Eur J Med Chem (1996) 31, 797–803

Using multivariate adaptive regression splines to QSAR studies of dihydroartemisinin derivatives

V Nguyen-Cong[1], G Van Dang[2], BM Rode[1]*

[1]*Institute for General, Inorganic and Theoretical Chemistry, Theoretical Chemistry Division, University of Innsbruck, 52a Innrain, A-6020 Innsbruck, Austria;*
[2]*Faculty of Pharmacy, School of Medicine and Pharmacy, 41-43 Dinh Tien Hoang Street, District 1, Hochiminh City, Vietnam*

(Received 12 February 1996; accepted 1 April 1996)

Summary — QSAR models for analogs of antiplasmodial artemisinin compounds were established, based on atomic net charges by using multivariate adaptive regression splines (MARS) in comparison with some other methods such as multiple linear regression, alternating conditional expectations and projection pursuit regression. The established models were then evaluated by an Anova decomposition procedure so that the effects of each predictor (additive or interaction) could be viewed graphically, facilitating the interpretation of the underlying relationship. It was found that the QSARs derived from the MARS method are the most satisfactory predictive models, and that the artemisinin pharmacophore identification is in agreement with previous experimental findings.

artemisinin / multivariate adaptive regression spline / Anova decomposition / alternating conditional expectations / projection pursuit regression

Introduction

In recent years, artemisinin (quinghaosu, arteannuin), extracted from the plant *Artemisia annua*, and its derivatives (fig 1) have attracted worldwide attention due to their particular efficiency against chloroquine-, mefloquine- and multidrug-resistant strains of *Plasmodium falciparum*, one of four species of parasitic protozoa of the genus *Plasmodium* causing malaria, a serious endemic disease in many developing countries. Many quantitative structure–activity relationship (QSAR) studies have been conducted to explain the drug's mechanism of action and give guidelines for synthesizing new derivatives with improved efficiency and stability. Avery et al [1] have built a CoMFA model for C-9 analogs of artemisinin and 10-deoxoartemisinin and, recently, Suter et al [2] have also correlated the three-dimensional molecular electrostatic potentials, calculated in quantum mechanics and projected on two-dimensional surfaces, with the biological activity of some artemisinin derivatives by using neural networks. For compounds within a congeneric series, the use of atomic net charges as stereoelectronic structural descriptors is a simple but efficient approach in QSAR studies.

This work describes an experiment with the multivariate adaptive regression splines (MARS) method [3] along with the traditional approach, multiple linear regression (MLR) and two other nonparametric nonlinear methods, alternating conditional expectations (ACE) [4] and projection pursuit regression (PPR) [5, 6], on a series of diastereomeric dihydroartemisinin α-alkylbenzylic ethers. Descriptors used for building predictive models are atomic net charges evaluated on the basis of PM3 semiempirical molecular orbital calculations.

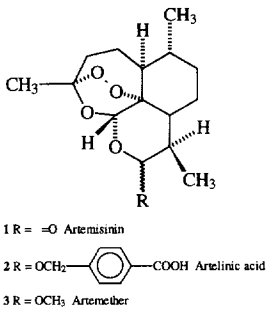


Fig 1. Structures of artemisinin, artelinic acid and artemether.

*Correspondence and reprints

In a recent QSAR study of pyridinium cephalosporins [7], it was shown that once an additive function was not sufficient for approximating the underlying relationship, the PPR approach could produce good predictive models due to its ability to model interactions between predictor variables. However, even for some simple functions, PPR might need a large number of terms for good approximation and does not flag interaction effects explicitly, resulting in ambiguous models. The MARS method allows for interactions more explicitly by separating additive contributions from interaction effects.

Application of MARS to chemical studies was introduced by De Veaux et al [8]. They compared both the accuracy and speed of MARS to those of artificial feedforward neural networks with sigmoid activation functions. In most cases, MARS was seen to be more accurate and much faster than neural networks. Rogers and Hopfinger [9] suggested a variant of MARS, replacing the statistical variable subset selection procedure in MARS by a genetic algorithm, and applied it in some QSAR/QSPR problems.

Materials and methods

Biological data

A series of 14 diastereomeric dihydroartemisinin α-alkylbenzylic ethers (table I) synthesized and tested by Lin and Miller [10] along with artemisinin, artelinic acid and artemether were used in this work. These compounds have been tested in vitro against two clones of human malaria, *P falciparum* D-6 (Sierra Leone clone, mefloquine-resistant) and W-2 (Indochina clone, chloroquine-resistant), and the average values of at least three experiments for each compound have been reported.

Molecular modeling

All the compounds were built by using the HyperChem molecular modeling software [11]. The template molecular model for structure building was the crystallographic X-ray structure of artemisinin [12] (fig 2) obtained from the Cambridge Structural Database [13]. All structures were then optimized using the semiempirical molecular orbital PM3 method implemented in the Gaussian92 software [14], and Mulliken atomic net charges (table II) as specified in figure 2 were used as predictors in statistical analyses.

Multivariate adaptive regression splines (MARS)

MARS [3] is a generalization of adaptive regression spline methods. It builds up a set of tensor product spline basis functions and fits the coefficients of these basis functions to the data by least squares. MARS models the true underlying function $f(\boldsymbol{x})$ by

$$\hat{f}(\boldsymbol{x}) = a_0 + \sum_{m=1}^{M} a_m \prod_{k=1}^{K_m} B_{km}(x_{v(k,m)})$$

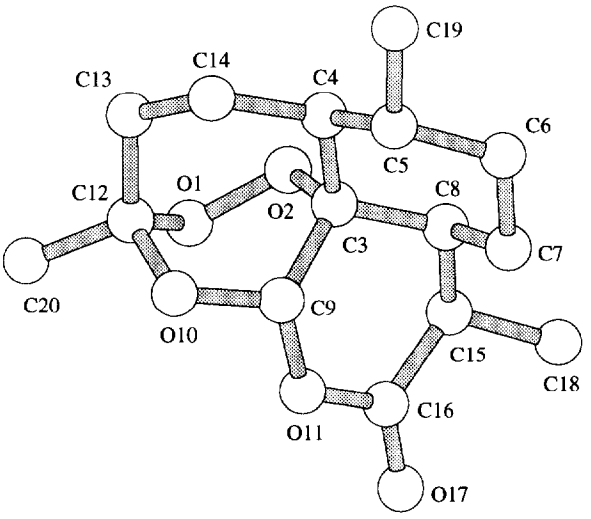


Fig 2. Crystallographic X-ray structure of artemisinin.

where $x_1, x_2, \ldots, x_p$ are predictor variables and $v(k, m)$ labels the predictor in the kth term of the mth product. K_m is a parameter that limits the order of interactions. For $K_m = 1$, the resulting model will be an additive one, pairwise interactions are allowed for $K_m = 2$, and the order of interactions is arbitrary when K_m is equal to the number of compounds (n). The basis functions B_{km} are first-order truncated power splines defined by

$$B_{km}(x) = \pm(x - t_{km})_+$$

where t_{km} is an observed value of the predictor x and

$$(x-t)_+ = \begin{cases} 0 & x \le t, \\ x-t & x > t. \end{cases}$$

The MARS algorithm can be summarized as follows.

1. *Initialize.* Start with the constant basis function in the model: $B_0(\boldsymbol{x}) = 1$. After the Mth iteration, there are $2M + 1$ functions in the model

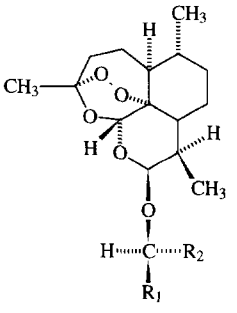
$$\{B_m(\boldsymbol{x})\}_{m=0}^{2M}.$$

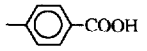
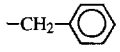
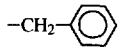
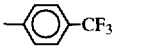
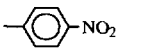
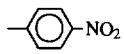
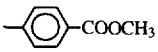
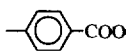
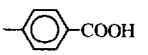
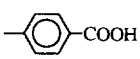
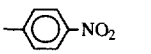
2. *Forward stepwise.* At each $(M + 1)$th iteration, two new basis functions that have strongest effect in decreasing the residual sum of squares are added to the current model at the same time

$$B_{2M+1}(\boldsymbol{x}) = B_{l(M+1)}(\boldsymbol{x})[-(x_{v(M+1)} - t_{M+1})]_+$$
$$B_{2M+2}(\boldsymbol{x}) = B_{l(M+1)}(\boldsymbol{x})[-(x_{v(M+1)} - t_{M+1})]_+$$

where $B_{l(M+1)}$ is one of the $2M + 1$ basis functions already chosen, $0 \le l(M + 1) \le 2M$, $v(M + 1)$ is one of the predictors not present in $B_{l(M+1)}$, and t_{M+1} is an observed value of that predictor. The interaction level of $B_{2M+1}(\boldsymbol{x})$ and $B_{2M+2}(\boldsymbol{x})$ should be $\le K_m$.

Table I. In vitro antimalarial activity against *P falciparum*.



No.		R_1	R_2	IC_{50}(ng/mL) W-2	D-6
1		Artemisinin		1.1050	2.3020
2		-H		2.0818	4.8659
3		-H	-H	0.3008	0.8689
4	(R)	$-CH_2CH_2CH_3$		1.9158	3.0490
5	(S)		$-CH_2CH_2CH_3$	0.6968	0.9722
6	(R)		$-COOCH_2CH_3$	0.0938	0.2872
7	(S)	$-COOCH_2CH_3$		0.2265	0.5093
8	(R)	$-CH_3$		1.1475	1.4580
9	(R)	$-CH_2COOCH_2CH_3$		0.2134	0.4957
10	(S)		$-CH_2COOCH_2CH_3$	0.1437	0.4593
11	(R)	$-CH_2COOCH_2CH_3$		0.2297	0.5629
12	(S)		$-CH_2COOCH_2CH_3$	0.0487	0.2463
13	(R)	$-CH_3$		0.3368	0.9210
14	(S)		$-CH_3$	0.3353	0.6921
15	(R)	$-CH_3$		2.5350	5.7720
16	(S)		$-CH_3$	1.2308	2.9360
17	(R)	$-CH_2COOH$		1.3470	2.8570

3. *Loop.* Repeat (2) until the maximum number of basis functions (M_{max}) has been reached.

4. *Backward stepwise.* The least important basis functions are eliminated one at a time. The cross-validation or generalized cross-validation criterion is used to select the best predictive model.

Once the final model is found, a procedure called ANOVA decomposition is applied to facilitate interpreting the underlying relationship. All the basis functions that involve only one predictor variable are grouped to represent the main effects, all the basis functions involving two predictors are grouped to represent second order interaction surfaces and so on

$$\hat{f}(\boldsymbol{x}) = a_0 + \sum_{K_m=1} f_i(x_i) + \sum_{K_m=2} f_{ij}(x_i, x_j) + \cdots$$

The representation of the MARS model by ANOVA decomposition gives explicitly the effects of each predictor (additive or interactive) in the final model and these effects can then be displayed graphically. References [3, 15] give more details about MARS.

The programs performing the ACE and MARS procedures are freely available from the StatLib archive [16] at Carnegie Mellon University and the program performing the PPR method called SMART is available from Friedman JH at Department of Statistics, Stanford University. A few minor modifications of these programs were made to fit some additional requirements.

Results and discussion

QSAR studies of dihydroartemisinin derivatives against P falciparum D-6

For a set of p given predictors, there are 2^p subsets to be estimated for finding out the best model, thus leading to a very time-consuming procedure. The 'leaps and bounds' algorithm [17] allows the best subset of a given size, based on maximal criterion of the Mallows Cp statistic, R^2 or adjusted R^2 value, to be found by evaluating only a fraction of the 2^p regressions. The 'leaps' procedure implemented in the S-PLUS software [18] with the R^2 criterion was used in this work. The leave-one-out cross-validation procedure was then applied to models with a high value of R^2 to estimate the best subset of predictors. It was observed that the MLR method needed a large number of predictors to give a good fit. The best fit out of linear models involving up to eight predictors gives a R^2 value of 0.721. Table III summarizes the results of analyses on the antimalarial activity against *P falciparum* D-6 using MLR and the three nonlinear methods, ACE, PPR and MARS. For nonlinear methods, all combinations involving up to five predictors were investigated to find satisfying predictive models. The MARS procedure was applied with $mi = 1$ and 2 ($K_m \leq 2$), where mi is the control parameter that allows the maximum number of variables to participate in inter-

Table II. Atomic net charges used in this study (see fig 2).

	O1	O2	C3	C4	C5	C6	C7	C8	C9	O10	O11	C12	C13	C14	C15	C16	O17	C18	C19	C20
1	-.132085	-.130579	.023690	-.168960	-.148088	-.235707	-.250707	-.157371	.138005	-.266715	-.246942	.214219	-.288234	-.247680	-.191522	.388319	-.347231	-.311324	-.314542	-.331566
2	-.136988	-.128774	.024225	-.168022	-.147381	-.234773	-.256999	-.141019	.148872	-.269064	-.266256	.217292	-.288287	-.247385	-.206045	.153586	-.292457	-.307273	-.313847	-.330668
3	-.135756	-.127076	.027282	-.169690	-.146174	-.236340	-.259299	-.143881	.151448	-.269596	-.286678	.217569	-.288775	-.247684	-.188845	.156638	-.292655	-.305693	-.314468	-.331284
4	-.136431	-.129385	.025579	-.168866	-.147230	-.234406	-.262841	-.134056	.146036	-.267812	-.257115	.216730	-.288325	-.247595	-.217280	.161194	-.295462	-.324812	-.314299	-.330772
5	-.136859	-.129337	.025063	-.168851	-.146910	-.233342	-.272248	-.133744	.146457	-.267480	-.257437	.216795	-.288297	-.247643	-.219786	.165093	-.291886	-.323721	-.314208	-.330841
6	-.136508	-.129430	.024699	-.167668	-.148670	-.234063	-.259250	-.142182	.143402	-.267569	-.259011	.216692	-.288103	-.248285	-.210734	.152575	-.275629	-.311242	-.313667	-.330985
7	-.134755	-.129189	.024493	-.168164	-.147712	-.235155	-.259278	-.139545	.144612	-.266572	-.253203	.216104	-.288470	-.247827	-.212450	.149059	-.271912	-.313360	-.314182	-.331002
8	-.136314	-.128788	.024750	-.168019	-.147655	-.235818	-.256174	-.141042	.146313	-.269117	-.259924	.216804	-.288361	-.247764	-.205461	.151359	-.283633	-.308995	-.314279	-.331015
9	-.136665	-.128720	.024514	-.167819	-.147944	-.235123	-.258018	-.140908	.146691	-.269143	-.261503	.216892	-.288270	-.247790	-.203961	.150504	-.278905	-.313597	-.313963	-.330900
10	-.135309	-.129827	.025452	-.168967	-.147571	-.238198	-.254089	-.135334	.146737	-.268058	-.270854	.216989	-.288402	-.247342	-.202513	.141040	-.300382	-.306789	-.314055	-.330867
11	-.135644	-.128207	.025435	-.168642	-.146641	-.236071	-.255703	-.143559	.148924	-.269659	-.269330	.216782	-.288468	-.247883	-.204437	.154274	-.286182	-.309865	-.314572	-.331205
12	-.134411	-.129616	.025170	-.168850	-.146990	-.238624	-.250594	-.140763	.147682	-.267886	-.267113	.216311	-.288623	-.247809	-.203235	.155662	-.289528	-.307484	-.314458	-.331305
13	-.136853	-.129059	.024619	-.167942	-.147708	-.235464	-.258165	-.140034	.146435	-.268770	-.259158	.216911	-.288224	-.247868	-.210520	.157634	-.283147	-.310032	-.314111	-.330846
14	-.135154	-.129962	.025247	-.168710	-.147514	-.236890	-.256169	-.137101	.145269	-.268032	-.256373	.216673	-.288436	-.247405	-.216701	.152484	-.285760	-.307361	-.314066	-.331165
15	-.136843	-.128598	.024715	-.168214	-.147176	-.235111	-.258939	-.141375	.148045	-.269047	-.264563	.217052	-.288289	-.247698	-.208251	.158122	-.286579	-.309989	-.314182	-.330899
16	-.135503	-.129753	.025203	-.168705	-.147504	-.236962	-.256138	-.137109	.145322	-.268209	-.256986	.216680	-.288411	-.247541	-.216628	.152844	-.285364	-.307504	-.314101	-.331047
17	-.135423	-.128640	.025236	-.168254	-.147381	-.236115	-.254405	-.143004	.146610	-.268933	-.261188	.216399	-.288419	-.248082	-.204457	.153063	-.280952	-.312203	-.314467	-.331182

action effects. The first value (mi = 1) corresponds to additive modeling, whereas the second allows interactions between at most two predictors.

Table III shows that except the linear model, all three nonlinear methods gave good fits but their predictive abilities expressed by the cross-validated R^2, noted as Q^2 [19], were different. The MARS model involving two-predictor interactions ($K_m \leq 2$) gave the best predictive ability (Q^2 = 0.896), whereas the PPR model with five predictors was slightly better than the additive ACE model. Table IV provides the ANOVA decomposition of the best MARS model. The first column labels the ANOVA function number, the second lists the standard deviation of the function, indicating its relative importance for the overall model. The third gives the number of basis functions forming

Table III. QSAR models for the activity against *P falciparum* D-6.

Method	*Predictors*	R^2	Q^2
MLR	O2, C3, C9, C11, C14, O17, C18, O20	0.721	0.327
ACE	C4, C6, O10, C14	0.957	0.419
PPR	C6, C9, O10, C13, C14	0.909	0.463
MARS (mi = 2)	O1, C4, C7, O11	0.949	0.896

Table IV. MARS ANOVA decomposition for the activity against *P falciparum* D-6.

Function	*Standard deviation*	*# Basis functions*	*Predictors*	
1	4.099	1	C7	
2	6.878	1	O11	
3	14.29	2	O1	
4	5.953	1	O1	C4
5	4.169	1	C4	C7
6	10.52	1	C7	O11
7	6.016	1	O1	C7

Table V. Relative predictor importance to the D-6 MARS model.

Predictor importance			
O1	*C4*	*C7*	*O11*
70.14	77.02	85.74	100.0

Table VI. QSAR models for the activity against *P falcipa-rum* W-2.

Method	*Predictors*	R^2	Q^2
MLR	O1, O2, C3, C5 C12, C14, O16, C18	0.871	0.633
ACE	O11, C12, C14, C15	0.955	0.606
PPR	C5, C6, C8, C9 O10	0.829	0.604
MARS (*mi* = 2)	O2, C4, C7, O11	0.955	0.872

Table VII. MARS ANOVA decomposition for the activity against *P falciparum* W-2.

Function	*Standard deviation*	*# Basis functions*	*Predictors*	
1	14.85	2	O11	
2	2.466	1	O2	
3	12.29	2	C7	
4	2.576	1	O2	C4
5	0.3292	1	C4	O11
6	17.49	1	C7	O11
7	4.834	1	O2	O11

Table VIII. Relative predictor importance to the W-2 MARS model.

Predictor importance			
O2	*C4*	*C7*	*O11*
64.48	89.69	75.04	100.0

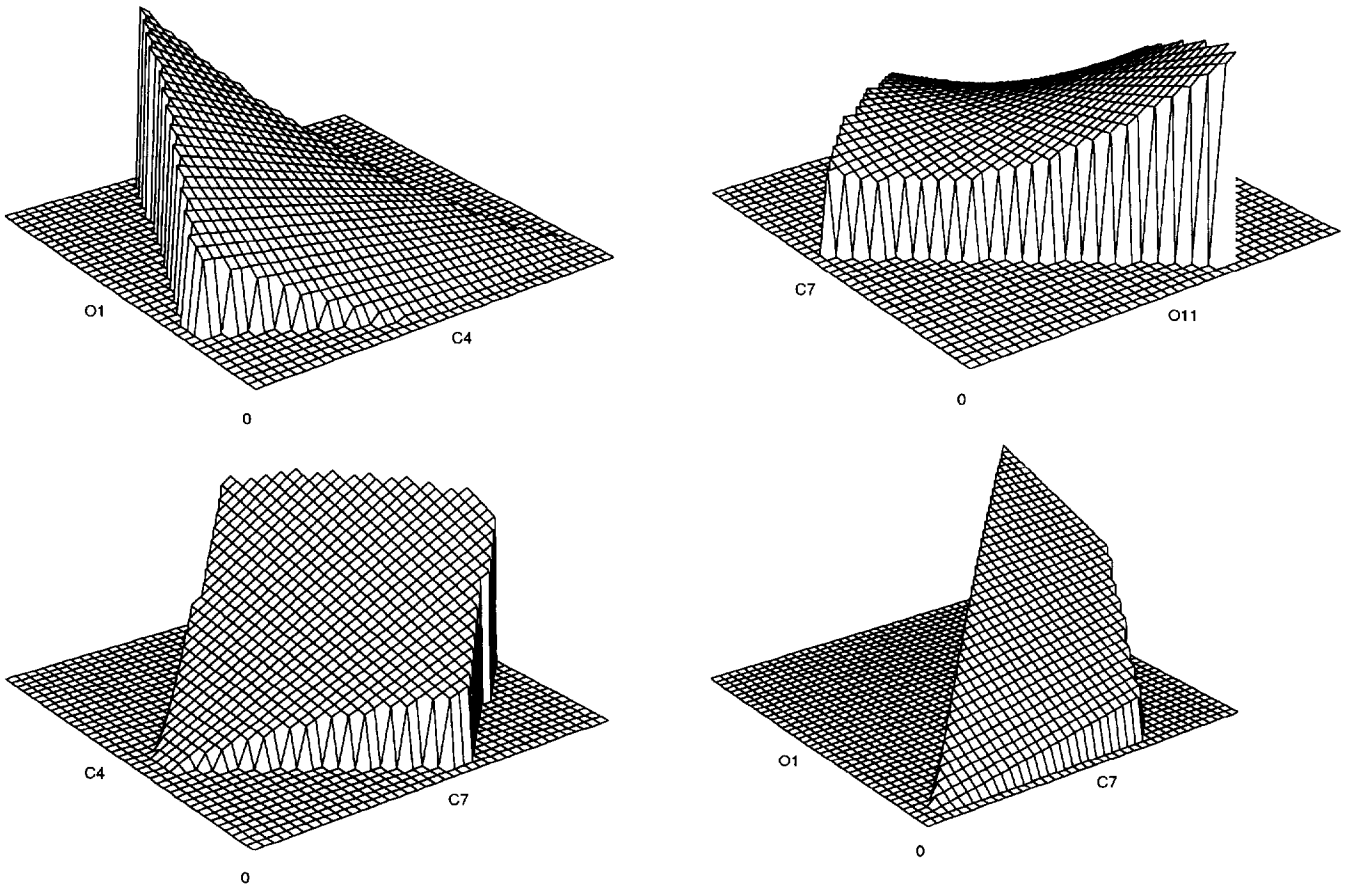


Fig 3. MARS ANOVA functions for the activity against *P falciparum* D-6. (**a**) Activity D-6 vs net charges O1 and C4. (**b**) Activity D-6 vs net charges C4 and C7. (**c**) Activity D-6 vs net charges C7 and O11. (**d**) Activity D-6 vs net charges O1 and O11.

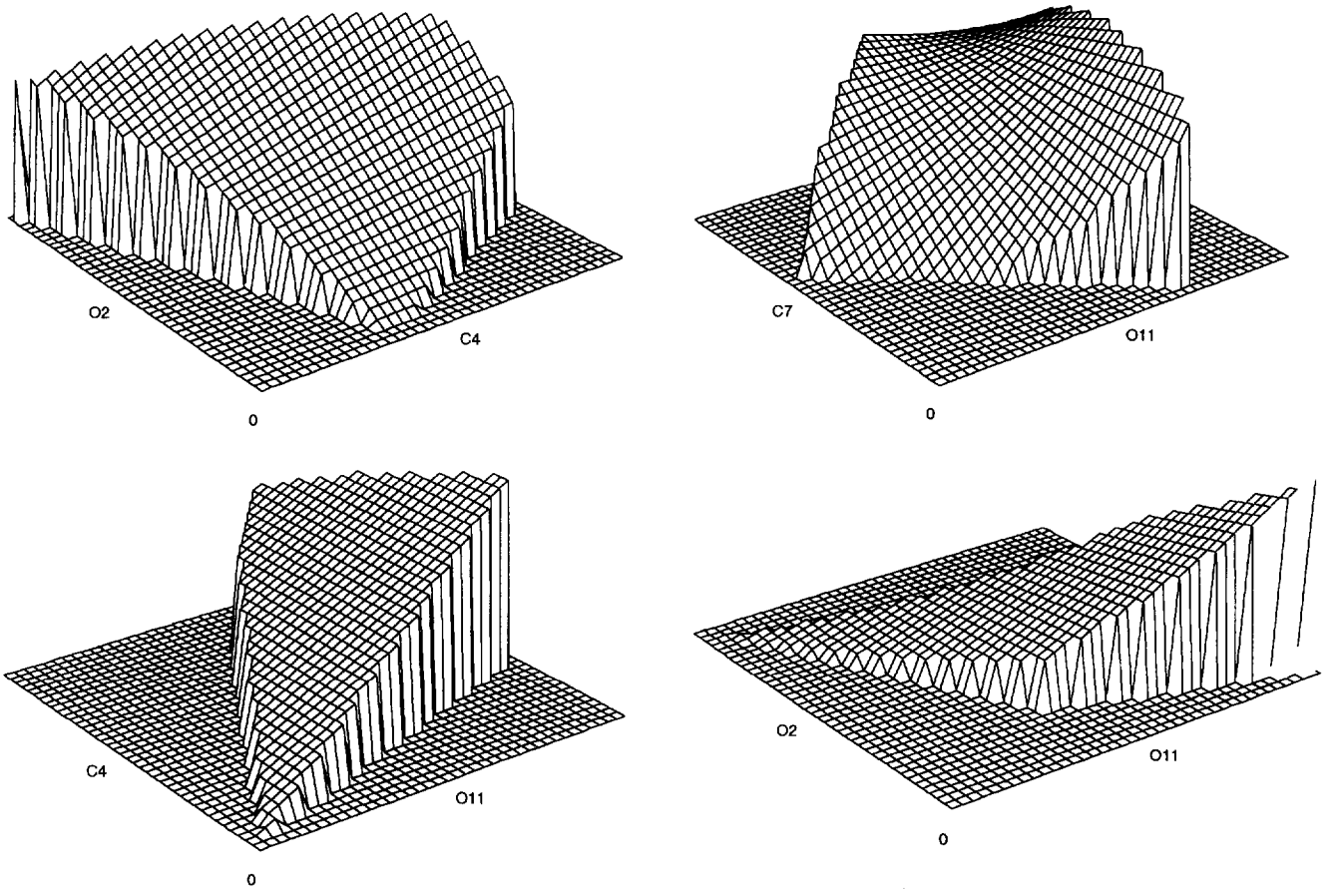


Fig 4. MARS ANOVA functions for the activity against *P falciparum* W-2. (**a**) Activity W-2 vs net charges O2 and C4. (**b**) Activity W-2 vs net charges C4 and O11. (**c**) Activity W-2 vs net charges C7 and O11. (**d**) Activity W-2 vs net charges O2 and O11.

the ANOVA function, and the last provides the particular predictors associated with the ANOVA function. It is seen that the best MARS model produced seven ANOVA functions, the first three of which involve one predictor, while the other four include two predictors. Examination of the second column shows that all of the ANOVA functions are important; removing any of them substantially degrades the fit. Figures 3a–d display three-dimensional perspective plots representing the joint dependence of the activity on the atomic net charges of various predictors. The figures reveal that an improvement of the activity can be made if the atomic net charges at positions C4 and C7 are kept at moderately positive values, and those at other positions (O1 and O11) are less negative. Table V gives the relative importance of each predictor in the MARS model. These values were standardized so that the most important predictor had a value of 100.

QSAR studies of dihydroartemisinin derivatives against P falciparum W-2

Table VI gives the summary of the best predictive models for the activity against *P falciparum* W-2. It appears that the MLR model used a large number of predictors to improve the fit and the predictive ability. Both ACE and PPR models involved five predictors and their predictive ability is comparable to the MLR one. However, the MARS model involving two predictor interactions gave the best Q^2 value (0.872) obtained from leave-one-out cross-validation, indicating that it is the best method for deriving a relationship between the antimalarial activity of the dihydroartemisinins and their molecular electronic structure. Tables VII and VIII give the ANOVA decomposition and the importance of predictors that entered the resulting MARS model, respectively. Graphical repre-

sentations of the joint contributions of predictors are provided in figures 4a–d. They suggest that the antimalarial activity against *P falciparum* W-2 increases with decreasing electron density on O11, keeping charges of O2, C4 and C7 at rather moderate values.

In both QSAR studies, the MARS method used almost the same set of predictors, {O1, C4, C7, O11}, for the activity against *P falciparum* D-6 and {O2, C4, C7, O11} for the activity against *P falciparum* W-2, to build good predictive models. This finding is not only in agreement with many previous statements about the vital role of the bridged endoperoxide group but also proposes that the electron densities at C4, C7 and O11 are very important for the antimalarial activity of this group of compounds. Furthermore, the graphical representation ability of MARS might be helpful in the search for new effective compounds. Hopefully, along with electronic structure parameters derived by quantum chemical calculations, MARS may contribute significantly to the development of computer-aided drug design.

Acknowledgments

Thanks are due to CB Roosen for sending us the reference [15] and for helpful correspondence. The authors would also like to thank The Institute for Statistics, University of Innsbruck, for their kind cooperation. A grant from the Austrian Federal Government for V Nguyen-Cong is gratefully acknowledged.

References

1 Avery MA, Gao F, Chong WKM, Mehrotra S, Milhous WK (1993) *J Med Chem* 36, 4264–4275
2 Suter HU, Maric DM, Weber J, Thomson C (1995) *Chimia* 49, 125–127
3 Friedman JH (1991) *Ann Statist* 19, 1–141
4 Breiman L, Friedman JH (1985) *J Am Stat Assoc* 80, 580–619
5 Friedman JH, Stuetzle W (1981) *J Am Stat Assoc* 76, 817–823
6 Friedman JH (1985) Technical Report No 12, Department of Statistics, Stanford University
7 Nguyen-Cong V, Rode BM (1995) *Eur J Med Chem* 31, 479–484
8 De Veaux RD, Psichogios DC, Ungar LH (1993) *Comp Chem Engng* 17, 819–837
9 Rogers D, Hopfinger AJ (1994) *J Chem Inf Comput Sci* 34, 854–866
10 Lin AJ, Miller RE (1995) *J Med Chem* 38, 764–770
11 HyperChem, Hypercube Inc, ON, Canada
12 Leban I, Golic L, Japelj M (1988) *Acta Pharm Jugosl* 38, 71
13 Allen FH, Davies JE, Galloy JJ et al (1991) *J Chem Inf Comp Sci* 31, 187–204
14 Gaussian92, Gaussian Inc, Pittsburgh, USA
15 Friedman JH, Roosen CB (1995) *Stat Meth Med Res* 4, 197–217
16 http://lib.stat.cmu.edu/general/ace and http://lib.stat.cmu.edu/general/mars3.5
17 Furnival GM, Wilson Jr RW (1974) *Technometrics* 16, 499–511
18 S-PLUS, StatSci Division, Seattle, USA
19 Crammer RD III, Patterson, DE, Bunce, JD (1988) *J Am Chem Soc* 110, 5959–5967