Journal of Computer-Aided Molecular Design, **15:** 787–810, 2001.
KLUWER/ESCOM

Free energy force field (FEFF) 3D-QSAR analysis of a set of *Plasmodium falciparum* dihydrofolate reductase inhibitors

Osvaldo A. Santos-Filho*, Rama K. Mishra & A. J. Hopfinger**
Laboratory of Molecular Modeling and Design (M/C-781), University of Illinois at Chicago, College of Pharmacy, 833 South Wood Street Chicago, Illinois 60612-7231, USA

Received 21 May 2001; accepted 3 July 2001

Key words: 3D-SQAR, molecular dynamics, *Plasmodium falciparum* DHFR, malaria

Summary

Free energy force field (FEFF) 3D-QSAR analysis was used to construct ligand-receptor binding models for a set of 18 structurally diverse antifolates including pyrimethamine, cycloguanil, methotrexate, aminopterin and trimethoprim, and 13 pyrrolo[2,3-d]pyrimidines. The molecular target ('receptor') used was a 3D-homology model of a specific mutant type of *Plasmodium falciparum* (Pf) dihydrofolate reductase (DHFR). The dependent variable of the 3D-QSAR models is the IC_{50} inhibition constant for the specific mutant type of *Pf*DHFR. The independent variables of the 3D-QSAR models (the descriptors) are scaled energy terms of a modified first-generation AMBER force field combined with a hydration shell aqueous solvation model and a collection of 2D-QSAR descriptors often used in QSAR studies. Multiple temperature molecular dynamics simulation (MDS) and the genetic function approximation (GFA) were employed using partial least square (PLS) and multidimensional linear regressions as the fitting functions to develop FEFF 3D-QSAR models for the binding process. The significant FEFF energy terms in the best 3D-QSAR models include energy contributions of the direct ligand-receptor interaction. Some changes in conformational energy terms of the ligand due to binding to the enzyme are also found to be important descriptors. The FEFF 3D-QSAR models indicate some structural features perhaps relevant to the mechanism of resistance of the *Pf*DHFR to current antimalarials. The FEFF 3D-QSAR models are also compared to receptor-independent (RI) 4D-QSAR models developed in an earlier study and subsequently refined using recently developed generalized alignment rules.

Introduction

Malaria is still a serious health problem in many regions of the planet, affecting mainly the tropical countries. 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 DHFR inhibitors as potential antimalarial drugs.

Three dimensional quantitative structure-activity relationship (3D-QSAR) analysis can be applied to a structure-activity training set in two different application modes. The first application mode is termed *receptor-independent 3D-QSAR* (RI 3D-QSAR) analysis and no receptor geometry is available. The other application of 3D-QSAR is termed *receptor-dependent 3D-QSAR* (RD 3D-QSAR) and the receptor geometry is used in computing potential QSAR independent variables.

This work is an application of the RD 3D-QSAR methodology proposed by Tokarski and Hopfinger called free energy force field (FEFF) 3D-QSAR analysis [2, 3]. Eighteen *Plasmodium falciparum* DHFR (*Pf*DHFR) inhibitors including pyrimethamine, cycloguanil, methotrexate, aminopterin and trimethoprim, and 13 pyrrolo[2,3-d]pyrimidines make up a training set for the development of a FEFF 3D-QSAR

*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 shoul be addressed. E-mail: hopfingr@uic.edu model. The molecular target ('receptor') used is a computational homology-based model of a specific mutant type of *Pf*DHFR.

Methods

The FEFF 3D-QSAR formalism [2, 3]

The ligand-receptor interaction can be expressed as

$$\begin{array}{ll} \mathrm{L \cdot M + R \cdot M} \overset{K}{\rightleftharpoons} \mathrm{(LR) \cdot M} & \\ \text{(unbound state)} \qquad \text{(bound state)} & \end{array} \tag{1}$$

where L is the ligand, R is the receptor, M is the solvent medium, and *K* is the equilibrium, or binding constant. The difference in free energy between the bound and unbound states of a ligand, L, to a receptor, R, in a solvent medium, M, can be expressed as

$$\Delta \mathrm{G} = \mathrm{G_{LR}} - (\mathrm{G_L} + \mathrm{G_R}) = -\mathrm{RT} \ln K, \tag{2}$$

where ΔG is the binding free energy, G_{LR} is the free energy of the bound, or complex, state, G_L is the free energy of the unbound ligand, G_R is the free energy of the unbound receptor, R is the gas constant and T is the temperature of the system. The free energy of an enzyme-ligand complex can be approximately broken down into a set of component interactions as follows,

$$\begin{aligned} \mathrm{G_{LR}} = {} & [\mathrm{G_{LR}(LL) + G_{LR}(RR) + G_{LR}(MM)} + \\ & \mathrm{G_{LR}(LR) + G_{LR}(LM) + G_{LR}(RM)}], \end{aligned} \tag{3}$$

where G_{LR} (XY) refers to the interaction between X and Y where they each can be L, M or R.

The interaction terms can be divided into their respective enthalpy, H_{LR}, and entropy, S_{LR}, contributions.

$$\mathrm{G_{LR} = H_{LR} - TS_{LR}}. \tag{4}$$

At low solute concentration the enthalpy terms, $H_{LR}(XY)$, can be represented by their respective internal energies, $E_{LR}(XY)$,

$$\begin{aligned} \mathrm{H_{LR}} = {} & \mathrm{E_{LR}} = [\mathrm{E_{LR}(LL) + E_{LR}(RR) + E_{LR}(MM)} + \\ & \mathrm{E_{LR}(LR) + E_{LR}(LM) + E_{LR}(RM)}] \end{aligned} \tag{5}$$

and the entropy term, $S_{LR}(XY)$, contributions can be expressed in the same manner,

$$\begin{aligned} \mathrm{S_{LR}} = {} & [\mathrm{S_{LR}(LL) + S_{LR}(RR) + S_{LR}(MM)} + \\ & \mathrm{S_{LR}(LR) + S_{LR}(LM) + S_{LR}(RM)}]. \end{aligned} \tag{6}$$

The unbound ligand, G_L, and receptor, G_R, free energies have the following components

$$\mathrm{G_L} = [\mathrm{G_L(LL) + G_L(LM) + G_L(MM)}], \tag{7}$$

$$G_{\mathrm{R}} = [\mathrm{G_R(RR) + G_R(RM) + G_R(MM)}]. \tag{8}$$

The enthalpy contributions of L and R at low concentration, $H_L(XY)$ and $H_R(XY)$ can also be represented by their internal energies, $E_L(XY)$ and $E_R(XY)$ as in Equations 5 and 6. The complete set of contributions to the internal energy and entropy, and their representations, are given in Table 1.

The terms in Table 1 can be selected as the independent variables used in the FEFF 3D-QSAR analysis. However, the free energy of binding, ΔG, can also be represented by the individual free energy force field terms for L, R, and LR in Table 1.

$$\begin{aligned} \Delta \mathrm{G} = {} & \alpha_1 \Delta \mathrm{E_{stretch}} + \alpha_2 \Delta \mathrm{E_{bend}} + \alpha_3 \Delta \mathrm{E_{torsion}} + \\ & \alpha_4 \Delta \mathrm{E_{vdW}} + \alpha_5 \Delta \mathrm{E_{electrostatic}} + \\ & \alpha_6 \Delta \mathrm{E_{hydrogen\ bonding}} + \alpha_7 \Delta \mathrm{E_{solvation}} - \\ & \alpha_8 \mathrm{T} \Delta \mathrm{S}, \end{aligned} \tag{9}$$

where $\Delta E_{stretch}$ is the unbound to bound change in internal energy for bond stretching, ΔE_{bend} is the change in bond angle bending energy, $\Delta E_{torsion}$ is the change in torsional energy, ΔE_{vdW} is the change in van der Waals interaction energy, $\Delta E_{electrostatic}$ is the change in electrostatics interaction energy, $\Delta E_{hydrogen}$ bonding is the change in hydrogen bonding energy, $\Delta E_{solvation}$ is the change in solvation energy, and ΔS is the total change in the entropy of the L, R, M system which can be partitioned into component contributions.

The hydration shell model proposed by Hopfinger [4] was included in the potential energy function to calculate the solvation energies. Only the L and R components to the free energy of aqueous solvation can be extracted from this model. Thus, only these energy terms can be used as trial descriptors in building a FEFF 3D-QSAR model.

The *PfDHFR* ligands have limited conformational flexibility. It was, therefore, assumed that the ligand entropy contribution from a change in the ligand conformational change upon binding to the receptor is small, and nearly constant across the analog series. The change in the conformational entropy of the receptor upon ligand binding is also assumed to be constant across the analog series. Overall, conformational entropy terms have been explicitly neglected in this study.

The internal energy change upon ligand-receptor binding is given by,

$$\Delta \mathrm{E_X} = \mathrm{E_{LR,X}} - (\mathrm{E_{L,X}} + \mathrm{E_{R,X}}), \tag{10}$$

Table 1. Breakdown of the FEFF interaction terms, XY, for a ligand (L)-receptor (R) in a solvent medium (M)

Binding component(s)	Type of interaction energy, XY	Change in internal energy, symbols	Change in Entropy, symbols
Ligand L	Intramolecular ligand Conformational Energy LL	$\Delta E_L(LL) = E_{LR}(LL) - E_L(LL)$	$\Delta S_L(LL) = S_{LR}(LL)$ - $S_L(LL)$
Ligand L	Ligand solvation energy, LM	$\Delta E_L(LM) = E_{LR}(LM) - E_L(LM)$	$\Delta S_L(LM) = S_{LR}(LM) - S_L(LM)$
Solvent medium M	Solvent reorganizational energy, MM	$\Delta E_M(MM) = E_{LR}(MM) - [E_L(MM) + E_R(MM)]$	$\Delta S_M(MM) = S_{LR}(MM) - [S_L(MM) + S_R(MM)]$
Receptor R	Intramolecular receptor conformational energy, RR	$\Delta E_R(RR) = E_{LR}(RR) - E_R(RR)$	$\Delta S_R(RR) = S_{LR}(RR) - S_R(RR)$
Receptor R	Receptor solvation energy, RM	$\Delta E_R(RM) = E_{LR}(RM) - E_R(RM)$	$\Delta S_R(RM) = S_{LR}(RM) - S_R(RM)$
Ligand-receptor RL	Intermolecular ligand-Receptor energy, LR	$\Delta E_{LR}(LR) = E_{LR}(LR)$	$\Delta S_{LR}(LR) = S_{LR}(LR)$

where X represents each of the internal energy contributions as defined in Equation 9. The potential function parameters used to calculate the nonbonded, electrostatic, torsional, bond stretching, and bond angle bending energy terms of Equation 9 were taken from the AMBER force field [6]. Missing force field parameters (torsional, bond stretching, and bond angle bending) were scaled from a set proposed by Hopfinger [7] and the MM2 force field [8]. The most similar set of atoms to those of the missing AMBER parameter is identified for a parameter which has both AMBER and MM2 (or Hopfinger) values. The ratio of the known parameter from the AMBER and MM2 force fields is determined. The unknown AMBER parameter value is then scaled by the same ratio against the known MM2 value. This type of linear scaling approximation in force field parameterization is further compensated by the subsequent force field fitting process, which is central to the FEFF methodology.

Binding free energy can often be estimated on a relative basis from binding and inhibition constants. Consequently, *in vitro* measures of biological activity, such as IC_{50}'s, can be taken to reflect relative ligand binding strength (thermodynamics) and re-expressed on an energy scale if these measures are to be used as the dependent variable set in a FEFF 3D-QSAR study [3]. Such a scaling of *in vitro* activities is only valid for the training set and its close analogs.

Biological activity: the dependent variable

The 18 *Pf*DHFR inhibitors are given in Table 2 [9]. The ΔG values for this training set of compounds is not available, but it has been assumed that the ΔG scale to the measured 50% nanomolar inhibition values, IC_{50} (nM), which are reported in ref. [9]. The inhibition potencies of these compounds have been expressed in negative logarithmic units, $-\log IC_{50}$, and are given in Table 2. This particular representation of the biological activity measures follows from,

$$\Delta G = -2.303RT \log K, \qquad (11)$$

where the ligand-receptor binding constant, K, at some fixed temperature T, can be approximately replaced by the IC_{50} value for an analog set of inhibitors to a common enzyme.

The receptor

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 [10]. This enzyme has been successfully used as a target for the treatment of cancer, bacterial infections, and malaria. The interested reader can obtain more information about DHFR in the reviews and references cited [10–12].

*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 sequence of 94 amino acids [13]. 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 bifuncional [10].

The three-dimensional structure of *Pf*DHFR has not been determined by experimental methods. However, one of us and his coworkers have proposed a low-resolution computational homology model for this enzyme using homology-modeling techniques [14–16]. The homology model has been constructed for a specific mutant type of PfDHFR (Asn51 → Ile51, Cys59 → Arg59, Ser108 → Asn108, and Ile164 → Leu164).

In order to be able to construct this enzyme homology model, crystal structures of DHFR (human [17], chicken liver [18, 19], and *Escherichia coli* [20, 21]) from the Protein Data Bank (PDB) [22, 23] were used as structural templates. Among these enzymes whose crystal structures were known, chicken liver DHFR was chosen as the primary template because this DHFR enzyme has the highest degree of homology (35%) with the *Pf*DHFR. It is important to mention here that in the construction of the *Pf*DHFR homology model the natural substrate (folate) and the reduced form of its cofactor (NADPH) were docked into their known respective sites on the enzyme.

One validation test of the *Pf*DHFR homology model was performed using the PROCHECK program [24]. The result of this validation study shows 95.3% of all amino acid residuals, and 100% of the active site amino acid residues, respectively, are in favorable regions of the Ramachandran plot [25], as is shown in Figure 2. The details of this enzyme homology modeling are being prepared for publication elsewhere.

Table 2. Chemical structures and the inhibition potencies of the training set compounds

Compounds	R_1	5–6	R_2	R_3-R_4	– $\log IC_{50}$ (nM) DHFR (specific mutant)
P-1	NH		CH_2	Glu	5.51
P-2	NH		$(CH_2)_2$	Glu(Gly)-OH	6.57
P-3	NH		$(CH_2)_2$	Glu	6.38
P-4	NH	H_2	$(CH_2)_2$	Glu	4.89
P-5	NH	H_2	$(CH_2)_3$Ph-3,4,5-$(OMe)_3$		4.70
P-6	NH		$(CH_2)_2$	Glu(NHPh-4-COOH)-OH	7.01
P-7	NH		$(CH_2)_2$	Glu(NHPh-4-CN)-OH	6.13
P-8	NH		$(CH_2)_3$	Glu	5.10
P-9	NH	H_2	$(CH_2)_2$-O-	Glu	4.85
P-10	NH		$(CH_2)_2$	Glu(Phe)-OH	6.87
P-11	O		$(CH_2)_2$	Glu	4.62
P-12	NH		$(CH_2)_2$	Glu(NH-Tet)-OH	6.94
P-13	NH	H_2	$(CH_2)_2$-S-	Glu	4.80
PYR*					4.54
CYC*					4.74
MTX*					7.31
AMP*					7.47
TMP*					<3.82

*See Figure 1 for the chemical structures.

Building and docking the ligands

The geometry of the folate-*Pf*DHFR homology complex was used as the starting structure for docking the training set inhibitors given in Table 2 to the enzyme active site. Inhibitors were intitally docked using the *HyperChem* program [26]. The initial docking conformations and alignment of the inhibitors were those identified in the prior 4D-QSAR analysis of this same training set [27]. Bad steric contacts between a bound inhibitor and side chains of active site residues were relieved by direct energy minimization of the inhibitor-active site receptor complex.

Enzyme model size determination

The *Pf*DHFR homology model, without substrate and cofactor, contains 3827 atoms including protons. The large size of this enzyme would require excessive computational resources in order to perform meaningful molecular dynamics simulations (MDS's) of the enzyme and enzyme-inhibitor complex. Thus, the *Pf*DHFR homology model was scaled down in size to make the requisite MDS practical to carryout. The largest inhibitor (compound 9 of Table 2) was first selected to be docked into the enzyme active site in order to determine the minimum effective size of a scaled down enzyme model needed to perform a reliable FEFF 3D-QSAR analysis.

The determination of the scaled down enzyme binding model was performed using the 'pruning' method of Tokarski and Hopfinger [2, 3]. Spherical enzyme models of 12, 10, and 8 Å radii, centered around the docked ligand, compound 9, see Figure 3, were constructed by retaining only those enzyme atoms within each sphere as atoms and neglecting the remainder of the enzyme. The pruned enzyme models essentially consist of amino acid residues clustered around the active site containing the inhibitors. Enzyme residues that had at least one non-hydrogen atom within the pruning sphere were included in the corresponding pruned enzyme model. The pruning process usually results in an enzyme model consisting of a number of nonbonded (unconnected) peptide residue clusters. Peptide residue clusters separated by less than five intervening amino acid residues were 'connected' by including the intervening amino acid residues. This scheme is intended to retain local geometric integrity of the enzyme active site model for the pruning process.

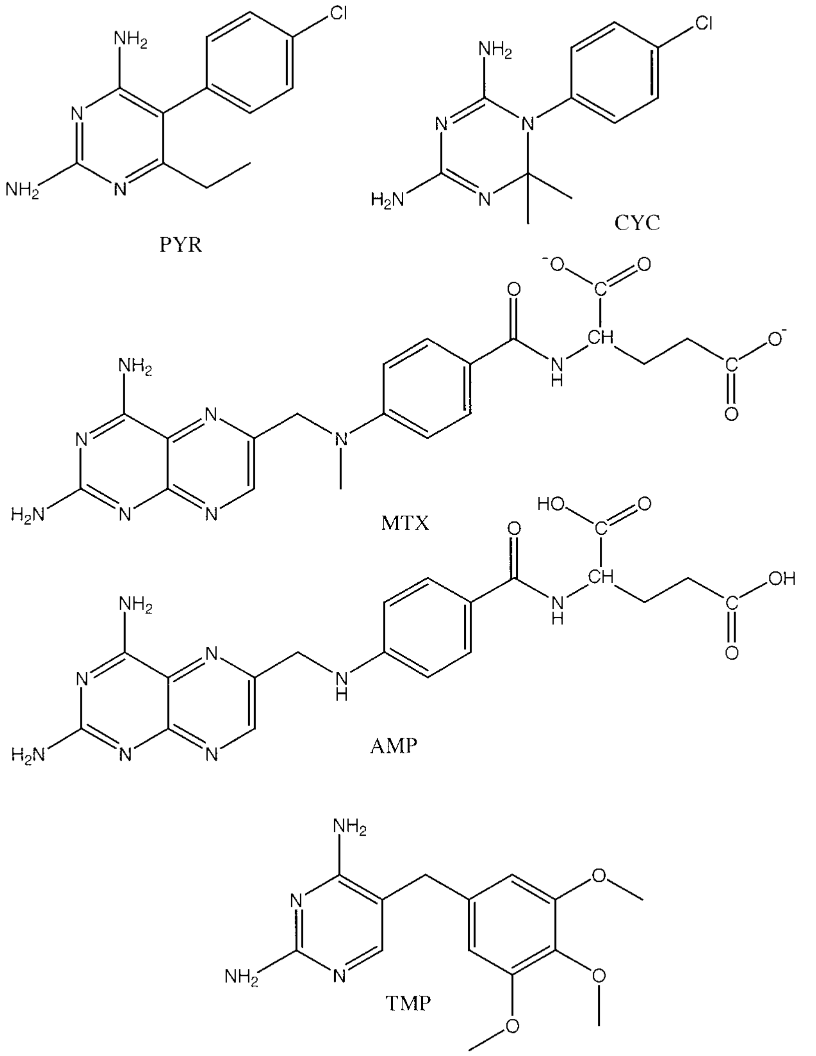
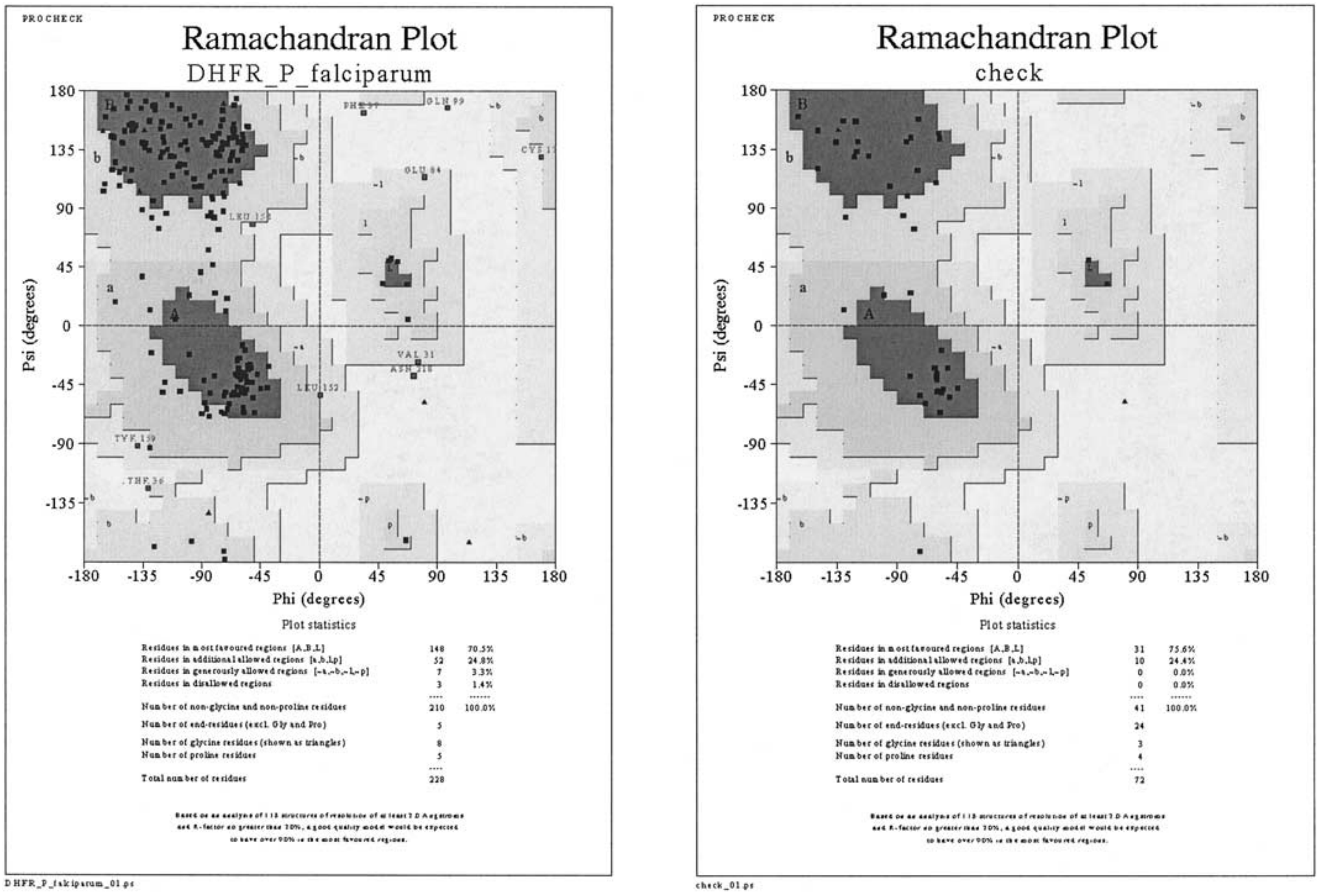


Figure 1. Structures of PYR, CYC, MTX, AMP, and TMP. The three-alignment atoms used in the 4D-QSAR models are the two amino nitrogens and the ring nitrogen 'between' the amino groups.

Each pruned enzyme active site - inhibitor model was evaluated for both overall conformational stability and to estimate the enzyme-inhibitor binding energy by performing MDS. A MDS of 20 ps at 310 K was first performed on each of the model enzyme active sites with compound 9 bound using a nonbonded interaction cutoff distance of 15 Å in the MDS runs. The molecular dielectric was fixed at 3.5, and 310 K was selected as the initial simulation temperature because this was the same used temperature in the assay for *Pf*DHFRs activity by Sano and coworkers [28], and in the previous 4D-QSAR analysis of this training set [27].

The pruned enzyme models contain unconnected peptides groups which could wander about in space over a MDS. However, assigning fictitious high masses to all of the atoms in a pruned enzyme model creates 'momentum reservoirs' which hold the receptor atoms near their initial locations in the parent enzyme structure, but do allow some flexibility of positioning to accomodate a ligand. The use of fictitious masses is virtually the same as using Cartesian constraints, particularly when the masses are chosen to be very large. Fictitious masses provides a convenient way to 'tune in' MDS motions in the pruned enzyme model that are similar to those of the complete parent enzyme by selectively varying the mass values [3].



Plot statistics

Residues in most favoured regions [A,B,L]	148	70.5%
Residues in additional allowed regions [a,b,l,p]	52	24.8%
Residues in generously allowed regions [~a,~b,~l,~p]	7	3.3%
Residues in disallowed regions	3	1.4%
	----	------
Number of non-glycine and non-proline residues	210	100.0%
Number of end-residues (excl. Gly and Pro)	5	
Number of glycine residues (shown as triangles)	8	
Number of proline residues	5	

Total number of residues	228	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

Plot statistics

Residues in most favoured regions [A,B,L]	31	75.6%
Residues in additional allowed regions [a,b,l,p]	10	24.4%
Residues in generously allowed regions [~a,~b,~l,~p]	0	0.0%
Residues in disallowed regions	0	0.0%
	----	------
Number of non-glycine and non-proline residues	41	100.0%
Number of end-residues (excl. Gly and Pro)	24	
Number of glycine residues (shown as triangles)	3	
Number of proline residues	4	

Total number of residues	72	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

Figure 2. A Ramachandran plot of the (ϕ, ψ) of *Pf*DHFR; (a) the entire enzyme, and (b) the active site.

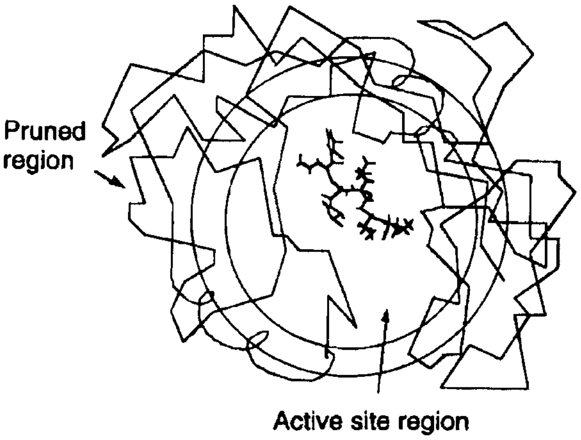


Figure 3. Schematic representation of the pruning sphere geometry used to determine the size of the pruned active site enzyme model. Two possible distance cutoffs are shown.

Simulation sampling of binding

The scheme of molecular dynamics simulation sampling of ligand-receptor binding interactions as a function of temperature, developed by Tokarski and Hopfinger [3], was used in this study. In this approach, a model of the ligand docked in the active site of the receptor is selected and used as the starting point of the ligand-receptor binding simulation sampling. A MDS of the complex at an initial temperature, and for a specific simulation time is performed. The lowest energy structure sampled during the MDS is identified and then used as the starting geometry for a subsequent MDS at a lower selected temperature. The lowest energy structure from this second MDS is then selected as the starting point of a subsequent simulation run at yet a lower simulation temperature. This process of performing structure-coupled MDS at progressively lower temperatures is repeated until a user-defined final temperature of evaluation is reached.

The lowest energy structure from the final ligand-receptor complex MDS is 'broken up' into its individual two components, namely, the ligand and the receptor. The separated bound ligand and bound receptor structures are subjected to independent MDS simulations, starting at the lowest simulation temperature of evaluation, to begin the process of modeling these two molecules in their respective unbound states. The lowest energy structure found for both the unbound ligand and unbound receptor from the MDS at the lowest simulation temperature are then used as starting structures for a MDS performed at the next higher simulation temperature. This process of 'warming up' the unbound ligand and unbound receptor is repeated until the highest temperature selected in the process is reached. The lowest energy structure found at the highest simulation temperature for both the ligand and receptor are then employed as respective starting points in a series of structure-coupled MDS each at a successively lower temperature identical to that done for the ligand-receptor complex. This type of 'cooling down' process allows the unbound ligand and the unbound receptor to equilibrate at each of the simulation temperatures. This general process of structure-coupled MDS sampling as a function of simulation temperature is shown in Figure 4.

In this particular study, the lowest energy geometries, and the corresponding energy terms, *from each of the simulation temperatures sampled* during the 'cooling down' process (for the each of the complexes, E_{LR}, unbound ligands, E_L, and unbound receptor, E_R) were used, to construct corresponding FEFF 3D-QSAR models. The set of energy terms considered are the composite set of those given in Table 1 and Equation 9, see Table 3. Thus, FEFF 3D-QSAR models were developed for each simulation temperature sampled. The best FEFF 3D-QSAR model and the most significant simulation temperature for this particular ligand-receptor system was selected as that QSAR model and temperature yielding the statistically most significant correlation relationship.

Trial descriptors not directly determined from the force field energy terms were also included as possible FEFF 3D-QSAR features. These descriptors are the homo, lumo, hbd, hba, and log P properties of the inhibitor (ligand) which are defined at the bottom of Table 3. It was thought these descriptors might compensate for some of the limitations of the force field representation in describing ligand-solvent (water) and solvent-solvent interactions.

Construction of the FEFF 3D-QSAR models

The non-scaled FEFF energy terms, as defined in Table 3, were used as the trial pool of descriptors (independent variables) in the construction of the FEFF 3D-QSAR models. Model optimization was carried out using the *genetic function approximation* (GFA) [29, 30]. A combination of partial least squares regression, PLS [31], and multidimensional linear regression were used to fit enzyme inhibition measures to the descriptor values in the FEFF 3D-QSAR models evolved in a GFA optimization. The robustness of each model was tested by evaluating statistical measures

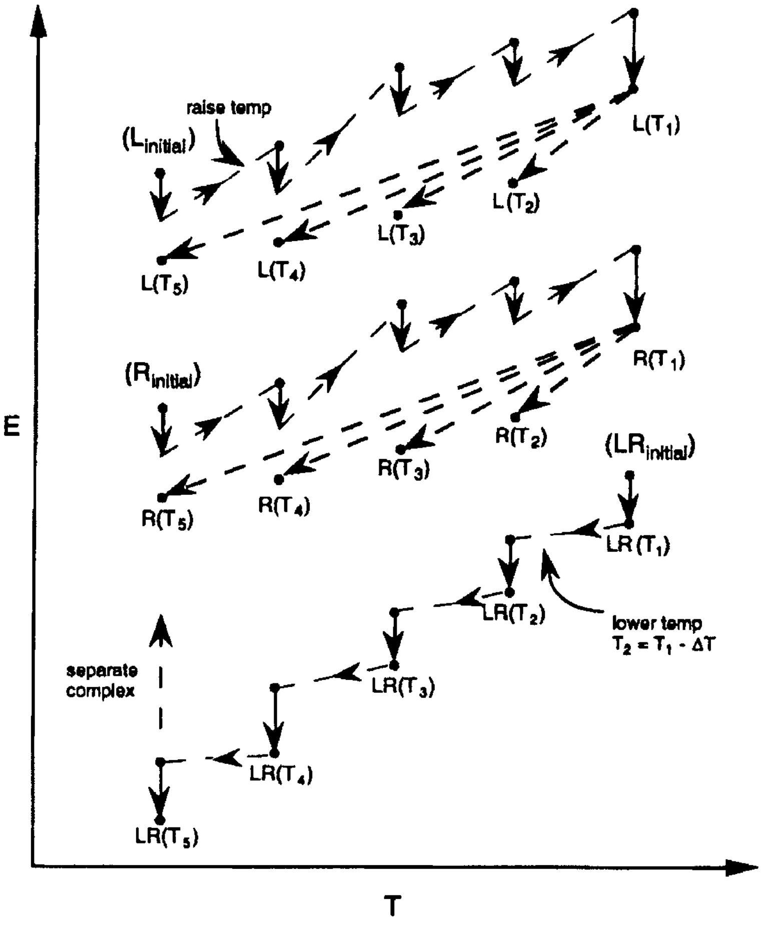


Figure 4. Representation of the MDS sampling scheme as function of temperature (T).

of fit which included the correlation coefficient (r^2), leave-one-out cross-validation correlation coefficient (xv$-r^2$), least-square error of fit (LSE), the single variable cross-correlation matrix and Friedman's lack of fit (LOF) [29, 32].

The GFA optimization scheme permits ranking of the models generated in the optimization process with respect to a measure of model significance. In this study the top five models with the highest r^2 values at each simulation temperature were retained for additional inspection. In order to determine if the top five FEFF 3D-QSAR models provide common, or distinct, structure-activity information, the cross-correlation coefficients of the residuals of fit (observed activity – predicted activity) between all pairs of the top five models were computed. Equivalent models are expected to have highly similar distributions in the residuals of error and, therefore, high cross- correlation coefficients. Distinct models should have poorly correlated patterns in their residuals of fit and low corresponding cross-correlation coefficients.. This type of analysis has been suggested by Rogers [33, 34] as a diagnostic to determine the subset of distinct models among a set of good models realized in a GFA analysis.

Results

FEFF 3D-QSAR models

The size of the pruned enzyme model used in the FEFF 3D-QSAR MDS was selected based on a combination of geometric stability of the ligand-enzyme model complex and the corresponding non-scaled ligand–pruned enzyme model binding energy. Spherical pruned enzyme models having of radii, R, of 8, 10 and 12 Å were explored. The root-mean-square, RMS, difference in non-proton atom positions for each

Table 3. Definition of the FEFF terms used in constructing the FEFF 3D-QSARs models

$\Delta E_{stretch}$	change in the stretching energy upon binding
ΔE_{bend}	change in bending energy upon binding
$\Delta E_{torsion}$	change in torsional energy upon binding
ΔE_{vdW}	change in van der Waals energy upon binding
$\Delta E_{electrostatic}$	change in electrostatic energy upon binding
$\Delta E_{E1,4}$	change in 1,4 interaction energy upon binding
ΔE_{hb}	change in hydrogen bonding energy upon binding
ΔE_{solv}	change in solvation energy upon binding
$\Delta E_{stre+bend}$	sum of changes in stretching and bending energies
$\Delta E_{stre+bend+tor}$	sum of changes in stretching, bending and torsion energies
ΔE_{el+hb}	sum of changes in electrostatic and hydrogen bonding energies
$\Delta E_{el+hb+E1,4}$	sum of changes in electrostatic, hydrogen bonding, and 1,4 interaction energies
E_{LR}(LL,RR,LR)	ligand-receptor complex energy
E_{LR}(LR)	intermolecular ligand-receptor energy
$E_{LR,vdW}$	van der Waals intermolecular ligand-receptor energy
$E_{LR,electrostatic}$	electrostatic intermolecular ligand-receptor energy
$E_{LR,hb}$	hydrogen bonding intermolecular ligand-receptor energy
$E_{LR,el+hb}$	sum of electrostatic and hydrogen bonding intermolecular ligand-receptor energies
$E_{LR,el+hb+vdW}$	sum of electrostatic, hydrogen bonding, and van der Waals intermolecular ligand-receptor energies
ΔE_L(LL)	change in intramolecular ligand energy upon binding
E_{LR}(LL)	intramolecular energy of bound ligand
E_L(LL)	intramolecular energy of unbound ligand
ΔE_R(RR)	change in intramolecular receptor energy upon binding
E_{LR}(RR)	intramolecular energy of bound receptor
E_R(RR)	intramolecular energy of unbound receptor
E_{LR}(LRM)	ligand-receptor complex solvation energy
ΔE_L(LM)	change in ligand solvation energy upon binding
E_{LR}(LM)	bound ligand solvation energy
E_L(LM)	unbound ligand solvation energy
ΔE_R(RM)	change in receptor solvation energy upon binding
E_{LR}(RM)	bound receptor solvation energy
E_R(RM)	unbound receptor solvation energy
Homo	highest occupied molecular orbital energy of the ligand
Lumo	lowest unoccupied molecular orbital energy of the ligand
Hbd	number of hydrogen bonding donor atoms of the ligand
Hba	number of hydrogen bonding acceptor atoms of the ligand
Log P	logarithm of the octanol/water partition coefficient for the ligand

of the three pruned models from those of the parent enzyme geometry were estimated from a 20 ps, 300 K MDS of each ligand-enzyme model complex. In order to promote pruned enzyme models to remain close to that of the parent enzyme structure, it was necessary to assign a fictitious mass of 5000 amu to all backbone atoms of the peptide residues composing a pruned enzyme model. The corresponding r.m.s. values for these heavy mass assignments for each of the pruned enzyme-ligand models are (R = 8, r.m.s. = 1.80 Å), (R = 10, r.m.s. = 1.60 Å) and (R = 12, r.m.s. = 1.94 Å). The corresponding non-scaled ligand-pruned enzyme model binding energies, E_{LR}(LR), are (R = 8, E_{LR}(LR) = −36.7 kcal/mole), (R = 10, E_{LRs}(LR) = −37.3 kcal/mole) and (R = 12, E_{LR}(LR) = −37.6 kcal/mole). Based on the values of RMS atomic positions and the apparent convergence of the ligand-binding energy, the pruned enzyme model of 10 Å was selected for all FEFF MDS as a compromise between structural and binding integrity and computational efficiency. All energy minimizations and MDS were done using the MOLSIM package [35].

The binding simulation sampling scheme [3], which is similar to a coupled MDS simulated annealing-energy minimization procedure, was initiated by a MDS of 20 ps, using a time step of 0.5 fs, on the ligand-enzyme complex model. An initial average temperature of 310 K [27] was held constant during the simulation by coupling the system to a temperature bath with a relaxation time [36] of 0.01 ps. The sampled geometry of the complex, and its associated energy terms, were retained every 0.2 ps over the MDS trajectory. The geometry of lowest total potential energy for the simulation at 310 K was next used as the starting state for a MDS at 200 K consisting of 10 ps of MDS using a time step of 0.5 fs. The sampled geometry of the complex, and its associated energy terms, were retained every 0.1 ps over the MDS trajectory. This scheme was repeated for structure-coupled 100, 50, 25 and 10 K MDS. The lowest energy structure, and corresponding energy terms, from each simulation temperature sampled were retained as the *representative* data for the complex at each temperature and used in the construction of a FEFF 3D-QSAR model for each simulation temperature considered in the MDS heating and cooling sampling scheme.

The sampling scheme for the unbound state of ligand-enzyme binding began with the bound conformations of the individual ligand and corresponding receptor geometry from the lowest energy complex state sampled during the 10 K MDS. The bound ligand was then separated from the enzyme. Individual MDS at 10 K were performed on the dissociated ligand and corresponding pruned enzyme model employing the same conditions as used for sampling the complex. Because of the potential added flexibility available to the unbound ligands [3], a 50 ps at 1 fs steps MDS was performed on each unbound ligand in order to fully explore its conformational space. However, for the unbound receptor the same conditions used for the ligand-enzyme complexes MDS, i.e., 10 ps MDS and time step of 0.5 fs., were employed. Both the unbound ligand and unbound enzyme were heated up in a series of structure-coupled MDS as described in the *Methods* section for 25, 50, 100, 200 and 310 K.

Finally, the heating/cooling sampling cycle presented in the *Methods* section was completed for each unbound ligand and unbound enzyme model by using the lowest energy structural states of the unbound ligand and unbound enzyme obtained at 310 K as the starting points for the subsequent ‘cooling down’ process. Cooling structure-coupled MDS for the unbound ligand and unbound enzyme models were then run at 200, 100, 50, 25, and 10 K. The MDS time for the enzyme models was 5 ps, whereas for the unbound ligands the MDS sampling time was again 50 ps.

In the same manner adopted in the ligand-enzyme complex temperature sampling, the lowest energy structures of each unbound ligand and of each unbound enzyme model, and the respective energy terms of each structure, were chosen as the *representative* conformations of the unbound ligand and enzyme model at a given simulation temperature. Thus, the energy contributions from both the intermolecular and intramolecular interaction contributions to ligand-receptor binding were computed and retained during the set of MDS at each simulation temperature. The pool of descriptor energy terms used for constructing a FEFF 3D-QSAR model, at a given temperature, are ensemble averaged energies derived from the states sampled in the corresponding set of MDS.

The trial FEFF 3D-QSAR models were constructed using the GFA as programmed in the WOLF software package [30]. A smoothing factor of 0.7 in the LOF statistical fitting measure was found to give, in general, the smallest FEFF 3D-QSAR models (number of independent variables) with a high measure of statistical fit. The FEFF terms used as the trial descriptor pool in the GFA optimization to generate a FEFF3D-QSAR model are given in Table 3. The terms that represent energy changes upon binding (see Table 3) were determined using Equation 10. The non-force field descriptors homo, lumo, hbd, hba and log P, all computed for the ligand, were determined using the *Cerius 2* software package [37].

The statistical measures, *and the number of descriptors employed*, for the top ten FEFF 3D-QSAR models for each sampling simulation temperature are shown in Table 4. Care must be taken in the interpretation and acceptance of these *preliminary* results. According to the data in Table 4, all of the simulation sampling temperatures provide good models ($\mathrm{xv}-\mathrm{r}^2 > 0.80$), but the number of descriptors used in the various models ranges from 3 to 8. The seemingly obvious route to take for the selection of the very best models from all those given in Table 4 would be to pick those models with the least number of descriptors [38] and the highest statistical fit measures. In particular, $\mathrm{xv}-\mathrm{r}^2$, is the most sensitive measure of statistical

Table 4. Statistical measures of the top ten FEFF 3D-QSAR models for each MDS temperature (training set analysis)

Temperature	Model	FEFF Terms	R^2	LSE	xv-r^2
310 K	1	4	0.93	0.44	0.84
	2	4	0.92	0.42	0.86
	3	4	0.92	0.45	0.84
	4	4	0.92	0.43	0.85
	5	4	0.92	0.43	0.85
	6	3	0.89	0.49	0.81
	7	4	0.92	0.43	0.85
	8	4	0.92	0.45	0.83
	9	4	0.91	0.45	0.83
	10	4	0.91	0.44	0.84
200 K	1	7	0.99	0.33	0.90
	2	8	0.99	0.32	0.91
	3	8	0.99	0.38	0.87
	4	8	0.99	0.35	0.89
	5	7	0.99	0.22	0.96
	6	8	0.99	0.40	0.86
	7	7	0.99	0.25	0.95
	8	7	0.99	0.27	0.94
	9	8	0.99	0.34	0.90
	10	8	0.99	0.34	0.90
100 K	1	5	0.95	0.76	0.48
	2	5	0.95	0.41	0.86
	3	5	0.95	0.44	0.83
	4	7	0.98	0.72	0.37
	5	6	0.97	0.38	0.88
	6	6	0.97	0.40	0.87
	7	6	0.97	0.40	0.86
	8	5	0.95	0.77	0.47
	9	6	0.96	0.75	0.49
	10	4	0.93	0.44	0.84
50 K	1	6	0.97	0.29	0.93
	2	4	0.93	0.39	0.88
	3	4	0.93	0.40	0.87
	4	6	0.97	0.31	0.92
	5	5	0.95	0.36	0.89
	6	4	0.93	0.39	0.87
	7	6	0.96	0.91	0.37
	8	5	0.95	0.34	0.90
	9	4	0.93	0.40	0.87
	10	5	0.95	0.38	0.88
25 K	1	5	0.94	0.39	0.88
	2	5	0.94	0.39	0.87
	3	4	0.92	0.44	0.84
	4	4	0.92	0.44	0.84
	5	4	0.91	0.44	0.84
	6	4	0.91	0.44	0.84
	7	4	0.91	0.45	0.83
	8	7	0.97	0.30	0.93
	9	3	0.88	0.48	0.81
	10	4	0.91	0.46	0.82

Table 4. Continued.

Temperature	Model	FEFF Terms	R^2	LSE	xv-r^2
10 K	1	8	1.00	0.14	0.98
	2	6	0.98	0.23	0.96
	3	8	0.99	0.18	0.97
	4	7	0.99	0.23	0.96
	5	7	0.99	0.21	0.96
	6	6	0.98	0.28	0.93
	7	8	0.99	0.17	0.98
	8	7	0.99	0.22	0.96
	9	7	0.99	0.27	0.93
	10	7	0.99	0.26	0.94

fit for this data set and has been used as the principal measure of model selection. Under these constraints, the best FEFF 3D-QSAR model at each temperature using the model numbers of Table 4, and the number of descriptors in the model, are: (310 K, Model 2, 4 descriptors), (200 K, Model 5, 7 descriptors), (100 K, Model 10, 4 descriptors), (50 K, Model 2, 4 descriptors), (25 K, Model 9, 3 descriptors) and (10 K, Model 6, 6 descriptors). The predicted versus observed –log(IC_{50}) values for each of these best FEFF 3D-QSAR models are shown in Figure 5. From the plots in Figure 5 it would appear that each best model for each simulation temperature fits the observed $-\log(IC_{50})$ values quite well except for the model at 310 K. The standard error for 310 K model is given in Table 4 and it is larger than the standard errors, again see Table 4, for the other best models. Thus, the FEFF 3D-QSAR model derived for a simulation temperature for what is meant to be the 'actual' temperature of the inhibitor-enzyme system, 310 K, is the poorest model to fit the observed activities of the training set.

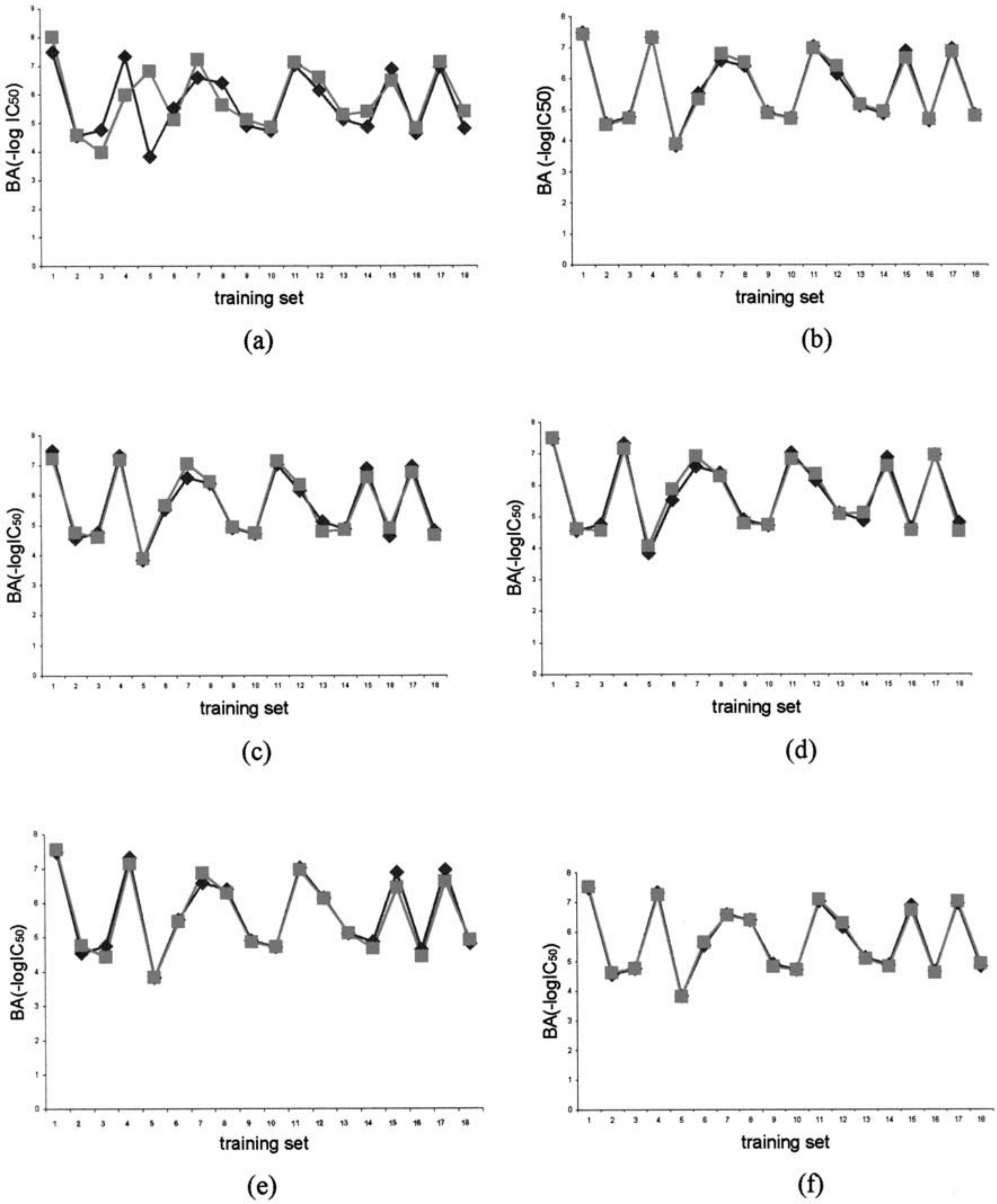
In order to further search for the 'optimum' FEFF 3D-QSAR model over the temperatures sampled, a test set of inhibitors was extracted from the original training set. This test set was generated by arbitrarily selecting five compounds from the original training set (Table 2) subject to the constraint that these compounds largely span the inhibitory potency ($-\log IC_{50}$) range of the original training set. These compounds are P-1, P-4, P7, P10, and MTX. The remainder of the compounds in the original training set were used to construct new FEFF 3D-QSAR models using only those descriptors found among the top ten models at each temperature for the original training set (Table 4) in the GFA descriptor pool. Further, all new FEFF 3D-QSAR models were further constrained to contain three descriptors. These new three-term FEFF 3D-QSAR models are, overall, subset models to those described in Table 4.

Figure 6 show the predicted versus observed $-\log(IC)$ values for the five compounds in the test set using the best three-term FEFF 3D-QSAR models at each simulation temperature. It is reasonably clear from Figure 6 that the best predictions are for the 50 K model followed by the 100 K model. The best three-term model at 50 K is a sub-model of the best model at 50 K for the original training set (Model 2 of Table 4). By sub-model it is meant that each of the descriptor terms of the three-term model are found in Model 2 of Table 4 and the corresponding regression coefficients are similar. This analysis of the best FEFF 3D-QSAR models generated at different simulation temperatures for different representations of the training set data has led to the selection of the four-term model at 50 K as being the best FEFF 3D-QSAR model identified in this study. The complete form of this FEFF 3D-QSAR model is,

$$\begin{aligned} -\log(IC_{50}) = {} & 0.012\Delta E_{el+hb+E1,4} - 0.094\text{lumo} + \\ & 0.048E_L(LM) - 0.014E_{LR,el+hb} - \\ & 9.88 \end{aligned} \tag{12}$$

$$N = 18 \quad r^2 = 0.93 \quad xv - r^2 = 0.88 \quad LSE = 0.39.$$

The linear cross-correlation matrix of the descriptors in Equation 12 are reported in Table 5. $\Delta E_{el+hb+E1,4}$ and $E_{LR,el+hb}$ are the only terms in Equation 12 highly cross-correlated to one another (0.82). These two descriptors taken together describe



(a)

(b)

(c)

(d)

(e)

(f)

Figure 5. Predictions of the best FEFF 3D-QSAR model of the training set at each MDS sampling temperature: (a) 310 K; (b) 200 K; (c) 100 K; (d) 50 K; (e) 25 K; (f) 10 K. The squares indicate observed - $\log(IC_{50})$ values and the diamonds are the corresponding predicted values. The abscissa lists the inhibitors by the numbering scheme – 1, 2, 3, 4, 5, 6, . . . , 18 are compounds AMP, CYC, PYR, MTX, TMP, P-1, . . . , and P-13, respectively.

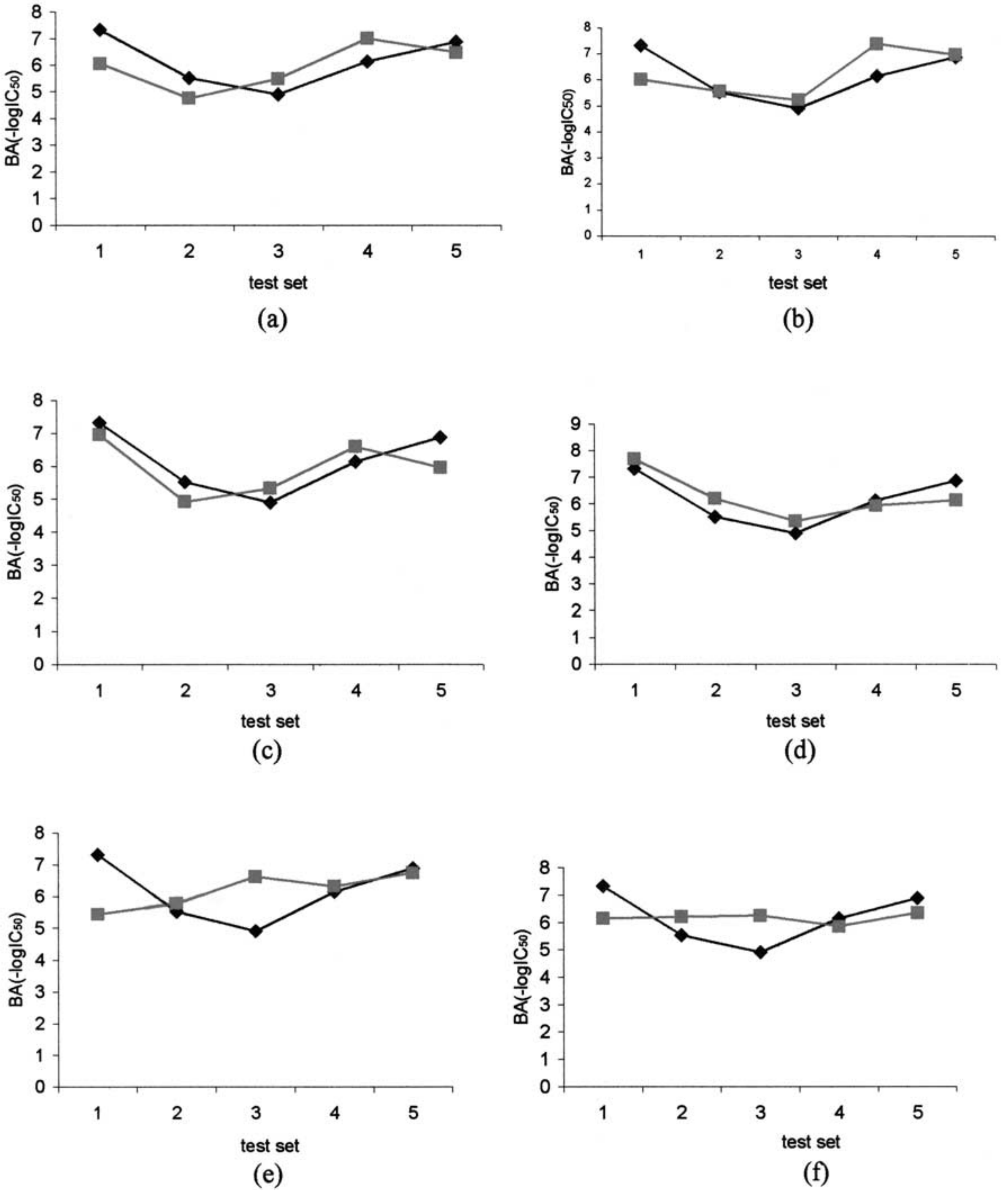


Figure 6. ‘Predictivity’ of the test set using the best FEFF 3D-QSAR model at each simulation temperature: (a) 310 K; (b) 200 K; (c) 100 K; (d) 50 K; (e) 25 K; (f) 10 K. The same plotting scheme has been used here as in Figure 5.

the energetic behavior of the ligand and receptor for the binding process. Thus, it is not surprising that these two terms are highly correlated, but neither would be statistically significant by itself. None of the other possible cross-correlations of the FEFF terms in Equation 12 are seen to be significant suggesting each corresponding term is contributing unique information to explaining the inhibition data.

The relative significance of each of the individuals FEFF descriptors most often employed in the GFA optimization process leading to Equation 12 are shown by the crossovers versus descriptor usage plot in Figure 7. This plot also demonstrates that the GFA model evolution has converged. Convergence is achieved when descriptor usage does not change as a function of an increasing number of crossovers (a horizontal line is observed). From an inspection of this plot, the most often used FEFF descriptors in the generation of Equation 12 are hbd and $E_{LR,el+hb}$. There is also considerable usage of the $\Delta E_{LR,el+hb+E1,4}$ and homo descriptors. $\Delta E_{LR,el+hb}$, $E_L(LM)$, hba, log P, lumo, $E_{LR}(LM)$, and $E_{LR,,el+hb+vdW}$ descriptors are moderately selected and $\Delta E_R(RR)$ is not often used. All of the most frequently used descriptors in a GFA model optimization may not show up in the final best model because;

1. Two, or more, frequently used descriptors are equivalent and explain the same variance in the data. They are competing for survival and one ends up dominating and is found in the final model.
2. A *combination* of one frequently used descriptor and one moderately used descriptor may explain more variance than the *combination* of two frequently used descriptors owing to fitting a complex surface described by the multi-dimensional regression analysis.

While it is the magnitude of the lumo that appears as a descriptor term in Equation 12, the spatial properties of the lumo are also intrinsically present in this FEFF 3D-QSAR model. The most likely reason lumo is found in Equation 12 is because its relative binding *location* on each inhibitor is nearly the same across the training set. Thus, the interaction of the lumo with the active site is *geometrically* similar for the set of inhibitors. Of course, if the structure of an inhibitor corresponds to a change in the location of the lumo relative to the binding alignment used to develop Equation 12, the inhibitor is likely to be an outlier to the FEFF 3D-QSAR model.

Figure 8 shows a 10A *Pf*DHFR active site model with PYR bound. The geometry shown in Figure 8 corresponds to one of the training set ligand-receptor complexes used to construct Equation 12.

Comparison of FEFF-3D-QSAR and RI-4D-QSAR models

Recently, we reported receptor-independent (RI) 4D-QSAR models derived from the same training set of ligands binding to the same mutant form of DHFR [27] as used in this study. The alignment with the imino nitrogen of the six-membered ring, along with the two nitrogen centers of the amino groups of the side chains, as defined in Figure 1 using the PYR structure, yielded the best four-term mutant RI-4D-QSAR model [27] which is given by

$$\begin{aligned}-\log(\mathrm{IC}_{50}) = {} & -10.07\mathrm{GC1(any)} + 24.8\mathrm{GC2(p+)} + \\ & 14.91\mathrm{GC3(p-)} - 5.27\mathrm{GC4(np)} + \\ & 5.29 \qquad (13)\end{aligned}$$

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

Each of the four grid cell occupancy descriptors, GCX(A)s, in Equation 13 can be associated with three of the four descriptors, the non-lumo descriptors, in Equation 12. The A in GCX(A) are as follows; ‘any’ refers to any type of atom occupying the grid cell, ‘p+’ indicates a polar atom with a positive residual charge, ‘p-’ is the same as p+, but for negative partial charge density and ‘np’ refers to a nonpolar atom type. GC1 likely reflects a steric interaction between the enzyme and the inhibitor, and GC4 suggests this region of the enzyme active site has steric and/or polar repulsive interactions with the corresponding groups of the inhibitor. GC2 and GC3 likely reflect favorable enzyme-inhibitor and/or inhibitor-solvent interactions.

Since the development of Equation 13 a new approach to establishing the optimum RI-4D-QSAR alignment, based on the intermolecular accessibility concepts used in the bimolecular model [39], has been used to derive mutant RI-4D-QSAR models. On the basis of the general three ordered–atom match alignment rule currently implemented in 4D-QSAR analysis, the inhibitor atoms with maximum degrees (δ_{max}) are selected to align each molecule. The δmax of a center (an atom) is the number of valence electrons of that center in a hydrogen suppressed molecular graph of the whole molecule. In practice, two adjacent atoms with maximum degrees are chosen, and the third one is selected by constraints to have the maximum degree, and, at the same time, to be bonded to another atom

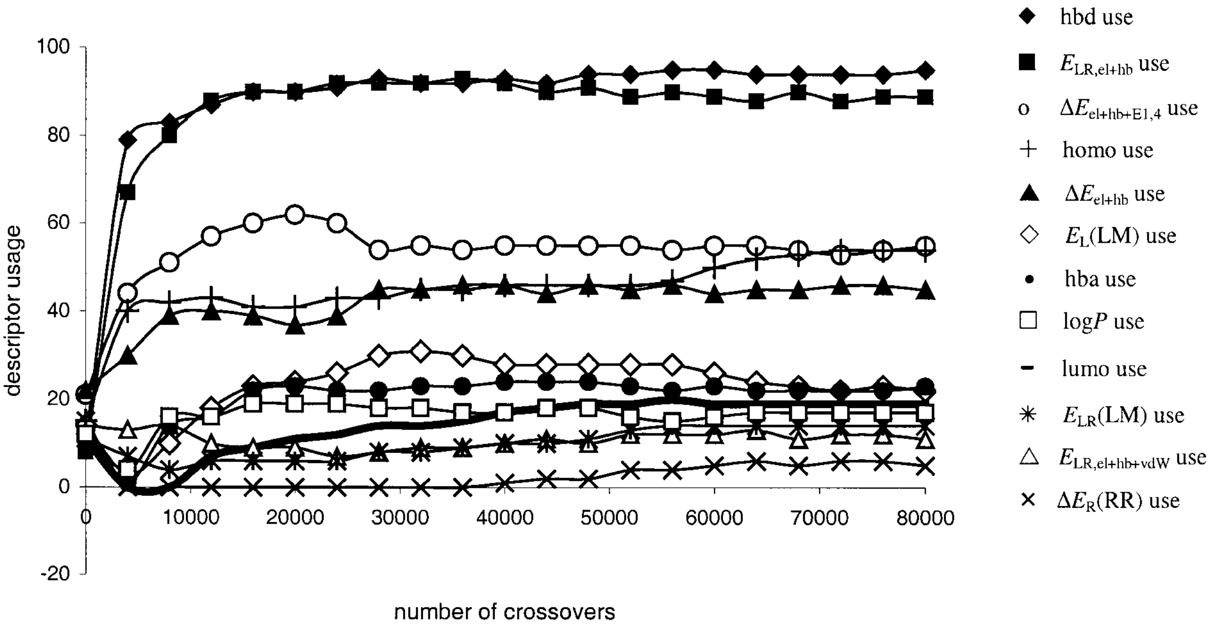


Figure 7. Descriptor usage as a function of the number of crossovers in the GFA FEFF-3D-QSAR model optimization.

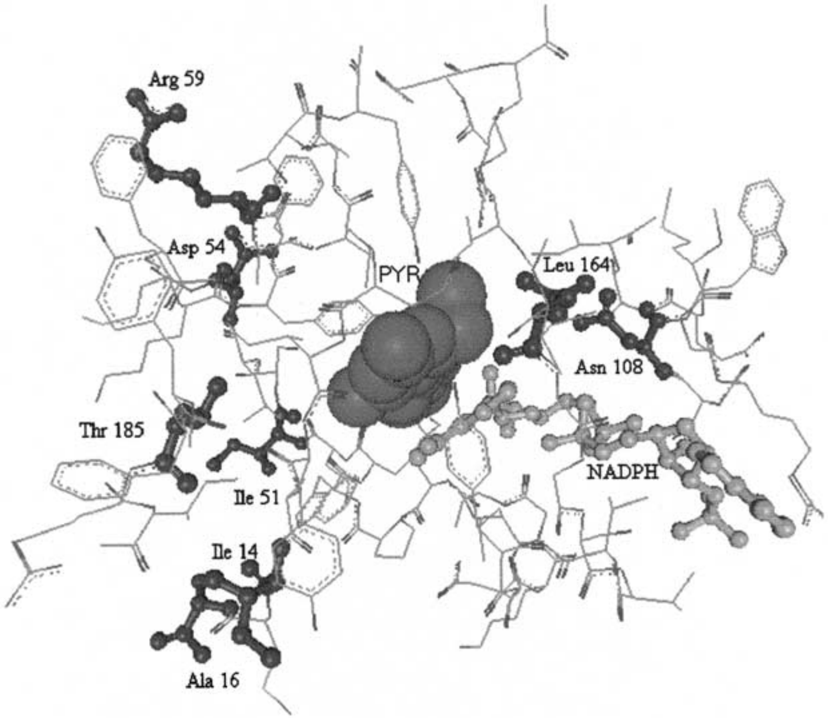


Figure 8. The 10A pruned binding site model of *Pf*DHFR with key residues of the binding site indicated by ball-stick models in dark shading, NADPH is in light shaded ball-stick representation and the PYR inhibitor is shown in a space-filling shaded representation.

Table 5. Linear cross-correlation matrix of the descriptors in the best FEFF 3D-QSAR model - Equation 12 determined as a function of the descriptor pool and simulation temperatures sampled

	$\Delta E_{\mathrm{el+hb+E1,4}}$	lumo	$E_L(\mathrm{LM})$	$E_{\mathrm{LR,el+hb}}$
$\Delta E_{\mathrm{el+hb+El,4}}$	1.00			
lumo	−0.06	1.00		
$E_L(\mathrm{LM})$	0.15	0.18	1.00	
$E_{\mathrm{LR,el+hb}}$	0.82	−0.01	0.18	1.00

with a high degree. Using this set of three-ordered atom trial alignments, two highly significant and independent four-term RI-4D-QSAR models have been constructed for the mutant structure-activity data of Table 2:

$$-\log(\mathrm{IC}_{50}) = 10.36\mathrm{GC1(np)} + 18.01\mathrm{GC2(np)} + 24.49\mathrm{GC3(any)} + 16.00\mathrm{GC4(Aro)} + 4.29 \qquad (14)$$

$$N = 18 \quad r^2 = 0.94 \quad xv - r^2 = 0.91 \quad \mathrm{LSE} = 0.08.$$

$$-\log(\mathrm{IC}_{50}) = 18.58\mathrm{GC1(np)} + 10.76\mathrm{GC2(p-)} + 26.93\mathrm{GC3(any)} - 1.64\mathrm{GC4(any)} + 4.75 \qquad (15)$$

$$N = 18 \quad r^2 = 0.95 \quad xv - r^2 = 0.92 \quad \mathrm{LSE} = 0.06.$$

The xv$-r^2$ values in Equations 14 and 15 are somewhat better than the values obtained by FEFF 3D-QSAR analysis Equation 12 and about the same as Equation 13. In Equation 14 all the GCXs have positive coefficients while in Equation 15 GC4 has a negative regression coefficient. Occupancy of GCXs by the correct atom-types enhances binding for positive value regression coefficients, and diminishes inhibition for negative coefficients. GCXs with nonpolar atom types are present in both Equations 14 and 15. GCXs 1 and 3 of the two RI-4D-QSAR models are the same. However, the other two GCXs of the two models are different from one another. Hence, the intrinsic characteristics of the two RI-4D-QSAR models are somewhat different. The predicted active conformations for two inhibitors using both Equations 14 and 15 are shown in Figures 9 and 10, respectively. The two representative inhibitors are aminopterin (AMP), which is large in size and with multiple functional groups, and has the maximum inhibition potency in the training set, and P-10 which possesses a relatively common core structure across the training set.

GC1 and GC2 of Equation 14 have nonpolar atom types in close proximity to the two carbon centers having maximum δ values which are used in the alignment. These two GCXs may account for van der Waals interactions with nonpolar residues Ile-51 and Leu-164 of the mutant form of *Pf*DHFR. GC3 is close to the bicyclic group of the two ligands. Interactions between Ile-14, Ala-16 and Ile-164 with the GC3 atoms may stabilize the formation of the inhibitor-enzyme complex. GC4 has an aromatic atom type and is in the close proximity of the aromatic ring of bicyclic ring of the inhibitor. The aromatic character of the ligand may also stabilize the inhibitor-enzyme complex by forming a corresponding π-complex.

An inspection of Equation 15 and Figure 10 provides insight to the nature of the second RI-4D-QSAR model. In this model GC4 with an 'any' atom type has a negative regression coefficient whereas the other three GCODs have positive regression coefficients. A polar negative atom type for GC2 suggests a stablizing electrostatic interaction with the appropriate residues is possible. GC2 is near the oxygen atom of the carbonyl group for almost all the ligands except PYR and CYC. For PYR and CYC GC2 is near the Cl group. GC1 has the nonpolar atom type and is close to a highly branched carbon center without any adjacent heterocenter. GC3 is in the proximity of the alignment carbon having the maximum degree, but with minimum nonpolar character. In fact, GC3 is near the carbon center of the carboxylic acid functional group. Hence, the atom type may not be too nonpolar in nature.

The descriptors present in the FEFF-3D-QSAR model, Equation 12, and the GCXs of the 4D-QSAR models, Equations 13–15 have been compared. The FEFF 3D-QSAR descriptors in Equation 12 can be partitioned into two classes characteristic of the molecular energetics The first set of descriptors, $\Delta E_{\mathrm{el+hb+E1,4}}$ and $E_{\mathrm{LR,el+hb}}$, contain electrostatic, hy-

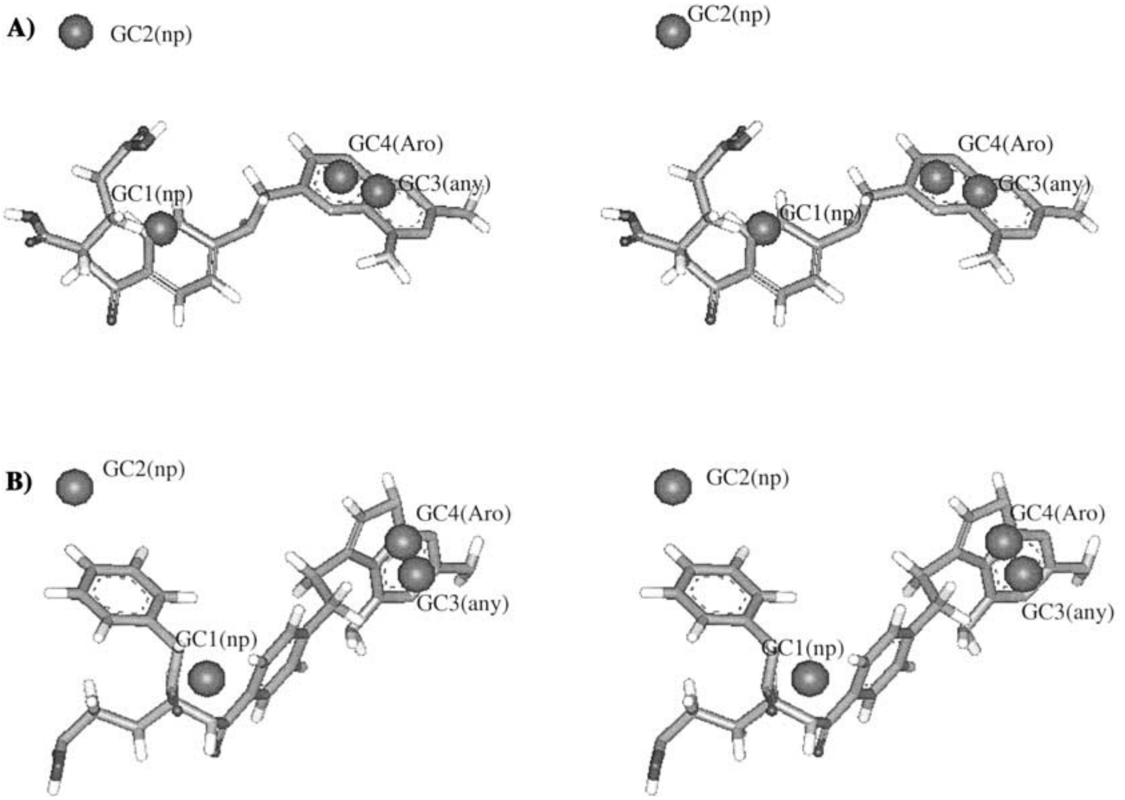


Figure 9. Graphical representations, in stereo format, of (A) AMP and (B) P-10 in their respective predicted active conformations using Equation 14 and the GCXs of Equation 14. Inhibition activity enhancing grid cells are shown as dark spheres, and grid cells which diminish inhibition potency are shown in lighter spheres.

drogen bonding and 1,4 van der Waals energies of the ligand and the receptor responsible for the formation of the inhibitor complex. Correspondingly in Equation 14 GC1 and GC2, each with ‘np’ atom types, represent van der Waal interaction energies while GC3 with ‘any’ is not inconsistent with electrostatic and hydrogen bonding type interactions with the enzyme. The other two descriptors in Equation 12, lumo and E_L(LM), reflect the electron affinity of the ligand and the aqueous solvation energy of the unbound ligand, respectively. In Equation 14, GC4 with an aromatic atom type corresponds to electron-electron interactions between the ligand and aromatic residues and/or polar residues of the enzyme. In the RI-4D-QSAR model given by Equation 15, GC2 has a polar negative atom type which likely incorporates electrostatic interactions between the ligand and the receptor.

Three GCX of Equation 13 parallel three GCX of Equation 15. GC3(p-) of Equation 13 is close to GC2(p-) of Equation 15. Likewise, GC4(np), Equation 13 parallels GC4(any), Equation 15 and GC2(p+), Equation 15 seems consistent with GC3(any) of Equation 13. Only GC1(any), Equation 13 and GC1(np) of Equation 15 appear to be different.

It should not be too surprising that there are multiple good and independent QSAR models to fit the structure – activity data set given in Table 2. A large number of trial descriptors are available to fit a relatively few (18) compounds data set. Hence, there are multiple models consistent with the structure-activity data. The observed, predicted and residual IC_{50} values of the training set inhibitors, based upon each of the QSAR models, Equation 12–15 are given in Table 6. There are no outliers, and the distribution of residuals across the four QSAR models exhibit no obvious patterns. Thus, all four QSAR models are explaining the training set structure-activity data about equally well, but in somewhat different ways. Additional structure-activity data is needed to resolve which QSAR model is most representative of describing enzyme inhibition.

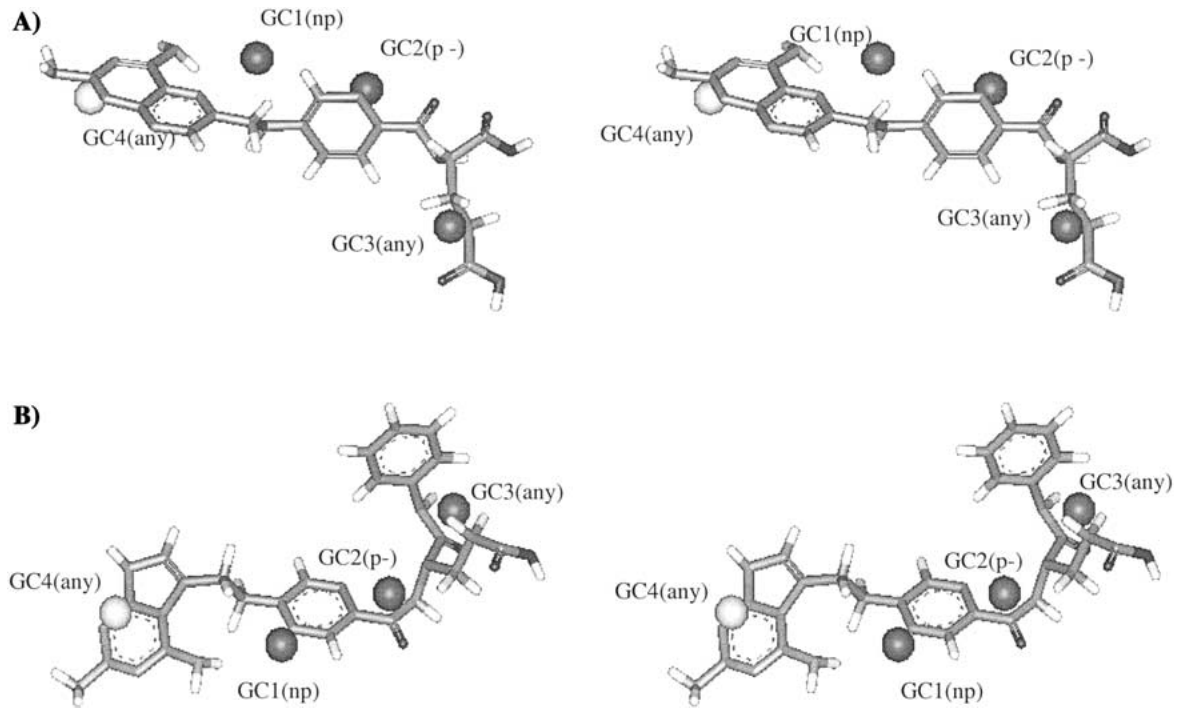


Figure 10. Same as Figure 9 but for Equation 15.

Table 6. The observed, predicted and residual fit $-\log(\mathrm{IC}_{50})$ values of the inhibitors of the training set using the QSAR models given by Equations 12 through 15

Compd.	Obs.	Pred. 12	Res. 12	Pred. 13	Res. 13	Pred. 14	Res. 14	Pred. 15	Res. 15
P-1	5.51	5.91	−0.40	5.27	0.24	5.68	−0.17	5.76	−0.25
P-2	6.57	6.95	−0.38	6.50	0.07	6.19	0.38	6.75	−0.18
P-3	6.38	6.17	0.21	6.71	−0.33	6.13	0.25	6.05	0.33
P-4	4.89	4.63	0.26	4.55	0.34	4.37	0.52	4.75	0.14
P-5	4.70	4.73	−0.03	4.48	0.22	4.74	−0.04	4.70	0.00
P-6	7.01	6.80	0.21	6.69	0.32	7.00	0.01	6.98	0.03
P-7	6.13	6.49	−0.36	6.60	−0.47	6.13	0.00	6.56	−0.43
P-8	5.10	5.55	−0.45	5.07	0.03	5.26	−0.16	5.18	−0.08
P-9	4.85	5.23	−0.38	4.68	0.17	4.78	0.07	5.05	−0.02
P-10	6.87	6.50	0.37	6.53	0.34	6.87	0.00	6.34	0.55
P-11	4.62	4.39	0.23	5.01	−0.39	5.05	0.43	4.83	−0.21
P-12	6.94	7.05	−0.11	6.97	−0.03	6.92	0.02	6.46	0.48
P-13	4.80	4.47	0.33	5.03	−0.23	4.45	0.35	4.67	0.13
PYR	4.54	4.31	0.23	4.77	−0.23	4.29	0.25	4.63	−0.09
CYC	4.74	4.76	−0.02	4.38	0.36	5.27	−0.53	4.75	−0.01
MTX	7.31	7.08	0.23	7.72	−0.42	7.33	0.02	7.41	−0.10
AMP	7.47	7.55	−0.08	7.16	0.31	7.49	−0.02	7.59	−0.12
TMP	<3.82	4.13	−0.31	3.55	>0.27	4.29	<0.47	3.77	>0.05

Finally, the predicted active conformations from the RI-4D-QSAR models are among the highly sampled conformations found in the MDS used in the construction of the FEFF 3D-QSAR model given by Equation 12. This observation suggests a high level of consistency between the 4D-QSAR and FEFF 3D-QSAR models.

Discussion

It is important to identify all the approximations inherent to performing FEFF 3D-QSAR analysis, their impact on the quality of the modeling process and what is done to minimize their possible negative impact. Table 7 reports these approximations, their impact and possible remedies, and is based on an analysis of the FEFF 3D-QSAR paradigm [3]. The study reported in this paper has emphasized the exploration of the role of MDS temperature on FEFF 3D-QSAR model significance and quality.

Figure 8 shows the 10A pruned enzyme active site model with the PYR inhibitor bound in the lowest energy MDS state for 50 K. The location of the NADPH cofactor in the homology and pruned enzyme model is very close to the position and alignment seen for the cofactor in many DHFR crystal structures. Thus, the approximations made in the pruning process as well as those used in the homology modeling seem to be consistent with known structural data.

Equation 12 is the best FEFF 3D-QSAR model. However, it is also noteworthy to point out the common qualitative features of the FEFF descriptors found in the better models at all simulation temperatures. Specific change in energy terms upon ligand-recpetor binding ($\Delta E_{el+hb+E1,4}$, ΔE_{el+hb}, and $\Delta E_{ER}(RR)$), ligand-specific descriptors (hbd, hba, homo, lumo, and log P), ligand-receptor complex descriptors ($E_{LR,el+hb}$, $E_{LR}(LM)$, and $E_{LR,el+hb+vdW}$), and a ligand solvation term ($E_L(LM)$) dominate the best FEFF 3D-QSAR models at all MDS temperatures sampled. These terms can be interpreted in terms of specific ligand-receptor binding, and thereby confirm the 'mechanistic nature' associated with the FEFF 3D-QSAR methodology. An important component of this 'mechanistic nature' of FEFF 3D-QSAR analysis is due to incorporating the dynamic features of the chemical system. Instead of a static picture of the ligand-receptor binding process (the 'lock and key' model), the molecular flexibility (the 'hand in the glove' view) is extensively embedded in performing a FEFF 3D-QSAR analysis.

The descriptor terms in Equation 12 can be interpreted as follows;

1. $\Delta E_{el+hb+E1,4}$ and $E_{LR,el+hb}$ must be considered together because they 'work' against one another in the binding process. The first term is the total change in the electrostatic, hydrogen bonding and the 1,4 van der Waals energies of the ligand and the receptor upon ligand-receptor binding. This term is a measure of how much electrostratic, hydrogen bonding and 1,4 van der Waals energy the isolated ligand and the isolated receptor sacrifice to realize the bound ligand-receptor state. The more positive this term, that is the more energy sacrificed, the lower the value of $-\log(IC_{50})$. This loss in energy must be compensated for in some way. $E_{LR,el+hb}$, is the electrostatic and hydrogen bonding ligand-receptor binding energy, and is the source of energy compensation to $\Delta E_{el+hb+E1,4}$. The ligand and the receptor, in composite, give up energy in exchange for a greater stablizing ligand-receptor binding energy.
2. Lumo and $E_L(LM)$ are taken together because as a pair they correlate significantly higher to $-\log(IC_{50})$ than each by themselves. $E_L(LM)$ is the aqueous solvation energy of the unbound ligand and becomes more negative as the compound becomes more soluble. Thus, the negative regression coefficient in Equation 12 for $E_L(LM)$ suggests that the inhibition potency should increase as an inhibitor becomes more aqueous soluble. In general, the aqueous solvation energy of an inhibitor should possess a parabolic relationship to enzyme inhibition potency. If the inhibitor does not have a favorable aqueous solvation energy it will be insoluble, and there will be no available concentration of inhibitor to bind to the enzyme. On the other hand, if the inhibitor is too soluble in the aqueous phase, then it will not partition toward, and bind to, the enzyme. Thus, there is an intermediate, but optimum, aqueous solvation energy with respect to maximizing inhibition potency. Equation 12 suggests the inhibitors of the training set fall on the 'left' side of the parabola where increasing aqueous solubility increases inhibition potency. However, this behavior is tempered by lumo. If increasing aqueous solubility corresponds to also increasing lumo, there is a negative impact from lumo on inhibition potency.

Table 7. The FEFF 3D-QSAR molecular modeling approximations, their impact on modeling and the approaches used in this study to minimize approximation impact

Approximations	Impact on modeling	Approaches to minimize impact
1. FEFF representation		
(a) Solvation energies estimated using a hydration shell models.	Incorrect balance between solvation energy and the rest of the FEFF during MDSs.	Consider only the solvation energies for low-energy states, or the conformer state used to construct the QSAR.
(b) Explicit water molecules are not included.	Hydrogen bonding of ligand through water molecules (water bridges) to active site residues.	Examine outliers within context of missing explicit water interactions.
(c) Entropic contributions are considered to be constant for the inhibitor analog series.	Neglect of conformational flexibility on binding.	Entropy contributions can be estimated, if necessary, by a group additive model and scaled with respect to temperature in the FEFF 3D-QSAR fitting procedure.
2. The LR, L and R are modeled as being neutral.	Multiple protonation states are possible and could influence electrostatic energetics.	The protonation state held constant for the entire training set, so error should be ‘constant’ over the training set. A neutral state approximates solvation and counterion effects on FEFF interactions.
3. Scaled down receptor model.	The scaled down receptor geometry can deviate from the crystal geometries over a long MDS and some RR and LR interactions are eliminated.	Heavy masses assigned to each of the atoms of the scaled down model to model missing momentum reservoir of the rest of the enzyme.
4. MDS temperature	Balance the enthalpy and entropic contributions to ΔG.	The preferred MDS temperature corresponds to the best FEFF 3D-QSAR model of a subset of the training set.
5. Sampling schemes used to explore the geometry-energy states of the LR, R, and L,	The sampling schemes may be incomplete with respect to sampling bound and unbound ligand conformations, and monitoring the change in geometry of the receptor for the bound and unbound states.	Use experimental data for bound ligand alignment and ligand-receptor geometry for defining the bound and unbound ligand reference states.

Overall, the interpretation of Equation 12, as expressed above, leads to an enzyme-inhibition model which fits the classic view of such a process. Inhibition potency is directly proportional to the concentration of the inhibitor [lumo and $E_L(LM)$], and to the strength of enzyme-inhibitor binding [$\Delta E_{el+hb+E1,4}$ and $E_{LR,el+hb}$]. Moreover, Equation 12 is not inconsistent with any of the three best RI-4D-QSAR models given by Equations 13–15 suggesting a consensus model/interpretation of enzyme inhibition has been determined. The RI-4D-QSAR models are different from one another in terms of specific binding interactions, but they do share common features. The variability among the QSAR models is very likely a result of the limited size of the training set.

Finally, perhaps the most novel molecular modeling methodology component to the work reported here has been the exploration of the form and quality of FEFF 3D-QSAR models as a function of simulation temperature. A surprising finding is that the best FEFF 3D-QSAR models are found at simulation temperatures less than 300 K which is the approximate temperature at which the binding experiments are carried out. In fact, the FEFF 3D-QSAR models at 300K are among the worst models in terms of statistical significance! The best model is found for a simulation temperature of 50 K and is arrived at by employing a specific evaluation methodology developed as part of this study.

Three plausible reasons can be put forward as to why the form and quality of the FEFF 3D-QSAR models are sensitive to simulation temperature;

1. The various terms in the MDS force field are not correctly weighted relative to one another. The simulation temperature serves as a means of calibrating the force field terms to one another and to inhibition potency measures.
2. Damping of vibrational motions, and introducing changes in the corresponding energies and effective molecular sizes, of the ligand and receptor by solvent molecules is underrepresented by the MDS model system. Lowering of simulation temperature compensates by reducing available kinetic energy.
3. The higher temperature MDS produce high levels of distortion into the molecular geometries of the ligand, receptor and their complex. MDS at lower simulation temperatures sample more realistic structures and energies of the molecular system.

In the next paper in this series we will compare the FEFF 3D-QSAR models reported in this work with those developed for 'wild-type' inhibition by the same set of inhibitors. We will also develop a library of inhibitors that explore hypotheses gleaned from the FEFF 3D-QSAR models regarding the sources of the resistance mechanism of *P. falciparum* DHFR.

Acknowledgements

Osvaldo A. Santos-Filho 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. Dr. 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. Tokarski, J.S. and Hopfinger, A.J., J. Chem. Inf. Comput. Sci., 37 (1997) 779.
3. Tokarski, J.S. and Hopfinger, A.J., J. Chem. Inf. Comput. Sci., 37 (1997) 792.
4. Hopfinger, A.J., In Conformational Properties of Macromolecules, Academic Press, New York, NY, 1973, p. 71.
5. Koehler, M.G. and Hopfinger, A.J., Polymer, 30 (1989) 116.
6. Weiner, S.J., Kollman, P.A. and Nguyen, D.T.A., J. Comput. Chem., 7 (1986) 230.
7. Hopfinger, A.J. and Pearlstein, R.A., J. Comput. Chem., 5 (1984) 486.
8. Allinger, N.L., J. Am. Chem. Soc., 99 (1977) 8127.
9. Brobey, R.K.B., Iwakura, M., Itoh, F., Aso, K. and Horii, T., Parasitol. Int., 47 (1998) 69.
10. 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.
11. 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.
12. Miller, G.P. and Benkovic, S.J., J. Chem. Biol., 5 (1998) R105.
13. Bzik, D.J., Li, W.-B., Horii, T. and Inserburg, J., Proc. Natl. Acad. Sci. USA., 84 (1987) 8360.
14. Burkhard, R. and Sander, C., Annu. Rev. Biophys. Biomol. Struct., 25 (1996) 113.
15. Šali, A., Curr. Opin. Biotechnol., 6 (1995) 437.
16. Johnson, M.S., Srinivasan, N., Sowdhamini, R. and Blundell, T.L., Crit. Rev. Biochem. Mol. Biol., 29 (1994) 1.
17. Davies II, J.F., Delcamp, T.J., Prendergast, N.J., Ashford, V.A., Freisheim, J.H. and Kraut, J., Biochemistry, 29 (1990) 9467.
18. Matthews, D.A., Bolin, J.T., Burridge, J.M., Filman, D.J., Volz, K.W. and Kraut, J., J. Biol. Chem., 260 (1985) 392.
19. Volz, K.W., Matthews, D.A., Alden, R.A., Freer, S.T., Hansch, C., Kaufman, B.T., Kraut, J., J. Biol. Chem., 257 (1982) 2528.

20. Bystroff, C., Oatley, S.J. and Kraut, J., Biochemistry, 29 (1990) 3263.
21. Sawaya, M.R. and Kraut, J., Biochemistry, 36 (1997) 586.
22. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Schimanouchi, T. and Tasumi, M., J. Mol. Biol., 112 (1977) 535.
23. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E., Nucl. Acids Res., 28 (2000) 235. (http://www.rcsb.org/pdb/)
24. Laskowski, R.A., McArthur, M.W., Moss, D.S. and Thornton, J.M., J. Appl. Cryst., 26 (1993) 283.
25. Ramachandran, G.N. and Sassiekharan, V., Adv. Prot. Chem., 28 (1968) 283.
26. HyperChem Program Release 5.01 for Windows, Hypercube, Inc.; 1996.
27. Santos-Filho, O.A. and Hopfinger, A.J., J. Comput. Aid. Mol. Des., 15 (2001) 1.
28. Sano, G.-I., Morimatsu, K. and Horii, T., Mol. Biochem. Parasitol., 63 (1994) 265.
29. Rogers, D. and Hopfinger, A.J., J. Chem. Inf. Comput. Sci., 34 (1994) 854.
30. Rogers, D. WOLF Reference Manual Version 5.5, Molecular Simulation Inc., 1994.
31. Glen, W.G., Dunn, W.J., III and Scott, D.R., Tetrahedron Comput. Methods, 2 (1989) 349.
32. Rogers, D., In The Proceedings of the Fourth International Conference on Genetic Algorithms, San Diego, 1991, p. 38.
33. Rogers, D., personal communication, 1996.
34. Rogers, D., In Proceedings of the Seventh International Conference on Genetic Algorithms, East Lansing, MI, Morgan-Kaufmann, San Francisco, CA, 1997.
35. Doherty, D.C.; MOLSIM – User Guide; The Chem21 Group; 1780 Wilson Dr., Lake Forest, IL 60045, 1994.
36. Berendsen, H.J.C., Postman, J.P.M., van Gunsteren, W.F.; Di Nola, A. and Haak, J.R., J. Chem. Phys., 81 (1984) 3684.
37. Cerius 2 version. 3.0. Molecular Simulations Inc., 9685 Scranton Road, San Diego, CA 92121-3752, USA.
38. Albuquerque, M.G., Hopfinger, A.J. Barreiro, E.J. and de Alencastro, R.B., J. Chem. Inf. Comput. Sci., 38 (1998) 925.
39. Kier, L.B. and Hall, L.H., J. Chem. Inf. Comput. Sci., 40 (2000) 792.