Journal of Computer-Aided Molecular Design, 9 (1995) 396–406
ESCOM

J-CAMD 309

A new procedure for improving the predictiveness of CoMFA models and its application to a set of dihydrofolate reductase inhibitors

Romano T. Kroemer* and Peter Hecht**

Sandoz-Forschungsinstitut, Brunnerstrasse 59, A-1235 Vienna, Austria

Received 28 December 1994
Accepted 19 July 1995

Keywords: Comparative molecular field analysis; PLS; Dihydrofolate reductase inhibitors; Realignment; Field fit

Summary

A new automated procedure to improve the predictive quality of CoMFA models for both training and test sets is described. A model of greater consistency is generated by performing small reorientations of the underlying molecules for which too low activities are calculated. In order to predict activities of test compounds, the most similar molecules in the previously optimized model are identified and used as a basis for the prediction. This method has been applied to two independent sets of dihydrofolate reductase inhibitors (80 compounds each, serving as training sets), resulting in a significant increase of the cross-validated r^2 value. For both models, the predictive r^2 value for a test set consisting of 70 compounds was improved substantially.

Introduction

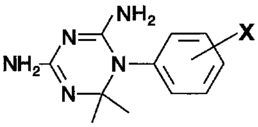
Comparative molecular field analysis (CoMFA) is a relatively new 3D QSAR method. Since its first publication [1], a number of applications in the field of medicinal chemistry have become known [2–7; Refs. 2 through 7 represent a more or less arbitrary collection of publications and do not cover all work done in this field]. The method aims to establish a relationship between the biological activities and the steric and/or electrostatic properties of a set of compounds. After definition of a superposition rule for the 3D representation of these molecules, the steric and electrostatic interaction energies between a probe atom with a given charge and each of the structures are calculated at the surrounding points of a predefined grid. The outcome of this procedure are matrices having many more columns than rows. In order to derive linear equations from these highly underdetermined matrices, a regression method called partial least squares (PLS) is applied [8–10]. As it operates with latent variables, this method is not sensitive to possible colinearity of the underlying descriptor matrix. PLS analyses are usually performed in combination with cross-validation [1,11], using a 'leave-one-out' procedure, in order to check for consistency of the model under consideration.

By default, the steric interaction energies are calculated using a Lennard-Jones 6–12 potential, characterized by a very steep increase in energy at short distances [12]. This may lead to significantly different energy values at a given grid point comparing two identical molecules that are not perfectly superimposed. Therefore, the definition of a very accurate alignment rule for the compounds with respect to the individual conformations as well as the relative orientation of the structures is of crucial importance. Poorly superimposed molecules would lead to rather random field values at the lattice points close to the structures. As these points have the highest variance in energy values, the statistics of the PLS analysis becomes significantly influenced. A user-specified reorientation of such 'outliers' might increase the consistency of the resulting model [13]. Nevertheless, the novel alignment rule has to be applicable to test compounds as well.

In the following, we present an automated procedure which, in a first step, systematically reorientates the compounds in a training set in order to improve the predictive quality of the corresponding CoMFA. Subsequently,

*To whom correspondence should be addressed at: Physical Chemistry Laboratory, University of Oxford, South Parks Road, Oxford OX1 3QZ, U.K.
**New address: Tripos GmbH, Martin-Kollar-Strasse 15, D-81829 Munich, Germany.

TABLE 1
IDs, STRUCTURES AND BIOLOGICAL ACTIVITIES OF THREE SUBSETS[a]



ID	X	–log(IC$_{50}$)	ID	X	–log(IC$_{50}$)
Subset A					
1	2,5-Cl$_2$	3.43	127	3-O(CH$_2$)$_3$OC$_6$H$_4$-4'-NHCOCH$_2$Br	7.55
5	2-Cl	4.15	131	3-Cl-4-OCH$_2$C$_6$H$_4$-4'-SO$_2$F	7.58
11	4-C$_6$H$_5$	4.70	134	3-CH$_2$NHCONHC$_6$H$_4$-3'-SO$_2$F	7.62
12	2-F	4.74	141	3-Cl-4-O(CH$_2$)$_4$NHCOC$_6$H$_4$-4'-SO$_2$F	7.66
15	4-CH=CHCONHC$_6$H$_4$-4'-SO$_2$F	5.19	145	3-(CH$_2$)$_4$C$_6$H$_3$-3'-Cl-4'-SO$_2$F	7.70
16	3-OCH$_2$CONMe$_2$	5.44	147	4-(CH$_2$)$_4$C$_6$H$_4$-4'-SO$_2$F	7.70
19	4-CH=CHCONHC$_6$H$_4$-3'-SO$_2$F	5.89	150	3-Cl-4-OCH$_2$C$_6$H$_4$-3'-CONMe$_2$	7.72
20	3-CONHC$_6$H$_4$-4'-SO$_2$F	5.96	162	3-CH$_2$NHCONHC$_6$H$_4$-3'-CON(Me)$_2$	7.77
28	4-OCH$_2$CONMe$_2$	6.26	167	4-(CH$_2$)$_3$CONHC$_6$H$_4$-2'-SO$_2$F	7.80
40	3-CONHC$_6$H$_4$-3'-SO$_2$F	6.60	168	3-Cl-4-O(CH$_2$)$_2$NHCONHC$_6$H$_3$-3'-Me-4'-SO$_2$F	7.82
44	4-OCH$_2$CONEt$_2$	6.72	169	3-O(CH$_2$)$_2$OC$_6$H$_4$-4'-SO$_2$F	7.82
45	3-CH$_2$CH(CH$_2$NHCOCH$_2$Br)C$_6$H$_5$	6.72	171	3-Cl-4-(CH$_2$)$_2$C$_6$H$_4$-4'-SO$_2$F	7.85
46	4-Cl-3-O(CH$_2$)$_5$OC$_6$H$_4$-4'-SO$_2$F	6.72	172	3-Cl-4-(CH$_2$)$_2$C$_6$H$_3$-5'-Cl-2'-SO$_2$F	7.85
51	3-C$_6$H$_5$	6.85	173	3-Cl-4-(CH$_2$)$_2$C$_6$H$_3$-3'-Cl-4'-SO$_2$F	7.85
54	3-Cl-4-OCH$_2$C$_6$H$_4$-4'-CONHC$_6$H$_4$-4"-SO$_2$F	6.92	174	3-Cl-4-OCH$_2$CON(CH$_2$CH$_2$)$_2$O	7.85
56	4-CH$_2$CN	6.92	182	3-Cl-4-O(CH$_2$)$_2$NHCONHC$_6$H$_4$-4'-SO$_2$F	7.92
62	3-Cl-4-O(CH$_2$)$_3$OC$_6$H$_4$-4'-SO$_2$F	7.07	185	3-OC$_6$H$_4$-4'NHCOCH$_2$Br	7.92
63	3-NO$_2$	7.07	186	3-Cl-4-(CH$_2$)$_4$C$_6$H$_5$	7.92
65	3-(CH$_2$)$_4$COCH$_2$Cl	7.10	187	4-(CH$_2$)$_4$C$_6$H$_3$-2',4'-Cl$_2$	7.92
79	4-CH$_2$CH(Ph-2"-CH$_3$)CONHC$_6$H$_4$-4'-SO$_2$F	7.24	192	3-Cl-4-OCH$_2$C$_6$H$_3$-4'-Cl-3'-SO$_2$F	8.00
80	3-Cl-4-OCH$_2$C$_6$H$_4$-3'-CONHC$_6$H$_4$-3"-SO$_2$F	7.24	201	3-Cl-4-(CH$_2$)$_4$C$_6$H$_4$-3'-SO$_2$F	8.03
83	3-Cl-4-OCH$_2$C$_6$H$_3$-5'-Cl-2'-SO$_2$F	7.27	205	3-Cl-4-(CH$_2$)$_2$C$_6$H$_3$-4'-Cl-2'-SO$_2$F	8.05
86	3-Cl-4-O(CH$_2$)$_3$NHCONHC$_6$H$_4$-3'-SO$_2$F	7.28	206	4-CH$_2$C$_6$H$_5$	8.05
87	4-(CH$_2$)$_2$CONEt$_2$	7.28	208	3-Cl-4-O(CH$_2$)$_3$NHCONHC$_6$H$_3$-4'-Me-3'-SO$_2$F	8.06
90	4-CH(CH$_3$)CH$_2$CONHC$_6$H$_4$-4'-SO$_2$F	7.29	209	4-CH$_2$CONHC$_6$H$_4$-3'-SO$_2$F	8.06
93	4-(CH$_2$)$_2$CON(CH$_2$CH$_2$)$_2$O	7.32	211	3-Cl-4-OCH$_2$C$_6$H$_4$-4'-SO$_3$C$_6$H$_4$-3"-CF$_3$	8.09
94	4-O(CH$_2$)$_3$NHCONHC$_6$H$_4$-3'-SO$_2$F	7.32	219	3-O(CH$_2$)$_2$OC$_6$H$_4$-4'-NHCOCH$_2$Br	8.13
96	3-CH$_2$CONHC$_6$H$_4$-4'-SO$_2$F	7.34	220	3-Cl-4-OCH$_2$C$_6$H$_4$-3'-CONEt$_2$	8.14
97	4-CH$_2$NHCONHC$_6$H$_4$-4'-SO$_2$F	7.35	221	3-Cl-4-(CH$_2$)$_4$C$_6$H$_4$-4'-SO$_2$F	8.14
99	3-Cl-4-OCH$_2$C$_6$H$_3$-6'-Cl-3'-SO$_2$F	7.38	223	4-(CH$_2$)$_4$OC$_6$H$_4$-4'-SO$_2$F	8.14
101	3-Cl-4-S(CH$_2$)$_2$CONHC$_6$H$_4$-4'-SO$_2$F	7.39	226	3-Cl-4-OCH$_2$C$_6$H$_4$-4'-SO$_2$OC$_6$H$_5$	8.20
103	3-Cl-4-OCH$_2$C$_6$H$_4$-4'-CONHC$_6$H$_4$-3"-SO$_2$F	7.41	231	4-(CH$_2$)$_4$OC$_6$H$_5$	8.24
104	4-(CH$_2$)$_2$NHSO$_2$C$_6$H$_4$-4'-SO$_2$F	7.41	232	3-Cl-4-OCH$_2$C$_6$H$_4$-4'-SO$_3$C$_6$H$_3$-3",4"-Cl$_2$	8.25
112	3-(CH$_2$)$_4$C$_6$H$_3$-2',4'-Cl$_2$	7.45	235	3-Cl-4-(CH$_2$)$_2$C$_6$H$_3$-3'-Cl-2'-SO$_2$F	8.30
113	3-Cl-4-O(CH$_2$)6OC$_6$H$_4$-4'-SO$_2$F	7.46	240	3-(CH$_2$)$_4$C$_6$H$_3$-4'-Cl-3'-SO$_2$F	8.37
115	3-Cl-4-OCH$_2$CON(CH$_3$)C$_6$H$_4$-4'-SO$_2$F	7.47	241	3-(CH$_2$)$_4$C$_6$H$_4$-4'-NHCOCH$_2$Br	8.38
117	3-Cl-4-OCH$_2$C$_6$H$_4$-4'-SO$_2$NMe$_2$	7.48	244	3-Cl-4-OCH$_2$C$_6$H$_4$-4'-SO$_3$C$_6$H$_4$-4"-F	8.40
123	4-SCH$_2$CONHC$_6$H$_4$-4'-SO$_2$F	7.52	253	3-Cl-4-OCH$_2$C$_6$H$_4$-4'-SO$_3$C$_6$H$_4$-2"-CON(CH$_3$)$_2$	8.63
125	3-CH$_2$NHCONHC$_6$H$_5$	7.52	256	3-Cl-4-OCH$_2$C$_6$H$_4$-4'-SO$_3$C$_6$H$_4$-3"-CON(CH$_3$)$_2$	8.76
Subset B					
2	2-OCH$_3$	3.68	132	4-(CH$_2$)$_2$CONC$_6$H$_4$-4'-SO$_2$F	7.60
4	2-CH$_3$	4.00	133	3-Cl-4-(CH$_2$)$_2$CONHC$_6$H$_4$-4'-SO$_2$F	7.62
7	2,4,5-Cl$_3$	4.38	136	3-Cl-4-OCH$_2$CONEt$_2$	7.64
10	4-COCNHC$_6$H$_4$-3'-SO$_2$F	4.68	139	3-O(CH$_2$)$_2$OC$_6$H$_4$-3'-NHCOCH$_2$Br	7.66
22	3-CH$_2$NHCONEt$_2$	6.11	140	3-Cl-4-SCH$_2$CONHC$_6$H$_4$-3'-SO$_2$F	7.66
25	4-CH$_2$CH(CH$_2$CH$_2$Ph)CONHC$_6$H$_4$-4'-SO$_2$F	6.20	142	4-(CH$_2$)$_3$CONHC$_6$H$_4$-4'-SO$_2$F	7.66
26	3-COCH$_2$Cl	6.21	146	3-Cl-4-(CH$_2$)$_4$C$_6$H$_3$-3'-Cl-4'-SO$_2$F	7.70
30	3-Cl-4-OCH$_2$C$_6$H$_{10}$CH$_2$OC$_6$H$_4$-4'-SO$_2$F	6.37	151	3-Cl-4-OCH$_2$C$_6$H$_4$-4'-SO$_3$C$_6$H$_4$-3"-Cl	7.72
33	4-COCH$_2$Cl	6.45	152	4-OCH$_2$CONHC$_6$H$_4$-4'-SO$_2$F	7.72
36	2,3-Cl$_2$	6.52	153	3-Cl-4-OCH$_2$CONHC$_6$H$_4$-3'-SO$_2$F	7.72
37	2-Cl-4-(CH$_2$)$_4$C$_6$H$_5$	6.54	156	4-CH$_2$NHCONHC$_6$H$_3$-3'-Me-4'-SO$_2$F	7.72
42	4-OCH$_2$CON(CH$_2$)$_4$	6.66	157	4-(CH$_2$)$_2$CONHC$_6$H$_4$-3'-SO$_2$F	7.74
43	3-OCH$_2$CON(Me)C$_6$H$_5$	6.68	158	3,5-Cl$_2$-4-OCH$_2$CONHC$_6$H$_4$-4'-SO$_2$F	7.74

TABLE 1
(continued)

ID	X	–log(IC$_{50}$)	ID	X	–log(IC$_{50}$)
Subset B (continued)					
50	3-OCH$_2$CONHC$_6$H$_5$	6.85	164	3-Cl-4-(CH$_2$)$_4$C$_6$H$_3$-2'-Cl-4'-SO$_2$F	7.77
52	4-CH$_2$CH(Ph)CONHC$_6$H$_4$-3'-SO$_2$F	6.89	165	3-Cl-4-CH$_2$NHCONHC$_6$H$_3$-3'-Me-4'-SO$_2$F	7.80
53	3-Cl-4-OCH$_2$C$_6$H$_4$-3'-CONHC$_6$H$_4$-4"-SO$_2$F	6.92	166	4-O(CH$_2$)$_2$OC$_6$H$_4$-4'-SO$_2$F	7.80
55	3-OCH$_2$CONHC$_6$H$_4$-4'-SO$_2$F	6.92	178	4-OCH$_2$CONHC$_6$H$_5$	7.89
58	3-OCH$_2$C$_6$H$_4$-3'-NHCOCH$_2$Br	6.92	179	4-(CH$_2$)$_2$C$_6$H$_5$	7.89
59	4-CH$_2$CON(Me)C$_6$H$_5$	7.00	183	4-(CH$_2$)$_3$CONHC$_6$H$_4$-3'-SO$_2$F	7.92
60	4-(CH$_2$)$_2$CONMe$_2$	7.05	188	3-Cl-4-(CH$_2$)$_4$C$_6$H$_5$	7.96
61	3-Cl-4-(CH$_2$)$_4$C$_6$H$_3$-5'-Cl-2'-SO$_2$F	7.06	190	3-(CH$_2$)$_4$C$_6$H$_3$-5'-Cl-2'-SO$_2$F	7.96
64	3-(CH$_2$)$_2$COCH$_2$Cl	7.10	191	4-(CH$_2$)$_4$C$_6$H$_3$-2'-Cl-4'-SO$_2$F	7.96
67	4-CH$_2$CON(CH$_2$CH$_2$)$_2$O	7.12	196	3-CH$_2$C$_6$H$_5$	8.00
69	3-Cl-4-OCH(CH$_3$)CONHC$_6$H$_4$-4'-SO$_2$F	7.13	198	3-Cl-4-OCH$_2$C$_6$H$_4$-3'-CON(CH$_2$)$_5$	8.02
70	4-CH$_2$CH(Ph)CONHC$_6$H$_4$-4'-SO$_2$F	7.13	200	4-(CH$_2$)$_2$CONHC$_6$H$_3$-4'-Me-3'-SO$_2$F	8.02
71	3-Cl-4-O(CH$_2$)$_2$O(CH$_2$)$_2$OC$_6$H$_4$-4'-SO$_2$F	7.14	203	4-CH$_2$NHCONHC$_6$H$_4$-3'-SO$_2$F	8.04
72	3-Cl-4-O(CH$_2$)$_3$CONHC$_6$H$_4$-4'-SO$_2$F	7.15	204	4-(CH$_2$)$_2$CON(Me)C$_6$H$_4$-4'-SO$_2$F	8.04
73	3-Cl-4-OCH$_2$CONMe$_2$	7.16	210	4-(CH$_2$)$_2$CONHC$_6$H$_3$-6'-OMe-3'-SO$_2$F	8.08
75	4-Cl-3-O(CH$_2$)$_4$OC$_6$H$_4$-4'-SO$_2$F	7.17	213	3-(CH$_2$)$_4$C$_6$H$_4$-4'-SO$_2$F	8.10
82	3-Cl-4-O(CH$_2$)$_4$CONHC$_6$H$_4$-4'-SO$_2$F	7.24	217	3-Cl-4-(CH$_2$)$_4$C$_6$H$_3$-4'-Cl-3'-SO$_2$F	8.11
88	3-Cl-4-OCH$_2$CON(CH$_2$)$_4$	7.29	222	3-Br-4-OCH$_2$CONHC$_6$H$_4$-4'-SO$_2$F	8.14
91	4-CH$_2$CON(Me)CH$_2$C$_6$H$_5$	7.30	229	4-(CH$_2$)$_2$CONHC$_6$H$_3$-4'-OMe-3'-SO$_2$F	8.24
92	4-(CH$_2$)$_2$CON(Me)CH$_2$C$_6$H$_5$	7.31	230	3-Cl-4-OCH$_2$C$_6$H$_4$-4'-SO$_3$C$_6$H$_4$-3"-CN	8.24
95	3-Cl-4-O(CH$_2$)$_3$NHCOC$_6$H$_4$-4'-SO$_2$F	7.34	233	3-(CH$_2$)$_2$C$_6$H$_4$-4'-NHCOCH$_2$Br	8.26
102	4-(CH$_2$)$_2$C$_6$H$_4$-4'-SO$_2$F	7.41	234	3-Cl-4-(CH$_2$)$_2$C$_6$H$_3$-4'-Cl-3'-SO$_2$F	8.27
105	3-Cl-4-SCH$_2$CONHC$_6$H$_4$-4'-SO$_2$F	7.42	237	3-Cl-4-OCH$_2$C$_6$H$_4$-4'-SO$_3$C$_6$H$_4$-2"-CF$_3$	8.33
106	3-Cl-4-OCH$_2$C$_6$H$_3$-3'-Cl-2'-SO$_2$F	7.42	238	3-(CH$_2$)$_4$OC$_6$H$_5$	8.35
110	3-Cl-4-OCH$_2$C$_6$H$_3$-2'-Cl-4'-SO$_2$F	7.44	247	3-Cl-4-OCH$_2$C$_6$H$_4$-4'-SO$_3$C$_6$H$_4$-3"-CH$_3$	8.44
122	3-Cl-4-OCH$_2$C$_6$H$_5$	7.52	248	3-Cl-4-OCH$_2$C$_6$H$_4$-4'-SO$_3$C$_6$H$_4$-3"-F	8.46
130	3-Cl-4-O(CH$_2$)$_5$OC$_6$H$_4$-4'-SO$_2$F	7.57	254	3-Cl-4-OCH$_2$C$_6$H$_4$-4'-SO$_3$C$_6$H$_4$-3"-CN	8.70
Subset TEST					
6	2-Br	4.25	128	4-(CH$_2$)$_2$CON(Me)C$_6$H$_5$	7.56
8	2-I	4.62	129	3-Cl-4-O(CH$_2$)$_4$OC$_6$H$_4$-4'-SO$_2$F	7.57
14	4-CN	5.14	135	4-(CH$_2$)$_2$NHSO$_2$C$_6$H$_4$-3'-SO$_2$F	7.64
23	3-OCH$_3$	6.17	137	3-O(CH$_2$)$_3$OC$_6$H$_4$-3'-NHCOCH$_2$Br	7.64
24	4-OCH$_2$CON(Me)C$_6$H$_5$	6.17	138	3-O(CH$_2$)$_2$OC$_6$H$_4$-2'-NHCOCH$_2$Br	7.66
31	3-CH(CH$_2$NHCOCH$_2$Br)(CH$_2$)$_3$C$_6$H$_5$	6.37	143	3-Cl-4-O(CH$_2$)$_3$NHCONHC$_6$H$_4$-4'-SO$_2$F	7.68
32	3-CH$_2$NHCON(CH$_2$CH$_2$)$_2$O	6.43	144	3-Cl-4-O(CH$_2$)$_4$NHCONHC$_6$H$_4$-3'-SO$_2$F	7.70
35	4-CH(CH$_2$NHCOCH$_2$Br)(CH$_2$)$_3$C$_6$H$_5$	6.52	148	4-CH$_2$-CONHC$_6$H$_4$-4'-SO$_2$F	7.70
38	3-Cl-4-O(CH$_2$)$_4$OC$_6$H$_4$-4'-SO$_3$C$_6$H$_4$-4"-Cl	6.55	149	4-O(CH$_2$)$_2$OC$_6$H$_4$-4'-NHCOCH$_2$Br	7.70
39	3-CH$_2$NHCOCH$_2$Br	6.58	154	3-Cl-4-OCH$_2$C$_6$H$_4$-3'-SO$_2$F	7.72
41	4-CH$_2$CONMe$_2$	6.63	155	3-Cl-4-OCH$_2$C$_6$H$_3$-6'-Cl-2'-SO$_2$F	7.72
48	4-Cl-3-(CH$_2$)$_4$C$_6$H$_4$-4'-SO$_2$F	6.77	161	3-Cl-4-OCH$_2$C$_6$H$_4$-4'-SO$_3$C$_6$H$_4$-4"-Cl	7.77
66	4-OCH$_2$CON(CH$_2$)$_5$	7.12	163	3-Cl-4-(CH$_2$)$_4$C$_6$H$_4$-2'-SO$_2$F	7.77
68	4-(CH$_2$)$_6$C$_6$H$_4$-4'-SO$_2$F	7.12	170	3-Cl-4-(CH$_2$)$_4$C$_6$H$_3$-4'-Cl-2'-SO$_2$F	7.82
74	3-Cl-4-O(CH$_2$)$_3$CONHC$_6$H$_4$-3'-SO$_2$F	7.17	175	3-Cl-4-OCH$_2$C$_6$H$_4$-3'-CON(CH$_2$CH$_2$)O	7.85
76	4-CH$_2$CH(Ph-3"-Me)CONHC$_6$H$_4$-4'-SO$_2$F	7.17	176	3-Cl-4-OCH$_2$C$_6$H$_4$-3'-CON(CH$_2$)$_4$	7.85
77	3-(CH$_2$)$_2$CONHC$_6$H$_4$-4'-SO$_2$F	7.19	177	3-Cl-4-OCH$_2$CON(Me)C$_6$H$_5$	7.89
78	4-CH$_2$CH(Ph-4"-Me)CONHC$_6$H$_4$-4'-SO$_2$F	7.24	180	4-(CH$_2$)$_2$CONHC$_6$H$_3$-3'-Me-4'-SO$_2$F	7.89
81	3-Cl-4-OCH$_2$C$_6$H$_4$-2'-CONHC$_6$H$_4$-4"-SO$_2$F	7.24	189	3-O(CH$_2$)$_3$OC$_6$H$_4$-4'-SO$_2$F	7.96
84	4-Cl-3-O(CH$_2$)$_2$OC$_6$H$_4$-4'-SO$_2$F	7.27	193	3-(CH$_2$)$_4$C$_6$H$_3$-2'-Cl-4'-SO$_2$F	8.00
89	4-OCH$_2$CON(CH$_2$CH$_2$)$_2$O	7.29	195	3-Cl-4-OCH$_2$C$_6$H$_4$-3'-CONHC$_6$H$_5$	8.00
98	4-(CH$_2$)$_2$CON(C$_3$H$_7$)$_2$	7.35	197	4-(CH$_2$)$_4$C$_6$H$_5$	8.00
100	3-Cl-4-OCH$_2$C$_6$H$_3$-2'-CH$_3$-4-SO$_2$F	7.38	199	3-CH$_2$NHCONHC$_6$H$_4$-3'-OCH$_3$	8.02
107	3-Cl-4-OCH$_2$CONHC$_6$H$_4$-4'-SO$_2$F	7.43	202	3-(CH$_2$)$_4$C$_6$H$_3$-2',4'-Cl$_2$	8.03
108	3-Cl-4-OCH$_2$C$_6$H$_4$-2'-SO$_2$F	7.43	215	3-(CH$_2$)$_2$C$_6$H$_4$-4'-SO$_2$F	8.10
109	3-Cl-4-OCH$_2$C$_6$H$_3$-3'-Cl-4'-SO$_2$F	7.43	218	3-Cl-4-OCH$_2$C$_6$H$_4$-3'-CON(Me)C$_6$H$_5$	8.12
111	3-Cl-4-O(CH$_2$)$_2$OC$_6$H$_4$-4'-SO$_2$F	7.44	224	3-(CH$_2$)$_2$C$_6$H$_5$	8.19
114	4-(CH$_2$)$_2$CONHC$_6$H$_3$-3'-OMe-4'-SO$_2$F	7.46	227	3-Cl-4-(CH$_2$)$_4$C$_6$H$_3$-3'-Cl-2'-SO$_2$F	8.20
116	3-Cl-4-OCH$_2$CON(CH$_2$)$_5$	7.47	236	3-Cl-4-(CH$_2$)$_2$C$_6$H$_3$-2'-Cl-4'-SO$_2$F	8.33
118	3-Cl-4-OCH$_2$C$_6$H$_3$-2'-Cl-3'-SO$_2$F	7.49	239	3-(CH$_2$)$_4$C$_6$H$_5$	8.35

TABLE 1
(continued)

ID	X	$-\log(IC_{50})$	ID	X	$-\log(IC_{50})$
Subset TEST (continued)					
119	$3\text{-}O(CH_2)_4OC_6H_4\text{-}4'\text{-}SO_2F$	7.49	243	$3\text{-}Cl\text{-}4\text{-}OCH_2C_6H_4\text{-}4'\text{-}SO_3C_6H_4\text{-}4''\text{-}OCH_3$	8.40
120	$4\text{-}Cl\text{-}3\text{-}O(CH_2)_3OC_6H_4\text{-}4'\text{-}SO_2F$	7.51	245	$3\text{-}Cl\text{-}4\text{-}OCH_2C_6H_4\text{-}4'\text{-}SO_3C_6H_4\text{-}2''\text{-}OCH_3$	8.40
121	$3\text{-}Cl\text{-}4\text{-}OCH_2C_6H_4\text{-}3'\text{-}CN$	7.51	249	$3\text{-}Cl\text{-}4\text{-}OCH_2C_6H_4\text{-}4'\text{-}SO_3C_6H_4\text{-}3''\text{-}OCH_3$	8.52
124	$3\text{-}Cl\text{-}4\text{-}OCH_2C_6H_3\text{-}4'\text{-}Cl\text{-}2'\text{-}SO_2F$	7.52	251	$3\text{-}Cl\text{-}4\text{-}OCH_2C_6H_4\text{-}4'\text{-}SO_3C_6H_4\text{-}2''\text{-}Cl$	8.62
126	$4\text{-}CH_2CH(Me)CONHC_6H_4\text{-}4'\text{-}SO_2F$	7.55	255	$3\text{-}Cl\text{-}4\text{-}OCH_2C_6H_4\text{-}4'\text{-}SO_3C_6H_4\text{-}2''\text{-}F$	8.74

[a] The three subsets correspond to Table 1 of Ref. 15.

test molecules are predicted after a fit to the structures in the training set. As an example, the classical QSAR data set of Hansch and co-workers was used [14]. From this ensemble of 256 dihydrofolate reductase inhibitors, two training sets consisting of 80 structures each and a test set of 70 compounds were randomly chosen.

Methods

General

All modelling work was performed with the program SYBYL, v. 6.0 or 6.04 [15], run on Silicon Graphics workstations. Table 1 lists the molecule identities (IDs) contained in the three data sets, which were randomly chosen from the compounds listed in Table 1 of Ref. 15. The IDs used in this paper are identical to the original ones. The structures were built according to the rules depicted in Fig. 1. For each compound, the five atoms of the triazine-phenyl moiety indicated in Fig. 1 were initially fitted onto molecule 256 (the most active one) as the reference structure by minimizing the rms difference between the respective atoms. Energy minimizations were performed with the Tripos standard force field [16] without inclusion of electrostatics, using the POWELL minimization technique [17]. The convergence criterion was defined as an energy change ≤ 0.05 kcal/mol between subsequent minimization steps. Partial charges were calculated with MOPAC 5.0 [18] by means of the MNDO method [19]. For the CoMFA models, the magnitude of the regions was defined to extend the ensemble of superimposed molecules by 4.0 Å along the principal axes of a Cartesian coordinate system. The standard grid spacing of 2.0 Å was chosen. An sp^3 carbon with a charge of +1.0 served as a probe atom. In all the CoMFA models, both steric and electrostatic fields were included and CoMFA standard scaling was applied [20]. The standard deviation threshold for exclusion of columns from the PLS analysis was set to 2.0. Cross-validation was performed by means of the 'leave-one-out' technique. The number of components used in the PLS analysis was selected to be that at which the difference in the cross-validated r^2 value (r^2_{cv}) to the next one was less than 0.02. This procedure was chosen since using the number of components at which the initial steep increase of the r^2_{cv} starts to level off has proven to give better predictive CoMFA models [21]. Throughout this paper, this number of components is indicated by an integer behind the r^2_{cv} (e.g., 0.582/4 means four components chosen with a corresponding r^2_{cv} of 0.582). All operations for improvement of the r^2 values were performed by means of macros written in SPL (SYBYL Programming Language). The term rigid-body field fit implies that the rms difference in the sum of steric and electrostatic interaction energies between a compound and some template molecule, averaged across all lattice points, is minimized with respect to the six rigid-body degrees of freedom [1]. In this context, minimization is performed by the Simplex method [22,23], a minimization procedure that requires only function evaluations, not derivatives. The similarity between two given fields is determined by a lattice point-by-lattice point comparison and yields a

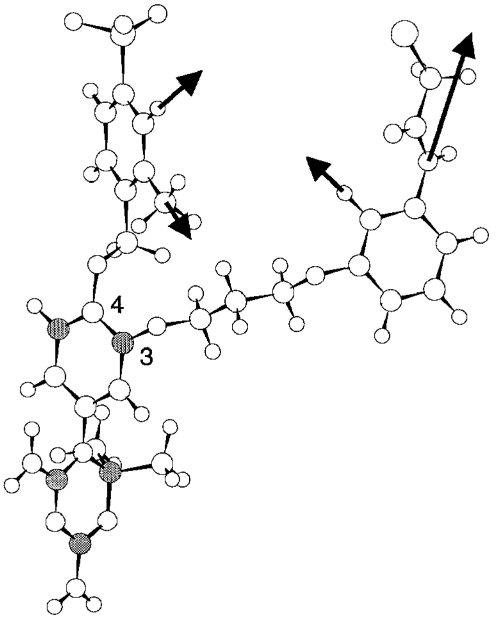
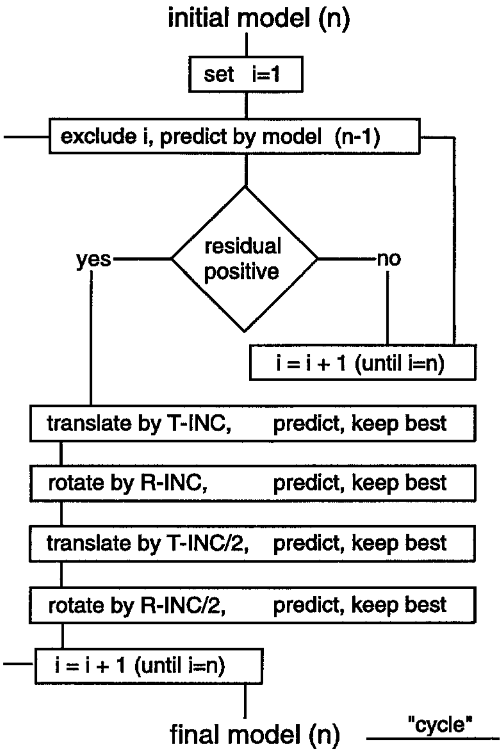


Fig. 1. Representation of the alignment points (shaded grey) and the overall conformation of the compounds. The molecule displayed is a 'hybrid' of two representatives of the major structural modifications in the series (the 3-side chain of compound 137 is linked to structure 100). The arrows indicate the directionality of 'non-para' substituents on the phenyl rings in the 'side chains'. If the number of chemical groups at a phenyl moiety in 'non-para' position is >1, the arrows indicate the position of the largest one. The 'side chains' were built in a staggered conformation.


initial model (n)
set i=1
exclude i, predict by model (n-1)
residual positive
yes
no
i = i + 1 (until i=n)
translate by T-INC, predict, keep best
rotate by R-INC, predict, keep best
translate by T-INC/2, predict, keep best
rotate by R-INC/2, predict, keep best
i = i + 1 (until i=n)
final model (n)
"cycle"

Scheme 1. Flow chart for improvement of the consistency of the training set. The data set consists of n compounds. i refers to the ID of the molecule to exclude, $i \in 1...n$. T-INC and R-INC are the increments for translation and rotation, respectively.

correlation coefficient r taking up values in the range from –1.0 (for complementary fields) to 1.0 (for identical fields). Dissimilar fields will have an r close to 0.0.

Cross-validated r^2 value (r^2_{cv}), training set

The procedure for improving the cross-validated r^2 value of the training set, consisting of n molecules, is outlined in Scheme 1. Each compound is excluded once and its activity is predicted by the CoMFA model derived from the remaining compounds. The residual is defined as:

$$\text{residual} = \text{activity}_{\text{real}} - \text{activity}_{\text{predicted}} \qquad (1)$$

If a molecule has a positive residual, it is translated along the three principal axes of a Cartesian coordinate system by a user-specified increment (all permutations of ±T-INC along x, y and z), resulting in 26 new orientations located at the points making up a cube with the initial position of the compound in its center (Fig. 2). After prediction of these 27 alignments (including the original one), the one with the smallest residual is kept. This provides the starting point for the next 26 reorientations of the molecule, generated by rotation around the three axes of the coordinate system by a predefined value (all permutations of ±R-INC around x, y and z). Subsequently, the increments for rotation and translation are set to half of the original value, and the translation followed by the rotation procedure are continued. The final (re)orientation of the molecule is then included in the data set. The whole process, referred to as a 'cycle', may be repeated several times for the entire set, leading to the final model. In the present work, the values for the translation increment (T-INC) were set to 0.1 Å and those for the rotation increment (R-INC) to 1.0°. Two cycles were performed, thus making up a maximum translation of 0.3 Å along one direction and a maximum rotation of 3.0° around one axis.

Based on the assumption that the compounds bind with an identical mechanism to the same binding site, they can be considered as members of a 'congeneric series'. Consequently, for each compound in a given model, a certain orientation exists where it exhibits optimum activity. Reorientations of molecules with a positive residual (i.e., their activities are predicted too low) may be interpreted as a search for such a local minimum [24]. On the other hand, in order to minimize the negative residual of a given compound, its activity has to be lowered. For such a task, no well-defined procedure exists; any arbitrary translation or rotation may result in decreased activity. Therefore, those molecules having a negative residual were not optimized.

Predictive r^2 value (r^2_{pred}), test set

The calculation of r^2_{pred} was based solely on the molecules in the test set; this parameter is defined in analogy to the cross-validated r^2 [1,11]:

$$r^2_{\text{pred}} = \frac{\text{SD} - \text{PRESS}}{\text{SD}} \qquad (2)$$

where SD is the variance of the biological activities of the molecules in the test set around the mean activity of the training set molecules. 'PRESS' represents the sum of the

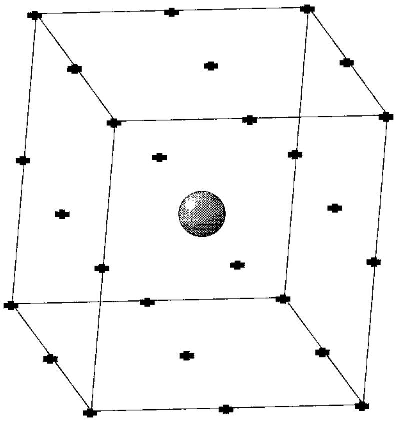
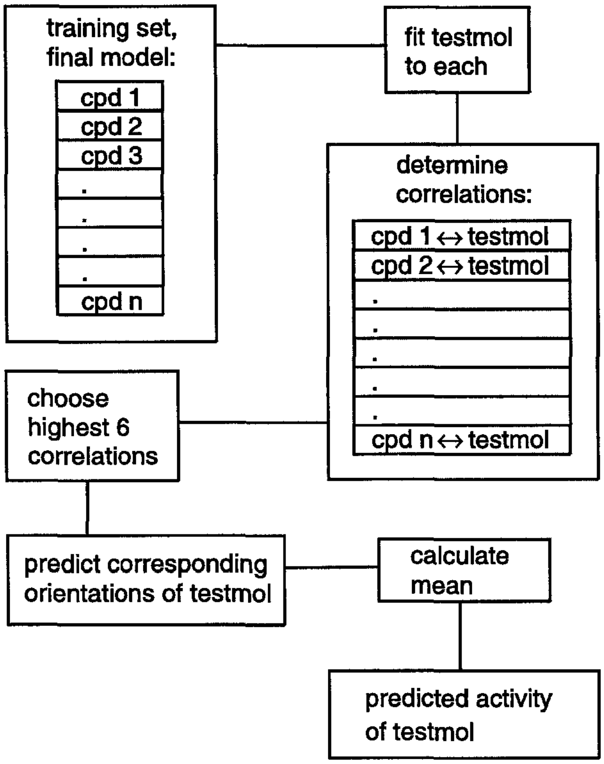


Fig. 2. Schematic representation of the starting point, situated at the center of the cube, and the 26 new locations of a molecule in the translation procedure.



Scheme 2. Representation of the procedure for predicting the activity of a molecule (testmol). The fit may be a point fit or a rigid-body field fit. The correlations are determined using a weighting field, as defined in Eq. 3, and are scaled by the steric and electrostatic contributions of the model (normalized to 1.0, Eq. 4).

squared differences between predicted and actual target property values for every compound in the test set. A likewise negative r^2_{pred} reflects a complete lack of predictive ability of the model under consideration.

Prediction of test molecules

As some of the molecules in the training set were reorientated during the refinement of the CoMFA model derived from the training set, the selection of a reference molecule for a compound to be predicted becomes of crucial importance. Therefore, an automated procedure for the identification of appropriate reference molecules in the training set, alignment and subsequent prediction of the test molecules was designed (Scheme 2). This method can be divided into three main parts:

(1) In the first step, each molecule in the training set is used as a reference molecule for the fitting of a given test compound. This fitting is done either by a rigid-body field fit ('field-fitted') or through a minimization of the sum of pairwise atom–atom distances ('point-fitted'). For the latter fit, the same atoms were used as in the definition of the original alignment rule (Fig. 1).

(2) In the second step, the similarities with respect to the molecular fields between the resulting orientations of the test compound and the corresponding template molecules are determined. In order to compare the molecules only at those positions (grid points) for which the QSAR equation has been defined, the individual fields are multiplied with a weighting field derived from the coefficients (electrostatic or steric) of the respective analysis:

$$\text{weighting field}_{\text{ster or elec}} = \frac{\text{coefficient field}_{\text{ster or elec}}}{\text{coefficient field}_{\text{ster or elec}}} \qquad (3)$$

Division of the coefficient field by itself yields a new field (the weighting field) containing either '1.0' or 'MISSING' as entries. The term 'MISSING' implies that at the respective grid position no coefficient has been determined because the corresponding column was excluded due to insufficient variance before the PLS analysis. This corresponds to a weight of 0.0 at these positions. As the molecules are going to be compared with respect to steric and electrostatic properties, two weighting fields – a steric and an electrostatic one – are generated.

After multiplication of the steric and electrostatic fields of the compounds with this weighting field, the steric and electrostatic correlation coefficients (r_{ij}^{ster} and r_{ij}^{elec}) are determined. Subsequently, a total correlation coefficient r_{ij}^{tot}, describing the similarity between the two respective molecules under investigation, is calculated by the sum of these two correlation coefficients:

$$r_{ij}^{tot} = r_{ij}^{ster} * \text{contrib}_{ster} + r_{ij}^{elec} * \text{contrib}_{elec} \qquad (4)$$

As the steric and electrostatic descriptors have a different impact on the previously derived CoMFA model, the correlation coefficients are multiplied by *contrib*, which represents the steric and electrostatic contributions to the underlying analysis, normalized to 1.0.

In other words, in order to determine the similarity between the two molecules (the respective molecule i from the test set, i.e., 'testmol' in Scheme 2, and one of the reoriented molecules j of the training set) two 'filters' are applied: firstly the molecules are only compared at grid points where the respective QSAR equation is defined and secondly, the overall similarity comprises the sum of steric and electrostatic similarity, each contributing according to the relation of these field types in the underlying CoMFA.

(3) Finally, the six highest correlations r_{ij}^{tot} between molecule i of the test set and the molecules j of the training set (the template molecules) are selected. This number of six correlations was more or less arbitrarily chosen and corresponds to 7.5% of the total amount of correlation pairs (in fact, we wanted to select the highest 5–10% correlations). The activities of the six respective orientations of test molecule i (resulting from the fit described in step 1) were predicted and the 'uncorrected' mean activity act_i was calculated:

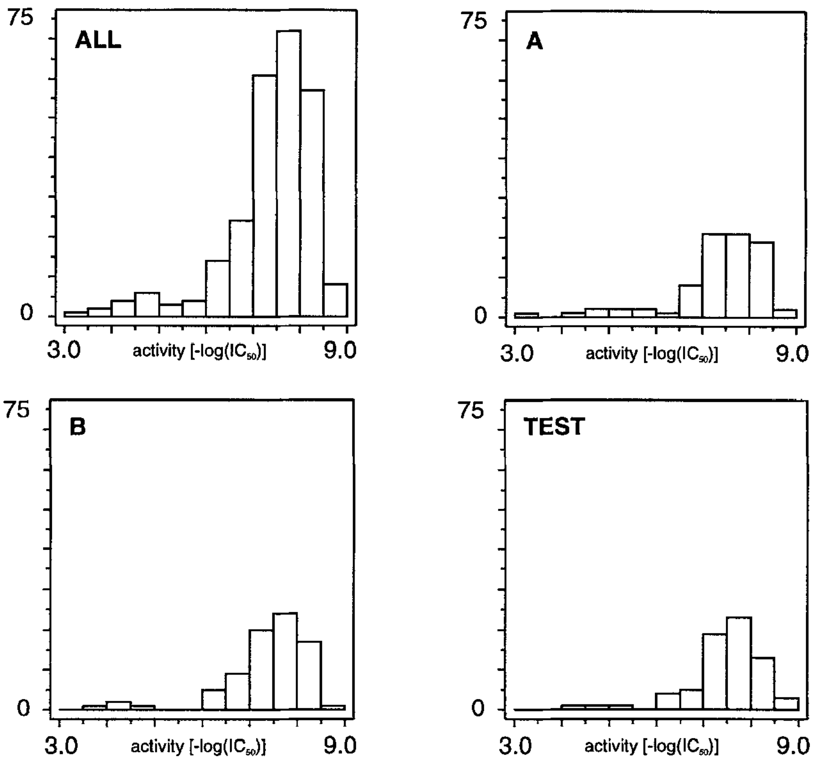


Fig. 3. Histograms for the distribution of the activities ($-\log(IC_{50})$) in the entire data set (ALL), the training sets (A and B) and the test set (TEST).

$$\mathrm{act}_i = \frac{1}{6}\sum_{j=1}^{6} \mathrm{act}_{i\ \mathrm{fitted\ to}\ j} \left(\frac{6}{\sum_{j=1}^{6} r_{ij}^{tot}} r_{ij}^{tot} \right) \qquad (5)$$

In order to consider the different degree of similarity between the test compound i and the six selected training set members (j), the expression in parentheses was introduced. This represents a weighting of the individual activities by the respective correlations, where r_{ij}^{tot} are the correlation coefficients for the similarity of the molecules as defined in Eq. 4.

Additionally, the inaccuracy of the underlying CoMFA may be introduced as well by consideration of the residuals of the six respective template molecules (residual$_j$, scaled by the correlation coefficients r_{ij}^{tot}); we consider this the 'corrected' act$_i$:

$$\mathrm{act}_i = \frac{1}{6}\sum_{j=1}^{6} \left(\mathrm{act}_{i\ \mathrm{fitted\ to}\ j} + \left(r_{ij}^{tot} * \mathrm{residual}_j\right)\right) \left(\frac{6}{\sum_{j=1}^{6} r_{ij}^{tot}} r_{ij}^{tot} \right) \qquad (6)$$

In summary, the combinations of two different fitting procedures ('point-fitted' and 'field-fitted') with two methods for calculating the mean activity (cf. Eqs. 5 and 6) of the test molecule lead to four different options.

TABLE 2
EFFECT OF DIFFERENT GRID ORIENTATIONS ON r^2_{cv}

Offset (Å)[a]			r^2_{cv}, model	
x	y	z	A	B
0.0	0.0	0.0	0.582/4	0.328/4
0.0	0.0	−1.0	0.563/3	0.394/5
0.0	−1.0	0.0	0.576/4	0.319/3
−1.0	0.0	0.0	0.569/4	0.315/3
−1.0	0.0	−1.0	0.639/4	0.388/4
−1.0	−1.0	0.0	0.544/3	0.250/3
0.0	−1.0	−1.0	0.617/4	0.345/3
−1.0	−1.0	−1.0	0.587/4	0.261/3

[a] The offset is defined with respect to the automatically defined standard region.

Results and Discussion

Out of a data set of 256 dihydrofolate reductase inhibitors, already published by Hansch et al., three groups were randomly selected: two training sets consisting of 80 compounds each (referred to as A and B), and a test set containing 70 molecules (TEST). The distributions of activities in these subsets (A, B and TEST) are very similar to that of the whole ensemble (ALL) and are shown in Fig. 3. Thus, biased results caused by differences in the activity distribution of the subsets can be ruled out.

In order to investigate the sensitivity of the analyses to the magnitude of the calculated interaction energies [1,25], the lattice of the automatically defined region was shifted

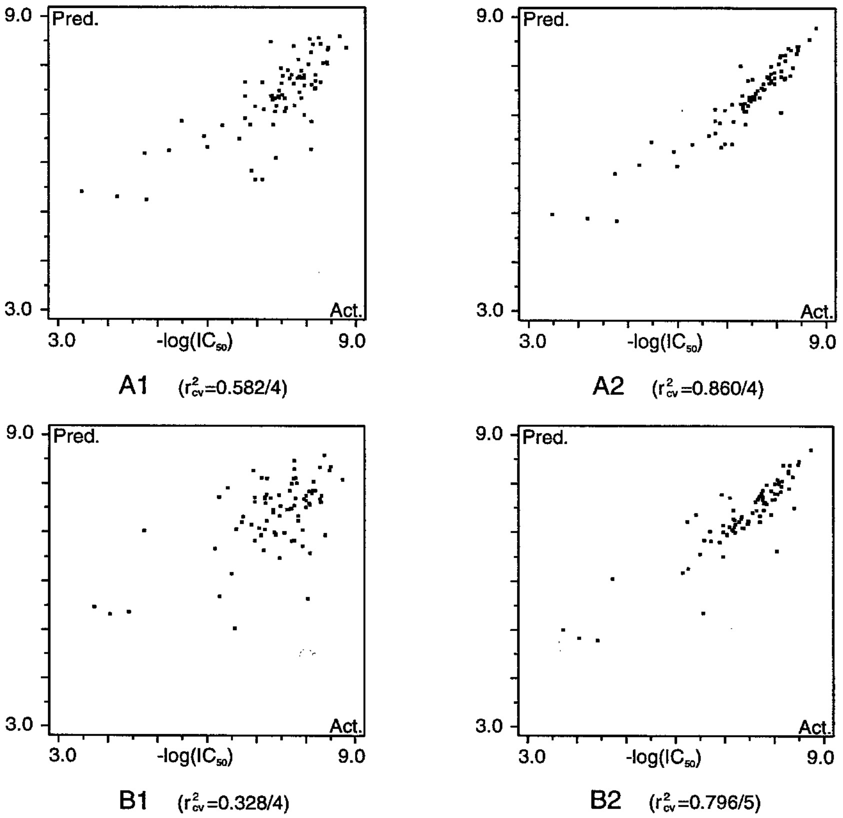
TABLE 3
STATISTICAL PARAMETERS OF THE CoMFA MODELS[a]

	A		B		A'		A"	
	r^2_{cv}	r^2_{pred}	r^2_{cv}	r^2_{pred}	r^2_{cv}	r^2_{pred}	r^2_{cv}	r^2_{pred}
Analysis								
initial	0.582/4	0.444	0.328/4	0.546	–0.099/1		–0.174/2	
final	0.860/4		0.796/5		0.624/5		0.596/5	
Prediction methods								
point-fitted, uncorrected		0.484		0.604		–0.447		–0.405
point-fitted, corrected		0.492		0.625		–0.456		–0.419
field-fitted, uncorrected		0.569		0.626				
field-fitted, corrected		0.598		0.645				

[a] The cross-validated r^2 values are obtained with the leave-one-out method.

by 1.0 Å along the x-, y- or z-axis of a Cartesian coordinate system, including all possible permutations. Especially for set B, this led to significantly different cross-validated r^2_{cv} values for the derived CoMFA; the changes in r^2_{cv} are rather high in comparison to the magnitude of the values (Table 2).

The statistical parameters of the models before and after the realignment procedure are listed in Table 3. The r^2_{cv} values for both CoMFAs (sets A and B) were dramatically improved by the realignment process involving two cycles (upper part of Table 3). The improvement of the models is also evident from the actual-versus-predicted plots shown in Fig. 4. Inspection of the residuals before and after the realignment procedure (Table 4) reveals that in many cases also the residuals of the overpredicted compounds (those compounds having a negative residual; these were not subjected to the reorientation process) have been reduced. This is related to the fact that by the reorientation of the underpredicted molecules, the variance at the individual lattice intersections is modified. As a consequence, the non-reoriented compounds become influenced as well.



A1 (r^2_{cv}=0.582/4)

A2 (r^2_{cv}=0.860/4)

B1 (r^2_{cv}=0.328/4)

B2 (r^2_{cv}=0.796/5)

Fig. 4. Actual versus predicted plots for the models before (A1 and B1) and after (A2 and B2) the realignment procedure.

TABLE 4
RESIDUALS BEFORE AND AFTER THE REALIGNMENT PROCEDURE

ID	Before	After	ID	Before	After	ID	Before	After	ID	Before	After
Subset A											
1	−1.98	−1.52	79	−1.23	0.05	127	0.43	0.04	201	−0.50	0.05
5	−1.14	−0.72	80	−0.12	−0.13	131	−0.30	−0.01	205	−0.21	0.18
11	−1.48	−1.08	83	−0.11	−0.14	134	−0.10	0.52	206	1.78	0.18
12	−0.50	−0.08	86	−0.03	−0.16	141	−0.14	0.53	208	1.22	0.01
15	−1.05	−0.77	87	0.00	−0.12	145	−0.68	0.69	209	0.46	0.06
16	−1.42	−0.99	90	0.51	−0.15	147	0.09	0.24	211	−0.33	0.29
19	−0.64	−0.34	93	0.00	−0.06	150	0.55	0.12	219	0.62	0.18
20	−0.35	0.03	94	0.27	0.34	162	0.04	0.09	220	0.47	−0.04
28	−0.50	−0.12	96	1.25	0.15	167	−0.31	0.25	221	0.34	1.00
40	0.11	−0.15	97	0.19	0.16	168	0.05	0.54	223	0.61	0.28
44	−0.19	−0.38	99	0.04	−0.08	169	0.38	0.29	226	−0.35	0.17
45	−0.65	−0.17	101	0.02	0.11	171	−0.03	−0.32	231	0.59	0.40
46	−0.93	−0.14	103	0.08	0.16	172	−0.02	0.20	232	−0.17	0.04
49	0.04	−0.75	104	−0.07	0.15	173	0.12	0.10	235	0.27	0.39
51	1.02	0.02	112	−0.48	0.05	174	0.53	0.20	240	0.31	0.44
54	−0.23	−0.12	113	−0.16	0.11	182	0.34	0.14	241	0.36	0.12
56	1.27	0.17	115	0.08	0.13	185	0.94	0.12	242	0.07	−0.01
62	−0.57	−0.24	117	−0.30	0.27	186	0.13	0.18	244	0.06	0.07
63	1.43	0.09	123	0.47	0.11	187	0.19	0.20	253	0.05	0.01
65	0.01	0.07	125	0.20	0.09	192	−0.01	0.00	256	0.42	−0.01
Subset B											
2	−1.78	−1.31	61	−1.05	0.02	132	0.15	−0.01	190	0.47	0.07
4	−1.31	−0.82	64	0.48	0.04	133	−0.22	0.01	191	0.22	0.05
7	−0.98	−0.39	67	0.08	0.02	136	0.17	0.05	196	2.37	0.15
10	−2.33	−1.36	69	−0.55	0.00	139	0.19	0.23	198	0.44	0.22
22	−0.55	−0.06	70	−0.58	0.25	140	−0.32	−0.09	200	0.20	0.04
25	−1.51	−1.00	71	−0.45	−0.05	142	0.84	0.17	203	1.47	0.54
26	0.53	−0.97	72	−0.61	−0.08	146	−0.40	0.26	204	0.49	0.17
30	−1.53	−0.30	73	−0.05	−0.13	151	−0.74	0.26	210	0.29	0.29
33	0.30	−0.35	75	−0.93	0.00	152	0.19	−0.04	213	0.26	1.38
36	1.50	−0.14	82	−0.45	0.29	153	−0.57	−0.10	217	0.08	−0.06
37	−0.50	−0.88	88	−0.10	0.24	156	0.89	1.18	222	0.30	0.04
42	−0.53	−0.21	91	−0.15	0.20	157	0.06	0.42	229	0.57	0.01
43	−0.62	−0.07	92	0.33	−0.05	158	−0.36	0.06	230	−0.08	0.16
50	−0.29	0.06	95	0.40	0.10	164	−0.23	0.23	233	0.65	0.07
52	−1.37	−0.64	102	−0.34	0.26	165	0.58	0.27	234	0.53	0.03
53	−0.78	−0.13	105	0.10	−0.05	166	0.48	0.20	237	−0.24	0.37
55	−0.69	−0.32	106	0.95	−0.07	178	0.86	0.13	238	1.41	−0.10
58	0.09	0.05	110	−0.08	−0.16	179	1.17	0.17	247	0.18	0.86
59	−0.07	−0.08	122	0.55	−0.02	183	0.25	0.33	248	0.12	0.06
60	0.12	0.18	130	0.12	−0.23	188	0.32	0.46	254	0.62	0.01

Although cross-validation is a very stringent test for the predictive power of a given CoMFA, the leave-one-out method might lead to r^2_{cv} values that overestimate the internal consistency of the respective model. Therefore, analyses with two cross-validation groups were performed as well. These groups consisted of 50% of the compounds and were randomly selected, the first serving as the training set and the second as a test set. As the random selection might have an impact on the results, this kind of analysis was repeated 100 times for the initial (A1 and B1) and the final (A2 and B2) models, with identical sets of cross-validation groups. The resulting mean r^2_{cv} values (Table 5) were slightly lower compared to the values obtained by the leave-one-out method. In all cases, a few analyses with a poor r^2_{cv} could be obtained, indicating a certain degree of inconsistency in the underlying data set. However, the enhanced quality of the models is also apparent from these values. Although lower than the r^2_{cv} values obtained with the leave-one-out method, there is a distinct improvement of the internal consistency of the models with regard to the mean values as well as to the lowest r^2_{cv}.

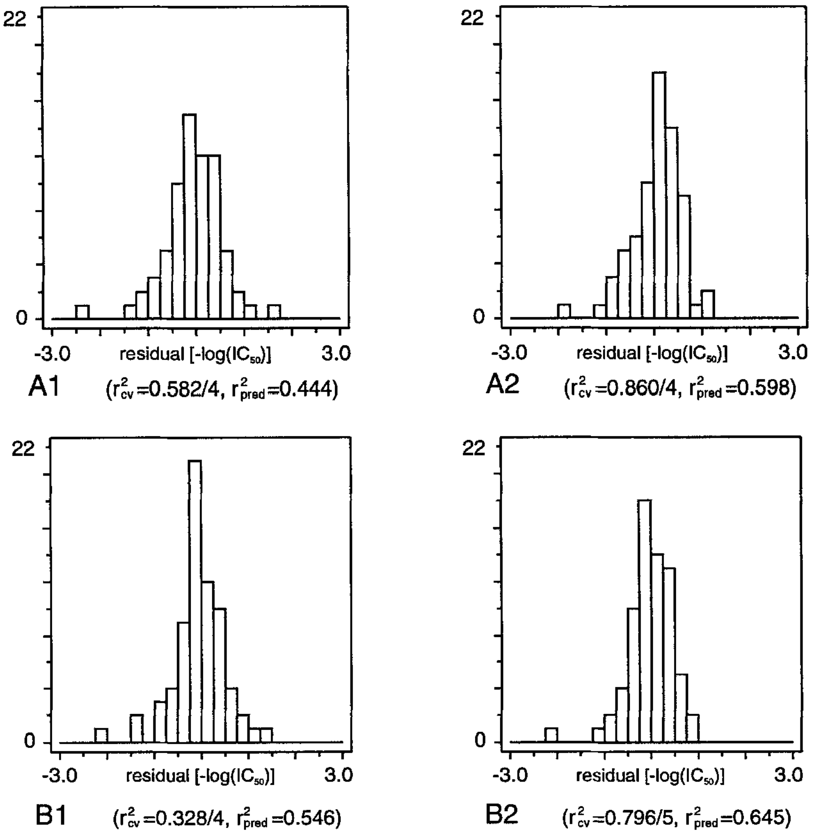


Fig. 5. Histograms for the distribution of the residuals in the predicted activities of the test set before (A1 and B1) and after (A2 and B2) the refinement of the underlying CoMFA models.

The r^2_{pred} values resulting from application of the four different options in the prediction process are listed in the lower half of Table 3. The rigid-body field fit of the structures in the test set to the template molecules gave better results than the point-fitting method. In both cases, a correction of the six respective predicted activities of the molecules in the test set by the residuals of the template molecules before calculating the mean ('corrected' act_i, Eq. 6) yielded higher r^2_{pred} values. The histograms in Fig. 5 reveal that the general accuracy of prediction increases for both models and is, therefore, not only related to the accommodation of certain 'outliers'. However, the increase in r^2_{pred} is relatively small compared to the improvement of r^2_{cv} in both models. It appears of interest that the r^2_{cv} and r^2_{pred} values do not necessarily correlate. For set A, r^2_{cv} is higher and r^2_{pred} is lower than for set B.

TABLE 5
SUMMARY OF THE CoMFAs WITH NUMBER OF CROSS-VALIDATION GROUPS = 2[a]

Parameter	A1	A2	B1	B2
Mean[b]	0.473	0.753	0.251	0.621
Standard deviation	0.143	0.103	0.148	0.141
Highest r^2_{cv}	0.643	0.870	0.444	0.775
Lowest r^2_{cv}	−0.007	0.394	−0.354	0.068

[a] The highest r^2_{cv} within the first five components extracted is always considered.
[b] Mean of 100 runs with randomly selected cross-validation groups.

The overall characteristics of the models before and after the reorientation of the underlying molecules have changed only moderately. This can be shown by a pointwise comparison of the coefficients of the initial (before the refinement) and the final QSAR. For training set A, the correlation of the electrostatic coefficients is 0.89 (172 points included) and for the steric ones it is 0.92 (at 334 points). The corresponding values for B are 0.90 (for 150 points) and 0.93 (for 282 points), respectively. Further indication for these moderate changes is given by an inspection of the steric and electrostatic contributions to the initial and final QSAR: in the case of model A, the steric contributions have been reduced from 59.7% to 58.2%; with respect to model B this change is from 56.0% to 54.4%.

Finally, the possibility for a chance correlation was investigated [26]. This was done twice by randomizing the activities of set A (i.e., randomly interchanging the existing activity values of the molecules, yielding models A' and A"). Subsequently, the CoMFAs were derived and the reorientation procedure was applied with the same increments and number of cycles as for sets A and B. In both cases, the initially negative r^2_{cv} improved to values higher than 0.5 (Table 3). Nevertheless, the r^2_{pred} remained negative.

Conclusions

A new procedure for improving the predictive ability of CoMFA models has been developed and successfully applied to a set of dihydrofolate reductase inhibitors. It is well recognized that the relative orientation of the molecules (the alignment) is the crucial point for a CoMFA. A major reason appears to be the steep potential wall for the steric interaction energies due to the r^{12}-term in the Lennard-Jones potential. In combination with a rather wide spacing of the grid points,* this might lead to the assignment of random numbers between 0 and 30.0 kcal/mol (the standard cutoff) at a 'repulsive' grid point: a difference of 0.1 Å in the distance between a lattice point and a compound may influence the corresponding interaction energy significantly. Evidence for this phenomenon is indicated by two facts: (i) the experiments in which the whole lattice is shifted by 1.0 Å along different directions, resulting in significantly different r^2_{cv} values and (ii) the ability to improve the r^2_{cv} for a given model by methodical and slight reorientations of some of the underlying molecules.

The major purpose of a CoMFA is to derive a linear equation by which the activities of hypothetical molecules can be predicted. Therefore, the main task is not to improve the r^2_{cv} but the r^2_{pred} value. For the given data set, this was accomplished by a combination of field fitting, determination of correlations and correction of the calculated target properties by the residuals of the respective template molecules. Nevertheless, further studies will have to be undertaken in order to find out whether this method works also for more heterogeneous data sets. For very diverse data sets, the fitting procedure might lead to unexpected orientations of the test molecules, consequently leading to a high variation in the predicted activities. Furthermore, additional investigations should aim at optimizing the prediction procedure with respect to r^2_{pred} as well as reducing the computational effort. For example, instead of fitting each test molecule to all the compounds in the training set, a 'supertemplate' could be defined to which each structure has to be fitted only once.

From the fact that the realignment procedure also for the data sets with randomized activities (A' and A") led to rather good r^2_{cv} values, the following question may arise: How can one, solely on the grounds of r^2_{cv}, determine whether a CoMFA result is predictive for compounds not included in the training set? We are aware that this refinement procedure is able to create a previously nonexistent consistency for a model derived from a given data set. This accounts especially for data sets in which the lack of consistency is not related to an alignment problem but to other factors, such as the source of the biological data. Therefore, we suggest to apply such a realignment procedure only to analyses with initially reasonable r^2_{cv} values. A further test may be the exclusion of a subset of compounds in order to create a test set and consequent monitoring of the statistical parameters (r^2_{cv} and r^2_{pred}); this will enable us to distinguish between a real improvement of the predictiveness or the creation of a 'pseudo-consistency' for a given model, as done in this investigation.

*A narrower spacing of the lattice intersections does not represent a solution to this problem, as the 'noise', to which PLS is particularly sensitive, increases substantially [26].

References

1 Cramer III, R.D., Patterson, D.E. and Bunce, J.E., J. Am. Chem. Soc., 110 (1988) 5959.
2 Avery, M.A., Gao, F. and Chong, W.K.M., J. Med. Chem., 36 (1993) 4264.
3 Horwitz, J.P., Massova, I., Wiese, T.E., Besler, B.H. and Corbett, T.H., J. Med. Chem., 37 (1994) 781.
4 Waller, C.L., Oprea, T.I., Giolitti, A. and Marshall, G.R., J. Med. Chem., 36 (1993) 4152.
5 Waller, C.L. and Marshall, G.R., J. Med. Chem., 36 (1993) 2390.
6 DePriest, S.A., Mayer, D., Naylor, C.B. and Marshall, G.R., J. Am. Chem. Soc., 115 (1993) 5372.
7 Debnath, A.K., Hansch, C., Kim, K.H. and Martin, Y.C., J. Med. Chem., 36 (1993) 1007.
8 Wold, S., Ruhe, A., Wold, H. and Dunn III, W.J., SIAM J. Sci. Stat. Comput., 5 (1984) 735.
9 Wold, S., Albano, C., Dunn III, W.J., Edlund, U., Esbenson, K., Geladi, P., Hellberg, S., Johannson, E., Lindberg, W. and Sjörström, M., In Kowalski, B. (Ed.) Chemometrics: Mathematics and Statistics in Chemistry, Reidel, Dordrecht, 1984, pp. 17–95.
10 Stahle, L. and Wold, S., Prog. Med. Chem., 25 (1988) 292.
11 Cramer III, R.D., Patterson, D.E. and Bunce, J.D., Quant. Struct.–Act. Relatsh., 7 (1988) 18.
12 Thibaut, U., Folkers, G., Klebe, G., Kubinyi, H., Merz, A. and Rognan, D., Quant. Struct.–Act. Relatsh., 13 (1994) 1.
13 Nicklaus, M.C., George, W.A. and Terrence Jr., R.B., J. Comput.-Aided Mol. Design, 6 (1992) 487.
14 Silipo, C. and Hansch, C., J. Am. Chem. Soc., 97 (1975) 6849.
15 SYBYL Molecular Modelling Package, v. 6.04, TRIPOS Associates, Inc., St. Louis, MO, 1993.
16 Vinter, J.G., Davies, A. and Saunder, M.R., J. Comput.-Aided Mol. Design, 1 (1987) 31.
17 Powell, M.J.D., Math. Program., 12 (1977) 241.
18 Stewart, J.J.P. and Seiler, F.J., MOPAC (v. 5.00), QCPE Program No. 455, Quantum Chemistry Program Exchange, University of Indiana, Bloomington, IN, 1989.
19 Dewar, M.J.S. and Thiel, W., J. Am. Chem. Soc., 99 (1977) 4899.
20 SYBYL Molecular Modelling Software version 6.0 Command Manual, Tripos Associates, Inc., St. Louis, MO, 1992, p. 1259.
21 SYBYL Molecular Modelling Software version 6.0 Theory Manual, Tripos Associates, Inc., St. Louis, MO, 1992, p. 2225.
22 Nelder, J.A. and Mead, R., Comput. J., 7 (1965) 308.
23 Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T., Numerical Recipes, Cambridge University Press, Cambridge, 1987, pp. 289–293.
24 Johnson, M., Lajiness, M. and Maggiora, G., Prog. Clin. Biol. Res., 291 (1989) 167.
25 Kim, K.H., Greco, G., Novellino, E., Silipo, C. and Vittoria, A., J. Comput.-Aided Mol. Design, 7 (1993) 263.
26 Clark, M. and Cramer III, R.D., Quant. Struct.–Act. Relatsh., 12 (1993) 137.