Targacept Active Conformation Search: A New Method for Predicting the Conformation of a Ligand Bound to Its Protein Target

Josef Klucik,* Yun-De Xiao, Phillip S. Hammond, Rebecca Harris, and Jeffrey D. Schmitt

Targacept Inc., 200 East First Street, Suite 300, Winston-Salem, North Carolina 27101

Received April 8, 2004

Targacept active conformation search (TACS) is a novel variation of well-established three-dimensional quantitative structure−activity relationship methodologies that seeks to determine probable conformation(s) of ligands bound to their protein targets. A combination of affinity or activity data and energetically accessible conformational ensembles, each conformer described by three-dimensional (3-D) sensitive descriptors, forms the basis of the TACS data model. Recursive pruning is used to reduce the size of both the conformational ensemble and the descriptor space until the TACS data model contains just enough information to determine probable conformation(s) of ligands bound to their protein targets. The TACS algorithm is comprised of five components: (1) conformational ensemble generation, (2) 3-D sensitive descriptor calculation, (3) ensemble descriptor preprocessing, (4) model generation, and (5) prediction of bound conformation(s). Significantly, this method precludes the need for subjective or objective molecular alignment. We report the application of this technique to five benchmark protein−ligand couples where the conformation of a bound ligand has been previously established using X-ray crystallography: 9-*cis*-retinoic (**1**) and 9-*trans*-retinoic acid (**2**), both agonists for the retinoic acid receptor γ, compounds KH1060 (**3**) and MC1288 (**4**), which bind to the vitamin D3 receptor, and R04 (**5**), an inhibitor bound to human rhinovirus 14 thermolysin. The binding conformations predicted by TACS were compared to the crystallographic structures extracted from their respective binding sites using root-mean-squared deviation (rmsd) criteria. Three of the conformations found using TACS were within crystallographic error. 9-*cis*-Retinoic acid, 9-*trans*-retinoic acid, and MC1288, when superimposed on their crystallographic structures, gave rmsd values of 0.22, 0.17, and 0.34 Å, respectively. The rmsd values for KH1060 (1.54 Å) and R04 (1.01 Å) were larger but still reasonable.

Introduction

Most strategies for the design of new drug candidates fall into one of two categories: *structure based design,* where the three-dimensional (3-D) structure of protein target is known from either crystallographic or high-resolution NMR studies, or *ligand based design,* where the 3-D structure of the target is unknown. In the absence of 3-D protein structure, it is difficult to determine the bound conformation of ligands. Therefore, a method that generates the probable bound conformation of ligand(s) would enhance ligand-based drug design, pharmacophore hypothesis generation, and understanding of ligand−protein interactions.

Because it is a well-known fact that 3-D (spatial) properties of biological molecules govern their biological behavior, the combination of structure−activity relationship (SAR) data, conformationally sensitive 3-D descriptors, and a representation of energetically accessible conformational ensembles may contain information about the most likely bound conformation of compounds bound to their protein target. TACS has been designed to find this intrinsic information, if it exists.

A number of approaches have been developed for solving the problem of generating bound conformation hypotheses in the absence of protein structure. Some methods focus on using hypothetical receptor binding site topology as a basis for evaluating the "fit" and hence binding conformations of new chemical entities; methods in this category include active-site modeling,[1−5] receptor surface modeling,[6,7] and hypothetical active-site lattice modeling.[8,9] The problem with these approaches is that the generated receptor topologies often have little structural resemblance to their natural counterparts.[10]

Other methods focus on consensus ligand geometry, using either feature-based (pharmacophore) or atomistic representations. Pharmacophore mapping techniques include DISCO[11] and HipHop/HypoGen.[12] Examples of methods that use explicit representations of molecular structure include the active-site approach, based on Marshall's active analogue approach,[13] and ensemble distance geometry.[14,15] A notable example, bearing particular import to this paper, is molecular shape analysis (MSA) developed by Hopfinger[16] where the quantitative characterization of molecular shape and its relation to biological activity are used to form quantitative structure−activity relationship (QSAR) equations. The chief weakness of MSA is the assumption that the lowest-energy conformer of the most active (or highest-affinity) compound represents the bound conformation. Furthermore, if some a priori evidence exists that this is not the case, then the user must subjectively choose which conformation of the highest-affinity compound

* Author to whom correspondence should be addressed. Phone: (336) 480-2127. Fax: (336) 480-2107. E-mail: jklucik@targacept.com.

10.1021/jm049729z CCC: $27.50
Published on Web 11/26/2004

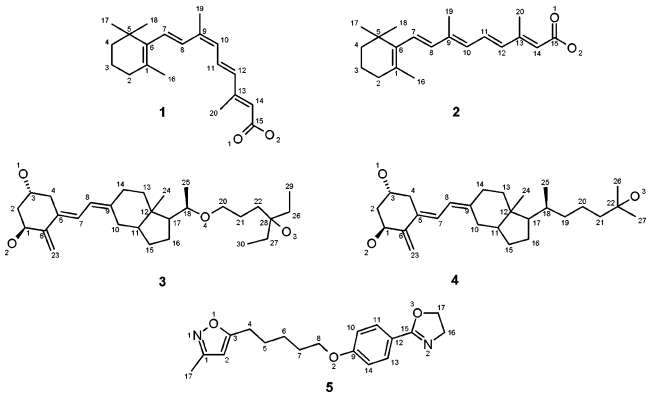

Figure 1. 9-*cis*-Retinoic (**1**), 9-*trans*-retinoic Acid (**2**), KH1060 (**3**), MC1288 (**4**), and R04 (**5**).

will serve as the shape analysis template. In this paper, we describe a methodology similar to MSA that generates predictive QSARs not as an end but a means to generating a robust hypothesis of bound ligand conformation(s). TACS has the added advantage of not requiring molecular alignment or subjective involvement of the investigator.

Because it is a well-known fact that 3-D (spatial) properties of biological molecules govern their biological behavior, the combination of SAR data, conformationally sensitive 3-D descriptors, and a representation of energetically accessible conformational ensembles may contain information about the most likely conformation of compounds bound to their protein target. The goal of TACS is to identify this intrinsic information and use it in identification of the most probable bound conformation of a ligand to a receptor.

Methods

Data Sets. Data sets obtained from various sources have been used for developing TACS models (Figure 1) and span three different biological domains. In total, five-benchmark protein−ligand couples, where the conformation of a bound ligand has been previously established using X-ray crystallography, were employed to demonstrate this method.

(i) Retinoic Acid Receptor γ. A set of 14 retinoic acid receptor (RARγ) agonists have been collected from the published work described by Dawson et al.,[17] with in vitro activities of RARγ expressed as EC_{50} values ranging from 2 to 2300 nM. They are SR-11004, SR-11202, SR-11215, SR-11224, SR-11201, SR-11332, SR-11225, SR-11247, SR-11245, SR-11251, SR-11249, SR-11269, 9-*trans*-retinoic acid (*t*-RA), and 9-*cis*-retinoic acid (*c*-RA). The first 12 compounds plus *t*-RA were used as a training set to predict the bound conformation of *c*-RA; conversely, the first 12 compounds plus *c*-RA were used to predict the bound conformation of *t*-RA. The crystallographic structures of *c*-RA[18] and *t*-RA,[19] used for evaluation of TACS predictive ability, were extracted from the cocrystallized protein structure of RARγ (PDB ID (resolution): 2LBD (2.0 Å) and 3LBD (2.4 Å)).

(ii) Vitamin D3 Receptor. The vitamin D3 (VD3) receptor training set used to predict the bound conformation of the novel ligand KH1060 is comprised of seven compounds (VD3, KH-1139, MC-1084, MC-1301, MC1292, MC1288, and MC1627) where ED_{50} values (nM), the effective dose required to reach 50% of maximal transcriptional activity, span the range of 0.0006−3 nM.[20−22] Likewise, the training set for MC1288 contained seven molecules (VD3, KH-1139, KH1060, MC-1084, MC-1301, MC1292, and MC1627). The crystallographic structures for KH1060 and MC1288 were extracted from cocrystallized vitamin D3 receptor[23] (PDB ID (resolution): 1IE8 (1.52 Å) and 1IE9 (1.40 Å)) and used as the basis of root-mean-squared deviation (rmsd) comparison with the TACS predicted conformation.

(iii) Human Rhinovirus 14 Thermolysin. The human rhinovirus 14 (HRV14) thermolysin training set consisted of eight compounds (R06, R07, RM2, RR1, RS1, RS3, RS5, and RS8) possessing an activity range of 30−2400 nM. In this case, the activity of a given compound is defined by the concentration required to reduce plaque count by a factor of 2.[24] The crystallographic structure of the test molecule, R04, was obtained by extraction from cocrystallized HRV14[25] structure (PDB ID (resolution): 2R04 (3.00 Å)) as described above.

In the following section, we describe in detail the five components of the TACS algorithm: (1) conformational ensemble generation and processing, (2) 3-D sensitive descriptor calculation, (3) ensemble descriptor processing, (4) training set assembly and model generation, and (5) prediction of bound ligand conformation. A schematic description of the TACS algorithm is given in Figures 2 and 3.

TACS Component 1: Conformational Ensemble Generation and Processing. All chemical structures were created using the Sybyl molecular modeling package[26] and assigned Gasteiger−Huckel charges. In this paper we refer to conformational ensembles using bracket notation, "⟨ ⟩", where $\langle \boldsymbol{M}_n \rangle_h$ refers to n conformations of molecule h (Figure 2). Each member of a TACS data set likely possesses different numbers of conformations (i.e., the n value for molecule 1 is not likely to equal the n value for molecule 2 and so on). The conformational ensemble for each compound ensemble $\langle M_n \rangle_h$ was generated by conducting two types of extensive conformational searches: (1) a simulated-annealing search with 10 cycles of heating to 1000 K for 1.5 ps, followed by annealing to 200 K over 1.0 ps, with snapshots taken every 10 fs at all temperatures. The resulting conformations are then minimized for 100 iterations using the Tripos force field, with default parameters, BFGS method and dielectric constant = 1, (2) an internal coordinate Monte Carlo search with Go−Scheraga ring deformation (implemented in Sybyl's random search) with the following parameters, max cycle = 4000, energy cutoff = 3 kcal, rms threshold = 0.1, convergence threshold = 0.005, max hit = 6, check chirality = on, maintaining the same energy minimization settings as described in method 1. All of the ligands were subjected to both types of conformational searches with exception of SR-11004, SR-11202, SR-11215, SR-11224, SR-11201, SR-11332, SR-11225, SR-11247, SR-11245, and SR-11251 from the RA training ensemble; due to the rigid nature of these compounds, only the random search was used. Conformations generated from both searches were combined for further processing.

A conformational filter was then applied to remove equivalent conformations (rmsd < 0.3) as well as high-

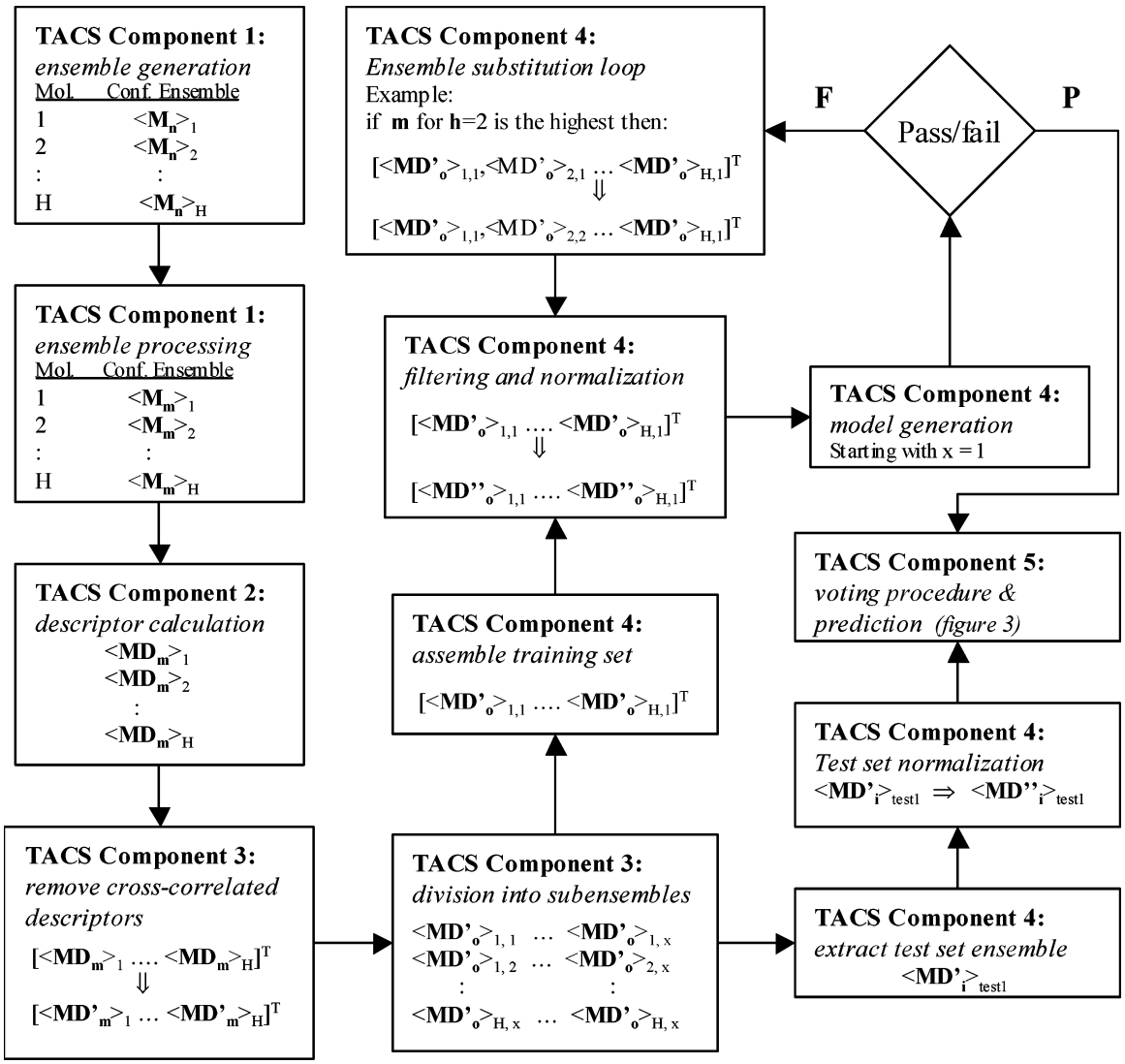

Figure 2. The TACS algorithm. ⟨M⟩ = conformational ensemble for a given molecule, H = number of molecules in dataset, h = molecule index, n = original number of conformations of molecule h, m = reduced number of conformations of molecule h, D = descriptor space, D′ = reduced descriptor space, $o \sim 10\%$ of m, x = subensemble (integer from 1 to 10), D″ = further reduced descriptor space, $\langle M_i \rangle_{test1}$ = ensemble for test molecule 1, i = number of conformations in the test set.

energy conformations using a conformational energy cutoff of 13 kcal/mol above the global minima. If the number of remaining m (i.e., $n \rightarrow m$) conformations for a given ligand is more then 7000, then further reduction is accomplished by removing conformations with rmsd < 0.5. In keeping with the notation established above, $\langle M_m \rangle_h$ represents the filtered conformational ensemble of molecule h containing m conformations (again where m is different among all molecules).

TACS Component 2: 3-D Sensitive Descriptor Calculation. Atom-centered partial charges are calculated for all m conformations in each ensemble via Mulliken population analysis using the AM1 semiempirical Hamiltonian.[27] A collection of 3-D sensitive descriptors are calculated using the following software packages: Volsurf,[28] Sybyl, Cerius2,[29] and QSARIS.[30] Descriptors include highest-occupied molecular orbital (HOMO) and lowest-unoccupied molecular orbital (LUMO) energies, Jurs and Shadow indices, comparative molecular moment analysis (CoMMA) descriptors, sums of absolute charge values, molecular dipole, largest positive charge, polarizibility, specific polarizibility (QSARIS), solvent excluded volume, and surface area, which represents the accessible surface (in Å^2) traced out by solvent.

To further describe the molecular shape of each conformation, the following custom descriptors are also calculated: the largest distance between two heavy atoms (Å), pseudotorsions between these largest-distance atoms (i.e., heavy atoms used for distance calculation and their immediate heavy atom neighbors), rugasity, and globularity.[31] In all, 128 3-D sensitive descriptors are generated for every conformation in each ensemble, leading to $\langle MD_m \rangle_h$, where MD refers to the conformational ensemble bioactivity and descriptor matrix. Each row vector contains an index denoting conformer number, a bioactivity value followed by descriptors.

TACS Component 3: Ensemble Descriptor Preprocessing. For a data set containing H molecules, individual conformational ensemble matrixes $\langle MD_m \rangle_h$ are then assembled to form a large matrix $[\langle MD_m \rangle_1 \cdots \langle MD_m \rangle_H]^T$. Then, to reduce descriptor space dimensionality in the concatenated ensembles, highly cross-correlated descriptors ($R^2_{XC} > 0.7$) are eliminated, giving rise to $[\langle MD'_m \rangle_1 \cdots \langle MD'_m \rangle_H]^T$, where D′ refers to the

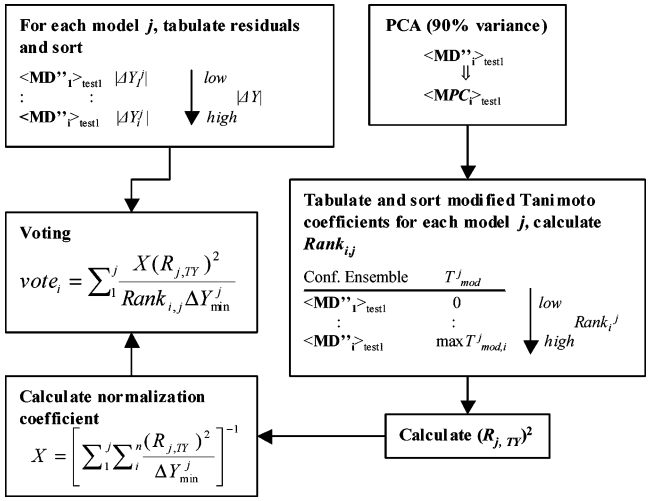

Figure 3. TACS voting scheme. Throughout j refers to the jth statistical method. $\Delta Y_i^j = Y_i^j - Y_{obs}$ = residuals, PCA = principal component analysis, $\langle MPC_i \rangle$ = principal component matrix for conformational ensemble, X = normalization coefficient, $Rank_{i,j}$ = rank of the ith conformation in the test ensemble, ΔY^j_{min} = smallest absolute value of a prediction residual, T^j_{mod} = modified Tanimoto coefficient, $(R_{j,TY})^2$ = squared correlation between T^j_{mod} and ΔY_i^j.

reduced descriptor space. The matrix $[\langle MD'_m\rangle_1 \cdots \langle MD'_m\rangle_H]^T$ is then separated into individual ensembles $\langle MD'_m\rangle_h$ and principal components are then generated for each $\langle MD'_m\rangle_h$ (explaining 90% variance). For this purpose, descriptors were temporarily normalized to unit variance and mean centered. A table of principal component (PC)-based Euclidian distances between pairs of conformations is generated and stored for later use in diversity analysis.[32] The conformational ensembles for each ligand are then evenly divided into a maximum of 10 subensembles $\langle MD'_o\rangle_{h,x}$ (x = subensemble index; when $x = 10$ the number of conformations in each subensemble is $o \approx m/10$) with decreasing diversity, where the subensemble with $x = 1$ is the most diverse. For very rigid ligands, for instance where $m < 10$, all conformations are assigned to one subensemble. Finally, ligands for which the bound conformation is sought ("test set" ligands) are processed differently; only one subensemble $\langle MD'_i\rangle_{test1}$ representing the 20% most diverse conformations is used. If the test ensemble contains <200 conformations, then only 50% of the most diverse conformations are used. Any number of test molecules (test1, test2, ...) can be evaluated.

TACS Component 4: Training Set Assembly and Model Generation. The creation of the training set is an iterative process designed to maximize information content and is implemented as follows:

(a) Join together the first most diverse ($x = 1$) subensembles $[\langle MD'_o\rangle_{1,1} \cdots \langle MD'_o\rangle_{H,1}]^T$ of all H ligands excluding the test set; this matrix is referred to as the "initial training set".

(b) Eliminate highly correlated descriptors ($R^2_{XC} > 0.7$) from the initial training set, followed by mean centering and normalization to unit variance. This process results in the reduced matrix of the initial training set $[\langle MD''_o\rangle_{1,1} \cdots \langle MD''_o\rangle_{H,1}]^T$.

(c) Using the three QSAR methodologies, build separate models that correlate biological activity/affinity and descriptor space in the $[\langle MD''_o\rangle_{1,1} \cdots \langle MD''_o\rangle_{H,1}]^T$ matrix. In this study, three statistical methods, partial least-squares (PLS, components = 3),[33] multiple linear regression,[34] and forward stepwise regression (maximum steps = 10 and $F = 4.00$)[35] are employed. Models are optimized by an iterative process, described below, using the following statistical termination criteria: if $\{R^2_{curr} > 0.5\}$ and $\{R^2_{curr} - \rho^2_{rnd} > 0.2\}$, then the algorithm terminates resulting in a "final model". Here, R^2_{curr} is the correlation coefficient for the most recently evaluated training set, and $\rho^2_{rnd} = [\sum_{j=1}^{19} R^2_{rnd,k}/19] + SD_{rnd}$ $R^2_{rnd,k}$ is the squared correlation coefficient of the kth randomization trial (randomization trials are conducted at the 95% certainty level, 19 repetitions), and SD_{rnd} is the standard deviation of the 19 trials.

(d) Evaluate individual multiple linear regression (MLR), forward stepwise regression (FSR), and PLS models using the termination criteria. *Note that different models may be derived from different training sets.* Any models not meeting the said criteria are subjected to the following iterative procedure: the subensemble of the molecule with the most conformations (m) is substituted with the second most diverse subensemble ($\langle MD'_o\rangle_{h,2}$) to form a new training set. For instance, if the second molecule of a given training set possesses the highest value of m, then $[\langle MD''_o\rangle_{1,1}, \langle MD''_o\rangle_{2,1} \cdots \langle MD''_o\rangle_{H,1}]^T$ becomes $[\langle MD''_o\rangle_{1,1}, \langle MD''_o\rangle_{2,2} \cdots \langle MD''_o\rangle_{H,1}]^T$. If the substitution does not lead to model(s) that satisfy the termination criteria, then revert to the original ensemble and the substitution continues as described above for the second, most conformationally populated ligand. If all options for ligands that have more then one subensemble are exhausted, then the procedure is repeated by the simultaneous substitution of two ligand subensembles and so on. If termination criteria are not met after all subensemble combinations are exhausted, the algorithm is aborted (never encountered in the author's experience).

TACS Component 5: Prediction of Bound Conformation(s). Once satisfactory models are generated, the descriptor space of each test ligand ensemble is normalized using the parameters generated during normalization of the training set that passed termination criteria during training, giving $\langle MD''_i\rangle_{test1}$. This procedure, although not guaranteeing consistent normalization across training and test descriptor space, has the practical advantage of facilitating continued use of successful models. Test molecules and their requisite descriptors are added to each of the three training sets; we refer to these as the "test sets". Each of the models is then used to predict the biological activity of every test molecule conformer in the test set ensemble, and the results are tabulated.

An evaluation procedure, which takes into account the *errors in biological activity prediction* and *the relationship of these errors to the descriptor space*, was devised to determine the most probable binding conformation of the test molecule(s) (Figure 3). The evaluation procedure is as follows:

(a) For every model, calculate the residuals between predicted and observed activity, $\Delta Y_i^j = Y_i^j - Y_{obs}$ for each conformation of each test set molecule. This equation is generalized to allow for using more than $j = 3$ statistical methods.

Table 1. Voting Results for Three Best Conformations from Each Test Set Ensemble

name	$rank_{PLS}$	$vote_{PLS}$	$rank_{MLR}$	$vote_{MLR}$	$rank_{FSR}$	$vote_{FSR}$	average vote	%vote	final rank
CIS00053	1	24	2	6.8	1	62.9	31.4	20.4	1
CIS00054	2	12	3	4.53	2	31.5	16.1	10.4	2
CIS00052	3	8.1	1	13.6	3	21	14.2	9.25	3
TRA00080	1	92.9	4	0.01	1	7.2	33.4	20.1	1
TRA00043	66	0.1	70	0	2	46.5	15.5	9.34	2
TRA00049	4	1.8	16	0	3	31	10.9	6.57	3
KH2_106000001	1	49	8	3.09	3	8.71	20.3	18.3	1
KH2_106000014	2	25	5	4.95	4	6.54	12	10.3	2
KH2_106000010	9	5.5	1	24.7	10	2.61	10.9	9.06	3
MC12882	1	58	4	1.28	1	36.6	32.1	25.7	1
MC2−128800010	2	29	10	0.51	4	9.14	12.9	10.4	2
MC2−128800008	3	19	8	0.64	2	18.3	12.8	10.3	3
R04_0052	40	0.9	13	2.11	1	38.3	13.8	14.7	1
R04_0029	1	34	40	0.69	49	0.78	11.9	8.05	2
R04_0026	18	1.9	1	27.5	24	1.6	10.3	7.43	3

(b) Assign a rank (denoted by $1 \geq \mathrm{Rank}_{i,j} \geq i$) to each of the i conformers in the test ligand ensemble, where the conformer with the lowest $|\Delta Y_i^j|$ value is set to 1; we refer to this as the base conformer.

(c) Generate principal components with 90% variance explained on the test ligand ensemble alone, leading to $\langle \mathrm{MPC}_i \rangle_{\mathrm{test1}}$.

(d) For each of the three statistical models, calculate a modified Tanimoto coefficient (T^j_{mod}) between each conformation in the test set and the base conformer in principal components space.

$$T^j_{\mathrm{mod}} = N_{i,\mathrm{base}}/(N_i + N_{\mathrm{base}} - N_{i,\mathrm{base}}) \qquad (1)$$

where

$$N_i = e^{\Sigma_k w_k \mathrm{PC}_{ik}}, \quad N_{\mathrm{base}} = e^{\Sigma_k w_k \mathrm{PC}_{\mathrm{base},k}}$$

and

$$N_{i,\mathrm{base}} = e^{\Sigma_k w_k \mathrm{PC}_{ik}} - e^{\Sigma_k w_k \mathrm{PC}_{\mathrm{base},k}}$$

Here, PC_{ik} refers to the kth principal component of the ith conformation; likewise $\mathrm{PC}_{\mathrm{base},k}$ refers to the base conformation; w_k is a normalization factor corresponding to the percent variance explained by the kth PC.

Then calculate the squared correlation coefficient, $(R_{j,\mathrm{TY}})^2$, between these *Tannimoto distances* and *the residuals* $\Delta Y_i^j = Y_i^j - Y_{\mathrm{obs}}$; tabulate $(R_{j,\mathrm{TY}})^2$. A normalization coefficient X is then defined as

$$X = \left[\sum_1^j \sum_i^n \frac{(R_{j,\mathrm{TY}})^2}{\Delta Y^j_{\mathrm{min}}} \right]^{-1} \qquad (2)$$

where $\Delta Y^j_{\mathrm{min}}$ is the value of the smallest residual.

(e) By use of the assigned rank, the TACS voting equation is defined as

$$\mathrm{vote}_i = \sum_1^j \frac{X(R_{j,TY})^2}{\mathrm{Rank}_{i,j} \Delta Y^j_{\mathrm{min}}} \qquad (3)$$

(f) Use eq 3 to assign vote_i to each test set conformation and identify the probable binding conformation based on the highest voting score. The voting equation assigns weight to each statistical method for its contribution to decision process.

Table 2. Superimposition of Heavy Atoms of TACS-Picked Conformations on Their Structural Counterparts

name	rmsd for heavy atoms (Å)	standard deviation	mean
c-RA	0.22303	0.15976	0.16
t-RA	0.17529	0.10131	0.14
KH1060	1.54597	0.79541	1.33
MC1288	0.34963	0.1826	0.3
R04	1.01321	0.46808	0.9

Results and Discussion

TACS Predictive Power. TACS results from the five test ligands are given in Table 1. Vote-winning conformations (i.e., most likely bound conformations) were superimposed on their respective structural counterparts extracted from cocrystallized ligand−protein assemblies. The heavy-atom rmsd of pairwise atomic distances ranges from 0.175 Å for RA to 1.55 Å for KH1060 (Figure 4, Table 2).

In the case of *t*-RA (**2**) (rmsd = 0.175 Å), TACS correctly identified the hydrophobic cyclohexene ring position and puckering, as well as the orientation of the carboxylate moiety with respect to the ring; the rmsd value is within the crystallographic error. Only slight deviations were noted in the unsaturated chain, where the chain in the crystal structure appeared to be flat with average torsion around the single bonds of 176.68°, as compared to 168.70° in the TACS structure. The distances between the carbonyl oxygen and C1 were 11.85 and 11.79 Å for the crystal structure and TACS structure, respectively.

Similar to the results above, the TACS predicted conformation of *c*-RA closely matches the crystallographic structure with an rmsd of 0.223 Å. Average torsion around the single bonds in the unsaturated chain of the crystal structure is 178.77° as compared to the 165.23° in the TACS conformation. The biggest contribution to the rmsd was the C13−C14−C15−O2 bond torsion, which erred by 16.43°. Compensating for this is the C1−O1 distance difference of only 0.08 Å that places the hydrophilic and hydrophobic moieties at the correct relative configuration.

The variety of structural features in the vitamin D3 ligand, MC1288, such as ring systems, conjugated alkene, and flexible saturated chain, did not present any problems for TACS conformational elucidation. The rmsd superposition of 0.349 Å is within the experimen-

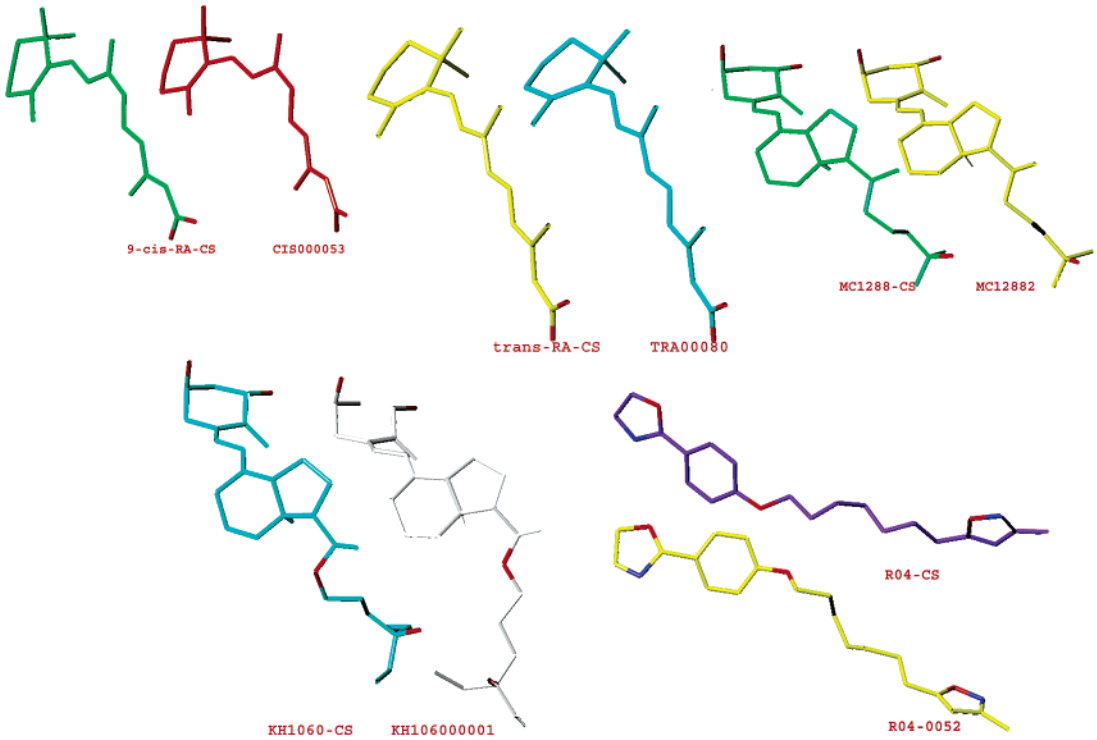

Figure 4. Superimposition of resulting TACS conformations on their respective cocrystallized structural counterparts. Extracted crystallographic structures are annotated with suffix −CS.

tal error for the crystallographic structure. Both ring systems (cyclohexane and bicyclo[4.0.3]) were predicted with correct configurations including correct spatial orientation. The majority of the rmsd arose due to deviations in the flexible saturated chain, specifically torsion around the C18−C19−C20−C21 bond, where the crystallographic structure "*cisoid*" arrangement was not correctly identified (torsions were −64.86° and −95.94° for crystal and TACS structures, respectively). The prediction of pharmacophoric features such as, O2 → O3 distance (crystal structure = 13.64 Å and TACS = 12.88 Å), O1−O2−O3 pseudoangle (crystal structure = 105.76° and TACS = 115.78°), and O1−O2−C22−O3 pseudotorsion (crystal structure = −71.34 and TACS = −99.49°) suggest the robustness of the TACS algorithm.

However, TACS applied to another vitamin D3 ligand, KH1060, resulted in a larger rmsd of 1.546 Å. Although TACS correctly identified the orientation of the ring systems and their substituents, the algorithm had marginal success in orienting the saturated chain. This was due to deviation of the C18−O4−C20−C21 bond torsion (24.84 and −170.45° for TACS and crystal structure, respectively), resulting in a different orientation that the hydrophobic constituents of the molecule undertake on the paths from O1 → O3 and O2 → O3 hydroxyl moieties. The protein−ligand hydrogen bonding interaction present in the crystal structure could be reproduced by docking the TACS structure in the binding pocket for most of the ligand's hydroxyl groups. However, the H bond of the aliphatic OH group could not be reproduced. The His397 moiety of the protein interacts with the aliphatic OH of the ligand's crystal structure, whereas the His308 interacts with the TACS ligand (data not shown).

TACS prediction of the HRV14 ligand, R04, resulted with overall orientation of the pharmacophoric elements (methylisoxazole and phenyloxazole) similar to the orientation of the crystal structure. TACS predicted distances between N1 → O3, and O1 → N2 were 16.76 and 14.83 Å, respectively, as compared to 16.10 and 15.64 Å in the crystal structure, observations well within the crystallographic structure resolution (3.00 Å). Deviations from the crystal structure arise from improper rotation at the C9−O2−C8−C7 torsion, resulting in the saturated chain pointing 129° away from the proper orientation (after the imidazole and phenyl groups were superimposed, data not shown). Docking the TACS ligand into the binding site resulted in moderate overlap with the R04 crystal structure orientation. In addition, TACS was able to capture the only H bond (N1−Asn219) present in the cocrystallized complex.

Assessment of Predictive Robustness. The TACS method correctly identified binding conformations in three out of five test case ligands (*c*-RA, *t*-RA, and MC1288) with very low rmsd, predicted correct orientation and critical pharmacophore interactions for R04, and performed only marginally for KH1060. The relative ability of TACS to predict bound conformations depends on (1) the robustness of the TACS algorithm, particularly as related to the statistical method(s) employed, (2) the D-space information content relative to moieties critical in molecular recognition, and (3) bias in conformational ensembles due to force-field inaccuracies.

A good indication of TACS performance may be the correlation coefficient between the above-described *Tanimoto distance* and the *biological activity residuals* $[(R_{j,\mathrm{TY}})^2]$. For example, in the *c*-RA, *t*-RA, and MC1288 activity predictions, at the least one statistical method

Table 3. Correlation Coefficients $(R_{J,TY})^2$

name	$(R_{PLS,TY})^2$	$(R_{MLR,TY})^2$	$(R_{FSR,TY})^2$
c-RA	0.51	0.272	0.621
t-RA	0.752	0.0037	0.4916
MC1288	0.5021	0.1225	0.3083
KH1060	0.037	0.0352	0.02352
R04	0.0485	0.051	0.0007

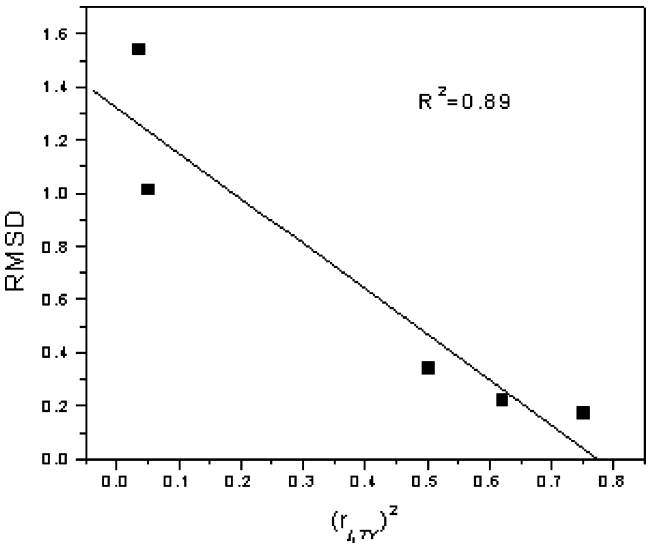

Figure 5. The relationship of the best correlation coefficient between Tanimoto distance and residuals $(R_{j,TY})^2$ versus the superimposition rmsd.

(PLS or FSR not MLR) yielded a correlation coefficient $(R_{j,TY})^2$ between the Tannimoto distance (D-space) and residuals ($\Delta Y = Y_t^j - Y_{obs}$) of 0.5 or better (Table 3). This was not observed in the KH1060 and R04 studies (the best KH1060 $(R_{j,TY})^2 = 0.037$, nor was it observed with the best R04 model $(R_{j,TY})^2 = 0.051$). Furthermore, there appears to be linear relationship ($r^2 = 0.8854$) between $(R_{j,TY})^2$ and rmsd (Figure 5). These observations indicate that more robust measures of TACS prediction, as well as confidence measures, can be developed in the future. Additionally, an intriguing unknown is the behavior of TACS in data sets characterized by multiple family specific binding modes that are known to exist.

Clearly with KH1060 and R04, the statistical methods failed to fully explain the descriptor space—activity relationship despite satisfying our reasonably conservative criteria for retention training set models ($R^2_{curr} > 0.5$ and $R^2_{curr} - \rho^2 > 0.2$, Table 4). We are currently investigating whether these global estimators of model quality will provide a threshold value, below which models should not be used in TACS. However, because certain regions of a molecule may be more significant than others in contributing to binding free energies, it is likely that more robust "local" estimators will eventually be required for TACS to perform best.

The more probable cause for TACS inaccuracy appears to be how descriptor space information is utilized. The rudimentary statistical methods employed in this study weight heavily those descriptors correlated to the biological activity or affinity. Therefore, the descriptors that describe regions of molecules that interact with the receptor without specific orientation (such as the hydrophobic chain of R04) may be underrepresented. Additional inaccuracies may arise from insufficient descriptor space description of critical pharmacophoric moieties. For instance, in the studies with KH1060 and R04, the inaccuracies may arise due to incomplete description of the oxygen atom and its environment in the hydrophobic saturated chain. We were intrigued by the observation that the KH1060 prediction was considerably worse than prediction for MC1288, given that these two compounds interact with the same binding site and are structurally similar. The training set for KH1060 contains MC1288 and vice versa; final training sets are composed of essentially equivalent conformations and descriptor spaces. In this case, only one training set ligand possesses similar structural features to KH1060 (specifically, an ether moiety, O−C, in the saturated chain) where the rest of the training set molecules have more structural similarity to MC1288. In certain cases, where a single moiety contributes significantly to the affinity or activity of a particular compound class, it may be necessary for the training set to contain more than one example of molecules containing that moiety.

In our attempts to determine the source of TACS error with KH1060 and R04, we observed these molecules contain a shared structural feature, the aliphatic ether. It so happens that the major contributor to the rmsd in both cases is the ether torsion (C−O−C−C) angle. Examination of the conformational ensemble for these compounds indicates that conformational space is adequately represented, in that numerous conformations similar to their respective crystallized counterparts (i.e., low rmsd) were found. We then focused on the possibility that descriptors relating specifically to the saturated chain C−O−C−C system are not properly represented (1) due to poor electrostatic force field parametrization *and* the subsequent effect on descriptor values or (2) statistical method failure to adequately emphasize these descriptors. Taken together, poor model performance (R^2_{curr}) and marginal rmsd in the KH1060 case, acceptable performance in R04, good rmsd for MC1288, and very low $(R_{j,TY})^2$ for either R04 and KH1060 would point to both factors as sources of error. Unfortunately, further work will be required to determine to what extent each of these are contributing factors.

Voting Mechanism. In early TACS development, we endeavored to find a single statistical method that

Table 4. Results from Training Set Construction and Randomization Trials[a]

	PLS				MLR				FSR			
name	R^2_{curr}	ρ^2	S R^2_{rnd}	$R^2_{curr} - \rho^2$	R^2_{curr}	ρ^2	S R^2_{rnd}	$R^2_{curr} - \rho^2$	R^2_{curr}	ρ^2	S R^2_{rnd}	$R^2_{curr} - \rho^2$
c-RA	0.86	0.27	0.18	0.77	0.99	0.69	0.06	0.30	0.81	0.30	0.06	0.51
t-RA	0.91	0.33	0.21	0.58	0.99	0.75	0.07	0.24	0.92	0.34	0.09	0.58
MC1288	0.91	0.22	0.16	0.69	0.92	0.73	0.14	0.19	0.92	0.36	0.09	0.56
KH1060	0.63	0.22	0.11	0.41	0.74	0.55	0.09	0.19	0.52	0.30	0.11	0.21
R04	0.50	0.31	0.07	0.19	0.74	0.54	0.10	0.20	0.99	0.32	0.10	0.67

[a] R^2_{curr} = correlation coefficient of the current model, ρ^2 = randomization correlation coefficient, $R^2_{rnd} = \rho^2$, S = standard deviation.

would be powerful enough for binding conformation elucidation. A more sophisticated statistical technique, genetic partial least squares (G/PLS), provided overall better results than any of the other single methods (MLR, PLS, FSR), but only provided moderate rmsd predictions. We observed a given statistical method would frequently find low rmsd conformations, but only for some of the test molecules, whereas alternate statistical method(s) would elucidate other members of the test sets. Because we were unable to develop a heuristic to assess the accuracy of a single model on given test molecules, we opted to develop a mechanism that makes use of multiple models. Inspired by current interest in the modeling community regarding the use of consensus data, based on multiple models, we developed the TACS voting method that seeks to extract relevant information for a given test conformation from multiple models.

The hypothesis guiding the voting algorithm's design was that if a given statistical method was able to distinguish between the noise and signal in the training set then residuals in biological activity predictions for the test set conformational ensemble should correlate with the descriptor space diversity of the test set ensemble $(R_{j,\mathrm{TY}})^2$. Please refer to section "TACS Component 4: Prediction of Bound Conformation(s)" above for a description of how variance of the descriptor space is represented as a one-dimensional data vector. Combined use of the $(R_{j,\mathrm{TY}})^2$ metric in combination with lowest residual metric (ΔY_{min}) led to a voting mechanism that properly identifies the under-performing statistical method(s).

For example, MLR gave better results (R^2_{curr}) than those of PLS in all cases and performed better than FSR in three cases (*c*-RA, *t*-RA, and KH1060), but the TACS voting mechanism identified MLR as having the weakest predictive power. The best $(R_{j,\mathrm{TY}})^2$ using MLR was seen with *c*-RA ($(R_{\mathrm{MLR,TY}})^2 = 0.27$), and based on the comparison of $(R_{\mathrm{PLS,TY}})^2 = 0.51$, $(R_{\mathrm{FSR,TY}})^2 = 0.62$ the vote weight was only 18.7%. Even better resolution was obtained by application of full-voting mechanism and the final contribution of $\mathrm{vote}_{\mathrm{MLR}}$ was only 13.6%. The trend where the MLR was identified as the poorest statistical performer was observed throughout training test set ensembles for all ligands.

Another example of how the voting mechanism was able to equate the predictive ability of given statistical methods is model generation for R04. The FSR model with $R^2_{\mathrm{curr}} = 0.99$ outperformed PLS ($R^2_{\mathrm{curr}} = 0.50$). However, the significance in contribution from each method in choosing the final conformation was about the same with $\mathrm{vote}_{\mathrm{FSR}} = 38.4$ and $\mathrm{vote}_{\mathrm{PLS}} = 34$.

Correlation between %vote and rmsd for heavy atoms of the top six vote receiving conformations is given in Table 5. These results demonstrate that the voting mechanism was able to distinguish the probable bound conformations from the rest of the conformations for three out of five test molecules. With KH1060 and R04, TACS failed to pick the conformation with lower rmsd for reasons stated above.

Next Steps. It is anticipated there exist four issues that, if properly investigated, will lead to an increase in the TACS predictive ability. The first is assessment of TACS sensitivity toward how training sets are constructed, especially as it relates to ensuring a high level of data density in the final training set. Second, the impact of conformational ensemble complexity, conformation removal criteria, and the effect of rigid ligands must be assessed. Third, we anticipate that better parametrization of the force field(s) used in conformational searching will increase TACS' predictive ability. And fourth, the use of alternative, perhaps unsupervised, statistical techniques may provide more uniform coverage of factors contributing to binding conformation, in that more emphasis may need to be placed on regions of the molecule that do not contribute significantly to binding free energy. We are currently investigating the robustness of TACS methodology on test cases where the binding site is interfacial or ill defined to assess the methods broader utility.

Table 5. Correlation of %Vote and rmsd for Heavy Atoms (Å) of the First Six Highest Vote-Receiving Conformations

name	%vote	rmsd for heavy atoms (Å)	standard deviation	mean	R^2 %vote vs rmsd
CIS00053	20.38	0.22	0.16	0.16	0.73
CIS00054	10.44	0.45	0.22	0.39	
CIS00052	9.25	0.51	0.30	0.41	
CIS00050	3.73	2.11	0.83	1.93	
CIS00507	2.89	2.76	1.30	2.43	
CIS00151	2.86	2.60	2.03	2.29	
TRA00080	20.09	0.18	0.10	0.14	0.87
TRA00043	9.34	0.97	0.62	0.75	
TRA00049	6.57	1.08	0.67	0.84	
TRA00043	5.80	1.17	0.68	0.96	
TRA00309	4.18	1.63	0.76	1.44	
TRA00033	2.70	2.02	1.00	1.76	
MC12882	25.70	0.35	0.18	0.30	0.94
MC2-128800010	10.40	2.31	1.27	1.93	
MC2-128800008	10.30	2.44	1.27	2.08	
MC2-128800111	7.24	2.37	1.09	2.15	
MC2-128800909	5.25	2.49	1.25	2.20	
MC2-128802014	3.94	2.69	1.05	2.47	
KH2_106000001	18.30	1.55	0.80	1.33	0.17
KH2_106000014	10.30	3.28	1.60	2.86	
KH2_106000010	9.06	3.03	1.40	2.70	
KH2_106003012	6.47	1.46	0.76	1.30	
KH2_106000318	5.72	2.77	1.45	2.35	
KH2_106001007	5.63	3.07	1.40	2.65	
R04_0052	14.70	1.01	0.47	0.90	0.14
R04_0029	8.05	3.53	1.04	3.38	
R04_0026	7.43	4.29	1.55	4.00	
R04_0512	5.45	3.55	1.80	3.04	
R04_0001	4.66	0.69	0.20	0.56	
R04_1138	4.06	3.63	1.70	2.91	

Conclusion

TACS is a promising new method that elucidates bound ligand conformational hypotheses in the absence of target protein structure. This methodology should provide much needed insight in ligand-based drug design or as a complement to structure-based drug design. The key innovations are that TACS requires only a small training set of molecules (and their respective activity or affinity toward the target) and no subjective user intervention (such as molecular alignment).

Note Added after ASAP Publication. This manuscript was released ASAP on 11/26/2004 with an incorrect ordering of the names in the authorship listing, with errors in the formatting of eq 1 and in the two

subsequent equations, with an error in a straddle heading of Table 4, and with minor text errors. The correct version was posted on 12/8/2004.

References

(1) DesJarlais, R. L.; Dixon, J. S.; Kuntz, I. D.; Sheridan, R. P.; Venkataraghavan, R. Docking flexible ligands to macromolecular receptors by molecular shape. *J. Med. Chem.* **1986**, *29*, 2149−2153.

(2) Kangas, E.; Tidor, B. Electrostatic complementarities at ligand binding sites: Application to chorismate mutase. *J. Phys. Chem. B.* **2000**, *105*, 880−888.

(3) Karplus, M.; Miranker, A. Functionality maps of binding sites: A multiple copy simultaneous search method. *Proteins: Struct. Funct. Genet.* **1991**, *11*, 29−34.

(4) Platt, D. E.; Silverman, D. Registration, orientation, and similarity of molecular electrostatic potentials through multipole matching. *J. Comput. Chem.* **1996**, *17*, 358−366.

(5) Stockman, G. Object recognition and localization via pose clustering. *Comput. Graph. Vis. Image Proc.* **1987**, *40*, 361−387.

(6) Hahn, M. A. Receptor Surface Models: 1. Definition and construction. *J. Med. Chem.* **1995**, *38*, 2080−2090.

(7) Hahn, M. A.; Rogers, D. Receptor Surface Models: 2. Application to quantitative structure activity relationship studies. *J. Med. Chem.* **1995**, *38*, 2091−2102.

(8) Doweyko, A. M. The hypothetical active site lattice approach to modeling active sites data on inhibitor molecules. *J. Med. Chem.* **1988**, *31*, 1396−1402.

(9) Block, J.; Doweyko, A. M.; Henry, D.; Magee, P. S. Bioactive mechanisms: Proof, SAR, and Prediction, a new tool for the study of structure activity relationships in three dimensions. HASL: The hypothetical active site lattice. *ACS Symp. Ser.* **1989**, *No. 413*.

(10) Dordrecht, A. A.; Folkers, G.; Kubinyi, H.; Martin, Y. C. *3D QSAR in Drug Design*; Kluwer Academic: Boston, MA, 1998; Vol. 6, pp 167−179.

(11) Bures, M.; Danahar, E.; Martin, Y.; A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonist. *J. Comput.-Aided Mol. Design.* **1993**, *7*, 83−102.

(12) Hoffman, R.; Li, H.; Sutter, J. Hypogen: An automated system for generating 3-D predictive pharmacophore models. In *Pharmacophore Perception, Development and Use in Drug Design*; Guner, O. F., Ed.; International University Line: La Jolla, CA, 2000; pp 178−189 (IUL Biotechnology Series).

(13) Bravi, G.; Elkins, S.; Wikel, J. H.; Wrighton, S. A. Three-dimensional quantitative strucure activity relationship (3D-QSAR) analysis of CYP3A4 substrates. *J. Pharmacol Exp. Ther.* **1999**, *291*, 424−433.

(14) Dantzig, A. H.; Elkins, S.; Kim, R. B.; Lan, L.-B.; Leake, B. F.; Scheutz, E.; Scheutz, J. D.; Shepard, R. L.; Winter, M. A.; Yasuda, K.; et al. Three-dimensional quantitative structure activity relationships of inhibitors of P-glycoprotein. *Mol. Pharmacol.* **2002**, *61*, 964−973.

(15) Marshall, G.; Mayer, D.; Motoc, I.; Naylor, C. A unique geometry of the active site of angiotension converting enzyme consistent with structure activity studies. *J. Comput.-Aided Mol. Design* **1987**, *1*, 3−16.

(16) Dixon, J.; Blaney, J. Distance geometry in molecular modeling. In *Reviews in Computational Chemistry*; Boyd, L., Ed.; Wiley-VCH: New York, 1994; Vol. 5, pp 299−355.

(17) Crippen, G. Chemical distance geometry: Current realization and future projection. *J. Math. Chem.* **1991**, *6*, 307−324.

(18) Hopfinger, A. J. *J. Med. Chem.* **1983**, *26*, 990−996.

(19) Cameron, J. F.; Chao, W. R.; Dawson, M. I.; Hobbs, P. D.; Jong, L.; Phall, R. Conformational effects on retinoid receptor selectivity 2. Effects of retinoid bridging group on retinoid X receptor activity and selectivity *J. Med. Chem.* **1995**, *38*, 3368−3383.

(20) Chambon, P.; Gronemeyer, H.; Klaholz, B. P.; Mitschler, A.; Moras, D.; Renaud, J. P.; Zusi, C. Conformational adaptation of agonists to the human nuclear receptor RAR gamma. *Nat. Struct. Biol.* **1998**, *5*, 199.

(21) Chambon, P.; Gronemeyer, H.; Moras, D.; Renaud, J. P.; Rochel, N.; Ruff, M.; Vivat, V. Crystal structure of the RARγ ligand-binding domain bound to all trans-retinoic acid. *Nature* **1995**, *378*, 681.

(22) Collins, E. D.; Liu, Y.-Y.; Norman, A. W.; Peleg, S. Differential interaction of 1,25-dihydroxyvitamin D_3 analogues and their 20-*epi* homologues with the vitamin D receptor. *J. Biol. Chem.* **1997**, *272*, 3336−3345.

(23) Bishop, J. E.; Collins, E. D.; Norman, A. W.; Peleg, S.; Sastry, M. Distinct conformational changes induced by 20-epi analogues of 1, 25-dihydroxyvitamin D_3 are associated with enhanced activation of the vitamin D receptor. *J. Biol. Chem.* **1995**, *270*, 10551−10558.

(24) Freedman, L. P.; Yang, W. 20-epi analogues of 1, 25-dihydroxyvitamin D_3 are highly potent inducers of DRIP co-activator complex binding to the vitamin D_3 receptor. *J. Biol. Chem.* **1999**, *274*, 16838−16845.

(25) Mitschler, A.; Moras, D.; Rochel, N.; Tocchini-Valentini, G.; Wurtz, J. M. Crystal structures of the vitamin D receptor complexed to superagonist 20-epi ligands. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 5491−5496.

(26) Badger, J.; Dianat, G. D.; Dutkot, F. J.; Fanchert, M.; Griffith, J. P.; Guerin, D. M.; Heinz, B. A.; Kremer, M. J.; Krishnaswamy, S.; Luo, M.; McKinlayt, Minor, I.; M. A.; Oliveira, M. A.; Rossman, M. G.; Rueckert, R. R.; Smith, T. J. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 3304−3308.

(27) Badger, J.; Minor, I.; M. A.; Oliveira, M. A.; Rossman, M. G.; Smith, T. J. *Proteins* **1989**, *6*, 1−19.

(28) *SYBYL*, version 6.8; Tripos Associates Inc.: 1699 South Hanley Rd., St. Louis, MO 63144-2913.

(29) Stewart, J. P. Optimization of parameters for semiempirical methods, I-method. *J. Comput. Chem.* **1989**, *10*, 209−220.

(30) *Volsurf*, version 3.0.9; Molecular Discovery Ltd.: 4 Chandos St.; London, W1A 3AQ.

(31) *Cerius*^{2}, version 4.9; Accelerys, 9685 Scranton Rd.; San Diego, CA 92121.

(32) *QSARIS*, version 1.1; SciVision, 200 Wheeler Rd.; Burlington, MA 01803.

(33) Torrens, F. Characterizing cavities in model inclusion. *Int. J. Mol. Sci.* **2001**, *2*, 72−88.

(34) van Ooyen, A. In *New Approaches for the Generation and Analysis of Microbial Typing Data*; Dijkshoorn, R., Ed.; Elsevier: Amsterdam, 2001; Vol. 1, pp 31−45.

(35) Glen, W. G.; Dunn, W. J.; Scott, D. R. Principal components analysis and partial least squares regression. *Tetrahedron Comput. Methodol.* **1989**, *349−376*.

(36) Hosmer, D. W., Jr.; Lemeshow, S. In *Applied Logistic Regression*; Wiley Series in Probability and Statistics—Applied Probability and Statistics Section; John Wiley & Sons: New York, 1989.

(37) Darlington, R. B. Multiple regression. *Psychological Bul.* **1968**, *2*, 161−182.

JM049729Z