Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism

A. J. Hopfinger,[†,‡,*] Shen Wang,[†,‡] John S. Tokarski,[†,‡] Baiqiang Jin,[†,‡] Magaly Albuquerque,[†,§] Prakash J. Madhav,[†] and Chaya Duraiswami[†,⊥]

Contribution from the Laboratory of Molecular Modeling and Design, M/C-781, College of Pharmacy, The University of Illinois at Chicago, 833 S. Wood St., Chicago, Illinois 60612-7231, and The Chem21 Group, Inc., 1780 Wilson Drive, Lake Forest, Illinois 60045

Received June 9, 1997. Revised Manuscript Received August 6, 1997[⊗]

Downloaded via UNIV FED DE MINAS GERAIS on January 26, 2025 at 15:47:05 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

Abstract: 4D-QSAR analysis incorporates conformational and alignment freedom into the development of 3D-QSAR models for training sets of structure−activity data by performing ensemble averaging, the fourth "dimension". The descriptors in 4D-QSAR analysis are the grid cell (spatial) occupancy measures of the atoms composing each molecule in the training set realized from the sampling of conformation and alignment spaces. Grid cell occupancy descriptors can be generated for any atom type, group, and/or model pharmacophore. A single "active" conformation can be postulated for each compound in the training set and combined with the optimal alignment for use in other molecular design applications including other 3D-QSAR methods. The influence of the conformational entropy of each compound on its activity can be estimated. Serial use of partial least-squares, PLS, regression and a genetic algorithm, GA, is used to perform data reduction and identify the manifold of top 3D-QSAR models for a training set. The unique manifold of 3D-QSAR models is arrived at by computing the extent of orthogonality in the residuals of error among the most significant 3D-QSAR models in the general GA population. Receptor independent (RI) 4D-QSAR analysis has been successfully applied to three training sets: (a) benzylpyrimidine inhibitors of dihydrofolate reductase, (b) prostaglandin $PGF_2\alpha$ antinidatory analogs, and, (c) dipyridodiazepinone inhibitors of HIV-1 reverse transcriptase (RT). Two general findings from these applications are that grid cell occupancy descriptors associated with the "constant" chemical structure of an analog series can be significant in the 3D-QSAR models and that there is an enormous data reduction in constructing 3D-QSAR models. The resultant 3D-QSAR models can be graphically represented by plotting the significant 3D-QSAR grid cells in space along with their descriptor attributes.

Introduction

Three-dimensional quantitative structure−activity relationship, 3D-QSAR, analysis is a major applications methodology in computer-assisted molecular design, CAMD. As part of the name implies (activity), the 3D-QSAR approach is most often used in pharmaceutical applications. However, the methodology is readily applicable to many chemical design problems and, within this general context, is referred to as three-dimensional quantitative structure−property relationship, 3D-QSPR, analysis.

Several schemes to doing 3D-QSAR analysis have been developed and are discussed in recent reviews.[1−3] Probably the most popular, and one of the first 3D-QSAR schemes, is comparative molecular field analysis, CoMFA.[4] The 4D-QSAR formalism described in this paper incorporates some CoMFA features including the spatial grid used to generate a basis set of QSAR descriptors. However, irrespective of the 3D-QSAR scheme, this CAMD methodology is receptor-independent. The data available in a 3D-QSAR analysis are a training set of compounds, usually analogs, and their measured biological activities in a common screen/assay. The geometry of the receptor is *not* available, and we will term this study a *receptor-independent* (RI) 3D-QSAR analysis. In contrast, *structure-based design* is a CAMD methodology for problems where the geometry of the receptor (usually an enzyme) *is* available along with corresponding ligand structure−activity data. When structure-based design is done to develop a quantitative model to forecast activity, we will term the approach *receptor-dependent* (RD) 3D-QSAR analysis. The 4D-QSAR scheme can be applied to both (RI) and (RD) problems. The receptor-independent formalism is presented and applied here.

(RI) 3D-QSAR analysis has three inherent problems to overcome. First is the identification of the *active* conformations/molecular shapes of flexible compounds in the training set. In the most straightforward interpretation, particularly for *in vitro* activity, the active conformation/shape of a ligand corresponds to the receptor-bound conformation/shape. Our working definition of active conformation/shape is that it is the one which optimizes the quantitative 3D-QSAR model. The second problem to be overcome is the specification of the basis for comparing molecules in constructing a 3D-QSAR which is referred to as the *molecular alignment*.[5] Finally, each molecule in the training set must be partitioned with respect to intermolecular (receptor) interactions. That is, different parts of each molecule can be expected to have different types of interactions

[†] The University of Illinois at Chicago.
[‡] The Chem21 Group, Inc.
[§] Permanent address: Instituto de Química, Universidade Federal do Rio de Janeiro, Ilha do Fundão, CT, Bl. A, Lab. 622, Rio de Janeiro, RJ, Brasil, 21949-900.
[⊥] Current Address: Skin Care Laboratories, Avon Products, Inc., Avon Place, Suffern, NY 10901.
[⊗] Abstract published in *Advance ACS Abstracts,* October 15, 1997.

(1) *3D-QSAR in Drug Design: Theory, Methods and Applications*; Kubinyi, H., Ed.; ESCOM Science Publishers: Leiden, The Netherlands, 1993.
(2) Green, S. M.; Marshall, G. R. *Tips* **1995** *16,* 285.
(3) Hopfinger, A. J.; Tokarski, J. S. 3D-QSAR. in *Practical Applications of Computer-Aided Drug Design*; Charifson, P. S., Ed.; Marcel Dekker: New York, Spring, 1997; p 106.
(4) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* **1988**, *16,* 5959.
(5) Hopfinger, A. J.; Burke, B. J.; Dunn, W. F., III *J. Med. Chem.* **1994**, *37,* 3768.

S0002-7863(97)01893-3 CCC: $14.00 © 1997 American Chemical Society

Table 1. The Ten Operational Steps in Performing a (RI) 4D-QSAR Analysis

step no.	description of the step operation
1.	Generate the reference grid and initial 3D models for all compounds in the training set.
2.	Select the trial set of interaction pharmacophores elements, IPEs.
3.	Perform a conformational ensemble sampling of each compound to generate its conformational ensemble profile, CEP.
4.	Select a trial alignment.
5.	Place each conformation of each compound in the reference grid cell space according to the alignment and record the grid cell occupancy profile, GCOP, for each IPE and choice in occupancy measure. The resulting composite set of grid cell properties constitute the set of grid cell occupancy descriptors, GCODs.
6.	Perform a PLS data reduction of the entire set of GCODs against the biological activity measures.
7.	Use the most highly weighted PLS GCODs and any other user-selected descriptors for the initial descriptor basis set in a GA analysis.
8.	Return to Step 4 and repeat Steps 4−7 unless all trial alignments have been included in the analysis.
9.	Select the optimum set of 3D-QSAR models with respect to alignment and any of the methodology parameters.
10.	Adopt the lowest-energy conformer state from the set sampled for each compound, which predicts the maximum activity using the optimum 3D-QSAR model as the "active" conformation (shape).

Table 2. Methodology Parameters of 4D-QSAR Analysis

parameter description	symbol
grid cell size [only cubic cells are allowed]	s
temperature of the molecular dynamics simulation, MDS	T
reference molecule	R
size of ensemble sampling (no. of distinct initial starting conformations in the sampling)	E_s(I)
no. of alignments	N_a
no. of descriptors in the GA initial basis set	N_d

with sites on a common receptor and/or in a common medium. This partitioned form of the molecule is called the *interaction pharmacophore*. The 4D-QSAR formalism has been developed to deal with each of these problems in constructing a 3D-QSAR model.

The fourth dimension of 4D-QSAR analysis is the "dimension" of ensemble sampling. The purpose of this paper is to describe the (RI) 4D-QSAR formalism and to demonstrate its utility in a series of applications. Throughout the paper *4D-QSAR* is used when discussing the joint RI and RD formalism, and *(RI) 4D-QSAR* is used when specifically referring to the receptor-independent version.

Methods

The ten steps involved in the current (RI) 4D-QSAR formalism are given in Table 1. **Step 1** is analogous to initiating a CoMFA 3D-QSAR analysis in that a reference grid cell space and a 3D structure for each compound in the training set are specified. The grid cell size is viewed in 4D-QSAR analysis as a *methodology parameter*, and the quantitative 3D-QSAR model can be optimized as a function of the methodology parameters. The set of methodology parameters in (RI) 4D-QSAR analysis are described in Table 2. CoMFA uses the input 3D structures as the (assumed) active conformations/shapes. In 4D-QSAR analysis each 3D structure is the initial starting point in a conformational ensemble sampling of the molecule. In principle, any 3D structure is acceptable to initiate the ensemble sampling. In practice, particularly for analog training sets, a common, low-energy (not necessarily an energy minimum) conformation, with respect to common torsion angles across the training set, should be selected. This provides a "reference point" over the course of the 4D-QSAR analysis. As an example, trimethoprim, one of the analogs in the dihydrofolate reductase, DHFR, inhibitor applications study is shown in its initial conformation in grid cell space in Figure 1.

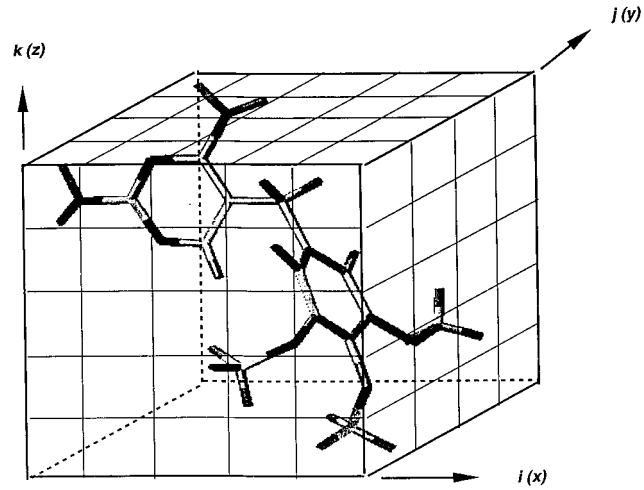


Figure 1. Trimethoprim, compound **1** of Table 4, shown in its initial CEP conformation, in 2 Å grid cell space.

Table 3. Interaction Pharmacophore Elements, IPEs, of (RI) 4D-QSAR Analysis

IPE description	symbol
all atoms of the molecule	IPE(a)
polar atoms of the molecule	IPE(p)
nonpolar atoms of the molecule	IPE(n)
hydrogen bond doners	IPE(hbd)
hydrogen bond acceptors	IPE(hba)
user defined IPE types	IPE(x)

The **Second Step** in (RI) 4D-QSAR analysis is the selection of the trial set of interaction pharmacophore elements, IPEs. The current default set of IPEs is listed in Table 3. In essence, the atoms of a molecule are partitioned into five classes: polar, nonpolar, hydrogen bond donor, hydrogen bond acceptor, and no differentiation-all atom occupancy. The last entry in Table 3, "user defined", provides an open-ended opportunity to test any, and all, IPEs. These user defined IPEs can range from modest subcharacterization, such as dividing polar atoms into specific types, to defining specific sets of atoms and/or molecular fragments.

The **Third Step** in the (RI) 4D-QSAR formalism, performing a conformational ensemble sampling of each compound in the training set, addresses the active conformation issue. The objective of this step is to Boltzmann sample the complete set of conformations available to each molecule. This objective is difficult to realize. If systematic conformational search is used, the number of conformations to explore may be too large for practical applications or the search done at a sampling resolution that leaves out crucial conformational states. Moreover, the data from a systematic search has to be rescaled, based upon conformational energetics, to a Boltzmann distribution. However, simulation sampling, in contrast to systematic sampling, has the problem of being open-ended. It is problematic to know when the simulation is large enough to accommodate all necessary states. A variety of approaches to conformational sampling have been proposed and are described and/or referenced in ref 6. In the (RI) 4D-QSAR application studies reported here, multilength and/or multitemperature molecular

dynamics simulation, MDS, was used to generate the conformational ensemble profiles, CEPs. To reiterate, the reason this 3D-QSAR scheme is titled **4D-QSAR** analysis is because ensemble sampling is considered a QSAR dimension. The current criteria for sampling convergence of the CEP are as follows: (1) The Boltzmann distribution becomes independent of the sampling size. (2) Different starting states lead to the same distribution. (3) Different sampling schemes produce the same optimized 3D-QSAR models. (4) For each sampling scheme the *average* rate of change of conformational energy with simulation time is effectively zero.

While these criteria are reasonable, they can be misleading in that the MDS can become trapped in a local minimum energy region of conformational space and appear to have converged. Ultimately, it is the predictive reliability of the quantitative 3D-QSAR model generated in the (RI) 4D-QSAR analysis which supports the scheme to construct the CEPs and, for that matter, the other components of the (RI) 4D-QSAR analysis.

MDSs have been done using the MOLSIM package[7] with an extended MM2 force field.[8] Partial atomic charges were computed using semiempirical molecular orbital methods. The specific partial charge calculation method is reported with the results of each application. The temperature, time, and CEP sampling schedule of the MDSs for each application is also given with the results.

Selection of the trial alignments is **Step 4** in a (RI) 4D-QSAR analysis. To be clear, 4D-QSAR analysis does not "solve" the alignment problem. Rather, the 4D-QSAR scheme permits a rapid evaluation of individual trial alignments. Consequently, the alignment problem can be treated as a search and sample operation analogous to conformational profiling. The ability to rapidly evaluate alignments in the (RI) 4D-QSAR algorithm is due to the complete decoupling of conformational analysis from alignment analysis and rapid descriptor estimation for each alignment. Only a single set of CEPs, one for each compound in the training set, is needed to evaluate an arbitrarily large number of alignments in terms of the significance of each corresponding 3D-QSAR model. Each alignment produces a unique grid cell occupancy distribution for a given CEP of a compound.

The current (RI) 4D-QSAR algorithm only considers unrestricted three-atom match alignment rules. However, there is no restriction in the algorithm to prevent the inclusion of any alignment rule. The rapid evaluation of any alignment rule is a necessary constraint for the practical application of the (RI) 4D-QSAR algorithm.

Each conformation from the CEP of each compound is placed in the reference grid cell space according to the trial alignment under consideration as part of **Step 5**. The grid cell occupancy profiles for each of the chosen IPEs are then computed and used as the basis set of trial 3D-QSAR descriptors. Three types of grid cell occupancy measures are considered for each IPE. The *absolute occupancy*, A_0, of grid cell (i, j, k), where i, j, and k define the *xyz*-coordinate location in Cartesian space of the cell, at time t in the MDS ensemble sampling for the IPE atoms of compound c is defined as

$$A_0(c, i, j, k, N) = \sum_{t=0}^{\tau} O_t(c, i, j, k) \tag{1}$$

τ is the time-length of the MDS ensemble sampling for Δt time steps, and $O_t(c, i, j, k) = 0$ if all IPE atoms of c are not in cell (i, j, k) at t, and $O_t(c, i, j, k) = m$ if m IPE atoms of c are in cell (i, j, k) at t. The "in"/"not in" a grid cell corresponds to the geometric center of the test atom residing anywhere within the grid cell. N is the number of sampling steps, ($\tau/\Delta t$).

Joint occupancy, J_0, of grid cell (i, j, k) for c with some reference compound, R, is defined as

$$J_0(c, R, i, j, k, N) = \sum_{t=0}^{\tau} O_t(c, i, j, k) \cap O_t(R, i, j, k) \tag{2}$$

Finally, self-occupancy, S_0, of grid cell (i, j, k) by c relative to R is given by

$$S_0(c, R, i, j, k, N) = \sum_{t=0}^{\tau} [O_t(c, i, j, k) - \{O_t(c, i, j, k) \cap O_t(R, i, j, k)\}] \tag{3}$$

Each of the occupancy measures can be "normalized" by dividing the sampling values by N. There are no obvious "rules" to apply in deciding which occupancy descriptors should be used in a particular study. The use of a reference compound, R, biases the 3D-QSAR model toward the template properties of R. This biasing can be useful when the training set is small. R can be selected to be a highly active member of the training set so that the resultant 3D-QSAR model is most highly influenced by high activity features. Conversely, if the training set is large, there is probably no reason to introduce any bias into the analysis, and the use of absolute occupancy descriptors is the preferred choice.

Step 6 is the data reduction step identical to that done in CoMFA. 4D-QSAR analysis, like CoMFA, generates an enormous number of trial QSAR descriptors because of the large number of grid cells and because of the IPEs and their three possible corresponding grid cell occupancy representations. Unlike CoMFA, however, the 4D-QSAR scheme intrinsically defines the complete set of grid cells to include in an analysis. The composite set of grid cells occupied at least once in constructing the CEP for the all atom IPE of each member of the training set defines the complete set of grid cells. CoMFA uses a distance cutoff in the evaluation of field potential to limit the set of descriptor grid cells.

A plot of an IPE CEP can be viewed as a Boltzmann average spatial distribution of molecular shape with respect to the IPE. Figure 2 is a difference plot of the all atom IPE CEPs, using A_0, eq 1, as a function of grid cell location for a rigid and inactive DHFR inhibitor subtracted from trimethoprim, compound **1** of Table 4, a potent DHFR inhibitor. In this plot grid cell location (i, j, k) is mapped into a single index, m^*. The plot in Figure 2 is referred to as a difference *molecular shape spectrum*, MSS, with respect to the all atom IPE. The working hypothesis in the 4D-QSAR formalism is that the observed difference in DHFR inhibition potency between the two compounds is related to the difference in their MSS. Data reduction is the tool used to facilitate the identification of this relationship, that is, the 3D-QSAR model. The MSS may be dependent on grid cell size and the absolute coordinate positioning of the compound. This possible dependence can be quickly evaluated by trying different positions and/or grid cell sizes.

(6) Huber, T.; Torda, A. E.; van Gunsteren, W. F. *Biopolymers* **1996**, *39*, 103.

(7) Doherty, D. *MOLSIM, Molecular Dynamics Simulation Software, User Guide, V.2.1*; The Chem21 Group, Inc., 1780 Wilson Drive, Lake Forest, IL 60045, 1994.

(8) (a) Allinger, N. L.; Yuh, Y. H. *Operating Instructions for MM2 and MMP2 Program - 1977*; Force Field Quantum Chemistry Program Exchange, Chemistry Dept., Indiana University, Bloomington, IN, 1980. (b) Hopfinger, A. J.; Pearlstein, R. A. *J. Comput. Chem.* **1984**, *5*, 486.

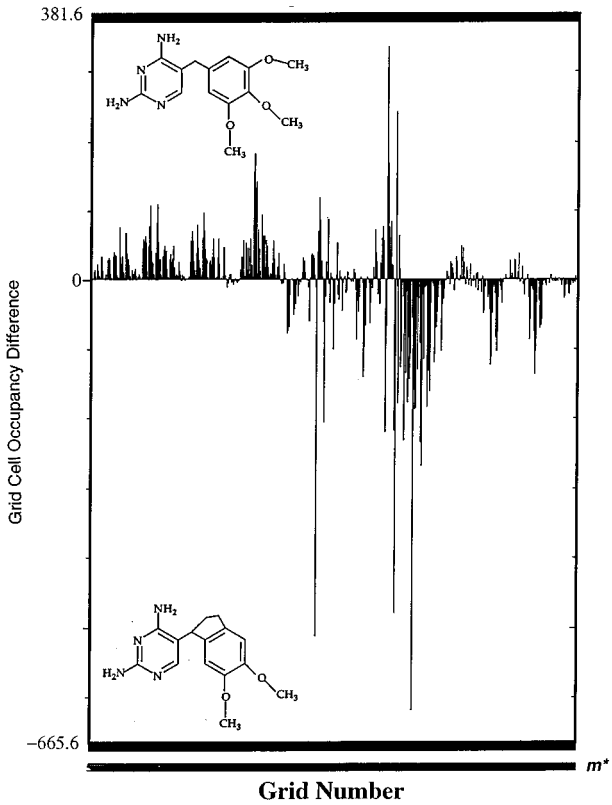


Figure 2. The IPE(a) difference molecular shape spectrum, MSS (grid cell occupancy as a function of grid cell number, m^*), for trimethoprim, an active DHFR inhibitor shown at the top of the plot, minus an inactive DHFR inhibitor shown in the bottom of the plot.

Partial least squares, PLS, regression[9] is used to perform the data reduction fit between the observed biological activities and the corresponding grid cell occupancy descriptor, GCOD, values. The resulting PLS regression fit is the quantitative 3D-QSAR model. The PLS weightings for the grid cell descriptors could be used, as done in CoMFA, to construct a graphical representation of the 3D-QSAR model. However, a general finding from the application of (RI) 4D-QSAR analysis (see **Results**) is that only a relatively small number of grid cell descriptors (<15 from many thousands) are significant in the quantitative 3D-QSAR model. Consequently, (RI) 4D-QSAR analysis offers the opportunity of providing an economical 3D-QSAR equation, in terms of the number of descriptors, reminiscent of those developed in classic 2D-QSAR studies.[10] These compact 3D-QSAR models are straightforward to explore and thereby provide substantial insight into the structure−activity information inherent to the training set.

The compact 3D-QSAR models are actually generated in **Step 7** as part of model building, optimization, comparison, and evaluation using a genetic algorithm, GA.[11] Currently, two GAs are being used in (RI) 4D-QSAR analysis: the *genetic function approximation*, GFA, developed by Rogers[12,13] and the GA of the 3D-QSAR method developed by Walters called GERM.[14]

(9) Glen, W. G.; Dunn, III, W. J.; Scott, D. R. *Tetrahedron Comput. Methods* **1989**, *2*, 349.

(10) Hansch, C.; Leo, A. *Exploring QSAR Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.

(11) Holland, J. *Adaptation in Artificial and Natural Systems*; University of Michigan Press: Ann Arbor, MI, 1975.

(12) Rogers, D. *The Proceedings of the Fourth International Conference on Genetic Algorithms*; San Diego, CA, 1991.

(13) Rogers, D.; Hopfinger, A. J. *J. Chem. Inf. Computer Sci.* **1994**, *34*, 854.

(14) Walters, D. E.; Hinds, R. M. *J. Med. Chem.* **1994**, *37*, 2527.

The M most highly weighted PLS descriptors are used to form the trial basis set for the GA analysis. Currently, $M = 200$, and only linear terms are used in multiple linear regression, MLR, fits in the GA optimization. Other descriptors, not derived from 4D-QSAR analysis, can be added into the trial basis set at the start of this step by the user.

Several diagnostic measures to analyze the resultant 3D-QSAR models are determined as part of the GA optimization. The diagnostic measures include descriptor usage as a function of crossover operation, linear cross-correlation among descriptors and/or biological activity measures, number of significant models, and measures of model significance including correlation coefficient, R, leave-one-out cross-validation correlation coefficient, xv-R, and Friedman's Lack of Fit, LOF.[12,13]

Steps 5−7 are performed for a fixed alignment. **Step 8** is a decision/selection operation to consider additional trial alignments in model construction or to proceed with an evaluation of the composite set of 3D-QSAR models generated by the repetitive application of steps 4−7. Presently, step 8 is carried out directly by the investigator, and treatment of alignment is by sampling and not by an optimization scheme.

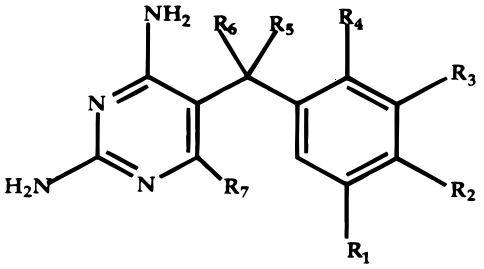
Once a desired set of trial alignments has been included in the 3D-QSAR model construction portion of the algorithm, the inspection and evaluation of the entire population of models is made. This is **Step 9**. The principal objective of step 9 is to identify the "best" 3D-QSAR with respect to alignment. However, this objective can be generalized to permit exploration and optimization of the 3D-QSAR with respect to not only alignment but also conformational sampling, IPE, and the methodology parameters listed in Table 2. Moreover, while a "best" 3D-QSAR model can be identified using one or more measures of significance of fit, added information comes from comparing a family of best models. In essence, the best and distinct set of 3D-QSAR models from GA analysis are used in composite to build a manifold 3D-QSAR model. One approach to identifying the set of distinct models from a GA analysis is given as part of the HIV-1 RT inhibition application.

The final step, **Step 10**, 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 being set at 2 kcal/mole. The resulting set of low-energy conformations are individually evaluated in the best 3D-QSAR equation by assigning grid cell occupancy to be zero or the maximum possible occupancy value consistent with the ensemble sampling scheme, depending on conformation and alignment. In essence, a CEP at $T = \mathrm{O}$ K, starting with the trial conformation, which will be the only MD conformation sampled at $T = \mathrm{O}$ K, is generated, and the corresponding grid cell occupancy descriptors are used to evaluate activity for the best 3D-QSAR model. The single conformation within ΔE, which predicts the highest activity, is selected as the active conformation of the compound. The postulated active conformations can be used as structure design templates, including their deployment as the molecular geometries in a corresponding CoMFA analysis. The preferred alignment found in the (RI) 4D-QSAR analysis can also be used in the corresponding CoMFA study.

Results

(RI) 4D-QSAR analysis has been applied to three training sets. The first training set to be discussed is a series of substituted 2,4-diamino-5-benzylpyrimidine inhibitors of *E. coli*

Table 4. The Training Set of Substituted 2,4-Diamino-5-benzylpyrimidine Inhibitors of *E. coli* DHFR and Their Corresponding Observed $\log(1/I_{50})$ Inhibition Measures[a]



no.	compd	R_1	R_2	R_3	R_4	R_5	R_6	R_7	log(1/150)
1	1	$-OCH_3$	$-OCH_3$	$-OCH_3$	−H	−H	−H	H	8.23
2	15	$-OCH_3$	$-OCH_3$	$-OCH_3$	$-CH_2-$	−H	−H	H	5.85
3	17*R*	$-OCH_3$	$-OCH_3$	$-OCH_3$	−H	−OH	$-CH_3$	H	4.00
4	17*S*								
5	22	$-OCH_3$	$-OCH_3$	$-OCH_3$	H	$=CH_2$		H	5.60
6	24*R*	$-OCH_3$	$-OCH_3$	$-OCH_3$	H	H	$-CH_3$	H	5.35
7	24*S*								
8	40	$-OCH_3$	-Br	$-OCH_3$	−H	−H	−H	H	8.53
9	55	$-OCH_3$	−OH	$-OCH_3$	−H	−H	−H	H	7.96
10	56	$-OCH_3$	−OH	$-OCH_3$	H	H	H	$-CH_3$	6.52
11	57	$-OCH_3$	$-OCH_3$	$-OCH_3$	H	H	H	$-CH_3$	7.00
12	59	−OH	−H	−OH	H	H	H	H	2.78
13	62	−H	−H	−H	H	H	H	H	5.71
14	68	$-CH_2OH$	−H	$-CH_3OH$	−H	−H	−H	H	5.83
15	81	−H	−H	Cl	H	H	H	H	6.14
16	84	−H	−Br	−H	H	H	H	H	6.30
17	92	$-OCH_3$	−H	−H	−H	−H	−H	H	6.40
18	102	$-OCH_3$	−H	$-OCH_3$	−H	−H	−H	H	7.75
19	118	$-CH_3$	−H	$-CH_3$	H	H	H	H	7.45
20	137	H	$-C_6H_5$	−H	−H	−H	−H	H	6.40

[a] Compound no. 1 ($R_1 = R_2 = R_3 = -OCH_3$) and ($R_4 = R_5 = R_6 = R_7 = H$) is trimethoprim.

dihydrofolate reductase, DHFR.[15] The principal reason for selecting this application is that the bound inhibitor−enzyme crystal complex geometry is available for one active inhibitor in the training set. Thus, a comparison of the 3D-QSAR model constructed from (RI) 4D-QSAR analysis, to the geometry of the inhibitor−enzyme complex, permits an assessment of the accuracy and information-content of the 3D-QSAR model.

The second applications study involves a set of prostaglandin $PGF_2\alpha$ analogs tested for the antinidatory effect in hamster[16] and rat.[17] These compounds are highly flexible and, consequently, explore how well (RI) 4D-QSAR analysis handles conformational freedom and alignment variability. Since an *in vivo* biological measure is used in the training set, this application also explores the utility of (RI) 4D-QSAR analysis for *in vivo* modeling applications.

The final application study reported here is for a training set of 2-substituted dipyridodiazepione inhibitors of both wild-type and mutant cysteine-181 HIV-1 reverse transcriptase (RT) enzymes.[18,19] The primary objective in choosing this training set is to determine if (RI) 4D-QSAR analysis can produce 3D-QSAR models which meaningfully differentiate between wild-type and cysteine-181 RT activities for a common set of analog inhibitors.

(RI) 4D-QSAR analysis has not been applied to the steroid data set used in the first reported CoMFA study[4] and has become something of a comparative "standard" in evaluating other 3D-QSAR approaches.[3] The steroid data set minimizes complications due to conformation and alignment. 4D-QSAR has been designed to handle conformation and alignment. Thus, the steroid data set offers little opportunity to evaluate the novel features and capabilities of 4D-QSAR analysis.

1. Dihydrofolate Reductase Inhibitors. The training set of 20 substituted 2,4-diamino-5-benzylpyrimidine inhibitors of *E. coli* DHFR considered in the (RI) 4D-QSAR application study are given in Table 4. The observed I_{50} values are also listed as $\log(1/I_{50})$. The second column in Table 4 labeled "compd" refers to a parent set of pyrimidine inhibitors that has been compiled over the last few years and is being constantly updated.[15] The 20 compounds in Table 4 are relatively diverse given the large number of sites of substitution[7] and the inclusion of two sets of isomer pairs [numbers 3, 4 and 6, 7]. Overall, some of these compounds have proven to be outliers in other QSAR studies in our laboratory. Thus, one "extra" goal of this study has been to see if (RI) 4D-QSAR can successfully model these compounds when other attempts have failed.

The (RI) 4D-QSAR analysis of the inhibitors given in Table 4 is summarized in Table 5. Part a of Table 5 lists the methodology parameters. Only four trial alignments, N_a, defined in Figure 3, were considered because these four alignments sample each alignment class: (1) the pyrimidine ring, (2) the "center" of the molecule including the methylene bridge between rings, (3) the benzyl ring, and (4) a combination of the pyrimidine ring, "center", and benzyl ring. The most active

(15) Duraiswami, C. *General Treatments of Conformation and Alignments in Quantitative Structure−Activity Relationships*; Ph.D. Thesis, Dept. Medicinal Chemistry & Pharmacognosy, University of Illinois at Chicago: Chicago, IL, 1996.

(16) Fletcher, D. G.; Gibson, K. H.; Moss, H. R.; Sheldon, D. R.; Walker, E. R. H. *Prostaglandins* **1976**, *12*, 493.

(17) Hayashi, M.; Arai, Y.; Wakatsuka, H.; Kawamura, M.; Konishi, Y.; Tsuda, T.; Matsumoto, K. *J. Med. Chem.* **1980**, *23*, 525.

(18) Cardozo, M. G., data supplied prior to submission for publication, 1996.

(19) Proudfoort, J. R.; Hargrave, K. D.; Kapadia, S. R.; Patel, U. R.; Grozinger, K.G.; McNeil, D. W.; Cullen, E.; Cardozo, M.; Tong, L.; Kelly, T. A.; Rose, J.; David, E.; Mauldin, S. C.; Fuchs, V. U; Vitous, J.; Hoermann, M.; Klunder, J. M.; Raghavan, P.; Skiles, J. W.; Mui, P.; Richman, D. D.; Sullivan, J. L.; Shih, C.-K.; Grob, P. M.; Adams, J. *J. Med. Chem.* **1995**, *38*, 4830.

Table 5. Description of the (RI) 4D-QSAR Analysis of the Training Set of DHFR Inhibitors of Table 4 and a Summary of the Best 3D-QSARs Found in the Study

(a)

methodology parameter, see Table 2	value	methodology parameter, see Table 2	value
(X_L, Y_L, Z_L) s	(20 Å, 20 Å, 20 Å) 1 Å	E_s(I)	25 000 (5)
T	300 K	N_a	4, see Figure 3
R	no. 1 of Table 4	N_d	205

(b) Interaction Pharmacophore Elements, IPEs

J_0 IPE(a)
S_0 IPE(a)

(c) Summary of Top Ten (RI) 4D-QSAR Models (3D-QSARs) Using Alignment 2 of Figure 3

range in R^2 and (xv-R^2) in the top ten GFA models	0.901−0.957 (0.790−0.885)
no. of unique outliers in all top ten GFA models	3
no. of unique grid cells in all top ten GFA models	8
no. of significant PLS components	3
no. of significant GFA descriptors in each of the top models	4
ΔR^2 and Δ(xv-R^2) for top GFA model	0.027 (0.034)
no. of non-GCODs in all of the top ten GFA models	0

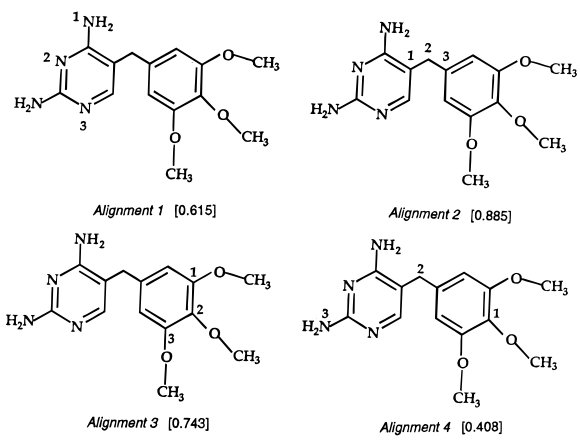


Figure 3. The four (N_a) alignment rules used in the DHFR inhibitor 4D-QSAR analysis. The numbers (1, 2, 3) define the alignment atoms. Each number in brackets is the xv-R^2 of the best 3D-QSAR found for the alignment.

analog in the training set, no. 1 of Table 4, trimethoprim, was used as the reference compound, R, to construct the J_0 IPE(a) and S_0 IPE(a) grid cell descriptor sets, see part b of Table 5. No other IPEs were considered in the analysis. Five thousand conformations, E_s, were sampled for each of five (I) different starting conformations, E_s(I), for each compound and led to virtually the same set of top-rated 3D-QSAR models for each starting conformation. The starting conformations were the five lowest minimum-energy conformations found in the free-space, intramolecular conformational analysis of trimethoprim.[20] Five nongrid cell occupancy descriptors (non-GCODs) found to be of possible importance in an earlier 3D-QSAR study[20] were used in addition to the 200 most highly weighted PLS GCODs to form the GFA analysis initial basis set, $N_d = 205$.

The most significant 3D-QSAR model, evaluated by (xv-R^2), is quite dependent upon choice of alignment. The (xv-R^2) for the best 3D-QSAR models found for each of the four test alignments are given in brackets in Figure 3. Alignment 2 (over the "center" of the molecule) is preferred, with alignment 3 ranking a marginal second best. Alignment 2 was adopted as the preferred manner of comparing analogs, and the results presented below are based upon using alignment 2 in Figure 3.

(20) Dunn, W. J., III; Hopfinger, A. J.; Catana, C.; Duraiswami, C. *J. Med. Chem.* **1996**, *39*, 4825.

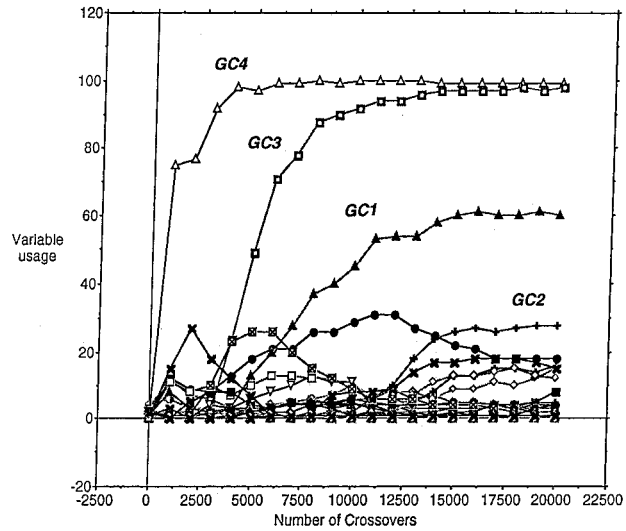


Figure 4. A plot of GCOD usage as a function of crossover operation number in the GFA analysis. The plateau curve associated with the usage of GCI ($I = 1-4$) indicates the evolutionary optimization has converged and GC1 to GC4 are the most often used independent variables in the best 3D-QSAR models.

The optimal 3D-QSAR model found in the 4D-QSAR analysis is

$$\log(1/I_{50}) = 0.0205GC1(J_0) - 0.0324GC2(J_0) + 0.1662GC3(J_0) + 0.1794GC4(J_0) + 5.85$$

$$N = 20 \quad R^2 = 0.957 \quad xv\text{-}R^2 = 0.885 \quad F = 83.1 \quad SD = 0.34 \quad (4)$$

Figure 4 is a plot of variable (descriptor) usage as a function of GFA crossover operation. In essence, Figure 4 portrays the relative significance of the GCODs in the optimum (RI) 4D-QSAR models. It is noteworthy that none of the non-GCODs (that is, descriptors not derived by the 4D-QSAR method) found important in an earlier study[20] survived as significant descriptors in any of the top ten 3D-QSAR models derived by (RI) 4D-QSAR analysis. The characteristic features of the top ten 3D-QSAR models from the combined PLS-GFA model building scheme used in (RI) 4D-QSAR analysis are given in part c of Table 5. The top ten models, in composite, include eight unique grid cells out of 1736 grid cells which were occupied. Of the eight grid cells, five are used in at least six of the top ten models. The "information loss", as measured by the decrease in both the R^2 and xv-R^2 values of eq 4, as compared to the complete PLS model, is given by ΔR^2 and Δ(xv-R^2) in part c of Table 5. Both ΔR^2 and Δ(xv-R^2) are small. Such a small "information loss" suggests that eq 4 captures the very large majority of information that can be extracted from the training set.

Figure 5a shows the GCODs of eq 4 in space. Two of the grid cells, GC1 and GC2, correspond to sites in the vicinity of the 4 and 5 substituents on the benzyl ring. One grid cell, GC4, is near the benzyl side of the methylene spacer, and the other grid cell, GC3, is near the 2-NH_2 group of the pyrimidine ring. This appears to be the first QSAR model for benzylpyrimidine inhibitors in which the "constant" portion of the inhibitor in the analog series, the 2,4-diaminopyrimidine unit, has a QSAR descriptor associated with it. (RI) 4D-QSAR analysis identifies a change in the all atom IPE population of GC3, as a function

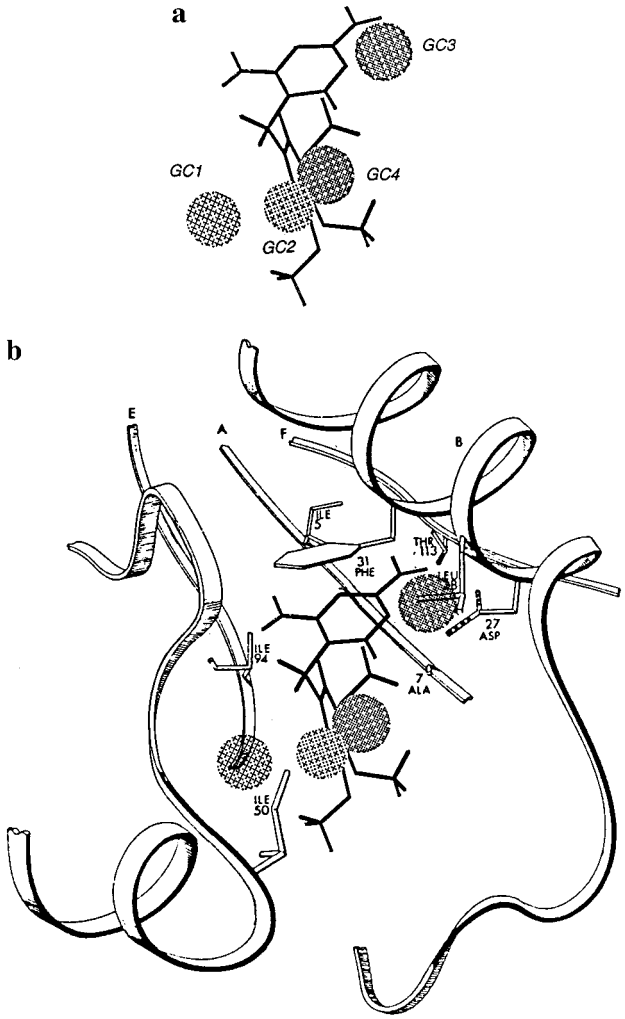


Figure 5. The GCODs of the best 3D-QSAR model, eq 4, in space. The grid cells are represented as spheres with 1 Å diameters: (a) isolated trimethoprim and the key grid cells of eq 4 and (b) trimethoprim bound to a schematic representation of *E. coli* DHFR reported in ref 29. The locations of the grid cells relative to key enzyme residues are shown.

of substituent modification elsewhere on the molecule under alignment 2 of Figure 3, as a significant GCOD. This change in grid cell occupancy for GC3 mainly involves the 2-NH_2. The positive regression coefficients for GC1, GC3, and GC4 in eq 4 indicate activity should increase with increasing ligand atom occupancy, while the opposite is true for GC2. This particular 3D-QSAR model can be rationalized by placing the model shown in Figure 5a into the active site of *E. coli* DHFR in the bound alignment, shown in schematic form in Figure 5b. GC1 corresponds to a hydrophobic region "between" Ile 94 and Ile 50, GC2 may be indicative of steric interactions with Ile50, and GC4 seems to be a hydrophobic space in the vicinity of Ala 7. GC3 presumably reflects favorable hydrogen bonding between the 2-NH_2 and the Asp-27 carboxyl group and, hence, its positive regression coefficient in eq 4. Overall, GC1 through GC4 reflect the extent of receptor probing realized by the 20 analogs of Table 4.

The active conformation of each analog was predicted using a $\Delta E = 2$ kcal/mole cutoff and eq 4. Figure 6 is a conformational energy plot of torsion angles "1" and "2" for trimethoprim as reported by Kuyper.[21] The locations of the bound conformations of trimethoprim to both *E. coli* and chicken DHFR enzymes are indicated. Also plotted in Figure 6 are the

(21) Kuyper, L. F. Inhibitors of Dihydrofolate Reductase. In *Computer-Aided Drug Design*; Perun, J. T., Propst, C. L., Eds.; Marcel Dekker: New York, 1989; p 327.

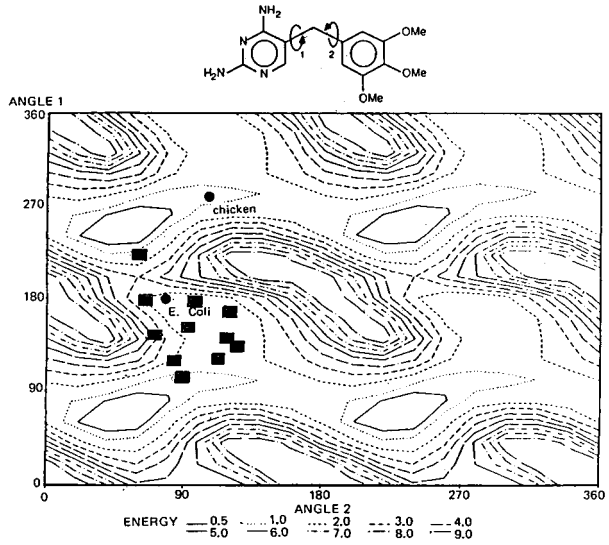


Figure 6. The conformational energy profile of trimethoprim relative to torsion angles 1 and 2 with the enzyme bound conformations to *E. coli* and chicken DHFR defined. The energy contours are in kcal/mol relative to the global minimum and defined below the profile. This plot is from ref 21. The locations of the postulated "active" conformations of the 11 inhibitors in Table 4 having $\log(1/IC_{50}) > 6$ are plotted as solid rectangles in this profile.

Table 6. Descriptions of the (RI) 4D-QSAR Analyses of the Training Set of $PGF_2\alpha$ Prostaglandin Analogs Given in Figure 7, and Summaries of the Corresponding Best 3D-QSAR Models Found in Each of the Two Studies.

(a)

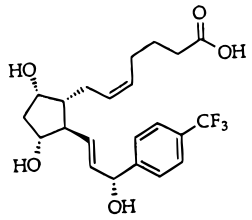
methodology parameter, see Table 2	value	methodology parameter, see Table 2	value
(X_L, Y_L, Z_L) s	(40 Å, 40 Å, 40 Å) study 1: 2 Å; study 2: 1 Å	E_s(I)	40 000 (1)
T	300 K	N_a	6
R	study 1: 26 of Figure 7; study 2: none	N_d	212

(b) Interaction Pharmacophore Elements

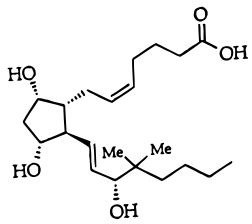
study 1	study 2	
J_0 IPE(a)	A_0 IPE(a)	A_0 IPE(n)
S_0 IPE(a)	A_0 IPE(p+)	A_0 IPE(hbd)
	A_0 IPE(p-)	A_0 IPE(hba)

(c) Summary of the Top Ten (RI) 4D-QSAR Models Using Alignment 3 of Table 7

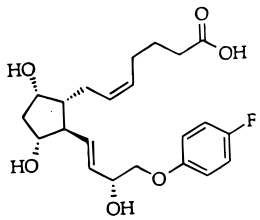
Study 1	
range in R^2 and (xv-R^2) in the top ten GFA models	0.726−0.855 (0.545−0.744)
no. of unique outliers in all top ten GFA models	5
no. of unique grid cells in all top ten GFA models	15
no. of significant PLS components	6
no. of significant GFA descriptors in each of the top models	7−10
ΔR^2 and $\Delta(xv\text{-}R^2)$ for the top GFA model	0.015 (0.021)
no. of non-GCODs in all the top ten GFA models	1, $[\log P]_\omega$
Study 2	
range in R^2 and (xv-R^2) in the top ten GFA models	0.719−0.752 (0.609−0.635)
no. of unique outliers in all top ten GFA models	5
no. of unique grid cells in all top ten GFA models	12
no. of significant PLS components	5
no. of significant GFA descriptors in each of the top models	6−8
ΔR^2 and $\Delta(xv\text{-}R^2)$ for the top GFA model	0.010 (0.018)
no. of non-GCODs in all the top ten GFA models	0



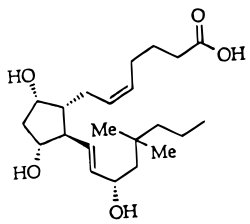
1 [1.699]



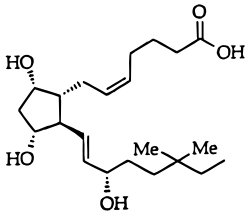
2 [0.081]



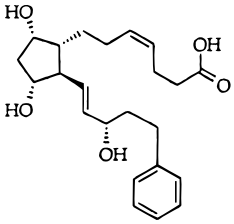
3 [2.301]



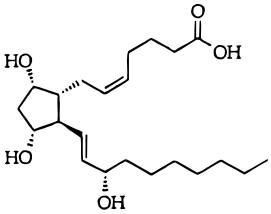
4 [0.279]



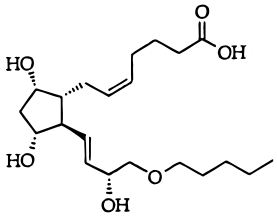
5 [0.664]



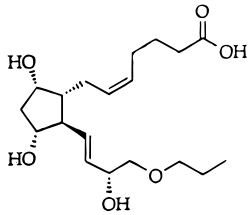
6 [2.301]



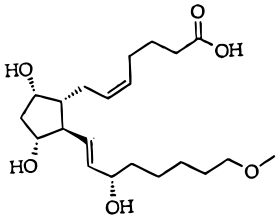
7 [0.699]



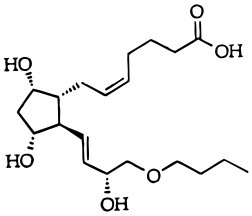
8 [1.000]



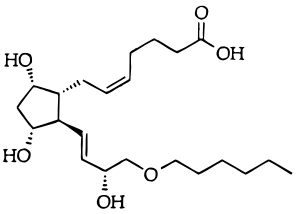
9 [0.398]



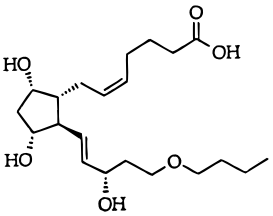
10 [0.398]



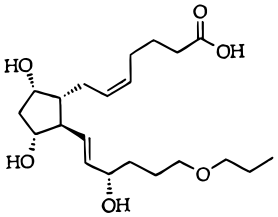
11 [1.000]



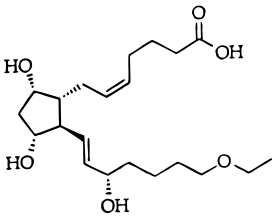
12 [0.699]



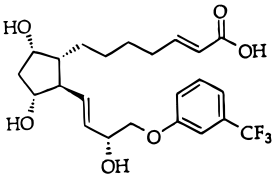
13 [-0.301]



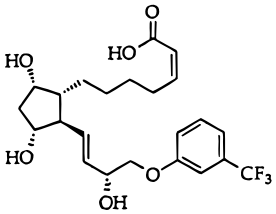
14 [-0.602]



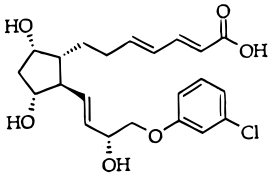
15 [0.699]



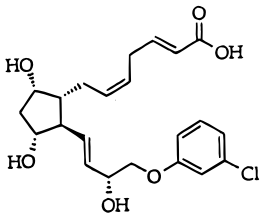
16 [0.602]



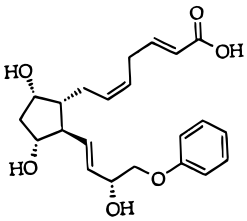
17 [0.482]



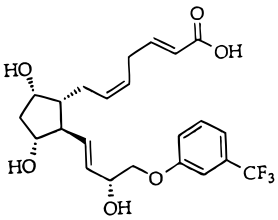
18 [0.777]



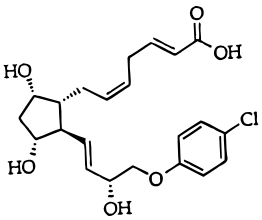
19 [2.301]



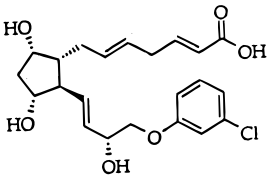
20 [2.481]



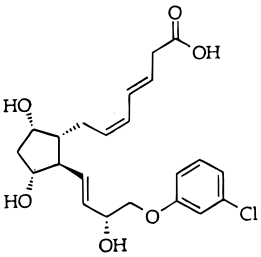
21 [2.000]



22 [2.000]



23 [2.000]



24 [2.886]

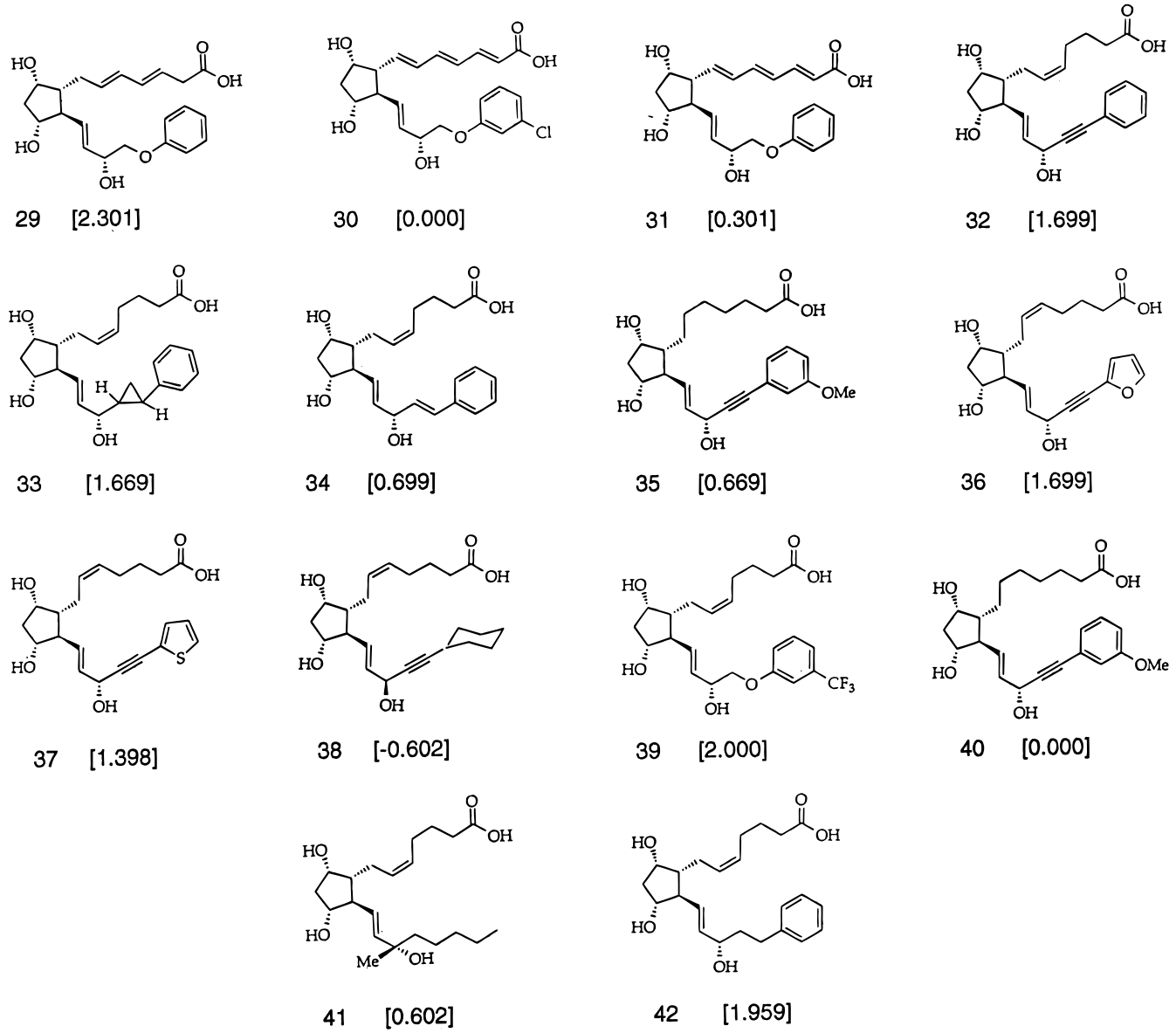


Figure 7. The structures of the $PGF_2\alpha$ analogs in the training set. The corresponding activities are given in brackets, [log(REL. ED_{50})].

predicted active conformations of the 11 analogs in Table 4 having $\log(1/IC_{50}) > 6$. The less active analogs in Table 4 adopt a relatively random pattern of "active" conformations, as compared to those in Figure 6. The active analog conformations cluster in the vicinity of the trimethoprim *E. coli* bound state.

2. $PGF_2\alpha$ Prostaglandins Tested for the Antinidatory Effect. The training set consists of 42 analogs of prostaglandin, $PGF_2\alpha$. The training set was assembled by investigators at Procter and Gamble Pharmaceuticals from compounds and corresponding biological activities reported in the literature.[16,17] The structures of the analogs are shown in Figure 7.

The biological activity measures are the ED_{50} values reported for the antinidatory effect in hamster[16] and rat.[17] Since the biological activity for some compounds are reported for hamster, while for others the reported activities are for rat, the activity measures have been scaled relative to the ED_{50} values of $PGF_2\alpha$, compound **40**, which was measured in both animals. For convenience, the logarithmic values of the ratio of the ED_{50} value of $PGF_2\alpha$ to the ED_{50} value of the test compound, log(REL. ED_{50}), are used as the dependent variables and are given in Figure 7.

The (RI) 4D-QSAR analysis is summarized in Table 6. In study 1 described in Table 6 the grid cell size was set to 2 Å because of the relatively large size of the compounds. A total ensemble sampling of 40 000 conformations per analog led to a constant CEP for each analog. Six different three-atom alignments were considered which span and sample the component "pieces" of the $PGF_2\alpha$ analogs. The six alignments are defined in Table 7 based upon the atoms specified in Figure 8. The xv-R^2 values of the best 3D-QSAR model for Study 1, see below, are also reported in Table 7. Alignment 3 stands out as the preferred choice and also is the preferred choice used in Study 2. The top ten 3D-QSAR models, summarized for the two studies in part c of Table 6, and discussed below, are for alignment 3 of Table 7.

Table 7. The Three Atoms Used in Each Alignment Rule, and the Cross-Validated Correlation Coefficient of the Best 3D-QSAR Model for Each Alignment[a]

alignment no.	atom 1	atom 2	atom 3	(xv-R^2) optimum
1	a	b	c	0.532
2	d	b	e	0.407
3	f	a	c	0.744
4	h	c	a	0.553
5	a	f	g	0.535
6	c	h	i	0.581

[a] The atoms are defined in Figure 8.

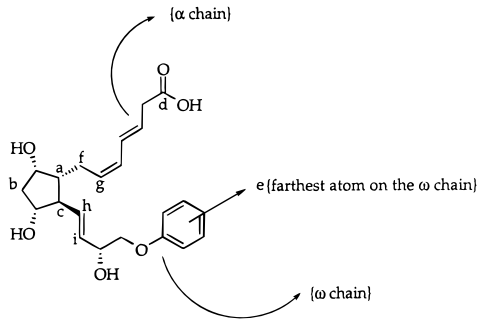


Figure 8. The atoms (a through i) used to define the six alignment rules used in study 1 of the (RI) 4D-QSAR analysis of the $PGF_2\alpha$ analog training set. The compound shown is number 26 in Figure 7.

Study 1 was performed using the highly active compound 26 of Figure 7 as the reference in determining J_0 and S_0 for only the IPE(a). The optimum 3D-QSAR model for all 42 analogs in the first study is

$$\log(\text{REL. ED}_{50}) = 2.64\text{GC1}(J_0) + 3.68\text{GC2}(J_0) + 0.46\text{GC3}(J_0) - 1.66\text{GC4}(J_0) + 0.57\text{GC5}(J_0) + 0.85\text{GC6}(J_0) + 3.98\text{GC7}(J_0) + 28.3\text{GC8}(S_0) - 9.11\text{GC9}(S_0) + 161.\text{GC10}(S_0) - 4.68$$

$$N = 42 \quad R^2 = 0.855 \quad xv\text{-}R^2 = 0.744 \quad F = 18.3 \quad \text{SD} = 0.36 \quad R = 26 \text{ of Figure 7} \quad (5)$$

Equation 5 has two significant outliers [their residuals are each larger in magnitude than 2.0 standard deviations of fit], compounds **37** and **38** of Figure 7. If these two compounds are removed in the construction of the 3D-QSAR model by application of the GFA, the optimum model is

$$\log(\text{REL. ED}_{50}) = 2.72\text{GC1}(J_0) + 3.61\text{GC2}(J_0) + 0.50\text{GC3}(J_0) - 1.69\text{GC4}(J_0) + 0.58\text{GC5}(J_0) + 0.82\text{GC6}(J_0) + 3.92\text{GC7}(J_0) + 29.5\text{GC8}(S_0) - 10.7\text{GC9}(S_0) + 164.\text{GC10}(S_0) - 4.69$$

$$N = 40 \quad R^2 = 0.899 \quad xv\text{-}R^2 = 0.811 \quad F = 25.9 \quad \text{SD} = 0.31 \quad R = 26 \text{ of Figure 7} \quad (6)$$

Figure 9 is a stick model of compound 26 of Figure 7 in its postulated active conformation and the grid cells of eq 6 plotted relative to compound **26**. Non-GCODs were included as part of the initial basis set of descriptors in the GFA analysis. These descriptors are listed and defined in Table 8. Only one of these non-GCODs is found among the top ten 3D-QSAR models. [log $P]_\omega$ is found in the numbers 5 and 8 of the top ten models. No other non-GCOD considered in Study 1 was found to be significant.

Some GCODs involve grid cells associated with a "constant" structure across the set of analogs similar to that found in eq 4 of the DHFR inhibitor study. For example, GC1(J_0) and GC10-(S_0) involve two grid cells near constant structure sites over the training set.

The active conformations of the $PGF_2\alpha$ analogs were determined using a $\Delta E = 2$ kcal/cutoff and eq 6. Figure 10 is a plot of the difference in predicted activity of the postulated active conformation and the observed activity for each analog. Most of the analogs have predicted active conformations with corresponding predicted activities greater than the observed activities used to construct eq 6. We think a positive residual in activity for an analog is an indirect, approximate measure of the loss in *intrinsic* activity of the analog due to its conformational entropy. The $PGF_2\alpha$ analogs only occupy activity-enhancing grid cells "part of the time" and also spend "part of the time" in grid cells that either decrease activity or do not contribute owing to their incipient conformational entropies. The predicted active conformation of an analog can be viewed as a rigid, $T = 0$ K conformation that locks the analog into a single geometry. Analogs, whose active conformations have predicted activities greater than observed, occupy grid cells which enhance activity to a greater extent than grid cells which diminish or contribute nothing to activity. The opposite is true for the active conformations of those few analogs which have the modest negative residuals shown in Figure 10.

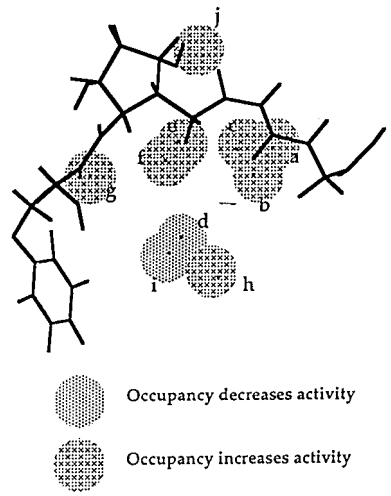


Figure 9. The GCODs of eq 6 plotted as spheres of 1 Å diameter relative to the postulated active conformation of compound **26** of Figure 7: (a) GC1(J_0), (b) GC2(J_0), (c) GC3(J_0), (d) GC4(J_0), (e) GC5(J_0), (f) GC6(J_0), (g) GC7(J_0), (h) GC8(S_0), (i) GC9(S_0), and (j) GC10(S_0).

Table 8. The Non-GCODs Used in the $PGF_2\alpha$ Analog Studies

descriptor description	symbol
log P of the α chain[a]	$[\log P]_\alpha$
log P of the ω chain[a]	$[\log P]_\omega$
log P of the entire molecule[a]	$[\log P]_M$
conformational entropy of the α chain[b]	e(α)
conformational entropy of the ω chain[b]	e(ω)
dipole moment of the terminal ω chain group[c]	d(ω)

[a] Computed from the MedChem software, ref 23. [b] Computed using TAU theory, ref 28. These entropy values are NOT computed as part of the conformational samplings used to construct the CEPs. [c] Extracted from the molecular dynamics simulation.

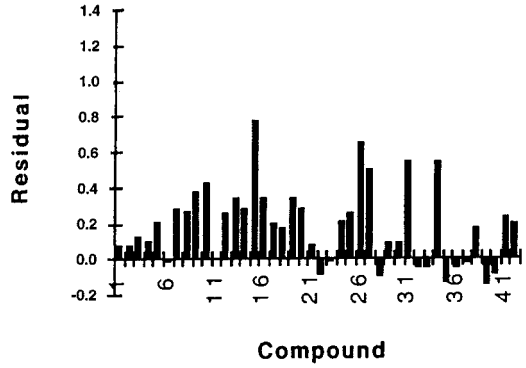


Figure 10. A plot of the difference, or residual, in predicted activity, using the postulated active conformation and eq 6 and the observed activity as a function of compound number.

Table 9. The Linear Correlation Matrix for the GCODs Used in Eq 6

	GC1(J_0)	GC2(J_0)	GC3(J_0)	GC4(J_0)	GC5(J_0)	GC6(J_0)	GC7(J_0)	GC8(S_0)	GC9(S_0)	GC10(S_0)	BA
GC1(J_0)	1.000										
GC2(J_0)	−0.262	1.000									
GC3(J_0)	**0.718**	0.026	1.000								
GC4(J_0)	−0.104	−0.191	−0.105	1.000							
GC5(J_0)	−0.336	−0.148	**−0.637**	0.191	1.000						
GC6(J_0)	−0.199	−0.338	−0.167	0.203	−0.024	1.000					
GC7(J_0)	−0.177	0.013	−0.043	−0.259	−0.017	−0.144	1.000				
GC8(S_0)	0.104	0.034	0.212	0.270	0.035	−0.033	−0.108	1.000			
GC9(S_0)	0.222	0.147	0.360	−0.031	−0.193	−0.187	−0.078	**0.565**	1.000		
GC10(S_0)	−0.147	−0.143	−0.242	0.459	0.316	0.194	−0.009	−0.034	−0.110	1.000	
BA	0.275	0.245	0.346	−0.163	0.067	0.089	0.321	0.317	0.133	0.177	1.000

Analogs with reasonably high observed activities, **and** large positive residuals in Figure 10, are good templates for new ligand design. That is, the compounds observed to be most active may not be the best design templates. Somewhat less active compounds, that are predicted to have high intrinsic activities in their predicted active conformations, may be better structure-design templates for characterizing specific ligand−receptor interactions.

Table 9 is the linear cross-correlation matrix of the GCODs of eq 6 and the activity measures. Two significant observations can be made from an inspection of Table 9. First, no individual GCOD has a high correlation with biological activity. Second, three pairs of GCODs are highly correlated to one another, and their R values are given in bold. A test was made to see if both members of each highly correlated pair of GCODs are needed in significant 3D-QSAR models. The results are given in Table 10. If one randomly selected member of each of the three pairs is deleted from the initial basis set, a poor model ($R^2 = 0.597$, xv-$R^2 = 0.400$), even with five outliers removed, results. Individual removal of GC5(J_0) cannot be tolerated, while removal of GC1(J_0), or removal of GC9(S_0), is marginally acceptable for models with four outliers. Overall, it appears that each member of a highly cross-correlated GCOD pair individually provides unique information to the 3D-QSAR model.

In the second (RI) 4D-QSAR analysis (Study 2) of the $PGF_2\alpha$ training set the J_0 and S_0 measures were replaced with A_0, and all five types of (but no-user selected) IPEs were employed. In addition, the polar atoms were subdivided into polar-positive (p+) charge and polar-negative (p−) charge subsets. Thus, a total of six IPEs were considered. A comparison of these two studies permits an assessment of the use of a reference compound in contrast to multiple IPEs and a smaller grid cell size. The top ten (RI) 4D-QSAR models found in Study 2 are summarized in part c of Table 6. The same methodology parameters, except the grid cell size, was reduced from 2 Å to 1 Å, and the same alignment (alignment 3 of Table 7) used in Study 1 was employed in Study 2. The best 3D-QSAR from this (RI) 4D-QSAR analysis is

$$\log(\text{REL. ED}_{50}) = 0.097\text{GC2(a)} - 0.008\text{GC4(n)} + 0.108\text{GC6(p+)} + 0.003\text{GC8(a)} + 0.038\text{GC10(p−)} - 0.043\text{GC11(hbd)} - 1.171$$

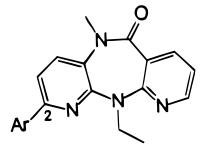
$$N = 42 \quad R^2 = 0.752 \quad xv\text{-}R^2 = 0.635 \quad F = 19.8 \quad \text{SD} = 0.44 \quad (7)$$

A comparison of eq 7 to eq 5 indicates that the use of A_0, multiple IPEs, and a smaller grid size reduces the number of GCODs from 10 to 6. The R^2 and xv-R^2 values are somewhat less for eq 7 than eq 5, but the F value is modestly higher. Five GCODs of eq 7 correspond to spaces within the grid cells used in eq 5. Only GC11(hbd) corresponds to a grid cell location not seen in eq 5. The added flexibility in characterizing descriptor space by using multiple IPEs is reflected in eq 7 where IPEs other than IPE(a) are found for four of the six GCOD terms. Somewhat surprisingly, reducing the grid cell size from 2 Å to 1 Å does not lead to a markedly better 3D-QSAR model from a statistical standpoint or a proliferation in the number of significant GCODs among the top ten models, see Table 6 part c for Study 2.

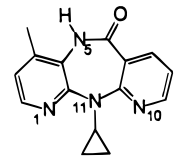
Table 10. R^2 and xv-R^2 Values for 3D-QSAR Models Generated by Excluding Highly Cross-Correlated GCODs from Eq 6

variables excluded	total no. of variables in model	no. of outliers	R^2	xv-R^2
none	10	2	0.855	0.743
GC1(J_0), GC5(J_0), GC9(S_0)	7	5	0.597	0.400
GC1(J_0)	9	4	0.801	0.644
GC5(J_0)	9	4	0.713	0.542
GC9(S_0)	9	4	0.833	0.710

3. 2-Substituted Dipyrididodiazepinone Inhibitors of HIV-1 Reverse Transcriptase (RT). The training set of inhibitors (I) are analogs to nevirapine (II) and are given in Table 11. This training set was constructed from literature compounds[19] and compounds provided by Boehringer Ingelheim Pharmaceuticals.[18] Inhibition activity, reported as $-\log(IC_{50})$, against both wild type, WT-RT, and the cysteine 181 mutant enzyme, Y181C-RT, are reported for each of the 40 compounds in Table 11. In those cases where the chemical structure has not been released for publication, no structure is given in Table 11, but the corresponding activities are listed. Some compounds in the training set can exist as tautomers. For each of these compounds both tautomers were individually considered in energy minimization calculations using the AM1 method.[22] The tautomer which gave the lowest energy geometry was retained for subsequent inclusion in the 4D-QSAR analysis.



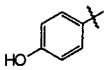
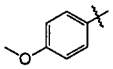
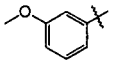
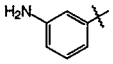
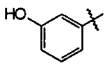
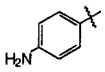
I



II

The (RI) 4D-QSAR analyses of the RT inhibitors given in Table 11 for both WT and Y181C activities are summarized in Table 12. Twenty alignments, described in Table 13, were considered in this 4D-QSAR analysis. The R^2 value for the best WT-RT inhibition GFA model realized for each alignment is also recorded in Table 13 along with the compound numbers (from Table 11) of the outliers of each best alignment 3D-

Table 11. The 2-Substituted (5-Methy-11-ethy1)dipyridodiazepinone (II) Training Set and Corresponding WT and Y181C RT Inhibition Measures, $-\log(IC_{50})$

No.	Ar	WT-RT - log (IC_{50})	Y181C-RT - log (IC_{50})	No.	Ar	WT-RT - log (IC_{50})	Y181C-RT - log (IC_{50})
1	—[a]	7.19	6.96	21		7.07	6.68
2	—[a]	6.23	5.83	22		6.57	6.07
3		7.17	6.77	23		7.68	7.02
4		7.26	7.13	24		6.74	5.79
5		7.15	6.69	25	—[a]	6.03	5.96
6		6.95	6.80	26		7.00	6.25
7		6.86	6.38	27	—[a]	6.87	6.51
8		7.48	7.34	28	—[a]	7.66	6.70
9	—[a]	6.85	7.35	29	—[a]	7.55	7.55
10	—[a]	6.87	6.35	30	—[a]	6.68	6.70
11	—[a]	7.02	6.39	31		7.01	6.27
12	—[a]	7.62	6.85	32		6.47	6.54
13	—[a]	6.39	6.36	33		6.87	6.72
14	—[a]	6.97	6.59	34		6.64	5.84
15	—[a]	6.86	6.40	35	—[a]	6.89	6.44
16	—[a]	6.82	6.60	36		6.08	5.60
17	—[a]	7.41	6.85	37		6.82	6.68
18	—[a]	6.60	7.07	38		7.13	6.25
19		6.76	6.01	39		7.00	6.74
20		6.48	5.69	40		7.40	6.94

[a] Structures not yet released for publication.

QSAR. Each best WT-RT inhibition 3D-QSAR model has at least one outlier, and the R^2 value is moderately sensitive to alignment rule ranging from 0.57 to 0.76. There is no apparent pattern among the outliers as a function of alignment although compounds **18**, **25**, and **36** occur as outliers most often.

Only A_0, using the six IPEs given in part b of Table 12, were used as trial GCODs. The IPE(p+) and IPE(p−) indicates that the polar atom types were divided into positive and negative charge density IPEs. Five non-GCODs were considered in the initial GFA trial basis set: molecular weight, HOMO, LUMO, dipole moment, and log *P*. HOMO, LUMO, and dipole moment were computed for the lowest-energy structure found using the AM1 scheme.[22] log *P* was determined using the MedChem Software package.[23]

The best 3D-QSAR model for WT-RT inhibition is

$$-\log(IC_{50}) = 6.87 - 0.40GC1(a) - 0.15GC2(p-) + 0.09GC3(p-) - 0.13GC4(a) + 0.19GC5(n) + 0.05GC6(n)$$

$$N = 40 \quad R^2 = 0.76 \quad xv\text{-}R^2 = 0.66 \quad F = 17.2 \quad SD = 0.26 \tag{8}$$

Table 12. Description of the (RI) 4D-QSAR Analyses of the WT- and Y181C-RT Inhibitor Training Sets Given in Table 11 and Summaries of the Corresponding Best 3D-QSAR Models Found in Each Study

WT-RT Inhibitors

(a)

methodology parameters, see Table 2	value/choice	methodology parameters, see Table 2	value/choice
(X_L, Y_L, Z_L) s	(25 Å, 25 Å, 25 Å) 1 Å	E_s(I)	20 000(1)
T	300 K	N_a	20
R	none	N_d	205

(b) Interaction Pharmacophose Elements, IPEs

A_0 IPE(a)	A_0 IPE(hba)
A_0 IPE(p+)[a]	A_0 IPE(hbd)
A_0 IPE(p-)[a]	A_0 IPE(n)

(c) Summary of the Top Five (RT) 4D-QSAR Models Using Alignment 4 of Table 13

Study 1	
range in R^2 and (xv-R^2) in the top five GFA models	0.67−0.76 (0.57−0.66)
no. of unique outliers in all top five GFA models	3
no. of grid cells in all top five GFA models	10
no. of significant PLS components	3
no. of significant GFA descriptors in each of the top models	4−6
ΔR^2 and $\Delta(xv\text{-}R^2)$ for the top GFA model	0.006 (0.009)
no. of non-GCODs in all the top five GFA models	0

Y181C-RT Inhibitors

(a)

methodology parameters, see Table 2	value/choice	methodology parameters, see Table 2	value/choice
(X_L, Y_L, Z_L) s	(25 Å, 25 Å, 25 Å) 1 Å	E_s(I)	20 000(1)
T	300 K	N_a	20
R	none	N_d	205

(b) Interaction Pharmacophose Elements, IPEs

A_0 IPE(a)	A_0 IPE(hba)
A_0 IPE(p+)[a]	A_0 IPE(hbd)
A_0 IPE(p-)[a]	A_0 IPE(n)

(c) Summary of the Top Five (RT) 4D-QSAR Models Using Alignment 13 of Table 13

range in R^2 and (xv-R^2) in the top five GFA models	0.62−0.71 (0.51−0.61)
no. of unique outliers in all top five GFA models	5
no. of grid cells in all top five GFA models	11
no. of significant PLS components	3
no. of significant GFA descriptors in each of the top models	4−5
ΔR^2 and $\Delta(xv\text{-}R^2)$ for the top GFA model	0.006 (0.009)
no. of non-GCODs in all the top five GFA models	0

[a] The polar atom types were divided into those with positive and those with negative charge densities.

None of the non-GCODs considered in the trial GFA basis set of descriptors are found to be significant among the five best 3D-QSAR models as indicated in part c of Table 12 for WT-RT inhibition. Part c of Table 12 also records that ten distinct GCODs are distributed over the five best 3D-QSAR models. The linear cross correlation analysis for the ten distinct GCODs indicates that only GC1 and GC1O are highly correlated ($R = 0.81$), but they appear in different 3D-QSAR models. In order to determine if the top five 3D-QSAR models are providing common, or distinct, structure−activity information, the correlation coefficients of the residuals in the error (observed activity − predicted activity) between pairs of models have been computed and are given in Table 14a. Equivalent models are expected to have identical distributions in residuals of the error. Distinct models should have noncorrelated patterns in their residuals of fit (error). This type of data analysis has been suggested by Rogers[24] as a diagnostic to determine the subset of distinct models among a set of good models realized in GFA analysis.

The residuals in error have $R > 0.75$ (high cross-correlation) for model pairs [2−3], [2−5], [3−5], and [4−5]. Hence, it appears there are two relatively distinct models among the top five, namely model 1 and any one of the other four. In other words, the *manifold* of distinct and significant WT-RT inhibition 3D-QSAR models consists of two models, models 1 and 2. Model 2, which is the best of the four correlated (similar) models, is

$$-\log(\mathrm{IC}_{50}) = 6.84 - 0.43\mathrm{GC1(a)} - 0.21\mathrm{GC2(p-)} + 0.08\mathrm{GC3(p-)} + 0.15\mathrm{GC7(p-)} + 0.12\mathrm{GC8(hbd)}$$

$$N = 40 \quad R^2 = 0.70 \quad xv\text{-}R^2 = 0.60 \quad F = 15.8 \quad \mathrm{SD} = 0.28 \tag{9}$$

Figure 11 is a plot of the manifold of distinct WT inhibition 3D-QSARs, eqs 8 and 9. Part a of Figure 11 shows compound **8**, an active inhibitor [$-\log(\mathrm{IC}_{50}) = 7.55$] relative to the composite GCODs of eqs 8 and 9. Part b of Figure 11 is the same plot as part a, but for the inactive inhibitor compound **36** [$-\log(\mathrm{IC}_{50}) = 6.08$]. It is clear from Figure 11 that atoms/groups of compound **8** occupy a grid cell, GC6(n), that enhances activity, while compound **36** has atoms/groups occupying grid cells GC1(a) and GC4(a) which are predicted to diminish inhibition potency. The conformations of compounds **8** and **36** are the predicted active conformations using step 10 of the 4D-QSAR methodology.

The number of significant PLS components, see part c of Table 12 for WT-RT inhibition models, is three, while five to six GCODs are found in each of the top five GFA 3D-QSAR models. The GFA models seem to incorporate most of the information in the training set since ΔR^2 and $\Delta(xv\text{-}R^2)$ are both very small.

Table 11 also contains inhibition activity for Y181C-RT, a major mutant. 4D-QSAR analysis, identical in form and procedure to that done using the WT-RT training set, was carried out, and the findings are summarized in part c of the Y181C-RT inhibitor section of Table 12. The best 3D-QSAR model is

$$-\log(\mathrm{IC}_{50}) = 6.81 - 0.07\mathrm{GC1(n)} + 0.53\mathrm{GC2(n)} - 0.08\mathrm{GC3(a)} + 0.10\mathrm{GC4(a)} - 0.35\mathrm{GC5(n)}$$

$$N = 40 \quad R^2 = 0.71 \quad xv\text{-}R^2 = 0.61 \quad F = 16.7 \quad \mathrm{SD} = 0.27 \tag{10}$$

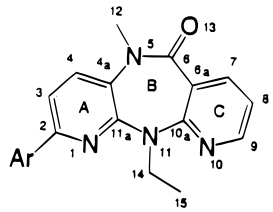
Equation 10 is based on alignment 13 of Table 13. This is significant to point out because the best Y181C-RT inhibition 3D-QSAR models are found for a different alignment from that for WT-RT inhibition. In other words, the same set of inhibitors are predicted to bind somewhat differently to the Y181C-RT enzyme than to the WT-RT enzyme. An attempt was made, in collaboration with colleagues at Boehringer Ingelheim, to map

(22) Stewart, J. J. P. *MOPAC Manual, MOPAC 6.0*; Frank J. Seiler Research Laboratory, United States Air Force Academy: 1990.

(23) Medicinal Chemistry Software, *Medchem Software Manual, Release 3.51*; Pomona College: Claremont, CA, 1987.

(24) (a) Rogers, D., private communication, 1996. (b) Rogers, D. Evolutionary Statistics: Using a Genetic Algorithm and Model Reduction to Isolate Alternate Statistical Hypotheses of Experimental Data. In *Proceedings of the Seventh International Conference on Genetic Algorithms*; East Lansing, MI; Morgan-Kaufmann, San Francisco, CA, 1997.

Table 13. Atom Numbering Used To Define the Three-Atom Alignments, Number of Occupied Cells for Each Alignment and Corresponding Number of GCODs for Each Alignment



alignment	atoms			cells	GCODs[a]	R^2 [b]	outliers
1	2	1	11a	664	3984	0.62	(28, 36)
2	13	6a	5	828	4968	0.71	(4, 18)
3	6a	10a	5	707	4242	0.57	(23)
4	9	10	11a	881	5286	0.76	(40)
5	1	11a	4a	622	3732	0.62	(16, 18)
6	5	10a	4a	641	3846	0.73	(12, 13, 18, 36)
7	8	9	10	974	5844	0.64	(25)
8	6	9	10a	694	4164	0.67	(13, 18)
9	5	6	6a	1155	6930	0.71	(25, 36)
10	5	6a	6	654	3924	0.61	(17, 18, 20)
11	2	6	6a	607	3642	0.58	(18, 23, 25, 36)
12	10a	11	9	716	4296	0.68	(20)
13	3	2	1	676	4056	0.68	(25, 36)
14	3	8	4	694	4164	0.65	(12)
15	8	3	9	653	3918	0.68	(9, 14, 25)
16	11a	5	8	686	4116	0.67	(15)
17	11a	10a	4a	832	4992	0.65	(13, 23, 40)
18	11	10	9	1112	6672	0.73	(13, 30)
19	11a	10a	8	835	5010	0.71	(27, 28)
20	10a	9	6	929	5574	0.75	(36, 40)

[a] Number of GCODs is the number of cells times six (the six types of occupancy). [b] The R^2 of the best WT 3D-QSAR model for each alignment is also given.

the locations of the grid cells of eqs 8 and 10 relative to the enzyme-bound inhibitor (compound **8**) geometries of WT- and Y181C-RT complexes. The resolution of the crystal structures is not sufficient at present to establish this mapping.

A residuals of error cross-correlation analysis of pairs of the five best Y181C-RT 3D-QSAR models was performed to identify the manifold of 3D-QSAR models. The cross-correlation matrix is given in Table 14b. Models [2−3] and [4−5] have high cross correlation coefficients. Therefore, the manifold of Y181C-RT 3D-QSAR models consists of three [1, 2, 4] of the top five models. The equations for models 2 and 4 are

$$-\log(\mathrm{IC}_{50}) = 6.35 - 0.07\mathrm{GC1(n)} + 0.23\mathrm{GC4(a)} + 0.09\mathrm{GC6(n)} + 0.45\,\mathrm{GC7(n)} - 0.05\mathrm{GC8(n)}$$

$$N = 40 \quad R^2 = 0.67 \quad xv\text{-}R^2 = 0.55 \quad F = 15.9 \quad \mathrm{SD} = 0.29 \tag{11}$$

[Model 2]

$$-\log(\mathrm{IC}_{50}) = 6.80 - 0.04\mathrm{GC1(n)} - 0.06\mathrm{GC3(a)} + 0.29\mathrm{GC9(p-)} - 0.45\mathrm{GC1O(n)} - 0.46\mathrm{GC11(n)}$$

$$N = 40 \quad R^2 = 0.68 \quad xv\text{-}R^2 = 0.53 \quad F = 14.4 \quad \mathrm{SD} = 0.29 \tag{12}$$

[Model 4]

The locations of the unique grid cells of eqs 10−12 relative to (a) compound **8** [$-\log(\mathrm{IC}_{50}) = 7.55$, active] and (b) compound **36** [$-\log(\mathrm{IC}_{50}) = 5.60$, inactive] are shown in Figure 12 and constitute the manifold set of GCODs. Figures 11 and 12 are identical in format except for the locations and orientations of the respective structures due to the two different alignments. There are additions, deletions and/or shifts in the grid cells between Figures 11 and 12. For example, grid cells GC1(a) and GC4(a) of Figure 11 merge and migrate "to become" GC3(a) in Figure 12. Grid cell GC5(n) appears in Figure 12 and has no equivalent in Figure 11. All of the GCODs near the 2-aryl substitutents for Y181C mutant activity have negative regression coefficients in the QSAR equations. Thus, increasing occupancy of any of these grid cells decreases activity, see Figure 12. One interpretation of this finding is that there is a "constant" binding interaction (a positive contribution to inhibition potency) in this region which is diminished by certain 2-aryl substitutents owing to their preferred average spatial locations.

Table 14. The Cross-Correlation Matrices of the Residuals in Error of the Five Best 3D-QSAR Models[a]

model	1	2	3	4	5
		(a) WT-RT			
1	1.00				
2	0.68	1.00			
3	0.70	0.88	1.00		
4	0.50	0.69	0.71	1.00	
5	0.68	0.95	0.93	0.77	1.00
		(b) Y181C-RT			
1	1.00				
2	0.45	1.00			
3	0.48	0.93	1.00		
4	0.52	0.42	0.35	1.00	
5	0.49	0.42	0.37	0.91	1.00

[a] Model 1 has the highest R^2 and xv-R^2 measures; model 5 the lowest of the top five models.

The number of significant PLS components and significant GCODs in the best GFA 3D-QSARs as well as ΔR^2 and $\Delta(xv\text{-}R^2)$ values for the Y181C-RT models are about the same as found for the WT-RT inhibition models. This information is in the Y181C-RT section of part c of Table 12.

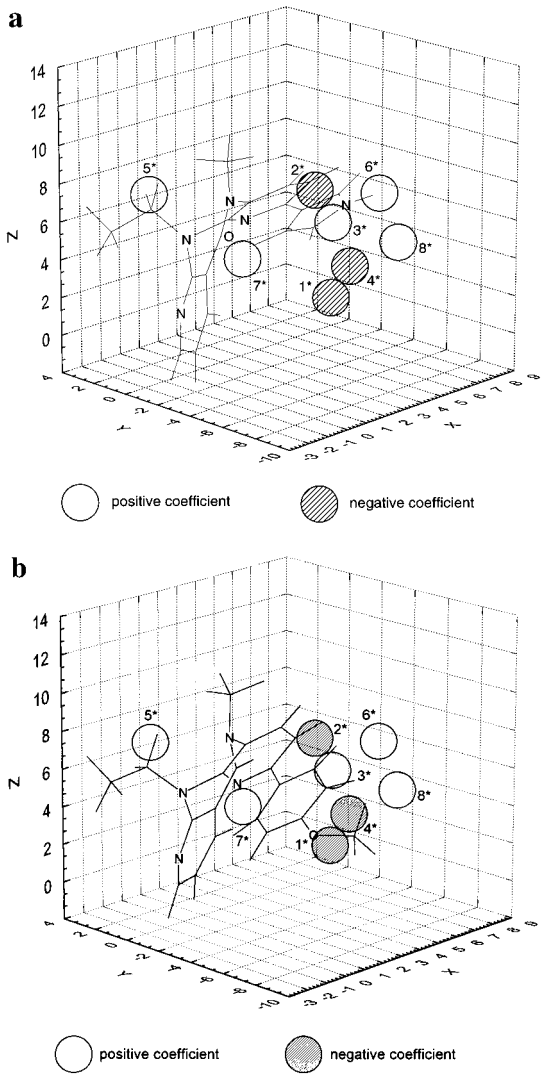


Figure 11. The distinct set of GCODs of the manifold 3D-QSAR model for WT-RT inhibition [eqs 8 and 9] plotted as spheres of 1 Å diameter relative to (a) compound **8** and (b) compound **36** in their respective postulated active conformations. The I* in each figure corresponds to GCI(x) in eqs 8 and 9. The open spheres correspond to grid cells where occupancy can enhance activity, while the hatched spheres correspond to grid cells where occupancy decreases inhibition potency.

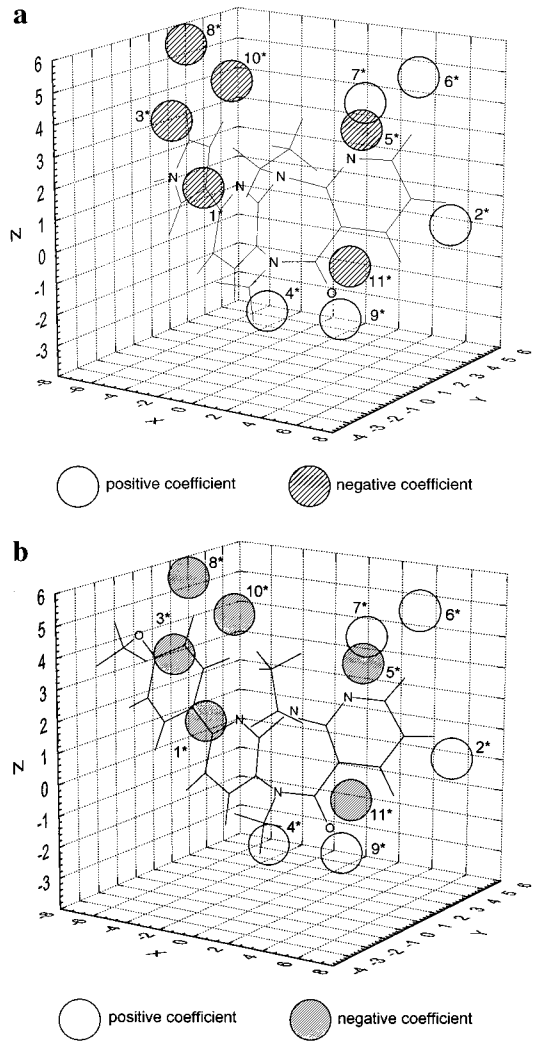


Figure 12. The same as Figure 11 but for the manifold 3D-QSAR model for Y181C-RT inhibition.

Discussion

Two unexpected findings from the (RI) 4D-QSAR application studies are (a) the enormous reduction in the number of independent variables (GCODs) realized in the generation of an optimized 3D-QSAR model and (b) that the optimum 3D-QSAR models can contain GCODs associated with "constant" part(s) in the structures of the analogs in a training set.

Data reduction by serial PLS and GFA analyses, Steps 6 and 7 in Table 1, reduces the number of GCODs used in any correlation equation from the 20 000−50 000 range to 10, or less, for the training sets studied. Moreover, the total number of unique GCODs found among the top ten 3D-QSAR models of any particular study is less than 20. The number of significant GCODs in the best 3D-QSAR models is found to be the same as, or slightly larger than, the number of significant components identified in the PLS phase of the data reduction. In other words, the GCODs in the optimized 3D-QSAR models capture about the same amount of information as is inherent to the significant PLS principal components. The small values of ΔR^2 and $\Delta(xv\text{-}R^2)$ in part c of Tables 5, 6, and 12 are indicative of this small loss in correlation information.

The number of significant 3D-QSAR GCODs increases as the flexibility and/or size of the molecules (alignment variability) in the training set increases. For example, the DHFR-benzylpyrimidine inhibitors have two principal torsion angle degrees of freedom plus some substituent torsion angles. The 3D-QSARs for this training set contains four to five GCOD terms. The $PGF_2\alpha$ analogs have eight to 12 torsion angle degrees of freedom, and the corresponding 3D-QSARs have six to ten GCOD terms.

The number of GCOD found in the best 3D-QSAR models in the three applications reported here each lead to a slightly oversubscribed problem given the respective number of analogs in each training set. Cross-validation readily deals with testing the robustness of the models and possible overfitting. This situation of marginal overfitting can be expected to persist in other 4D-QSAR applications. The number of essential GCODs is a characteristic of the chemistry and biology of the training set and only influenced by the number of observations (compounds) to the extent of the chemistry and biology sampled by the compounds.

All three of the (RI) 4D-QSAR studies reported in this paper have optimized 3D-QSAR models which include GCODs associated with the "constant" parts of the structures of the analogs. A grid cell near the 2-amino group of the pyrimidine ring, see eq 4 and Figure 4, is significant for the DHFR inhibitor 3D-QSAR; a grid cell near the 2-OH of the ring in the $PGF_2\alpha$ analogs, see Figure 9, is essential to a good 3D-QSAR, and

two grid cells associated with the dipyridodiazepinone tricycle and one with the 11-ethyl are found, see Figure 11, in the RT-WT inhibitor 3D-QSARs. This is the first time, as far we know, that a QSAR predicts a feature associated with the fixed structure of an analog series as being crucial to establishing the correlation equation. In essence, 4D-QSAR analysis can test if the independent binding site model[25] holds for a particular training set. In each of the three applications studies reported here, the independent binding site model does not rigorously hold. The variability in the GCODs associated with the constant part of an analog series is due to freedom in both conformation and alignment. Depending on the choice in the alignment rule, the variations in conformation of the constant chemical structure part of an analog can be shunted, or attenuated, relative to other analogs. In other words, 4D-QSAR analysis identifies slight differences among the binding modes of analogs as corresponding differences in both the location and/or occupancy values of the GCODs associated with constant chemical structures across the analogs.

The ensemble averaging of conformational behavior in constructing a 3D-QSAR using 4D-QSAR analysis permits an estimation of the effect of conformational entropy on activity. Currently, the entropy-activity interrelationship depends upon identifying a single active conformation of each analog from its CEP. This single active conformation is selected on the basis of the lowest-energy conformer state which maximizes the predicted activity using the best 3D-QSAR model. The role of conformational entropy on activity is then associated with the difference in predicted activity using the single, "active" conformation and the observed activity. The predicted activity of the single active conformer state is usually greater than the observed activity (see Figure 10 for the analysis of $PGF_2\alpha$ analogs). Thus, the predicted activity of the active conformation of an analog can be viewed as the *intrinsic activity* of the analog in the absence of its entropy or its *enthalpic activity*. In turn, it can be argued that analogs in the training set with the highest intrinsic activities may be better design templates to realize favorable ligand−receptor interactions than simply adopting analogs with the highest observed activities.

The predicted "active" conformation of each compound in a training set, and the corresponding preferred alignment realized from a 4D-QSAR analysis can be used to define the input for other 3D-QSAR methods. Thus, (RI) 4D-QSAR analysis could be used as a "preprocessor" to a CoMFA and provides the requisite input molecular geometries and alignment. Perhaps more exciting, the CoMFA field descriptors can be readily combined with the (RI) 4D-QSAR GCODs to generate an expanded pool of trial descriptors to use in constructing a 3D-QSAR model.

(RI) 4D-QSAR analysis, through the use of IPEs, allows each of the compounds in a training set to be partitioned into sets of structure types and/or classes with respect to possible interactions with a common receptor. In a 4D-QSAR analysis, sets of GCODs, defined by the IPEs, are simultaneously mapped into a common grid cell space. Thus, the combined PLS-GFA mapping of the GCODs to activity space includes not only attributes of the whole molecule over space but also spatial attributes of all the component IPEs of a compound. Hence, if some parts of a ligand are present only to correctly position and orient other ligand groups for receptor binding, 4D-QSAR analysis has the capacity to identify this feature.

Each application of (RI) 4D-QSAR analysis reported here has resulted in 3D-QSAR models with good statistics of correlation fit. However, it is still important to know how good these models are to QSARs constructed in other ways. The DHFR 3D-QSAR model reported here ($R^2 = 0.957$, xv-$R^2 = 0.885$, four-terms) is clearly superior to a model constructed for the same training set ($R^2 = 0.913$, xv-$R^2 = 0.816$, five-terms) using tensor-based 3D-QSAR.[20] For the $PGF_2\alpha$ training set, collaborators at Procter and Gamble Pharmaceuticals[26] have been able to generate a QSAR model of comparable statistical fit to that reported in eq 5 using parametric QSAR design. Cardozo[27] used a variation of molecular shape analysis[1,3] to construct a 3D-QSAR nearly comparable in statistical significance to the best model for WT-RT inhibition, eq 9, developed for the training set reported in Table 11.

Overall, 4D-QSAR not only would appear to yield 3D-QSAR models at least as good as can be generated using other methods but also provide added value information not realized by other methods.

Interpretation of multiple "good" models from a GA analysis remains problematic irrespective of whether or not the application is to a QSAR problem. Rogers' [24] approach of computing and comparing the linear cross-correlation coefficients of the residuals of the error between pairs of the good GA models provides a basis for organizing and exploiting information from multiple models. Two models that have about the same residual errors over the training set ($R \approx 1$) very likely express the same information. Conversely, two models with different residuals of the error over the training set ($R < 0.5$) may provide different information. Thus, the composite set of good models, whose residuals of the error are poorly correlated, are a *manifold model* of the structure−activity profile inherent to the training set.

Acknowledgment. This work was supported in part by an NSF SBIR Phase I Grant (No. DMI-9560439) to The Chem21 Group, Inc. Resources of The Laboratory of Molecular Modeling and Design were used to perform the reported work. M.A. is grateful to the CNPq of Brazil for fellowship support. The UIC investigators appreciate financial assistance from The Chem21 Group, Inc.

JA9718937

(25) Foye, W. O. *Principles of Medicinal Chemistry*, 2nd ed.; Lea and Febiger: Philadelphia, PA, 1989; Chapter 2.
(26) Stanton, D. T., private communication, 1996.
(27) Cardozo, M. G., private communication, 1996.
(28) Koehler, M. G.; Hopfinger, A. J. *Polymer* **1989**, *30*, 116.
(29) Baker, D. J.; Beddell, C. R.; Champress, J. N.; Goodford, P. J.; Norrington, F.E. A.; Smith, D. R.; Stammers, D. K. *FEBS Lett.* **1989**, *126*, 49.