Predictive Flip Regression: A Technique for QSAR of Derivatives of Symmetric Molecules

Brian W. Clare*,† and Claudiu T. Supuran‡

School of Biomedical and Chemical Science, The University of Western Australia, 35 Stirling Highway, Crawley, Western Australia 6009, Australia, and Laboratorio di Chimica Inorganica e Bioinorganica, Università degli Studi, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy

Received May 11, 2005

We have for the first time used the flip regression technique to predict the activity of unknown drugs given a substantial series of derivatives of symmetrical molecules such as benzene, with known activity. As descriptors we have used quantum theoretic parameters such as orbital energies and the orientation of π-orbital nodes and Mulliken charges of atoms. Flip regression is a technique for dealing with multiply substituted derivatives of symmetric molecules, such as phenethylamine or benzenesulfonamide. The method has been tried on two large data sets: thrombin inhibitors and carbonic anhydrase (CA) inhibitors. Multiple validation tests were run, randomly splitting the data into training and test sets and successfully predicting the activity of the test sets, with predictive R^2 of around 0.9 for the thrombin inhibitors and 0.7 for the CA inhibitors.

INTRODUCTION

When carrying out QSAR on flat, symmetrical, usually aromatic molecules, symmetry considerations often dictate that alternative orientations be examined. For phenethylamines, for example, there are five substitution sites on the benzene nucleus. If substituents with property P_i were introduced to site i ($i = 2-6$), an equation may be developed: $\log A = \sum_i C_i P_i + C_0$, where A is activity and C_i are constants to be determined by regression techniques. It must be remembered, however, that, for example, 2-methoxyphenethylamine is identical with 6-methoxyphenethylamine, so the equation must predict the same activity for both molecules. At first sight it may appear that C_2 must equal C_6 and C_3 must equal C_5, but that this is not the case becomes apparent when considering 2,3,4-trimethoxyamphetamine and 2,4,5-trimethoxyamphetamine. Simply equating the C values would predict the same activity for both these substances, but experimentally one of these is a potent hallucinogen and the other is inactive.[1]

The problem of symmetry seems to have been first recognized by Kishida and Manabe[2] in the context of bifunctional benzenesulfonamides. They handled it using a computer program that swapped descriptor values, in a way very similar to one of our procedures. Other authors have handled the symmetry problem by ignoring it altogether,[3] restricting the substitution pattern of the compounds studied[4−7] introducing indicator variables for crowding, or by introducing classical statistical interaction variables.[8] None of these options is entirely satisfactory. The first introduces ambiguity as to numbering the ring, the second severely restricts the compounds that can be studied, and the third and fourth can become complex in highly substituted cases. A similar problem arises when using the components of the dipole moment as a descriptor: how do we assign the sign. The same thing happens when we consider the orientation of the nodes in the π orbitals, a factor that we have found to be a very useful: which way around the ring do we consider positive. The problem has been solved using a procedure similar to that of Kishida and Manabe. We must realize that the physical counterpart of this mathematical problem is the

Table 1. Sulfonylguanidines ($RSO_2N{=}C(NH_2)_2$) **A1**−**A25** Prepared in the Present Study, with Their Inhibition Data against Human Thrombin

R	compound	K_I[a] (nM)
p-F−C_6H_4−	**A1**	240
p-Cl−C_6H_4−	**A2**	225
p-Br−C_6H_4−	**A3**	220
p-I−C_6H_4−	**A4**	210
p-CH_3−C_6H_4−	**A5**	290
p-O_2N−C_6H_4−	**A6**	180
m-O_2N−C_6H_4−	**A7**	190
o-O_2N−C_6H_4−	**A8**	320
3-Cl-4-O_2N−C_6H_3−	**A9**	160
p-AcNH−C_6H_4−	**A10**	195
p-H_2N−C_6H_4−	**A11**	95
m-H_2N−C_6H_4−	**A12**	107
C_6F_5−	**A13**	146
o-HOOC−C_6H_4−	**A14**	240
m-HOOC−C_6H_4−	**A15**	121
p-HOOC−C_6H_4−	**A16**	104
o-HOOC−C_6Br_4−	**A17**	225
p-CH_3O−C_6H_4−	**A18**	240
2,4,6-$(CH_3)_3$−C_6H_2−	**A19**	255
4-CH_3O-3-H_2N−C_6H_3−	**A20**	103
2-HO-3,5-Cl_2−C_6H_2−	**A21**	152
4-Me_2N−C_6H_4−N═N−C_6H_4−	**A22**	134
5-dimethylamino-1-naphthyl-	**A23**	120
1-naphthyl	**A24**	136
2-naphthyl	**A25**	132

[a] K_I values were obtained from Dixon plots using a linear regression program, from at least three different assays. Errors (data not shown) were ±5−10% of the shown values.

* Corresponding author phone: +61 8 6488 8562; fax: +61 8 6488 1005; e-mail: bwc@theochem.uwa.edu.au.
† The University of Western Australia.
‡ Università degli Studi.

10.1021/ci050191v CCC: $30.25
Published on Web 08/04/2005

Table 2. Derivatives (*p*-RNH−C_6H_4−SO_2N=$C(NH_2)_2$) **A26**−**A33** Obtained from Sulfaguanidine **A11** as Lead, with Their Inhibition Data against Human Thrombin[b]

R[c]	compound	K_I[a] (nM)
Cbz-D-Phe	**A26**	54
ts-D-Phe	**A27**	43
ts-L-Pro	**A28**	48
ts-D-PhePro	**A29**	12
Cbz-D-PhePro	**A30**	13
ts-GlyHis	**A31**	18
ts-*β*-AlaHis	**A32**	15
ts-L-ProGly	**A33**	21

[a] K_I values were obtained from Dixon plots using a linear regression program, from at least three different assays. Errors (data not shown) were ±5−10% of the shown values. [b] From ref 5a. [c] Cbz = $PhCH_2OCO$; ts = p-$MeC_6H_4SO_2NHCO$−; these groups acylate the amino-terminal H_2N moiety. When configuration is not specified, L-amino acid moieties were employed. The usual polypeptide formalism is used: the amino-terminal residue is written first (and it is always protected either by the Cbz or the ts moieties), whereas the carboxyterminal residue is acylating the sulfaguanidine N-4 amino group.

Table 3. Sulfonyl-*O*-Methyl-Isoureas (RSO_2N=$C(NH_2)$OMe) **B1**−**B25** Prepared in the Present Study, with Their Inhibition Data against Human Thrombin

R	compound	K_I[a] (nM)
p-F−C_6H_4−	**B1**	280
p-Cl−C_6H_4−	**B2**	266
p-Br−C_6H_4−	**B3**	265
p-I−C_6H_4−	**B4**	243
p-CH_3−C_6H_4−	**B5**	328
p-O_2N−C_6H_4−	**B6**	189
m-O_2N−C_6H_4−	**B7**	213
o-O_2N−C_6H_4−	**B8**	355
3-Cl-4-O_2N−C_6H_3−	**B9**	177
p-AcNH−C_6H_4−	**B10**	202
p-H_2N−C_6H_4−	**B11**	106
m-H_2N−C_6H_4−	**B12**	99
C_6F_5−	**B13**	153
o-HOOC−C_6H_4−	**B14**	325
m-HOOC−C_6H_4−	**B15**	197
p-HOOC−C_6H_4−	**B16**	133
o-HOOC−C_6Br_4−	**B17**	239
p-CH_3O−C_6H_4−	**B18**	276
2,4,6-$(CH_3)_3$−C_6H_2−	**B19**	348
4-CH_3O-3-H_2N−C_6H_3−	**B20**	96
2-HO-3,5-Cl_2−C_6H_2−	**B21**	170
4-Me_2N−C_6H_4−N=N−C_6H_4−	**B22**	169
5-dimethylamino-1-naphthyl-	**B23**	138
1-naphthyl	**B24**	154
2-naphthyl	**B25**	147

[a] K_I values were obtained from Dixon plots using a linear regression program, from at least three different assays. Errors (data not shown) were ±5−10% of the shown values.

Table 4. Derivatives (*p*-RNH−C_6H_4−SO_2N=$C(NH_2)$OMe) Obtained from Sulfanilyl-*O*-Methyl-Isourea **B11** as Lead and Their Inhibition Data against Human Thrombin

R[b]	compound	K_I[a] (nM)
Cbz-D-Phe	**B26**	62
ts-D-Phe	**B27**	60
ts-L-Pro	**B28**	63
ts-D-PhePro	**B29**	18
Cbz-D-PhePro	**B30**	16
ts-GlyHis	**B31**	21
ts-*β*-AlaHis	**B32**	19
ts-L-ProGly	**B33**	27

[a] K_I values were obtained from Dixon plots using a linear regression program, from at least three different assays. Errors (data not shown) were ±5−10% of the shown values. [b] Cbz = $PhCH_2OCO$; ts = p-$MeC_6H_4SO_2NHCO$-; these groups acylate the amino-terminal H_2N moiety. When configuration is not specified, it means that L-amino acid moieties were employed. The usual polypeptide formalism is used: the amino-terminal residue is written first (and it is always protected either by the Cbz or the ts moieties), whereas the carboxyterminal residue is acylating the sulfanilyl-*O*-methyl-isourea N-4 amino group.

Table 5. Sulfonylaminoguanidines (RSO_2NHN=$C(NH_2)_2$) **C1**−**C25** Prepared in the Present Study, with Their Inhibition Data against Human Thrombin

R	compound	K_I[a] (nM)
p-F−C_6H_4−	**C1**	225
p-Cl−C_6H_4−	**C2**	212
p-Br−C_6H_4−	**C3**	203
p-I−C_6H_4−	**C4**	177
p-CH_3−C_6H_4−	**C5**	270
p-O_2N−C_6H_4−	**C6**	166
m-O_2N−C_6H_4−	**C7**	170
o-O_2N−C_6H_4−	**C8**	324
3-Cl-4-O_2N−C_6H_3−	**C9**	154
p-AcNH−C_6H_4−	**C10**	172
p-H_2N−C_6H_4−	**C11**	91
m-H_2N−C_6H_4−	**C12**	88
C_6F_5−	**C13**	123
o-HOOC−C_6H_4−	**C14**	205
m-HOOC−C_6H_4−	**C15**	112
p-HOOC−C_6H_4−	**C16**	97
o-HOOC−C_6Br_4−	**C17**	213
p-CH_3O−C_6H_4−	**C18**	227
2,4,6-$(CH_3)_3$−C_6H_2−	**C19**	219
4-CH_3O-3-H_2N−C_6H_3−	**C20**	219
2-HO-3,5-Cl_2−C_6H_2−	**C21**	98
4-Me_2N−C_6H_4−N=N−C_6H_4−	**C22**	139
5-dimethylamino-1-naphthyl-	**C23**	130
1-naphthyl	**C24**	125
2-naphthyl	**C25**	129

[a] K_I values were obtained from Dixon plots using a linear regression program, from at least three different assays. Errors (data not shown) were ±5−10% of the shown values.

ability of the molecule to lay on the receptor surface, or in a pocket, in two ways, flipping about, in this case, the 1−4 axis.[9] Which way it does lay will be determined by energetic considerations. So how do we find which way the drugs in a series actually bind?

In the absence of structural data the only way in which we can proceed is to carry out regressions with every combination of each drug in both orientations and find which regression fits best. For N drugs this requires 2^N regressions. With N at all large this rapidly becomes impossible if we proceed directly. Using the iterative technique of Kishida and Manabe is fast but will usually find a local rather than a global minimum. The problem is one of combinatorial optimization, and the classic solution to such problems requires techniques such as genetic algorithms, evolutionary programming, or, in this case, simulated annealing. We refer to this procedure as flip regression, and we have discussed it in a previous publication.[9] To date however we have not discussed the very important problem of using the technique to predict the activity of untested compounds. This is the subject of the present contribution.

When we have the regression coefficients for a series of drugs we can use them directly to predict the activity of an unknown substance. But there are two predicted activities: that for the formula as written and that for the "flipped" version, with the ring numbered in the opposite sense. Which

Chart 1. Structures of the CA Inhibitors Considered—the Rings 1−25 and Tails A−F

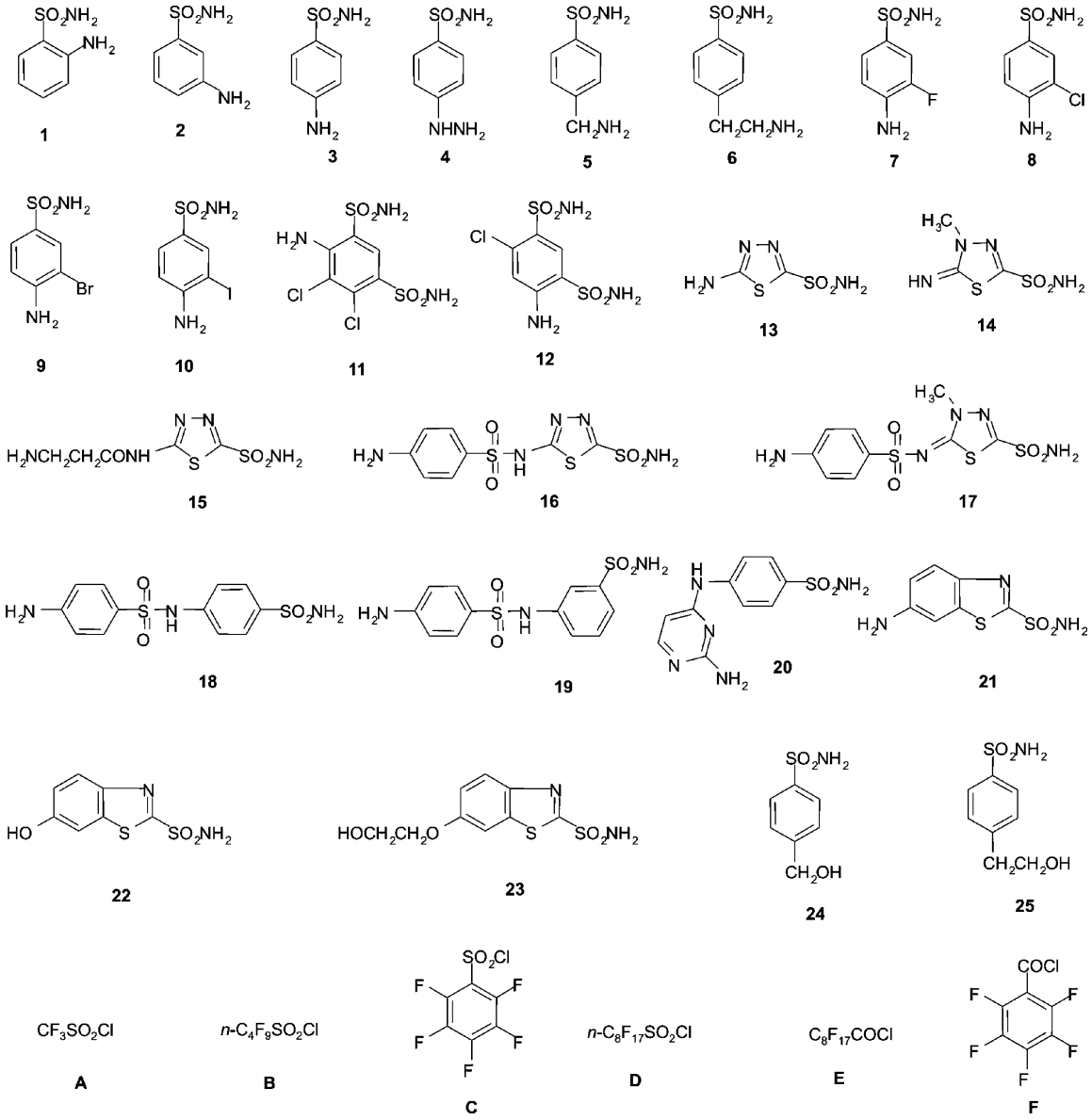

Table 6. Derivatives **C26**−**C33** (*p*-RNH−C_6H_4-SO_2NHN=$C(NH_2)_2$) Obtained from Sulfanilylaminoguanidine **C11** as Lead, with Their Inhibition Data against Human Thrombin[b]

R[c]	compound	K_I[a] (nM)
Cbz-D-Phe	**C26**	50
ts-D-Phe	**C27**	44
ts-L-Pro	**C28**	43
ts-D-PhePro	**C29**	10
Cbz-D-PhePro	**C30**	11
ts-GlyHis	**C31**	15
ts-β-AlaHis	**C32**	10
ts-L-ProGly	**C33**	15

[a] K_I values were obtained from Dixon plots using a linear regression program, from at least three different assays. Errors (data not shown) were ±5−10% of the shown values. [b] From ref 5a. [c] Cbz = $PhCH_2OCO$; ts = p-$MeC_6H_4SO_2NHCO$-; these groups acylate the amino-terminal H_2N moiety. When configuration is not specified, L-amino acid moieties were employed. The usual polypeptide formalism is used: the amino-terminal residue is written first (and it is always protected either by the Cbz or the ts moieties), whereas the carboxyterminal residue is acylating the sulfaguanidine N-4 amino group.

Table 7. Activities of the CA II Inhibitors of Chart 1

Cmp	K_I	Cmp	K_I	Cmp	K_I	Cmp	K_I	Cmp	K_I	Cmp	K_I
A01	20500	B01	250	C01	24	D01	96	E01	103	F01	35
A02	18700	B02	170	C02	10	D02	67	E02	69	F02	19
A03	10900	B03	160	C03	10	D03	62	E03	66	F03	17
A04	14800	B04	180	C04	16	D04	110	E04	107	F04	23
A05	5000	B05	150	C05	15	D05	60	E05	62	F05	20
A06	600	B06	150	C06	15	D06	54	E06	55	F06	17
A07	500	B07	98	C07	9	D07	125	E07	51	F07	15
A08	730	B08	425	C08	110	D08	550	E08	55	F08	125
A09	1040	B09	485	C09	125	D09	680	E09	69	F09	156
A11	430	B11	51	C11	15	D11	98	E11	72	F11	38
A12	90	B12	8	C12	5	D12	32	E12	35	F12	12
A13	100	B13	2	C13	0.3	D13	3	E13	5	F13	2
A14	24	B14	2	C14	0.3	D14	3	E14	6	F14	1.5
A15	13	B15	3	C15	0.4	D15	2	E15	2	F15	2
A16	3	B16	2	C16	1.0	D16	2	E16	1.5	F16	8
A17	5	B17	4	C17	1.5	D17	3	E17	2.0	F17	7
A18	21	B18	15	C18	8	D18	12	E18	11	F18	18
A19	23	B19	20	C19	8	D19	21	E19	16	F19	36
A20	25	B20	20	C20	11	D20	20	E20	22	F20	27
A21	0.9	B21	0.5	C21	0.2	D21	0.6	E21	1.0	F21	0.5
A22	0.9	B22	0.5	C22	0.3	D22	0.5	E22	3	F22	0.6
A23	1.0	B23	1.0	C23	0.5	D23	0.6	E23	2	F23	0.7
A24	5100	B24	460	C24	40	D24	540	E24	550	F24	54
A25	550	B25	385	C25	35	D25	440	E25	410	F25	50

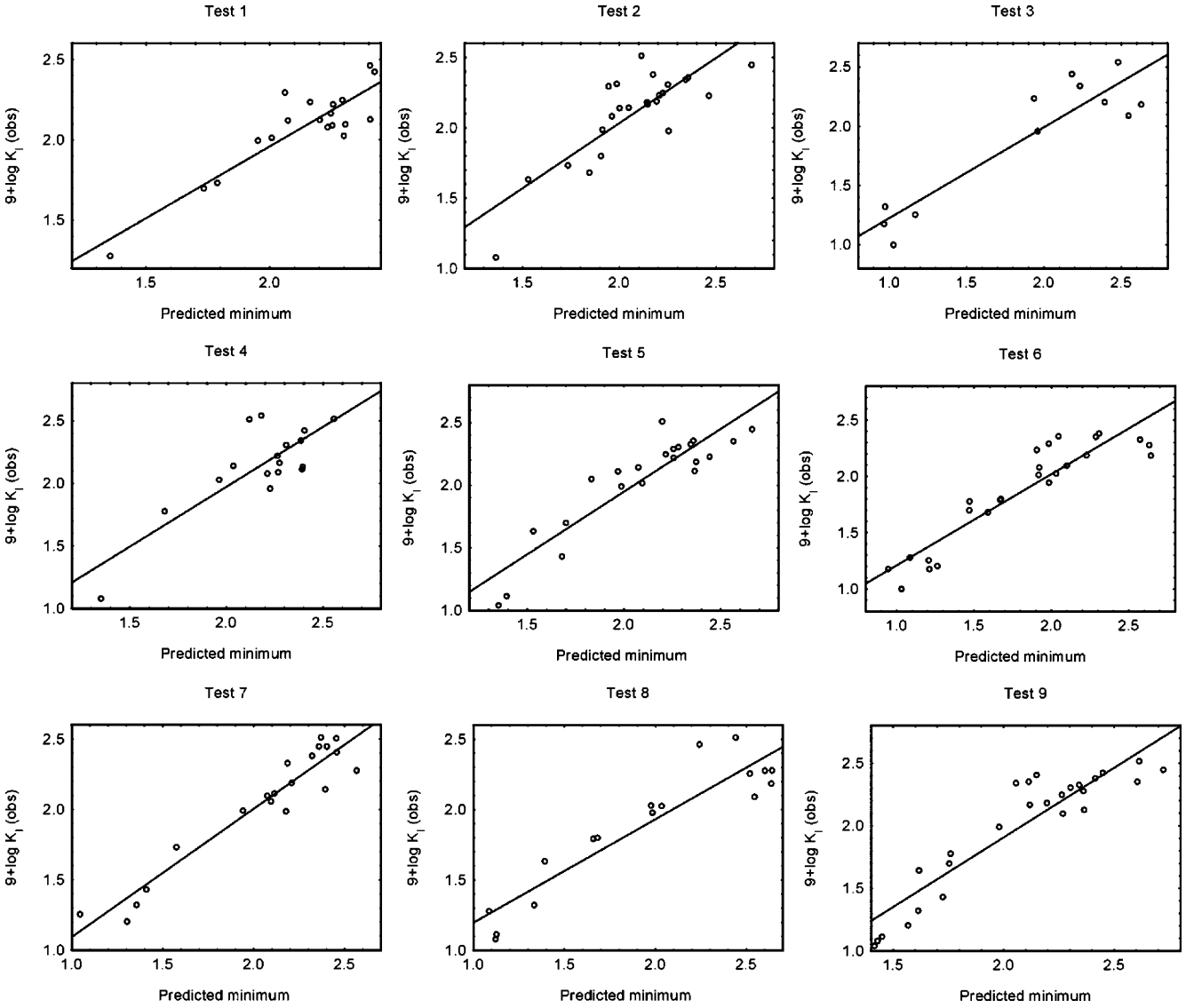

Figure 1. Plots of observed log K_I against the lower predicted log K_I for the thrombin inhibitor test sets.

Table 8. Descriptors Considered for Thrombin Inhibitors

symbol	*a*	descriptor
log P	*	lipophilicity − ClogP
I_A		indicator − 1 for isoureas, 0 otherwise
I_S	*	indicator − 1 for sulfonylaminoguanidines, 0 otherwise
Π_{xx}		polarizability component − longest axis
Π_{yy}	*	polarizability component − intermediate axis
Π_{zz}	*	polarizability component − shortest axis
E_{SH}	*	energy of second highest occupied π-like orbital
E_H	*	energy of highest occupied π-like orbital
E_L	*	energy of lowest unoccupied π-like orbital
E_{SL}	*	energy of second lowest unoccupied π-like orbital
D_I	*	local dipole index
ΔH_S	*	solvation energy from AM1/COSMO calculation
Θ_H	*	orientation of node in highest occupied π-like orbital
Θ_L	*	orientation of node in lowest unoccupied π-like orbital
Q_{NA}	*	potential-based charge on guanidine N
Q_{NB}		potential-based charge on sulfonamide N
Q_H		potential-based charge on sulfonamide H
Q_S		potential-based charge on sulfonamide S
Q_O		potential-based charge on sulfonamide O
Q_C		potential-based charge on C bearing sulfonamide

[a] *Variable retained after running FLIPSTEP.

Table 9. Descriptors Considered for the CA Inhibitors

symbol	*a*	descriptor
Φ_H	*	angle between node in highest occupied π orbital and SO_2NH_2 group, DFT (degrees)
Φ_L	*	angle between node in lowest unoccupied π orbital and SO_2NH_2 group, DFT (degrees)
E_H	*	energy of highest occupied π orbital, DFT (eV)
E_{SH}		energy of second highest occupied π orbital, DFT (eV)
E_L		energy of lowest unoccupied π orbital, DFT (eV)
E_{SL}	*	energy of second lowest unoccupied π orbital, DFT (eV)
Q_M		mean absolute Mulliken charge over all atoms, AM1
Q_N		Mulliken charge on sulfonamide N, DFT
Q_O	*	Mulliken charge on sulfonamide O, DFT
Q_S		Mulliken charge on sulfonamide S, DFT
Q_C	*	Mulliken charge on C attached to sulfonamide, DFT
Q_H		Mulliken charge on sulfonamide H, DFT
Π_{xx}		component of polarizability, longest axis, AM1 (Å^3)
Π_{yy}		component of polarizability, intermediate axis, AM1 (Å^3)
Π_{zz}		component of polarizability, shortest axis, AM1 (Å^3)
ΔH_S	*	difference between heats of formation with and without COSMO, AM1 (kcal)
π_{tail}	*	ClogP value for tail

[a] *Variable retained after running FLIPSTEP.

do we choose? That which gives the maximum value of the predicted activity, assuming that activity reflects the energy of binding. So we have written FLIPPRED, an addition to the authors MARTHA package. The program accepts as data

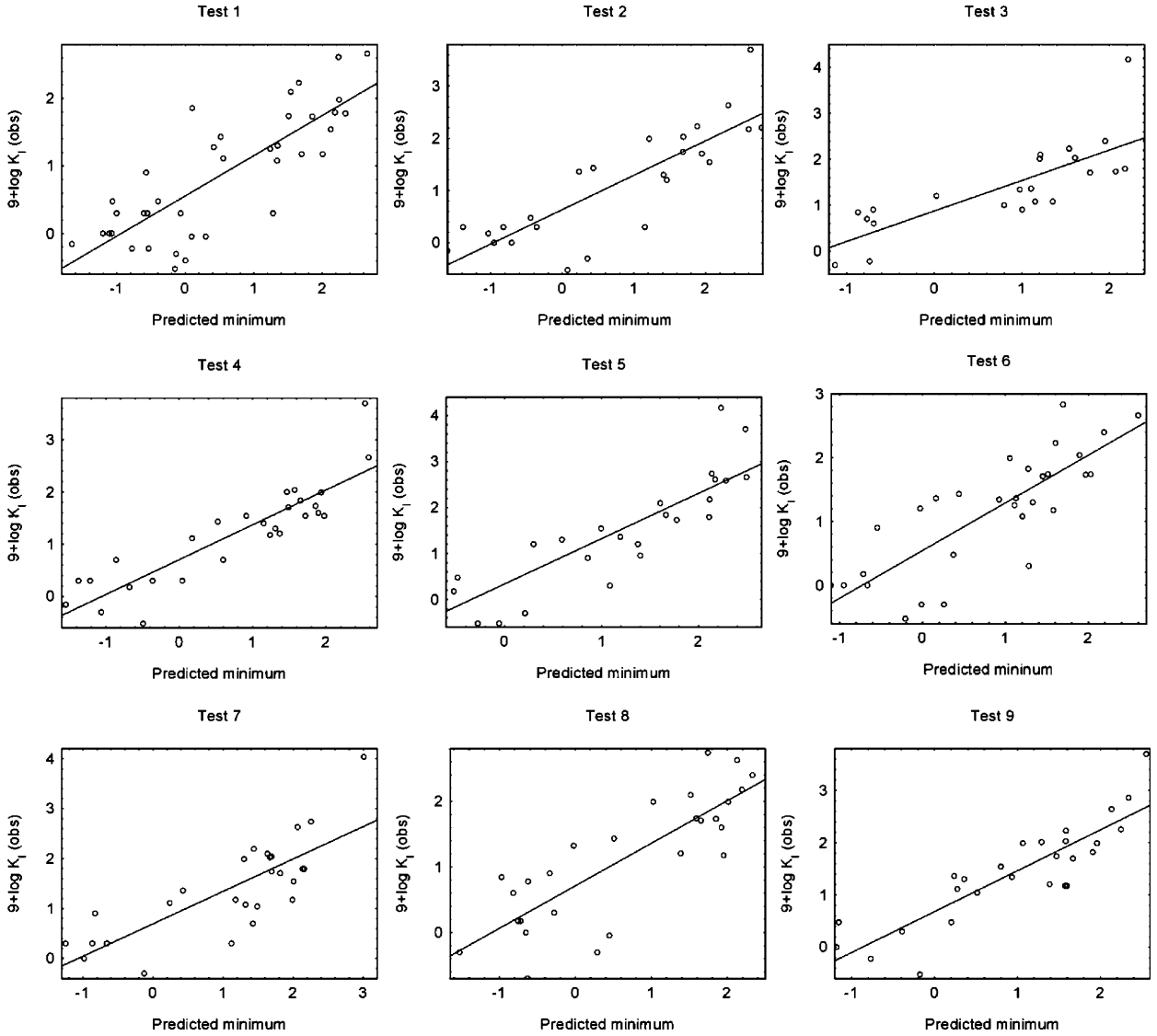

Figure 2. Plots of observed log K_I against the lower predicted log K_I for the CA inhibitor test sets.

a MARTHA file with activity as the last variable as a training set, carries out flip regression on it, then accepts another MARTHA file with identical variables as a test set and computes both a high and a low activity, and appends these to an output MARTHA file, along with a status variable that indicates whether the compound has or has not flipped in order to give the maximum activity. In practice the test and training sets are obtained from a preliminary regression using FLIPSTEP, a stepwise flip regression and backward stepwise variable elimination program.

We shall consider two series of drugs, both of which we have described previously. One is a series of inhibitors of human thrombin,[10] the structures and inhibition constants of which are provided in Tables 1−6. The other is a series of inhibitors of carbonic anhydrase,[11] the structures and inhibition constants of which are provided in Chart 1 and Table 7.

EXPERIMENTAL AND CALCULATIONS

The preparation of the drugs and measurement of their inhibitory properties have been described previously.[10,11] The calculation of the descriptors using MOPAC 93[12] and AM1[13] for the thrombin inhibitors, Gaussian 2003[14] for the CA inhibitors, and the authors program NODANGLE[15] for the angles has also been described previously, and the numbers used here are taken directly from those publications.[11,16] NODANGLE calculates the orientation of π-orbital nodes after a MOPAC or Gaussian calculation and is available free on the WWW. For thrombin the descriptors used were those listed in Table 8, and for CA II they are listed in Table 9. The calculated values are given in the Supporting Information.

The descriptors were reduced in number, and collinear variables and variables of poor significance were eliminated using the program FLIPSTEP,[11] which is part of the authors MARTHA package,[17] using the default parameters SIGMAX = 0.05 and VIFMAX = 30, as described previously. MARTHA is available free of charge on the WWW. FLIPSTEP is a classical backward-stepwise variable selection program adapted to handle flipping and written so that a flippable pair of variables is eliminated only if both members of the pair are unacceptable through either collinearity or poor statistical significance.

The variables selected by FLIPSTEP were isolated from the full data set with the MARTHA routine EDIT. Asterisks

Table 10. Thrombin Inhibition: Regression of Observed log K_I on Low and High Predicted Values, Giving Slope, Intercept, R^2, and Standard Error of Estimate

	low				high			
no.	R^2	s	slope	int	R^2	s	slope	int
1	0.83	0.11	0.75	0.57	0.79	0.13	0.71	0.46
2	0.83	0.13	0.89	0.33	0.85	0.13	0.85	0.17
3	0.87	0.21	0.91	0.28	0.87	0.21	0.91	0.07
4	0.78	0.16	0.89	0.29	0.76	0.17	0.85	0.17
5	0.89	0.14	0.98	0.12	0.87	0.15	0.96	−0.05
6	0.91	0.13	0.94	0.20	0.91	0.14	0.93	−0.02
7	0.90	0.14	0.97	0.21	0.85	0.17	0.95	0.02
8	0.90	0.13	0.93	0.25	0.89	0.14	0.83	0.02
9	0.92	0.15	1.02	0.04	0.91	0.15	1.01	−0.18

Table 11. CA II Inhibition: Regression of Observed log K_I on Low and High Predicted Values, Giving Slope, Intercept, R^2 and Standard Error of Estimate

	low				high			
no.	R^2	s	slope	int	R^2	s	slope	int
1	0.65	0.54	0.59	0.09	0.56	0.61	0.62	0.02
2	0.71	0.60	0.66	0.63	0.55	0.74	0.65	0.11
3	0.61	0.69	0.66	0.87	0.50	0.68	0.62	0.40
4	0.78	0.44	0.67	0.70	0.68	0.54	0.75	0.10
5	0.70	0.68	0.98	0.34	0.53	0.84	1.15	−0.58
6	0.64	0.56	0.73	0.57	0.52	0.64	0.69	0.05
7	0.62	0.60	0.65	0.69	0.55	0.66	0.71	0.17
8	0.70	0.60	0.71	0.72	0.62	0.67	0.76	0.10
9	0.76	0.48	0.78	0.68	0.68	0.55	0.85	−0.02

in Tables 8 and 9 mark the subsets of variables selected for inclusion. Each of the two reduced sets of data was divided into a training set (80%) and a test set (20%) using the MARTHA routine SPLIT. For each of the two data sets nine training and test sets were created. SPLIT accepts a probability of inclusion in the training set and splits on that basis. The actual number in each set varies from run to run. Each of the training sets was tested with FLIPSTEP. In most cases no further variables were rejected. Those training sets for which FLIPSTEP did reject further variables were discarded. Thus the training sets that were used has no single variables or complete variable pairs with a statistical significance greater than 0.05. This step is important as it was found that training sets containing nonsignificant variables can give bizarre predictions.

The new program FLIPPRED was applied to each training-set test-set pair. FLIPPRED carries out a standard flip regression on the training set to generate regression coefficients and then applied these to the test set to give two predicted values for each drug in the test set, corresponding to the flipped and unflipped orientation of the drug. The program sorted these predictions into a low and high value and appended these, together with a flip status value, to an output file.

It is assumed that because a low value of the inhibition constant indicates tight bonding of inhibitor to enzyme, the low value will be the relevant prediction. The observed inhibition constant was then regressed on the high and low predicted value using the MARTHA routine MULTLR. The results of these regressions are tabulated in Tables 10 and 11 for thrombin and CA II, respectively, and a plot of observed activity against low predicted activity for the nine test runs on the two enzymes in Figures 1 and 2.

RESULTS

The results of stepwise flip regressions on the two original data sets are given in the following two equations:

For CA II inhibition

$$\log K_I = 7.53(10.2)\, Q_M - 10.29(4.4)\, Q_C - 1.32(9.5)\, E_H - 0.709(4.8)\, E_{5L} + 0.8105(6.8)\, \pi_{tail} + 0.0354(8.6)\, \Delta H_S - 38.65(18.1)\, Q_O - 0.1660(2.1)\, \cos 2\Phi_H + 0.173(11.7)\, \sin 2\Phi_H - 0.8583(3.8)\, \cos 4\Phi_L - 0.1503(2.1)\, \sin 4\Phi_L + 53.41 \quad (1)$$

$$N = 144,\ R^2 = 0.876,\ Q^2 = 0.854,\ F = 85.06,\ s = 0.39,\ \alpha = 7 \times 10^{-61}$$

For thrombin inhibition:

$$\log K_I = 0.108(13.0) \log P + 0.105(8.5)\, I_S - 0.0024(13.6)\, \Pi_{yy} - 0.0022(12.2)\, \Pi_{zz} + 0.497(6.5)\, D_l + 0.0018(5.7)\, \Delta H_S + 0.159(5.2)\, E_{SH} - 0.115(5.2)\, E_H - 0.194(5.8)\, E_L + 0.099(3.9)\, E_{SL} + 0.904(5.5)\, Q_N - 0.030(2.1)\, \cos 2\Phi_H + 0.155(18.0)\, \sin 2\Phi_H + 0.044(2.6)\, \cos 4\Phi_L - 0.013(1.7)\, \sin 4\Phi_L + 3.46 \quad (2)$$

$$N = 96,\ R^2 = 0.986,\ Q^2 = 0.979,\ F = 365.6,\ s = 0.055,\ \alpha = 3 \times 10^{-152}$$

Here N is the number of compounds, R^2 is the square of the multiple correlation coefficient, Q^2 is the cross-validated value of the same, F is the Fisher variance ratio, s is the standard error of estimate, and α is the statistical significance based on a randomization test with 5000 randomizations. The numbers in parentheses are Students t values for the individual terms. A t greater than approximately 2 is indicative of statistical significance at the 0.05 confidence level, and higher values are indicative of more important terms.

CONCLUSIONS

It is apparent from eq 1 that by far the most important term for CA II inhibitors is Q_O, and from eq 2 that the most important term for the thrombin inhibitors is the angle Φ_H. Note also that one of the terms for thrombin inhibition is not significant at the 0.05 confidence level. Within a flippable pair of descriptors, when one term is significant, either both must be included, or both must be excluded.

Tables 10 and 11 list the results of nine splittings of the two reduced data sets into training and prediction sets and regressions of the observed prediction set activity on the low and high predicted activity. The observed activity in the prediction set has in no way been used in the prediction. Because a low value of K_I is indicative of strong binding it would be expected that the low value would correlate better, and that ideally R^2 would be near 1, s near zero, the slope close to unity, and the intercept close to zero. This indeed turns out to be the case. The correlation coefficient and slope are higher, and the standard error of estimate is lower for the low prediction than the high. Only the intercept goes against this trend. These results have in no way been imposed by the flip regression procedure. It follows directly from

thermodynamic considerations and is an independent confirmation of the assumptions of the procedure.

Figures 1 and 2 are plots of observed activity against the low predicted activity and illustrate the predictive ability of the procedure. It will be noted that the R^2 values in Tables 10 and 11 are not as good as those for eqs 1 and 2. In part this is because R^2 for a prediction set is always poorer than that for the training set, and in part because in the training process statistical significance becomes inflated by the swapping process. The very low α values for the randomization trials on the selection subsets of eqs 1 and 2 indicate that this inflation is not dominant, and the statistics of Tables 10 and 11 are entirely uncontaminated with any such inflation. The results confirm that quite good predictions can be obtained by the flip procedure, that the assumption that the lower of the two predicted values of log K_I is the relevant predictor, and that it corresponds to tightest binding is a valid one.

Supporting Information Available: Two tables containing values of all descriptors considered for the CA inhibitors and thrombin inhibitors. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Shulgin, A.; Shulgin, A. *PIHKAL: A Chemical Love Story*; Transform Press: Box 13675, Berkeley, CA, 1991; pp 864−869.

(2) Kishida, K.; Manabe, R. The role of the hydrophobicity of the substituted groups of dichlorphenamide in the development of carbonic anhydrase inhibition. *Med. J. Osaka Univ.* **1980**, *30*, 95−100.

(3) Buchbauer, G.; Klein, C. Th.; Wailzer, B.; Walschann, P. Threshold-based structure−activity relationship of pyrazines with bell-pepper flavor. *J. Agric. Food Chem.* **2000**, *48*, 4273−4278.

(4) Kakeya, N.; Aoki, M.; Kamada, A.; Yata, N. Structure−activity relation of sulfonamide carbonic anhydrase inhibitors. 1. *Chem. Pharm. Bull.* **1969**, *17*, 1010−1018.

(5) Hansch, C.; McLarin, J.; Klein, T.; Langridge, R. A quantitative structure−activity relationship and molecular graphics study of carbonic anhydrase inhibitors. *Mol. Pharmacol.* **1985**, *27*, 493−498.

(6) Carotti, A.; Raguseo, C.; Campagna, F.; Langridge, R.; Klein, T. E. Inhibition of carbonic anhydrase by substituted benzenesulfonamides. A reinvestigation by QSAR and molecular graphics analysis. *QSAR* **1989**, *8*, 1−10.

(7) DeBenedetti, P. G.; Menziani, M. C.; Frassineti, C. A quantum chemical QSAR study of carbonic anhydrase inhibition by sulfonamides. Sulfonamide carbonic anhydrase inhibitors: quantum chemical QSAR. *QSAR* **1985**, *4*, 23−28.

(8) Clare, B. W. Structure−Activity Correlations for Psychotomimetics. 1. Phenylalkylamines: Electronic, Volume, and Hydrophobicity Parameters. *J. Med. Chem.* **1990**, *33*, 687−702.

(9) Clare, B. W. QSAR of benzene derivatives: Comparison of classical descriptors, quantum theoretic parameters and flip regression, exemplified by phenylalkylamine hallucinogens. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 611−633.

(10) Clare, B. W.; Scozzafava, A.; Briganti, F.; Iorga, B.; Supuran, C. T. Protease Inhibitors Part 2: Synthesis and QSAR study of nonbasic thrombin inhibitors incorporating sulfonylguanidine moieties as S1 anchoring groups. *J. Enzyme Inhib.* **2000**, *15*, 235−264.

(11) Clare, B. W.; Supuran, C. T. A Physically Interpretable Quantum-Theoretic QSAR for some Carbonic Anhydrase Inhibitors with Diverse Aromatic Rings, Obtained by a New QSAR Procedure. *Bioorg. Med. Chem.* **2005**, *13*, 2197−2211.

(12) Stewart, J. J. P. *MOPAC 93.00 (1993)*; Fujitsu Ltd.: Tokyo, Japan. Also Stewart, J. J. P. *MOPAC93, Release 2. QCPE Bull.* **1995**, *15*, 13−14. (Copyright Fujitsu 1993, all rights reserved).

(13) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P.Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.

(14) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision B.04*; Gaussian, Inc.: Pittsburgh, PA, 2003.

(15) File nodangle.zip on site http://www.chem.biomedchem.uwa.edu.au/staff/homepages/BrianClare.

(16) Supuran, C. T.; Clare, B. W. A Quantum Theoretic QSAR of Benzene Derivatives: Some Enzyme Inhibitors. *J. Enzyme Inhib. Med. Chem.* **2004**, *19*, 237−248.

(17) File martha.zip on site http://www.chem.biomedchem.uwa.edu.au/staff/homepages/BrianClare.

CI050191V