Journal of Computer-Aided Molecular Design, **15:** 1–12, 2001.
KLUWER/ESCOM
© 2001 *Kluwer Academic Publishers. Printed in the Netherlands.*

A search for sources of drug resistance by the 4D-QSAR analysis of a set of antimalarial dihydrofolate reductase inhibitors

Osvaldo Andrade Santos-Filho* & Anton J. Hopfinger**
Laboratory of Molecular Modeling and Design (M/C-781), University of Illinois at Chicago, College of Pharmacy, 833 South Wood Street, Chicago, IL 60612-7231, U.S.A.

Received 1 May 2000; Accepted 18 July 2000

Key words: dihydrofolate reductase, malaria, mutation, pharmacophore, QSAR

Summary

A set of 18 structurally diverse antifolates including pyrimethamine, cycloguanil, methotrexate, aminopterin and trimethoprim, and 13 pyrrolo[2,3-d]pyrimidines were studied using four-dimensional quantitative structure-activity relationship (4D-QSAR) analysis. The corresponding biological activities of these compounds include IC_{50} inhibition constants for both the wild type, and a specific mutant type of *Plasmodium falciparum* dihydrofolate reductase (DHFR). Two thousand conformations of each analog were sampled to generate a conformational ensemble profile (CEP) from a molecular dynamics simulation (MDS) of 100,000 conformer trajectory states. Each sampled conformation was placed in a 1 Å cubic grid cell lattice for each of five trial alignments. The frequency of occupation of each grid cell was computed for each of six types of pharmacophore groups of atoms of each compound. These grid cell occupancy descriptors (GCODs) were then used as a descriptor pool to construct 4D-QSAR models. Models for inhibition of both the 'wild' type and the mutant enzyme were generated which provide detailed spatial pharmacophore requirements for inhibition in terms of atom types and their corresponding relative locations in space. The 4D-QSAR models indicate some structural features perhaps relevant to the mechanism of resistance of the *Plasmodium falciparum* DHFR to current antimalarials. One feature identified is a slightly different binding alignment of the ligands to the mutant form of the enzyme as compared to the wild type.

Introduction

In spite of early progress malaria remains one of the most serious health problems facing humanity. It affects 300–500 million people causing over 2.5 million deaths annually, the majority of which are children. Today the malaria threat is most significant in Latin America, Africa, Asia, and some regions of the South Pacific. The population at risk represents about 40% of the world's inhabitants [1].

Many strategies are used in developing malaria chemotherapy. One of them involves the use of dihydrofolate reductase inhibitors as potential antimalarial drugs. Dihydrofolate reductase [DHFR; 5,6,7,8-tetrahydrofolate-$NADP^+$ oxidoreductase (E.C. 1.5.1.3)] is an enzyme that catalyses the NADPH-dependent reduction of 7,8-dihydrofolate to 5,6,7,8-tetrahydrofolate, which is the precursor of the cofactors required for the synthesis of purine nucleotides, tymidylate, and several amino acids [2]. Thus, inhibition of DHFR can lead to disruption of DNA synthesis and death of rapidly proliferating cells [2, 3]. This enzyme has been successfully used as a target for the treatment of cancer, bacterial infections, and malaria. A vast amount of information about DHFR has been published and reviewed [2, 4, 5].

Plasmodium falciparum DHFR, *Pf*DHFR, exists as a domain of a bifunctional enzyme, DHFR-thymidylate synthase (TS; 5,10-methylenetetrahydrofolate: dUMP C-methyltransferase, E.C. 2.1.1.45). DHFR is linked to the TS domain by a junctional se-

*Current address: Departamento de Engenharia Química, Instituto Militar de Engenharia, Praça General Tibúrcio 80, Praia Vermelha 22290-270, Rio de Janeiro-RJ, Brazil.
**To whom correspondence should be addressed. E-mail: hopfingr@uic.edu

quence of 94 amino acids [6]. The DHFR domain of the enzyme includes 228 residues (residues 1–228), and the TS domain includes 286 residues (residues 323–608). There is evidence that the DHFR of other protozoa are similarly bifunctional [2].

The DHFR domain is the target for pyrimethamine (PYR), and cycloguanil (CYC, the active metabolite of proguanil) which have been the most widely used antimalarials. Owing to the rapid appearance of antifolate-resistant parasites, the efficacy of these drugs has been severely compromised. Consequently, efforts are now underway to search for new antifolates to overcome current drug resistance. The pyrimethamine and cycloguanil resistance in *P. falciparum* is associated with point mutations in the DHFR part of the DHFR-TS gene. DNA sequencing of several field isolates has revealed that a single Ser108→Asn108 mutation contributes moderately to the resistance to PYR [6–11], which increases when either Asn51→Ile51 or Cys59→Arg59 mutations are present [7,9–12], and becomes much higher with the inclusion of Ile164→Leu164 mutation [8,13]. In contrast, the simultaneous Ser108→Thr108 and Ala16→Val16 mutations have been associated with resistance only to CYC [11, 14], although moderate decreases in CYC response have been found to parallel the resistance to PYR [7, 8, 12].

In this paper we report the four-dimensional quantitative structure-activity relationship, 4D-QSAR, analysis of a set of 18 antifolates composed of pyrimethamine (PYR), cycloguanil (CYC), methotrexate (MTX), aminopterin (AMP), trimethoprim (TMP), and 13 substituted pyrrolo[2,3-d] pyrimidines, as described by Brobey and co-workers, see Table 1 [15]. These authors carried out functional testing on the wild-type, and on all of the naturally occurring *P. falciparum* dihydrofolate reductase mutants, by the characterization of their kinetic parameters and inhibitory responses to the set of inhibitors.

The 4D-QSAR analysis of this set of compounds employed the IC_{50} inhibition constants of each compound against the wild-type enzyme and against one of the seven enzyme mutants as reported in Reference 15 (Asn51→Ile51, Cys59→Arg59, Ser108→Asn108, and Ile164→Leu164). This specific mutant was chosen because it is resistant to both the PYR and CYC antimalarial agents.

Methods

Training set of the Pf*DHFR analogue inhibitors*

The training set of 18 *Pf*DHFR inhibitors is given in Table 1. Biological activities are the 50% nanomolar inhibition measures, IC_{50} (nM), which have been reported in Reference 15. The inhibition potencies of these compounds have been expressed in negative logarithmic units, $-\log IC_{50}$, in the construction of the 4D-QSAR models, and are given as part of Table 1.

The 4D-QSAR formalism applied to the training set

The 4D-QSAR scheme can be applied to both *receptor-dependent* (RD) and *receptor-independent* (RI) problems. In the first scheme, also called structure-based design, the geometry of the receptor (or molecular target) is available. In contrast, in the second scheme either the geometry of the receptor is not available, or it is neglected in the 4D-QSAR analysis because of uncertainty in the receptor geometry and/or ligand binding mode. The RI formalism is applied in this study.

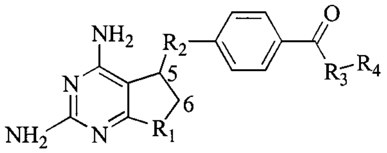
The current methodology formulation of 4D-QSAR analysis consists of ten operational steps which are given in Table 2 [16]. The implementation of this formalism for the analogues listed in Table 1 is described below.

Step 1. The three-dimensional structures of each of the 18 analogues (Table 1) in their neutral forms were constructed using the *HyperChem 5.01* software [18]. Each structure was energy-minimized using the *HyperChem 5.01* MM+ force field without any restriction. Partial atomic charges were computed using the MNDO semiempirical method [19], also implemented in the *HyperChem* program.

Step 2. Seven types of atomic groups, listed below, were used to define the interaction pharmacophore elements, IPEs, in this analysis; (a) any type of atom – *any (0)*, (b) nonpolar atoms – *np (1)*, (c) polar atoms of positive charge – *p+ (2)*, (d) polar atoms of negative charge – *p− (3)*, (e) hydrogen bond acceptor – *hba (4)*, (f) hydrogen bond donor – *hbd (5)*, (g) aromatic carbons and hydrogens – *ar (6)*.

Step 3. The minimized structures of each of the analogs were used as the initial structures in each MDS used to construct the conformational ensemble profile (CEP) of each analog. The *Molsim 3.0* program [20] was used to perform the MDS and to generate the trajectories for, in turn, deriving the CEP. The MDS protocol employed 100 000 steps for each compound,

Table 1. Chemical structures and the biological activities of the training set compounds



Compounds	R_1	5–6	R_2	R_3-R_4	−log IC_{50} (nM) DHFR type	
					Wild	Mutant
P-1	NH		CH_2	Glu	6.30	5.51
P-2	NH		$(CH_2)_2$	Glu(Gly)-OH	7.98	6.57
P-3	NH		$(CH_2)_2$	Glu	7.87	6.38
P-4	NH	H_2	$(CH_2)_2$	Glu	6.42	4.89
P-5	NH	H_2	$(CH_2)_3$Ph-3,4,5-$(OMe)_3$		6.30	4.70
P-6	NH		$(CH_2)_2$	Glu(NHPh-4-COOH)-OH	8.60	7.01
P-7	NH		$(CH_2)_2$	Glu(NHPh-4-CN)-OH	7.90	6.13
P-8	NH		$(CH_2)_3$	Glu	6.87	5.10
P-9	NH	H_2	$(CH_2)_2$-O-	Glu	6.70	4.85
P-10	NH		$(CH_2)_2$	Glu(Phe)-Oh	8.72	6.87
P-11	O		$(CH_2)_2$	Glu	6.66	4.62
P-12	NH		$(CH_2)_2$	Glu(NH-Tet)-OH	8.96	6.94
P-13	NH	H_2	$(CH_2)_2$-S-	Glu	7.49	4.80
PYR*					8.52	4.54
CYC*					7.60	4.74
MTX*					8.74	7.31
AMP*					8.33	7.47
TMP*					6.68	<3.82

*See Figure 1 for the chemical structures.

the step size was 0.001 ps and the simulation temperature was 310 K, the same temperature used in the assay for *Pf*DHFRs activity by Sano and co-workers [21]. An output trajectory file was saved every 50 simulation steps to generate a CEP consisting of 2000 conformations. 4D-QSAR analysis does not use a single conformation in constructing a 4D-QSAR model, but rather the intrinsic conformational flexibility of each compound is taken into account through its CEP.

Steps 4 and 5. Each conformation from the CEP, consisting of the 2000 conformations generated by MDS sampling for each compound, was placed in a reference grid cell space according to the trial alignment under consideration. In this study the size of the cubic grid cell was 1 Å, on its side, and the size of the overall grid cell lattice was chosen to enclose each analog of the training set.

The current (RI) 4D-QSAR methodology uses 3-ordered atom alignments to compare molecules of a training set. The alignments are selected to explore each major ‘part’ of a molecule, as well as all possible combinations of major parts of a molecule. The part, or parts, of a molecule which provide good alignment 4D-QSAR models in a first pass analysis can be explored in greater detail with respect to alignment selection since alignments can be rapidly evaluated in the (RI) 4D-QSAR methodology.

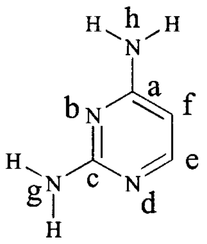
Five different 3-ordered-atom alignments were selected in this study which involve atoms of the 2,4-diamine pyrimidine ring, and the two nitrogens of the substituent amino groups. The atom numbers, and the corresponding letter sequence, for each alignment are listed in Table 3.

The normalized occupancy of each grid cell by each IPE atom type over the CEP for a given alignment forms a unique set of QSAR descriptors referred to as grid cell occupancy descriptors, GCODs. The GCODs were computed and used as the basis set of trial 4D-QSAR descriptors in the 4D-QSAR analyses.

Table 2. The ten operational steps in performing a (RI) 4D-QSAR analysis

Step	Description of the step operation
1	Generate the reference grid cell lattice and initial 3D models for all compounds in the training set
2	Select the trial set of interaction pharmacophore elements, IPEs
3	Perform a MDS [17] conformational ensemble sampling of each compound and generate the corresponding conformational ensemble profile, CEP
4	Select a trial alignment
5	Place each conformation of each compound in the reference grid all lattice according to the alignment, and record the grid cell occupancy descriptor, GCOD, for each IPE and choice in occupancy measure for the CEP
6	Perform a partial least-squares (PLS) data reduction of the entire set of GCODS against the biological activity measure
7	Use the most highly weighted PLS GCODs, and any other user-selected non-GCOD descriptors, for the initial basis set in a genetic algorithm (GA) 4D-QSAR model optimization
8	Return to STEP 4 and repeat STEPS 4–7 unless all trial alignments have been included in the analysis
9	Select the optimum 4D-QSAR model with respect to alignment and any of the methodological parameters
10	Select the low-energy conformer state, from the CEP set, for each compound which predicts the maximum activity using the optimum 4D-QSAR model as the 'active' conformation

Table 3. The set of trial alignments used in constructing the 4D-QSAR models



Alignment no.	1st atom	2nd atom	3rd atom
1	a	b	c
2	h	b	g
3	b	e	f
4	a	f	e
5	b	d	f

Table 4. Wild-type inhibition cross-validated correlation coefficient of the best 4D-QSAR model for each test alignment

Alignment no.	xv – r^2 (optimum)
1	0.80
2	0.80
3	**0.86**
4	0.62
5	0.72

Table 5. Cross-validated correlation coefficient of the best 4D-QSAR model for each alignment after removing the outliers for wild-type inhibition

Alignment no.	xv – r^2 (optimum)
1	**0.90**
2	0.81
3	0.81
4	0.68
5	0.70

Step 6. A 4D-QSAR analysis generates an enormous number of trial QSAR descriptors, GCODs, because of the large number of grid cells and the seven IPEs. Partial least squares (PLS) regression analysis [22] is used to perform a data reduction fit between

the observed dependent variable (in this study the observed biological activities, Table 1) measures and the corresponding GCOD values.

Step 7. The M (currently 200) most highly weighted PLS GCOD descriptors, generated in *step 6*, are used to form the trial basis set for the genetic algorithm (GA) analysis [23]. In this study the Genetic Function Approximation (GFA) [24] was employed in 4D-QSAR model building and optimization. The GFA optimizations were initiated using 200 randomly generated 4D-QSAR models. Mutation probability over the crossover optimization cycle was set at 10%. The smoothing factor (the variable that controls the number of independent variables in the models) was first set to 2.5, but then varied in order to determine the optimal number of descriptors in the 4D-QSAR models [25] on the basis of Friedman's lack-of-fit, LOF [26].

The diagnostic measures used to analyze the resultant 4D-QSAR models generated by the GFA include descriptor usage as a function of crossover operation, linear cross-correlation among descriptors and/or dependent variables (biological activity measures), number of significant models, and indices of model significance including the correlation coefficient, r^2, leave-one-out cross-validation correlation coefficient, $\mathrm{xv} - r^2$, and LOF [26]. Analogues of the training set are considered outliers when the differences in predicted and observed activities exceed 2.0 standard deviations, SD, from the mean of a model.

Step 8. Steps 4–7 were repeated until all (five) trial alignments were included in the 4D-QSAR analysis.

Step 9. The inspection and evaluation of the entire population of 4D-QSAR models is made in this step. The objective is to identify the 'best' 4D-QSAR models with respect to alignment.

Each alignment considered will lead to a particular best 4D-QSAR model *for that specific alignment*. The alignment corresponding to the 4D-QSAR model with the *overall highest* r^2 and $\mathrm{xv} - r^2$measures, for all alignments tested, is selected as the *best* alignment. For the best alignment, a cross-correlation matrix of the residuals in error (observed less predicted activities) between pairs of the top 10 4D-QSAR models, based on their $\mathrm{xv} - r^2$, is built. This is done to determine if the top 10 4D-QSAR models are providing common, or distinct, structure-activity information. In other words it is possible to identify the set of *unique* best 4D-QSAR models. Pairs of models with highly correlated residuals of fit are judged to be nearly the same model while pairs of models with poorly correlated residuals are distinct from one another. Also, the linear cross-correlation matrix of the GCODs for the best 4D-QSAR model for the best alignment is built to determine if these significant GCODs are correlated to one another.

Step 10. The final step of the 4D-QSAR formalism, is to hypothesize the '*active*' *conformation* of each compound in the training set. This is achieved by first identifying all conformer states sampled for each compound, one at a time, that are within ΔE of the global minimum energy conformation of the CEP. Currently, ΔE is set at 2 kcal/mol. The resulting set of low-energy conformations are individually evaluated using the correlation equation of the best 4D-QSAR model. The single conformation within ΔE which predicts the highest 'activity' is selected as the active conformation of the compound.

The hypothesized active conformations can be used as structure design templates, which includes their deployment as the molecular geometries of each ligand in a structure-based ligand-receptor binding study. This is the theme of our further research where the training set (Table 1) and the three-dimensional structures of both wild-type and mutant-type *Pf* DHFR will be used to generate a (RD) 3D-QSAR using the free energy force field (FEFF) 3D-QSAR ligand-receptor binding formalism, as proposed by Tokarski and Hopfinger [28, 29].

Results

Wild-type inhibitors study

For wild-type inhibition, alignment 3 (see Table 3) provides the best 4D-QSAR models as judged by the cross-validated correlation coefficient. The cross-validated correlation coefficients for the five alignments considered are given in Table 4. The best 4D-QSAR model for alignment 3 is,

$$\begin{aligned}-\log\mathrm{IC}_{50} =& -8.46\mathrm{GC1(np)} + 2.38\mathrm{GC2(ar)}\\ &+0.74\mathrm{GC3(np)} + 2.12\mathrm{GC4(p+)}\\ &+1.89\mathrm{GC5(any)} + 6.18\end{aligned} \tag{1}$$

$$N = 18 \quad r^2 = 0.94 \quad \mathrm{xv} - r^2 = 0.86 \quad \mathrm{LSE} = 0.05$$

Three analogues in the training set of Table 1, P-4, P-5, and PYR, are found to be outliers for Equation 1 based on their differences in predicted and observed activities exceeding two standard deviations from the mean. Figure 2 is a composite plot of the observed and the predicted biological activity for each compound of the training set according to Equation 1. The outliers

Table 6. Linear correlation matrix of the residuals of fit for the top ten 4D-QSAR models

Model no.	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	0.96	1.00								
3	0.96	1.00	1.00							
4	0.99	0.97	0.97	1.00						
5	0.98	0.91	0.91	0.97	1.00					
6	0.94	0.96	0.96	0.93	0.85	1.00				
7	0.96	1.00	1.00	0.97	0.91	0.96	1.00			
8	0.87	0.96	0.96	0.89	0.77	0.97	0.96	1.00		
9	0.87	0.96	0.96	0.89	0.77	0.97	0.96	1.00	1.00	
10	**0.54**	**0.50**	**0.50**	**0.54**	**0.56**	**0.48**	**0.50**	**0.44**	**0.44**	1.00

are flagged by name on the horizontal axis in Figure 2. The outlier PYR is predicted to be less active than observed. It is the only compound to have di-ortho substitution with respect to the phenyl ring that restricts the overall conformational freedom of this ligand. P-4 and P-5 are both predicted to be better inhibitors than observed. These are the only compounds in the training set to have H_2 at the 5–6 ring site and only *three* torsion degrees of freedom with respect to R_2. The other two analogs [P-9 and P-13] having H_2 at the 5–6 ring site have *four* degrees of torsion angle freedom in terms of R_2.

In order to improve the 4D-QSAR model, the three outliers were removed from the training set, and the 4D-QSAR analysis was repeated. This new analysis revealed that alignment 1 provides the best 4D-QSAR model, as defined by the highest cross-validated correlation coefficient (see Table 5), and alignment 3 matches with alignment 2 as the second best alignment.

In order to determine if the top ten 4D-QSAR models of the outlier-free data set are providing common, or distinct, structure-activity information among themselves, the cross-correlation coefficients of the residuals of fit between pairs of models were computed, and are reported in Table 6. The idea in determining the pair cross-correlations of these residuals of fit is that a pair of equivalent models will have near-identical residuals of fit, while a distinct pair of models should essentially have noncorrelated residuals [16, 27]. Analysis of Table 6 indicates that only two models, 1 and 10, do not have highly correlated residuals of fit, indicating that they constitute the only two distinct 4D-QSAR models in the top ten set.

The two distinct 4D-QSAR models, 1 (with 4 GCODs), and 10 (with 3 GCODs), were generated, respectively:

$$\begin{aligned}-\log \mathrm{IC}_{50} =& -9.37\mathrm{GC1(np)} + 4.50\mathrm{GC2(any)}\\ &+4.44\mathrm{GC3(np)} + 2.72\mathrm{GC4(any)}\\ &+6.20\end{aligned} \quad (2)$$

$$N = 15 \quad r^2 = 0.94 \quad \mathrm{xv} - r^2 = 0.90 \quad \mathrm{LSE} = 0.04$$

and

$$\begin{aligned}-\log \mathrm{IC}_{50} =& -10.57\mathrm{GC1(any)} + 2.97\mathrm{GC2(np)}\\ &+3.54\mathrm{GC3(hbd)} + 7.13\end{aligned} \quad (3)$$

$$N = 15 \quad r^2 = 0.79 \quad \mathrm{xv} - r^2 = 0.71 \quad \mathrm{LSE} = 0.15$$

In order to be certain that there are only distinct models, the two 4D-QSAR models given by Equations 2 and 3 were further explored. Plots of the predicted biological activities of each of the inhibitors of the training set for the two models as a function of the number of GCODs were constructed. The results are shown in Figure 3, which indicates that models 1 and 10 yield different biological predictions. Figure 4 contains plots of the observed and the predicted IC_{50} values for the two distinct models 1 and 10 (Equations 2 and 3), respectively.

The linear cross-correlation matrix of the GCODs of Equation 2 is given in Table 7. Two insightful observations can be made from an inspection of Table 7. First, no individual GCODs of Equation 2 are significantly correlated with the biological activity, BA. Second, three pairs of GCODs are moderately correlated to one another ($|r| > 0.5$), and their r values are given in bold print. The finding that no single GCOD is highly correlated to BA suggests multiple ligand-receptor sites are involved in tight binding. Moreover, the moderate cross-correlations observed for three pairs of the GCODS, taken with the inter-

Table 7. Linear cross-correlation matrix of the GCODs and inhibition potency for the optimal 4D-QSAR model (Equation 2)

	GC1	GC2	GC3	GC4	BA
GC1	1.00				
GC2	**0.51**	1.00			
GC3	−0.09	−0.39	1.00		
GC4	0.33	**0.58**	−**0.57**	1.00	
BA	−0.08	0.44	0.39	0.29	1.00

Table 8. Mutant inhibition study – the cross-validated correlation coefficient of the best 4D-QSAR model for each alignment

Alignment no.	xv − r^2 (optimum)
1	0.79
2	**0.92**
3	0.77
4	0.80
5	0.83

pretation that multiple ligand-receptor sites are needed for tight binding, is consistent with a cooperative set of intramolecular ligand motions/relaxations as part of the binding process to the enzyme.

The *predicted active conformations* for certain analogues are shown in Figure 5. These conformer states are generated for the best 4D-QSAR model (Equation 2). This model has been developed for a grid cell resolution of 1 Å, which is the diameter of the spheres used to portray the GCODs of the 4D-QSAR equation. The actual grid cells are cubes in space. In Figure 5 the GCODs which increase $-\log \mathrm{IC}_{50}$ are shown as lighter spheres, and the GCODs which diminish inhibition potency are shown as dark spheres. Some hydrogens have been deleted in order to better view the spatial representations of the 4D-QSAR models.

From an inspection of Equation 2 and Figure 5, it is possible to gain a perspective on the GCODs of the model. The positive regression coefficients for GC2, GC3, and GC4 indicate that $-\log \mathrm{IC}_{50}$ should increase with increasing appropriate ligand atom (IPE) occupancy, while the opposite is true for GC1 which has a negative regression coefficient. A hypothesis from an inspection of Figure 5 is that hydrogen bonding between the imino nitrogen of the pyrrole ring and the polar residues Asn51, or Asp54 would improve the ligand-enzyme interaction. Moreover, the GC2 descriptor also suggests hydrogen bonding between a ligand hydrogen acceptor and enzyme residues Ser108, or Thr185. This hydrogen bonding would correspond to the oxygen atom of the furan ring in compound P-11, and the nitrogen atoms of the pterin, pyrimidine, triazine, in compounds MTX and AMP, TMP, and CYC, respectively. The likely reason the IPE of GC2 is ‘any’, and not ‘hba’, is because there are not enough compounds in the training set having ‘hba’ IPE types, but rather a mixture of ‘hba’ and ‘p-’ which ‘forces’ the selection of the ‘any’ IPE.

It is noteworthy that the presence of nonpolar atom types in the grid cell of GC1 diminishes inhibitor binding (a negative regression coefficient). The ‘bottom’ of the DHFR active site is basically a hydrophobic cavity [2, 4, 5], so that hydrophobic interactions involving residues Ile14, Ala16, and Ile164 play important roles in the formation of the ligand-enzyme complex. According to Equation 2 and Figure 5, the cycloguanil chloride ion, and the ‘spacer-group’ between the phenyl ring and the pyrimidine or pyrrole/furan ring are not too important, and/or particularly involved, in the hydrophobic interactions which stabilize the inhibitor-enzyme complex.

Mutant-type inhibitors study

An identical 4D-QSAR study, as described in the last section for the wild-type enzyme, was done for the same training set (Table 1 and Figure 1), using the $-\log \mathrm{IC}_{50}$ of the same inhibitors against a specific mutant [15], as the dependent variable.

The best 4D-QSAR models have four and five GCODs, respectively. These 4D-QSAR models are defined by Equations 4 and 5, given below:

$$\begin{aligned} -\log\mathrm{IC}_{50} =& -10.07\mathrm{GC1(any)} + 24.85\mathrm{GC2(p+)} \\ & +14.91\mathrm{GC3(p-)} - 5.27\mathrm{GC4(np)} \\ & +5.29 \end{aligned} \tag{4}$$

$$N = 18 \quad r^2 = 0.95 \quad \mathrm{xv} - r^2 = 0.91 \quad \mathrm{LSE} = 0.07$$

and

$$\begin{aligned} -\log\mathrm{IC}_{50} =& -9.77\mathrm{GC1(np)} + 16.59\mathrm{GC2(p+)} \\ & +7.15\mathrm{GC3(hbd)} + 14.24\mathrm{GC4(hba)} \\ & -6.05\mathrm{GC5(np)} + 5.34 \end{aligned} \tag{5}$$

$$N = 18 \quad r^2 = 0.97 \quad \mathrm{xv} - r^2 = 0.92 \quad \mathrm{LSE} = 0.05$$

In this 4D-QSAR analysis using the mutant enzyme inhibition data, alignment 2 of Table 3 provides the best 4D-QSAR models according to the cross-validated correlation coefficients which are reported in

Table 9. Linear correlation matrix of the residuals of fit for the top ten 4D-QSAR models for mutant inhibition

Model no.	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	0.71	1.00								
3	0.87	0.62	1.00							
4	0.87	0.62	1.00	1.00						
5	0.87	0.85	0.75	0.75	1.00					
6	0.87	0.85	0.75	0.75	1.00	1.00				
7	0.88	0.81	0.90	0.90	0.88	0.86	1.00			
8	0.88	0.81	0.90	0.90	0.88	0.88	1.00	1.00		
9	0.87	0.83	0.75	0.75	1.00	1.00	0.87	0.87	1.00	
10	0.87	0.83	0.75	0.75	1.00	1.00	0.87	0.87	1.00	1.00

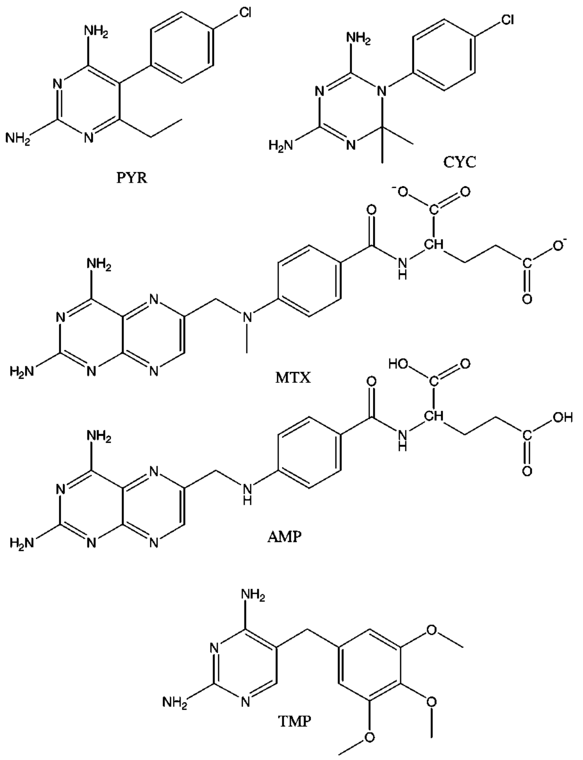


Figure 1. Structures of PYR, CYC, MTX, AMP, and TMP.

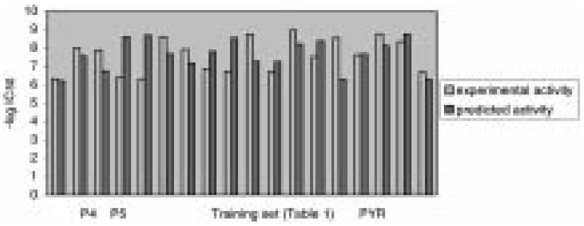


Figure 2. Observed and predicted inhibition potency for each compound using Equation 1.

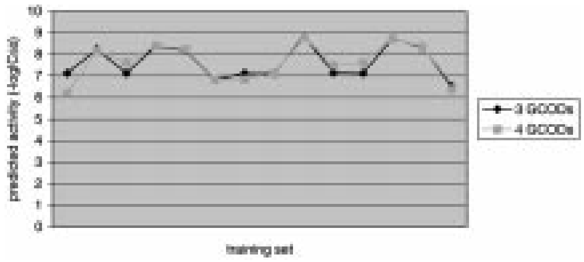


Figure 3. Predicted −log IC_{50} (as a function of the number of GCODs) for the two distinct 4D-QSAR models, defined by Equations 2 and 3, for the wild type inhibition.

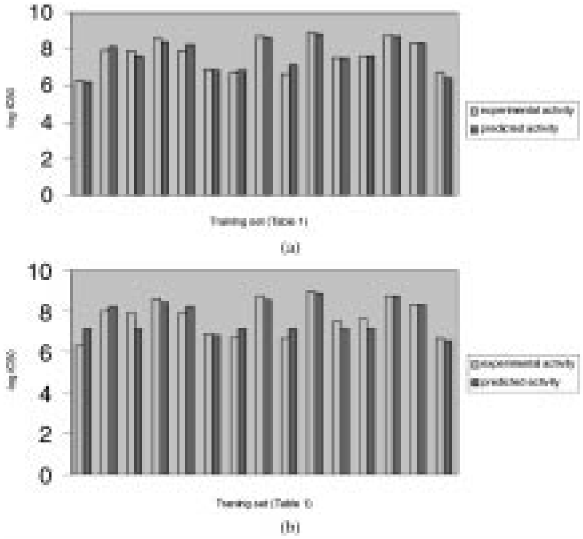


Figure 4. Observed and predicted −log IC_{50} of the training set for models 1 (a) and 10 (b), according to Equations 2 and 3, respectively.

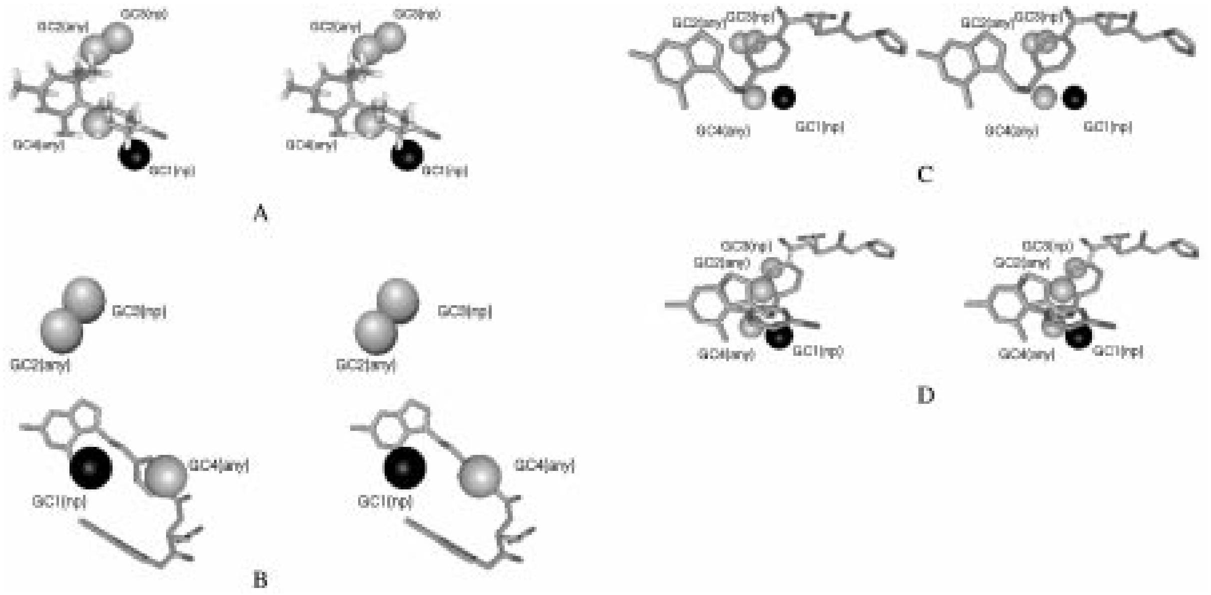


Figure 5. Graphical representations, in stereo format, of three analogues in their respective predicted active conformations using Equation 2. Biological activity-enhancing grid cells are shown as lighter spheres, and grid cells which diminish inhibition potency are shown as dark spheres. (A) CYC; (B) compound P-7; (C) compound P-12; (D) alignment between CYC and P-12.

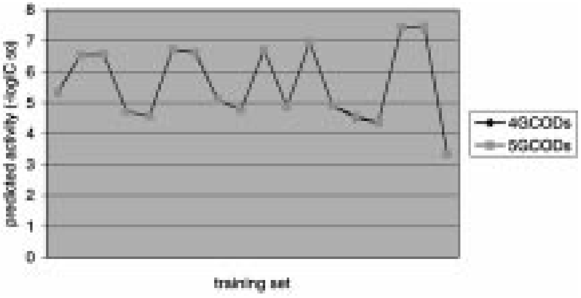


Figure 6. Predicted $-\log$ IC_{50} (as a function of the number of GCODs) for the 4D-QSAR models defined by Equations 4 and 5, for mutant inhibition.

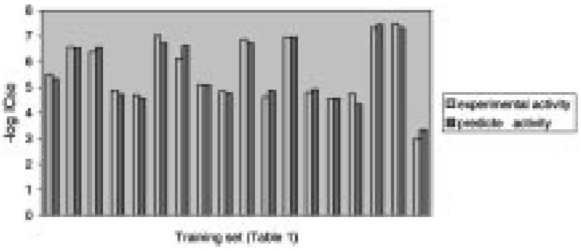


Figure 7. Observed and predicted $-\log$ IC_{50} of the training set for the best 4D-QSAR mutant model.

Table 10. The linear cross-correlation matrix of the GCODs and inhibition potency of the optimal 4D-QSAR model (Equation 4)

	GC1	GC2	GC3	GC4	GC5	BA
GC1	1.00					
GC2	**0.56**	1.00				
GC3	0.22	**0.82**	1.00			
GC4	−0.01	−0.39	−0.43	1.00		
GC5	−0.24	−0.06	0.29	−0.42	1.00	
BA	0.15	0.40	0.23	**0.51**	**−0.71**	1.00

Table 8. The cross-correlation coefficients of the residuals of fit of pairs of top ten models were calculated and are reported in Table 9. All of the top ten 4D-QSAR models have residuals of fit which are highly correlated to one another ($r > 0.70$) which indicates there is only a single 4D-QSAR model.

As in the wild-type 4D-QSAR study, plots of the predicted biological activities of each of the inhibitors of the training set were constructed for Equations 4 and 5 (that is, in a sense, as a function of the number of GCODs), and are shown in Figure 6. The plots in Figure 6 provide a re-confirmation of the results given in Table 10, that there is only one 4D-QSAR model. The model with the highest $\mathrm{xv} - r^2$ value, namely the 4D-QSAR model defined by Equation 5, was selected as the best of the top ten models. No outliers in the training set are found for this 4D-QSAR model. The predicted and the observed biological activities ($-\log$ IC_{50}) are plotted in Figure 7.

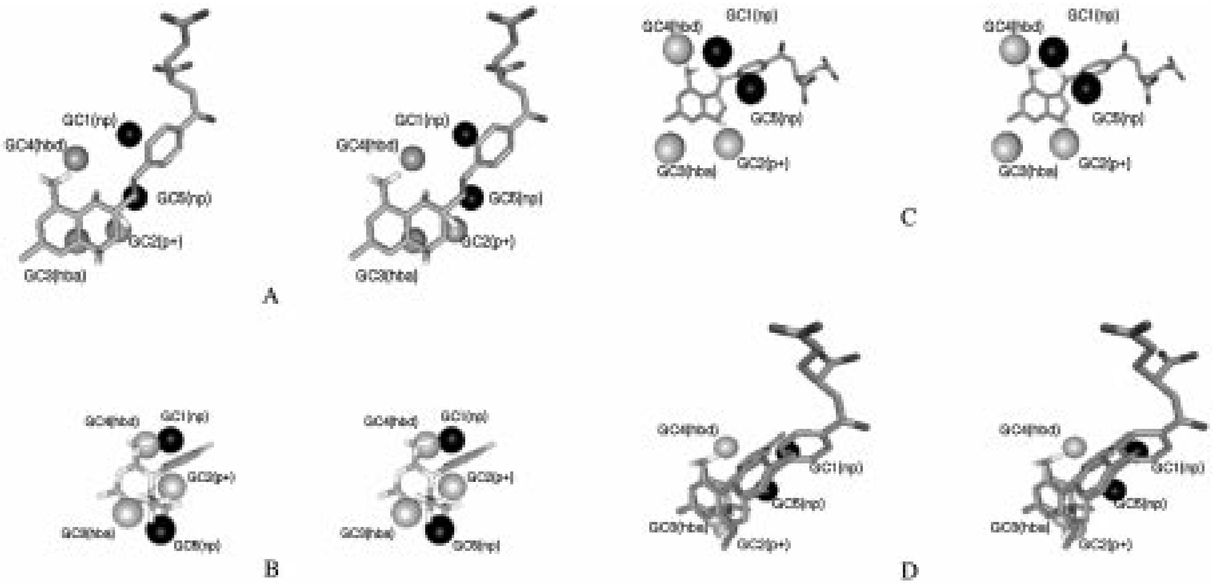


Figure 8. Graphical representations, in stereo format, of three inhibitors in their respective active conformations using Equation 5. Biological activity-enhancing grid cells are shown as lighter spheres, and grid cells which diminish inhibition potency are shown as dark spheres. (A) AMP; (B) CYC; (C) compound P-1; (D) alignment between CYC, PYR, and AMP.

In order to evaluate statistical significance of the mutant 4D-QSAR model (Equation 5), a linear cross-correlation matrix of the GCODs terms was built, see Table 10. As can be inferred from this table, GC4 and GC5 are reasonably correlated to one other, while GC1 and GC2, and GC2 and GC3 are significantly correlated to one another, respectively. Moreover, GC5 has a large positive individual correlation with $-\log \mathrm{IC}_{50}$ and GC4 is moderately self-correlated to inhibition potency.

The *predicted active conformations*, based on the best 4D-QSAR model (Equation 5), for certain analogues are shown in Figure 8 along with the GCODs of Equation 5. Figure 8 and Equation 5 suggest some possible features of specific interactions between individual analogues of the training set and the mutant-form of *Pf* DHFR.

Inhibition potency is predicted to increase with the ligand atom occupancy of grid cells of GC2, GC3, and GC4 by the appropriate IPE types, while the opposite is true for GC1 and GC5. It is noteworthy that the two GCs responsible for predicted decreases in inhibition potency correspond to occupancy by nonpolar atom-types. Either aromatic rings or aliphatic groups (the ‘spacer’) between the aromatic and the pyrimidine (or fused pyrimidine) rings are most often near these two GCODs. In this regard it is important to note that in the mutant-type *Pf* DHFR there are two residue changes that, perhaps, are responsible for the observed negative coefficients of the GC1 and GC5. These mutations are: Asn51→Ile51 (Asn is a polar residue, Ile is nonpolar), and Ile164→Leu164 (both nonpolar residues). Occupancy of GC5 by nonpolar ligand atoms is particularly detrimental to inhibition potency.

All GCs responsible for the increase of the biological activity are characterized by polar, or hydrogen bonding, atom IPEs. This observation indicates the relative ‘importance’ of electrostatic interactions for inhibition of the mutant enzyme by the training set inhibitors. A possible explanation for some of the differences between Equations 2 and 5 may be as follows; GC2(p+) of the mutant model is near an imino nitrogen of the pyrrole ring of some inhibitors. The mutant form of *Pf* DHFR has the mutation Ser108→Asn108. Asn108 can form hydrogen bonds with the imino nitrogens of these ligands. However, compounds P-11, MTX, CYC, AMP, and TMP do not have this requisite imino nitrogen. These compounds have either an oxygen (compoud P-11) or a pterin nitrogen (MTX and AMP). Still, it is possible to propose an alternate possible stabilizing ligand-enzyme interaction for these compounds through a proton transfer, involving another residue such as the Arg provided by the Cys59→Arg59 mutation. GC3 of Equation 5 is located near a nitrogen hydrogen bonding acceptor of the pyrimidine ring which is present in all of

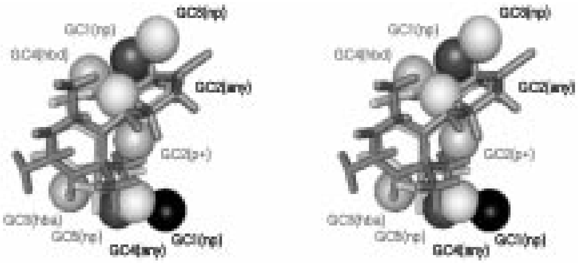


Figure 9. Graphical representations, in stereo format, of CYC in its predicted wild type and mutant active conformations, using Equations 2 and 5, respectively. The GCODs of the two 4D-QSAR models are also shown. The dark labeled GCODs correspond to Equation 2 while the light labeled GCODs are those of Equation 5. Biological activity-enhancing GCODs are shown as lighter spheres, and GCODs which diminish inhibition potency are shown as dark spheres.

the inhibitors of the training set. This GCOD is also likely related to an interaction with the Cys59→Arg59 mutation.

Discussion

There are some significant differences between the *Pf* DHFR wild-type and the mutant 4D-QSAR inhibition models. First, different alignments are found to provide the best 4D-QSAR models for the two forms of the enzyme. The alignment for the wild type enzyme involves three atoms of the pyrimidine ring, C-1, N-3, and C-4, while the alignment for the mutant enzyme uses just one of the atoms of this ring, the N-3. The other two alignment atoms are the two amino nitrogens bonded to the ring. This subtle, but significant, difference in alignment may be a possible hint of steric changes in the active site of the mutant *Pf* DHFR relative to the wild type. This 'hint', together with other complementary theoretical and/or experimental investigations, might prove useful in proposing a resistance mechanism of the mutant strains of the malarial enzyme for both PYR and CYC.

A second relevant difference observed between the wild type and mutant 4D-QSAR models is the 'predominance' of the polar character of the IPEs in the mutant enzyme 4D-QSAR model, while nonpolar IPEs are often found in the wild type inhibition models. This is potentially an important point because if this finding is confirmed in future work, it would be an important guide (a molecular template) for the synthesis and screening of potential drugs to overcome resistance of mutant malarial DHFR.

Figure 9 contains a molecular superimposition of the predicted active conformations of CYC using both Equations 2 and 5. The GCODs of Equation 2 (defined by dark print labels) and of Equation 5 (defined by light print labels) are also superimposed. For both 4D-QSAR models dark spheres denote actively decreasing GCODs. An inspection of Figure 9 suggests that GC3 (np) of the wild type 4D-QSAR model, Equation 2, corresponds to GC1(np) of the mutant model. However, G3 (np) of the wild type is predicted to enhance activity, while GC1(np) of Equation 5 decreases inhibition. GC1(np) and GC4(any) of Equation 2 (wild type) correspond to GC5 (np) of the mutant 4D-QSAR model. GC3(hba), GC4(hbd) and GC2(p+) of the mutant 4D-QSAR model have no corresponding GCODs in the wild type model. GC2(any) of the wild type model may have no corresponding GCOD in the mutant model or could possibly be related to GC4(hbd) and/or GC1(np). The two predicted active conformations are quite similar to one another, and do not differ at any particular sites from one another.

Finally it is very important to mention that this is a *receptor-independent* (RI) 4D-QSAR analysis of a training set whose inhibition activities span both the wild type and a specific mutant-type *Pf* DHFR. We are in the process of expanding this investigation by doing a *receptor-dependent* (RD) 3D-QSAR, using the free energy force field (FEFF) 3D-QSAR method [28, 29], using both homology modeled wild type and mutant type *Pf* DHFR, the same inhibitor training set, and the same respective best ligand alignments as found and employed, respectively, in this study. Moreover, the crystallographically determined structure of *Plasmodium falciparum* dihydrofolate reductase has been reported [30] and three sets of coordinates are on hold with the Protein Data Bank. Once these coordinates are released, we will use them in performing an additional FEFF 3D-QSAR modeling of this training set.

Acknowledgements

O.A.S.-F. is grateful to the CAPES Foundation, a federal scientific agency of Brazil, for scholarship support to study at UIC. Resources of the Laboratory of Molecular Modeling and Design were used in this research. Amit S. Kulkarni, of the Laboratory, provided many useful discussions and insights to us over the course of this work.

References

1. The World Health Organization Report; Who Publications, Geneva, 1997.
2. Blakley, R.L., In: Blakley, R.L. and Benkovic, S.J. (Eds) Folates and Pterins, Vol. 1; John Wiley & Sons, New York, NY, 1984, p. 191.
3. Brown, K.A. and Kraut, J., Faraday Discuss., 93 (1992) 217.
4. Kraut, J. and Matthews, D.A., In Jurnak, F. and McPherson, A. (Eds), Biological Macromolecules and Assemblies, Vol. 3. John Wiley & Sons, New York, NY, 1987, p. 1.
5. Miller, G.P. and Benkovic, S.J., J. Chem. Biol., 5 (1998) R105.
6. Bzik, D.J., Li, W.-B., Horii, T. and Inserburg, J., Proc. Natl. Acad. Sci. USA, 84 (1987) 8360.
7. Basco, L.K., DePécoulas, P.E., Wilson, C.M., LeBras, J. and Mazabraud, A., Mol. Biochem. Parasitol., 69 (1998) 135.
8. Cowman, A.F., Morry, M.J., Biggs, B.A., Cross, G.A.M. and Foot, S.J., Proc. Natl. Acad. Sci. USA, 85 (1988) 9109.
9. Peterson, D.S., Walliker, D. and Wellens, T.E., Proc. Natl. Acad. Sci. USA, 85 (1998) 9114.
10. Peterson, D.S., DiSanti, S.M., Povoa, M., Calvosa, V.S., DoRosario, V.E. and Wellens, T.E., Am. J. Trop. Med. Hyg., 45 (1991) 492.
11. Peterson, D.S., Milhouse, W.K. and Wellens, T.E., Proc. Natl. Acad. Sci. USA, 87 (1990) 3018.
12. Thaithong, S., Chan, S.-W., Songsomboon, S., Wilairat, P., Seesod, N., Sueblinwong, T., Goman, M., Ridley, R. and Beale, G., Mol. Biochem. Parasitol., 52 (1992) 149.
13. Snewin, V.A., England, S.M., Sims, P.F.G. and Hyde, J.E., Gene, 76 (1989) 41.
14. Foot, S.J., Galatis, D. and Cowman, A.F., Proc. Natl. Acad. Sci. USA, 87 (1990) 3014.
15. Brobey, R.K.B., Iwakura, M., Itoh, F., Aso, K. and Horii, T., Parasitol. Int., 47 (1998) 69.
16. Hopfinger, A.J., Wang, S., Tokarski, J.S., Jin, B., Albuquerque, M.G., Madhav, P.J. and Duraiswami, C., J. Am. Chem. Soc., 119 (1997) 10509.
17. van Gunsteren, W.F. and Berendsen, H.J.C., Angew. Chem., Int. Ed. Engl., 29 (1990) 992.
18. HyperChem Program Release 5.01 for Windows; Hypercube, Inc., 1996.
19. Dewar, M.J.S. and Theil, W., J. Am. Chem. Soc., 99 (1977) 4899.
20. Molsim User's Guide v.3.0, Molecular Mechanics and Dynamics Simulation Software, D.C. Doherty and The Chem21 Group, Inc., Lake Forest, IL, 1994.
21. Sano, G.-I., Morimatsu, K. and Horii, T., Mol. Biochem. Parasitol., 63 (1994) 265.
22. Glen, W.G., Dunn, W.J., III and Scott, D.R., Tetrahedron Comput. Methods, 2 (1989) 349.
23. Holland, J., Adaptation in Artificial and Natural Systems. University of Michigan Press, Ann Arbor, MI, 1975.
24. Rogers, D. and Hopfinger, A.J., J. Chem. Inf. Comput. Sci., 34 (1994) 854.
25. Rogers, D., WOLF Reference Manual Version 5.5, Molecular Simulation Inc., 1994.
26. Friedman, J., Multivariate Adaptative Regression Splines. Technical Report No. 102; Laboratory for Computational Statistics, Department of Statistics, Stanford University, Stanford, CA, November 1988 (revised August 1990).
27. 4D-QSAR Manual v.1.0, A.J. Hopfinger and The Chem21 Group, Inc., Lake Forest, IL, 1997.
28. Tokarski, J.S. and Hopfinger, A.J., J. Chem. Inf. Comput. Sci., 37 (1997) 779.
29. Tokarski, J.S. and Hopfinger, A.J., J. Chem. Inf. Comput. Sci., 37 (1997) 792.
30. Peterson, M.R., Hall, D.R., Berriman, M., Nunes, J.A., Leonard, G.A., Fairlamb, A.H. and Hunter, W.N., J. Mol. Biol., 298 (2000) 1230.