34 (1991) 2824.
4 Hansch, C. and Fukunga, J., CHEMTECH, (1977) 120.
5 Hansch, C. and Silipo, C., J. Med. Chem., 17 (1974) 661.
6 Minitab, release 7.2, VAX/VMS version, Minitab, Inc., Pennsylvania State University, Philadelphia, PA, 1989.
7 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 Artifical Intelligence, Ohmsha Press, Tokyo, 1990, pp. 368–381.
8 Silipo, C. and Hansch, C., J. Med. Chem., 19 (1976) 62.
9 So, S.-S. and Richards, W.G., J. Med. Chem., 35 (1992) 3201.