A Simple Clustering Technique To Improve QSAR Model Selection and Predictivity: Application to a Receptor Independent 4D-QSAR Analysis of Cyclic Urea Derived Inhibitors of HIV-1 Protease

Craig L. Senese and A. J. Hopfinger*

Laboratory of Molecular Modeling and Design (M/C-781), University of Illinois at Chicago, College of Pharmacy, 833 South Wood Street, Chicago, Illinois 60612-7231

Received August 4, 2003

A training set of 50 tetrahydropyrimidine-2-one based inhibitors of HIV-1 protease, for which the $-\log K_i$ values were measured, was used to construct receptor independent 4D-QSAR models. A novel clustering technique was employed to facilitate and improve model selection as well as test set predictions. Following the manifold model theory, five unique models were chosen by the clustering algorithm ($q^2 = 0.81-0.84$). The models were used to map the atom type morphology of the inhibitor binding site of HIV-1 protease as well as to predict the potencies ($-\log K_i$) of 10 test set compounds. The rank-difference correlation coefficient was used to evaluate the quality of the test set predictions, which was improved from 0.39 to 0.68 when the clustering technique was applied. The set of five models, collectively, identify the important binding characteristics of the HIV protease receptor site. This study demonstrates that the selected simple clustering technique provides a discrete algorithm for model selection, as well as improving the quality of test set, or unknown, compound prediction as determined by the rank-difference correlation coefficient.

INTRODUCTION

For nearly 15 years, the aspartyl protease from the human immunodeficiency virus type I (HIVPR) has been a major target for the drug design community.[1−6] The discovery that inhibition of this enzyme results in noninfectious progeny[7−9] has led to exhaustive efforts focused on generating increasingly potent drugs with favorable toxicity and bioavailability profiles. The need for an extensive array of inhibitors is compounded by the ability of HIVPR to rapidly mutate to produce resistance. In the interest of identifying the full potential of the various classes of HIVPR inhibitors, numerous QSAR studies have been performed to characterize and define ligand−receptor binding interactions.[5−6,10−11] Recent work in our laboratory has focused on utilizing the 4D-QSAR paradigm to explore the binding characteristics of this important enzyme.[12] The current study reports the results of a 4D-QSAR analysis of a set of cyclic urea derived inhibitors of HIV-1 protease, the tetrahydropyrimidine-2-ones (THP).

Cluster analysis has been utilized extensively in drug discovery.[13−16] Its versatility has been highly correlated with the advancement of computers and computing power. The incorporation of a simple clustering technique into the 4D-QSAR method provides a way to more fully exploit the information content of a data set.

Clustering is a means by which to divide objects into groups such that objects within a group are similar to each other while being different from the objects in other groups. There are many types and variations of clustering methods, each directed at providing a unique approach for varying types of data.[13] More important than the specific clustering algorithm, however, is the basis by which the objects are clustered. When the objects to be clustered are drugs or drug-like compounds, the criteria by which to group the compounds is infinite. Since the mid 1980s, extensive work has focused on finding an accurate and efficient basis for comparison of compounds so as to improve the results of a clustering analysis.

A majority of the clustering methods developed for drug discovery and design have targeted diverse sets of compounds.[17] The utility of these methods is to reduce a large database of compounds into subsets to facilitate property prediction or to further analyze the subsets by a particular QSAR method. Conversely, a clustering method has yet to be employed in the interest of augmenting a QSAR study applied to a set of analogue compounds. The current study illustrates how a simple clustering technique can be exploited to assist in the model selection and test set prediction steps in a 4D-QSAR analysis.

METHOD

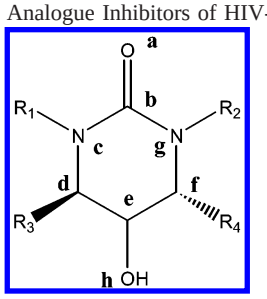
Training Set of THP Analogues. The training set of THP compounds used in this work is the same as employed in a previous computational study performed by Nair et al.[18] The core structure of the THP compounds as well as the atom coding used for alignment identification is given in Table 1. The 60 THP analogues used in this study are listed along with their corresponding $-\log K_i$ (pK_i) values in Table 2. From the total set of compounds, 11 were chosen as a test set and thereby not included in the training set for model construction. These test set compounds are numbered t1 through t11 in Table 2.

The inhibition constants, K_i, were determined by following the cleavage of a fluorescent substrate using HPLC, as described previously.[19,20] The range of $-\log K_i$ values for the training set compounds spans 5 orders of magnitude (6.01 to 11.00). 4D-QSAR models from this study should, therefore, reflect both subtle, as well as substantial structure−activity specificity to capture the vast differentiation of activity in the training set. To best represent and sample the

* Corresponding author phone: (312)996-4816; fax: (312)413-3479; e-mail: hopfingr@uic.edu.

10.1021/ci034168q CCC: $25.00
Published on Web 10/31/2003

Table 1. Three Ordered-Atom Alignments Used in the 4D-QSAR Analysis of the THP Analogue Inhibitors of HIV-1



alignment #	atom 1	atom 2	atom 3
1	b	d	f
2	c	e	g
3	a	b	e
4	a	d	f
5	d	e	f
6	h	g	c
7	c	b	g
8	h	e	b

wide range in activity, the set of test compounds, as noted by Nair et al., were chosen to include inhibitors exhibiting low, moderate, and high activity.

4D-QSAR Analysis. The 4D-QSAR method has been previously presented in detail[21] and will only be summarized here. For this study, the commercial version of the 4D-QSAR package (V3.0) was utilized.[22] The 4D-QSAR paradigm is currently being expanded to provide the ability to include receptor geometry. Therefore, there is now distinction made between receptor dependent 4D-QSAR (RD-4D-QSAR)[23] and receptor independent 4D-QSAR (RI-4D-QSAR) type methods. The work reported here is an RI-4D-QSAR study.

The first step in an RI-4D-QSAR analysis is to generate the grid cell lattice space, comprised of a set of one (normally) angstrom cubes, in which to place the three-dimensional structures of the training set compounds. Currently, the structures are generated and optimized in Hyperchem 6.0.[24] Structure optimization is achieved by first determining the preferred compound geometry using molecular mechanics with an MM+ force field, followed by assignment of partial atomic charges using the semiempirical AM1 method. However, in contrast to 3D-QSAR methods, the compounds will not be restricted to this initial starting conformation but rather a sampling of available thermodynamic conformer states is employed as part of the analysis.

The next step is to assign the *interaction pharmacophore elements*, or IPEs, defined in Table 3, and generate the *conformational ensemble profile* (CEP). In this step, thermodynamically accessible conformer states are determined for each of the training set compounds. A molecular dynamics simulation (MDS) is currently used to generate the CEP. The MOLSIM[25] software package is used to perform the MDS, with an extended MM2 force field. The simulation temperature was set at 300 K, and the molecular dielectric constant was set to 3.5. A total sampling time of 100 ps was used, with sampled conformations recorded every 0.1 ps. This results in the generation of 1000 "snapshots" or the thermodynamically sampled set of conformations that make up the CEP.

Molecular alignments are chosen next in the 4D-QSAR method. A three-ordered atom alignment rule was used to process this step. Alignments are typically chosen to span the common scaffold of the compounds. Atoms from all parts of the common scaffold of the compounds should be represented in the chosen alignments to ensure a thorough alignment analysis. By spanning the scaffold, information regarding the substituent properties of the compounds is obtained. This strategy is illustrated in the alignments used in the current study (Table 1). The small, relatively rigid scaffold of the THP analogues resulted in only eight alignments chosen for this analysis. Alignments 1, 2, 5 and 7 focus on the core six-membered ring. Alignments 3 and 4 combine the six-membered ring with the common carbonyl moiety, while alignments 6 and 8 consider the common hydroxyl combined with the six-membered ring.

All conformations from the CEP of each compound are next placed in a grid cell lattice according to the selected trial alignment. The occupancy of the grid cells by the different IPE types is then recorded to form the set of grid cell occupancy descriptors or GCODs. The GCODs constitute the pool of independent variables, or trial descriptors, used in the model generating and optimization process. The genetic function approximation,[26] or GFA, is employed to optimize the set of trial descriptors in a multidimensional linear regression equation in which the activity of the compounds is expressed as a function of the occupancy of certain grid cells by specific IPE types. It is the properties of these grid cells, namely location, IPE type, and regression coefficient sign/magnitude, that provide direct information regarding the binding features of the training set compounds and indirect information related to the binding features of the target receptor.

Several criteria are considered when choosing the *best* model in any QSAR analysis. In 4D-QSAR, the first model quality parameter considered is typically a statistical measure of fit. This usually consists of the leave-one-out cross-validated correlation coefficient or q^2. The q^2 value is influenced by the number of terms included in the model. To help choose the optimum number of model terms, it is typical to plot the number of model terms versus the q^2 value. The point in the plot where the cross-validated correlation-coefficient does not significantly increase with the inclusion of additional model term is chosen as the optimal number of model terms relative to the training set. A test set may also be utilized to evaluate the predictive power of the chosen models. The test set consists of a set of compounds, usually analogues of the training set compounds, which were not used to generate the models. The chosen *best* models are then used to predict the activity of the test set compounds.

It is often the case, however, that several 4D-QSAR models from an analysis meet, about equally well, the criteria of a good model. In this situation, a set of models, or a *manifold model*, may be needed to best represent the data in the training set. Models that are statistically significant, yet not highly correlated to one another, provide unique information regarding the training set. It is important to be able to identify the set of distinct individual 4D-QSAR models which form the manifold model in order to be sure that the QSAR analysis has made full use of the data available. The current study illustrates how a simple clustering method can be utilized to help identify the multiple components of a manifold model.

Once the best model, or models, are chosen, the active conformation of each of the training set compounds can be

Table 2. 49 Training Set and 11 Test Set Compounds and Their Measured Inhibition Constants ($-\log K_i$)

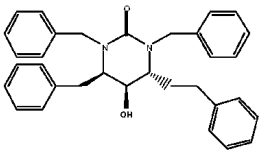
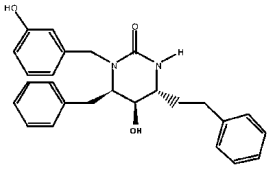
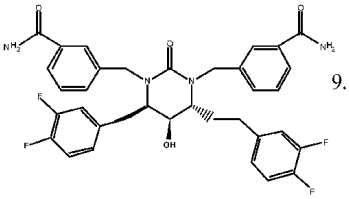
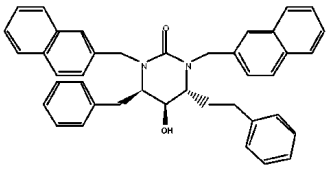
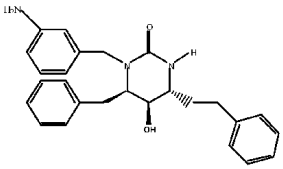
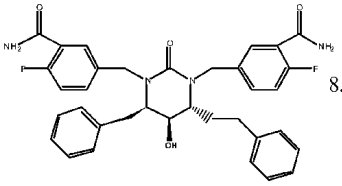
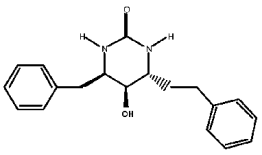
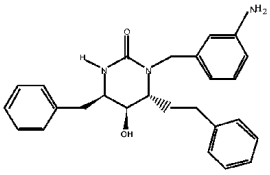
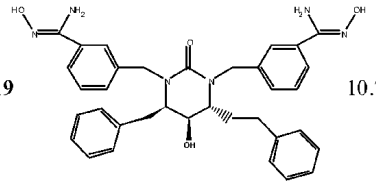
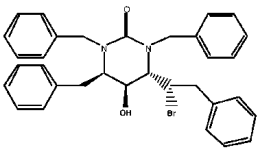
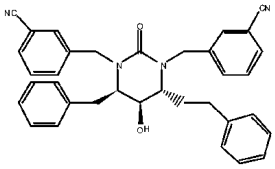
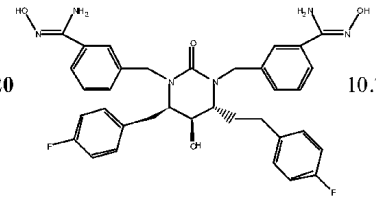
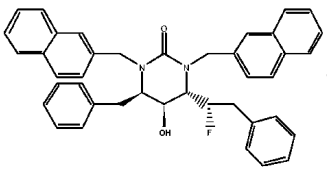
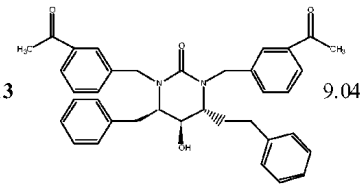
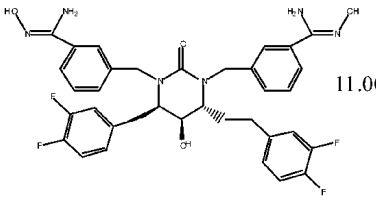
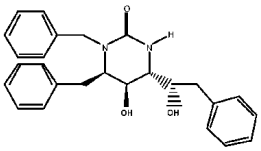
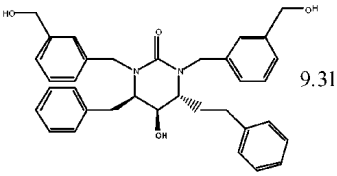
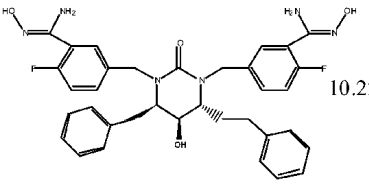
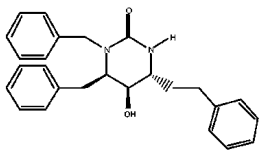
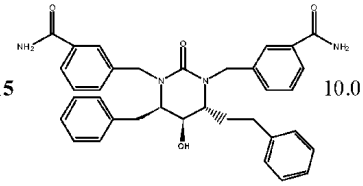
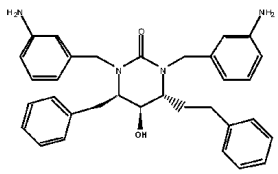
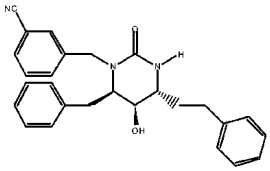
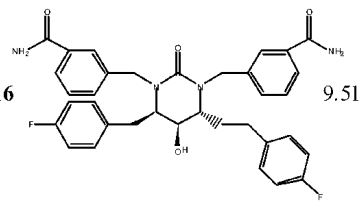
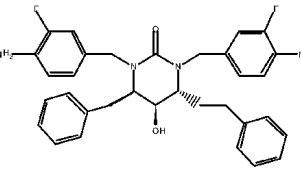
No.	Structure	pK_i	No.	Structure	pK_i	No.	Structure	pK_i
1		7.82	9		8.59	17		9.82
2		7.64	10		6.80	18		8.85
3		6.01	11		8.21	19		10.70
4		6.73	12		7.96	20		10.70
5		6.27	13		9.04	21		11.00
6		6.49	14		9.31	22		10.22
7		7.18	15		10.05	23		8.77
8		7.82	16		9.51	24		8.10

Table 2 (Continued)

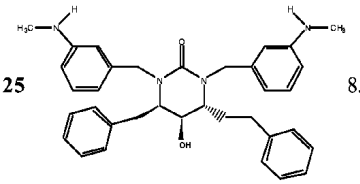
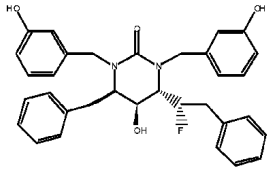
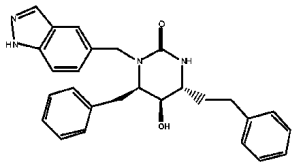
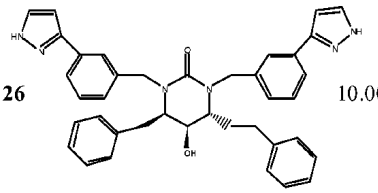
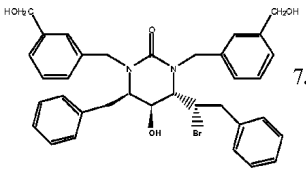
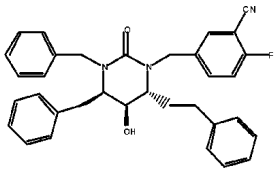
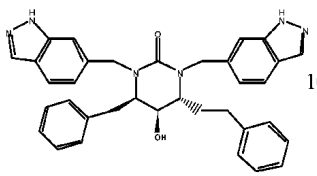
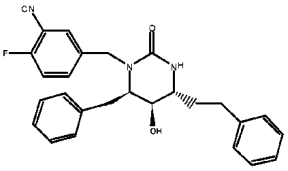
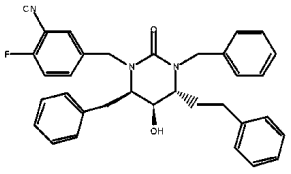
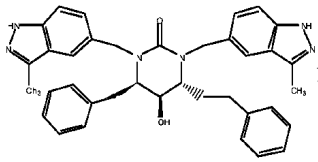
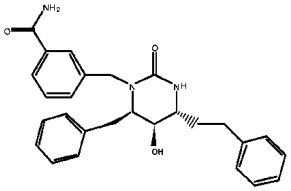
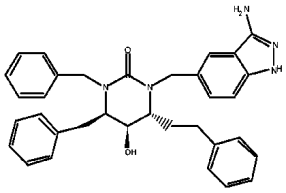
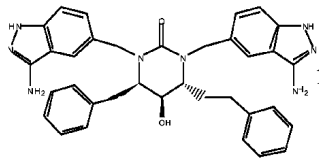
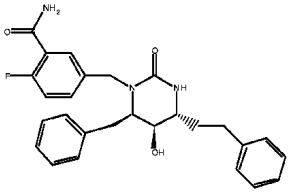
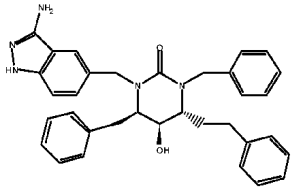
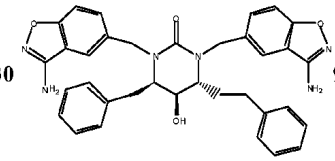
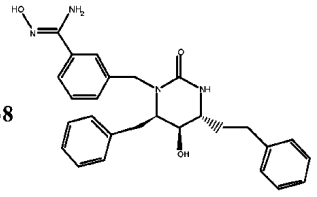
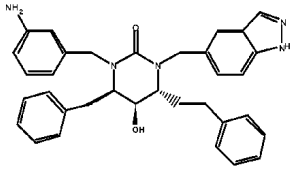
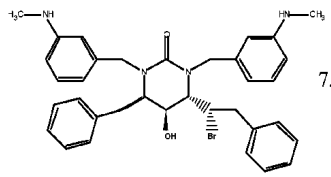
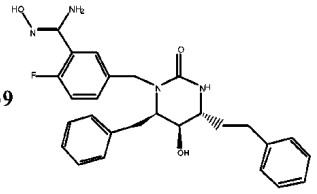
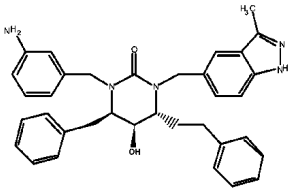
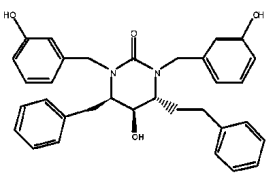
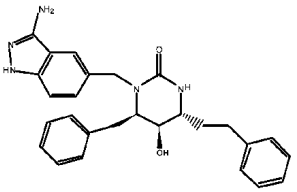
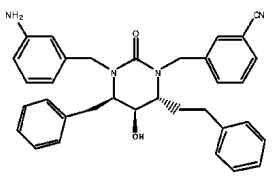
No.	Structure	pK_i	No.	Structure	pK_i	No.	Structure	pK_i
25		8.31	**33**		7.92	**41**		8.92
26		10.00	**34**		7.96	**42**		7.48
27		10.22	**35**		6.96	**43**		7.40
28		10.00	**36**		8.25	**44**		9.19
29		10.52	**37**		7.85	**45**		9.55
30		9.39	**38**		9.62	**46**		10.0
31		7.64	**39**		8.85	**47**		9.89
32		9.52	**40**		9.42	**48**		8.80

Table 2 (Continued)

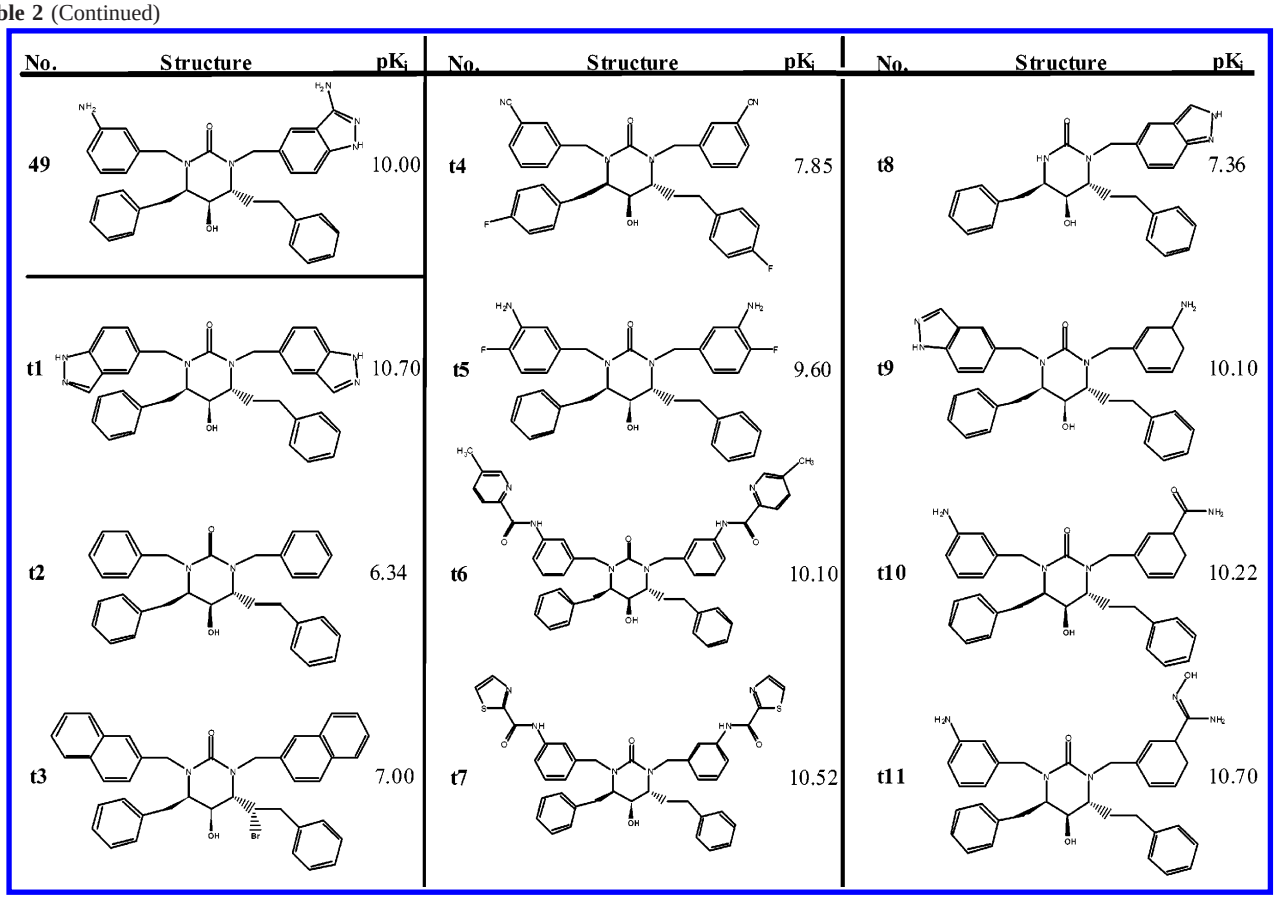


Table 3. Interaction Pharmacophore Elements, IPEs, Used in 4D-QSAR Analyses

IPE description (abbreviation)	IPE code
all atoms in the molecule (any)	0
nonpolar atoms (np)	1
polar (+) atoms (p+)	2
polar (−) atoms (p−)	3
hydrogen bond acceptor atoms (hba)	4
hydrogen Bond donor atoms (hbd)	5
aromatic atoms (a)	6
non-hydrogen atoms (hs)	7

postulated *relative to the model*. This step is carried out by initially determining the conformations of a compound that are within a threshold energy limit, i.e., only thermodynamically reasonable conformations are considered, and then determining which conformation within the threshold energy limit has the highest predicted activity by the model. The predicted active conformation can be used in various ways, such as serving as a structural template for rational drug design or providing a starting point for one of the QSAR methodologies that require a user postulated active conformation, such as CoMFA.[27]

Cluster Analysis. Much of the effort in applying cluster analysis to chemical systems has focused on subdividing structurally diverse compound libraries into more homogeneous groups with the goal of making property predictions.[17] In the current work, we introduce how simple clustering can be combined with a standard QSAR analysis of an analogue set, resulting in a more thorough study with improved property prediction. The application of the clustering technique described here is not specific to 4D-QSAR analyses and has been formulated to apply to most QSAR methods.

There are two specific points in a general QSAR study where a clustering technique may be applied to improve the quality of the overall analysis. The first point involves the selection of the *best* model or models. Typically, the best models are chosen based on statistical significance and/or a priori knowledge of ligand−receptor binding features. Sometimes, however, several models meet the specified criteria for selection as the best model. The disregard of a model that contains good model statistics, because it lacks the statistical significance of the *best* models as identified, say, by the highest r^2 and q^2 values, may result in information loss. It is possible that good models possessing acceptable statistical significance, yet not deemed a member of the *best* models, may contain information regarding the structure−activity relationship that is unique relative to the *best* models. Therefore, neglecting such good models in the overall analysis may result in important information being overlooked.

A simple clustering scheme can be employed in evaluating models based on the "uniqueness" of their information content as well as with respect to their statistical significance. First, all models that meet a predetermined minimum statistical significance are identified. Next, the residuals of fit (to the training set) vectors for all of the identified models are pair-correlated to produce a matrix of model-correlation. It is an accepted working assumption that models having highly correlated residuals of fit represent largely common information, while models that have a lesser degree of correlation represent more distinct information. The matrix of model-correlation is used instead of the actual residuals of fit to facilitate the interpretation of the clustering results.

Any one of a number of clustering techniques may be applied to identify groups, or clusters, within the matrix of model-correlation. A simple linkage, hierarchical clustering technique is used in the current study and described below. Clustering on the matrix of model-correlation produces groups of models that are similar in "model" space and the extent of similarity of the elements within a cluster is a parameter that may be altered to adjust cluster number and size. The result is that models within a cluster represent similar structure−activity information. The statistically most significant model from each cluster may now be chosen as an element of a manifold model, alleviating the possibility of any significant structure−activity information loss. Models may also be identified from the analysis that do not belong to any cluster and are referred to as "singletons". These are models that are completely unique in information content and may not have been otherwise identified as having a singleton nature.

The second operation in a general QSAR analysis that can be improved through use of a simple clustering technique is that of test set, or unknown compound, property prediction. It is commonly the case, especially when using a genetic algorithm for model exploration and optimization, that the result of a QSAR analysis is a set of models (a manifold model), rather than a single model. The individual models of the manifold model are different structure−activity representations of the training set. Therefore, the predicted activities for the training set compounds will vary depending on which equation from the manifold model is used to make the prediction. A corollary to the previous statement is certain models will predict better for certain compounds. It follows that this prediction behavior is also true for test set compounds as well as unknowns in general. It would be beneficial to know, a priori, which specific model predicts best for each member of a test set of compounds. Again, a simple clustering technique can be applied to help achieve this goal.

The general method to employ clustering is as follows; First, all of the compounds in the data set, both training and test set molecules, are clustered based on a predetermined means of comparison. The means, or basis, for comparison is dependent on the properties of the data set employed. Useful means of comparison may be inferred by the properties of the QSAR models generated for the training set. For example, if the models present in the manifold model indicate that hydrophobic interactions contribute significantly to inhibitor binding, then clustering based on some hydrophobic property may prove successful. Various clustering algorithms may also be employed, though in the current work a simple hierarchical linkage method proved quite adequate.

The result of clustering are groups of compounds that are similar within a group with respect to the basis of comparison used. The test set compounds will, therefore, belong to a group of training set compounds that are similar. For each group, or cluster, of compounds, a different equation from the manifold model will predict best for that group. To determine which equation may best predict for the test set compounds, the equation that has the minimum sum-squared error of residuals of prediction for the training set compounds of the same group as the test set compound is used. To fully explore this procedure, it may be necessary to group the compounds using different means and/or criteria of comparison.

Table 4. Statistical Quality, as Measured by r^2 and q^2, of the Five-, Six-, and Seven-Term Models for the Eight Alignments

alignment[a]	r^2 range	q^2 range
1	0.80−0.87	0.75−0.84
2	0.79−0.86	0.74−0.83
3	0.80−0.91	0.74−0.86
4	0.83−0.85	0.78−0.81
5	0.79−0.86	0.75−0.82
6	0.80−0.88	0.77−0.84
7	0.80−0.86	0.75−0.82
8	0.82−0.89	0.76−0.84

[a] For alignment definitions, see Table 1.

The benefit to this clustering method is 3-fold. First, if a test set compound, or analogue compound with an unknown activity, exists as a singleton in the optimized cluster grouping, this is an indication that this compound has minimal opportunity to be well predicted by the manifold model. Second, the basis for comparison that best optimizes the clustering procedure provides information regarding the important pharmacophore/binding features of the analogues and the target. Finally, the accuracy of prediction for new, untested analogue compounds can be improved by determining which of the optimized groups of training set compounds the new analogue belongs and then using the appropriate equation for prediction.

A linkage clustering method was used in the current study to establish grouping within the data. The purpose of linkage clustering is to produce a hierarchical, or multilevel, relationship between the objects. The levels in this clustering scheme are Euclidean distances between clusters, where lower level clusters are grouped into larger clusters at higher levels. Once a cluster is formed, the location in Euclidean space of that cluster is considered to be the centroid, or average, of the member elements of the cluster. The result is a cluster tree, or dendrogram, that graphically illustrates the grouping relationships between the objects. The hierarchical cluster tree can be divided at any level to increase, or decrease, the Euclidean distance between clusters and subsequently altering the number of clusters present in the data. The correct number of clusters, or level of division, is determined to be that which best fits the application and data. For example, if improving the predictions of unknown compounds is the goal of the cluster analysis, the results of test set predictions may be used to evaluate the clustering level.

RESULTS

Model Selection and Interpretation. The eight alignments chosen for the set of THP analogue inhibitors of HIV-1 protease are defined in Table 1. The statistical quality of the five-, six-, and seven-term models for the eight alignments are given in Table 4. Five, six- and seven-term models were chosen to fully represent the data set. From the statistical measures of fit, all alignments yield models with significant statistical properties ($q^2 > 0.8$). Therefore, choosing the best model, or models, based on highest statistical measure of fit would, in this case, eliminate quality models containing useful information. To avoid this loss of information, the models were evaluated for "uniqueness of their information content" as well as for their statistical significance. The best five-, six-, and seven-term models, identified by highest q^2 for a particular alignment, were first identified

Table 5. Cross-Correlated Residuals of Fit for the 24 Best 4D-QSAR Models from the Eight Alignments[a]

	1−5	1−6	1−7	2−5	2−6	2−7	3−5	3−6	3−7	4−5	4−6	4−7	5−5	5−6	5−7	6−5	6−6	6−7	7−5	7−6	7−7	8−5	8−6	8−7
1−5	1.00																							
1−6	0.76	1.00																						
1−7	0.76	0.96	1.00																					
2−5	0.75	0.57	0.59	1.00																				
2−6	0.73	0.61	0.65	0.93	1.00																			
2−7	0.67	0.57	0.64	0.68	0.64	1.00																		
3−5	0.59	0.64	0.59	0.52	0.56	0.54	1.00																	
3−6	0.57	0.51	0.48	0.53	0.51	0.53	0.78	1.00																
3−7	0.56	0.52	0.57	0.42	0.48	0.48	0.49	0.47	1.00															
4−5	0.61	0.52	0.50	0.51	0.47	0.51	0.63	0.57	0.35	1.00														
4−6	0.63	0.54	0.54	0.53	0.54	0.49	0.57	0.48	0.41	0.77	1.00													
4−7	0.56	0.45	0.48	0.51	0.54	0.55	0.45	0.52	0.44	0.63	0.56	1.00												
5−5	0.80	0.70	0.67	0.74	0.72	0.68	0.68	0.63	0.63	0.61	0.60	0.55	1.00											
5−6	0.59	0.60	0.58	0.58	0.64	0.51	0.61	0.57	0.57	0.53	0.56	0.56	0.71	1.00										
5−7	0.69	0.64	0.63	0.60	0.64	0.57	0.60	0.56	0.57	0.56	0.64	0.59	0.76	0.91	1.00									
6−5	0.53	0.55	0.59	0.55	0.62	0.46	0.61	0.58	0.56	0.46	0.65	0.52	0.64	0.67	0.69	1.00								
6−6	0.42	0.43	0.36	0.63	0.64	0.50	0.61	0.53	0.26	0.32	0.47	0.46	0.62	0.65	0.63	0.59	1.00							
6−7	0.39	0.35	0.29	0.60	0.63	0.49	0.54	0.54	0.22	0.32	0.49	0.49	0.57	0.67	0.65	0.59	0.93	1.00						
7−5	0.62	0.67	0.66	0.49	0.53	0.52	0.62	0.58	0.48	0.53	0.56	0.40	0.69	0.59	0.64	0.71	0.44	0.38	1.00					
7−6	0.64	0.67	0.67	0.53	0.57	0.52	0.60	0.59	0.52	0.52	0.55	0.44	0.65	0.59	0.64	0.76	0.45	0.42	0.94	1.00				
7−7	0.59	0.67	0.64	0.57	0.54	0.54	0.60	0.51	0.32	0.64	0.47	0.40	0.66	0.63	0.62	0.42	0.42	0.41	0.63	0.58	1.00			
8−5	0.49	0.46	0.46	0.60	0.67	0.57	0.66	0.57	0.49	0.57	0.61	0.54	0.72	0.60	0.61	0.69	0.71	0.67	0.58	0.55	0.47	1.00		
8−6	0.41	0.43	0.45	0.49	0.59	0.45	0.52	0.36	0.46	0.38	0.43	0.48	0.60	0.50	0.41	0.60	0.54	0.49	0.44	0.43	0.41	0.81	1.00	
8−7	0.43	0.44	0.46	0.49	0.57	0.46	0.57	0.40	0.49	0.45	0.48	0.52	0.63	0.54	0.46	0.65	0.57	0.51	0.46	0.46	0.43	0.86	0.98	1.00

[a] The column and row labels are the alignment number followed by the number of model terms (1−5 is a five-term model from alignment 1).

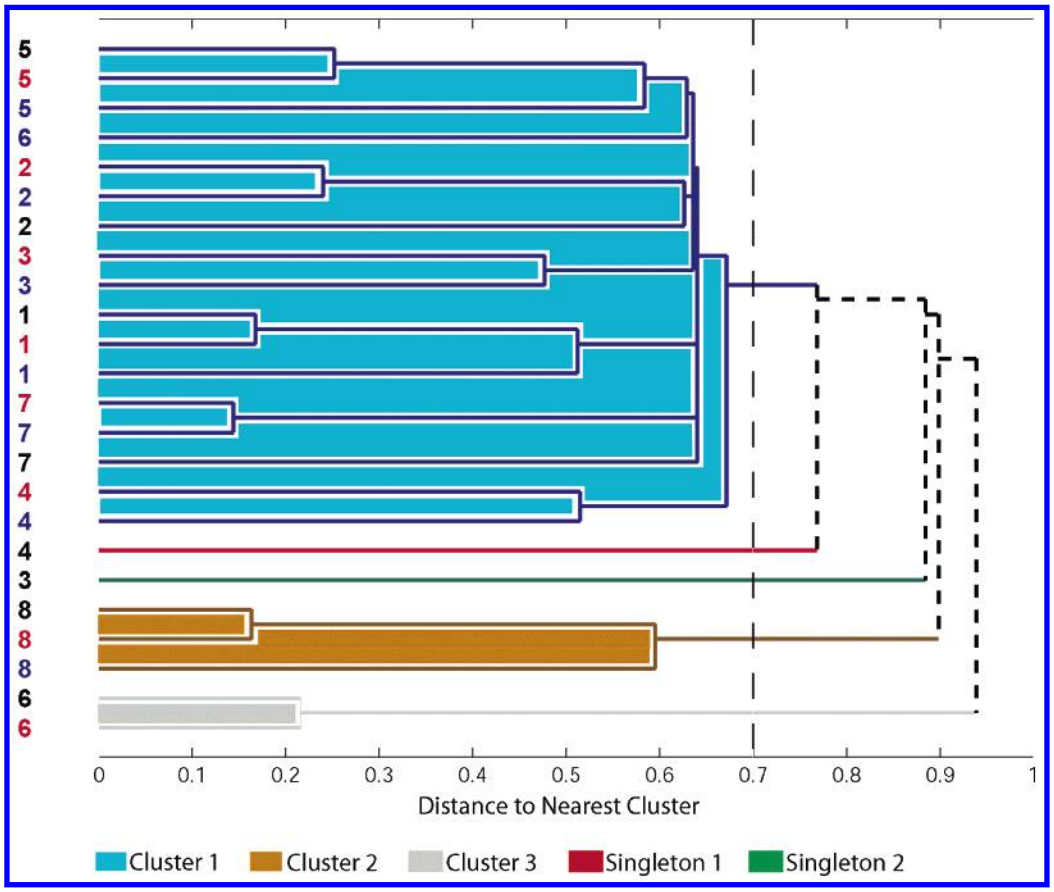


Figure 1. A dendrogram showing the 4D-QSAR model clusters in molecular similarity space. The 24 models are each represented by a number on the left of the cluster tree. This number corresponds to the model alignment, and the color of the number defines the number of model GCOD terms. Blue numbers are five-term models, red numbers are six-term models, and black numbers are seven-term models.

for the eight alignments, resulting in 24 good models. The residuals of fit of these 24 models were next pair-correlated to produce the matrix of model-correlation given in Table 5. The elements of the matrix of model-correlation were subsequently used as a data set on which the single linkage clustering method was applied. The cluster tree, or dendrogram, shown in Figure 1 was constructed from the clustering. The *distance between clusters* parameter was selected to be 0.7, indicated by a dashed line in Figure 1, and ensures that models with a correlation coefficient, or *r*-value, greater than 0.5 will appear in the same cluster. Employing this criteria, three clusters and two singletons were found for the 24 models, indicating that five models are required to fully represent the information present in this data set. The three

clusters in Figure 1 are color-coded blue, brown, and gray, while the singletons are colored red and green. The largest cluster contains 17 of the 24 models and indicates a high degree of redundant information. This cluster contains at least one model from seven of the eight alignments, only alignment 8 is not represented. The three models from alignment 8 belong to their own unique cluster, indicating this alignment is unique in its representation of the data. Two models from alignment 6 also belong to their own cluster, again pointing to the unique nature of the models from this alignment. The two singleton models, one each from alignments 3 and 4, illustrate how the clustering method can identify single models containing distinct information that may have been overlooked by a conventional model selection approach. These two models are statistically significant, but not the "best" models, as determined by q^2. Disregarding these models would result in the loss of valuable information provided by the data set.

The numerical representation of the best model from each of the three clusters, as determined by q^2, as well as the two singleton models are given by eqs 1 through 5.

$$-\log K_i = 9.0 - 6.56*\text{GC1 (any)} - 19.39*\text{GC2 (np)} - 21.22*\text{GC3 (any)} + 26.75*\text{GC4 (any)} + 12.97*\text{GC5 (any)} + 15.76*\text{GC6 (any)} + 15.56*\text{GC7 (np)}$$

$$r^2 = 0.87,\ q^2 = 0.84 \quad (1)$$

$$-\log K_i = 6.17 + 12.91*\text{GC1 (any)} + 13.45*\text{GC2 (np)} - 22.42*\text{GC3 (np)} - 28.75*\text{GC4 (any)} - 8.19*\text{GC5 (a)} + 45.78*\text{GC6 (any)} + 18.70*\text{GC7 (np)}$$

$$r^2 = 0.91,\ q^2 = 0.86 \quad (2)$$

$$-\log K_i = 6.16 - 6.90*\text{GC1 (any)} + 16.11*\text{GC2 (any)} - 8.56*\text{GC3 (np)} + 3.45*\text{GC4 (np)} + 15.83*\text{GC5 (hba)} + 17.45*\text{GC6 (any)} + 19.39*\text{GC7 (np)}$$

$$r^2 = 0.85,\ q^2 = 0.81 \quad (3)$$

$$-\log K_i = 5.99 + 13.02*\text{GC1 (np)} + 2.87*\text{GC2 (np)} - 4.72*\text{GC3 (any)} + 13.95*\text{GC4 (hba)} + 7.46*\text{GC5 (a)} + 17.38*\text{GC6 (hba)} + 12.35*\text{GC7 (any)}$$

$$r^2 = 0.86,\ q^2 = 0.82 \quad (4)$$

$$-\log K_i = 6.24 + 21.81*\text{GC1 (hba)} + 15.29*\text{GC2 (np)} + 29.78*\text{GC3 (any)} + 24.0*\text{GC4 (any)} - 9.45*\text{GC5 (np)} + 6.66*\text{GC6 (any)}$$

$$r^2 = 0.86,\ q^2 = 0.82 \quad (5)$$

Equation 1 represents the best model from the largest cluster. Equations 2 and 3 are the singleton, or outlier, models. Equation 4 is the best model from the smallest cluster, while eq 5 is the best model from the cluster solely occupied by models from alignment 8. One common feature across the five models is the dominance of nonpolar and steric effects (evidenced by presence of GCODs of IPE type "any" in the equations) on the inhibition constant, illustrated by the fact that 28 of the 34 grid cells in the five models are of IPE type (np) or (any).

Figure 2 shows the graphical representations, or 3-D pharmacophores, of the five 4D-QSAR models using compound 1, in its postulated active conformation for each model, as a reference structure. Each model appears to focus on a different region of the overall ligand−receptor interaction. Equations 1 and 2 focus on the R_1 and R_2 regions, respectively. Equations 3 and 5 both appear to concentrate on the entire "upper" portion of the molecules, encompassing the R_1 and R_2 regions shown in the reference structure of Figure 2. Equation 4 illustrates the important interactions in regions R_1 and R_4. Collectively, the five models define the important ligand−receptor interactions present for three of the four varied positions located around the cyclic core structure of the THP compounds. This model diversity, resulting in a comprehensive overview of ligand−receptor interaction, is realized by the clustering technique applied to identify models that are unique based on their structure−activity information content.

As might be expected, the position least varied in the training set, R_3, is not represented in any of the five models. The predicted active conformations by the five models are highly similar in the R_3 region of compound 1 and nearly identical in the R_4 region. There is consensus among the models regarding the preferred conformation of the phenyl ring joined to the core ring structure by a two-carbon linker, i.e., the R_4 region. Furthermore, four of the five models indicate that the presence of a nonpolar substituent on the carbon closest to the cyclic urea ring of this carbon linker enhances potency. This SAR feature is further illustrated by the fact that when a *polar* substituent is included at this position, inhibition potency is significantly reduced. This behavior is seen when comparing compound 32 to compound 33. The structures of compounds 32 and 33 are identical with the exception of a fluorine atom on carbon 1 of the linker in compound 33, leading to an activity 1.6 orders of magnitude less than compound 32, which contains a proton in the same position.

There is also consensus among the models as to the preferred conformation of the phenyl ring at R_1. The five models each contain an activity enhancing grid cell of type "any" at similar locations in the R_1 region. This GCOD suggests that the preferred conformation of an R_1 substituent is nearly perpendicular to the plane of the central cyclic urea ring, and, also, nearly parallel to the conformation of the phenyl ring located in the R_3 region. Additional support for this preferred conformation is given by the activity enhancing grid cell of IPE type "aro" in the R_1 region of eq 4, suggesting occupation of this grid cell by the aromatic phenyl ring enhances potency. Equations 1 and 3, although different in overall pharmacophore representation, share a similar potency reducing grid cell of type "np" in the R_2 region. This portion of the receptor site appears to exhibit a preference for electrostatic (polar) interactions. This adverse pharmacophore interaction is illustrated in Figure 3. Compounds 28 and 29 differ only by a single substituent in the R_1 and R_2 region, yet their inhibition potencies differ by more than a half order of magnitude ($\Delta[-\log K_i] = 0.52$). Both compounds are highly active (compound 28, $-\log K_i = 10.00$; compound 29, $\log K_i = 10.52$). Thus, a successful 4D-QSAR model is also able to distinguish subtle features

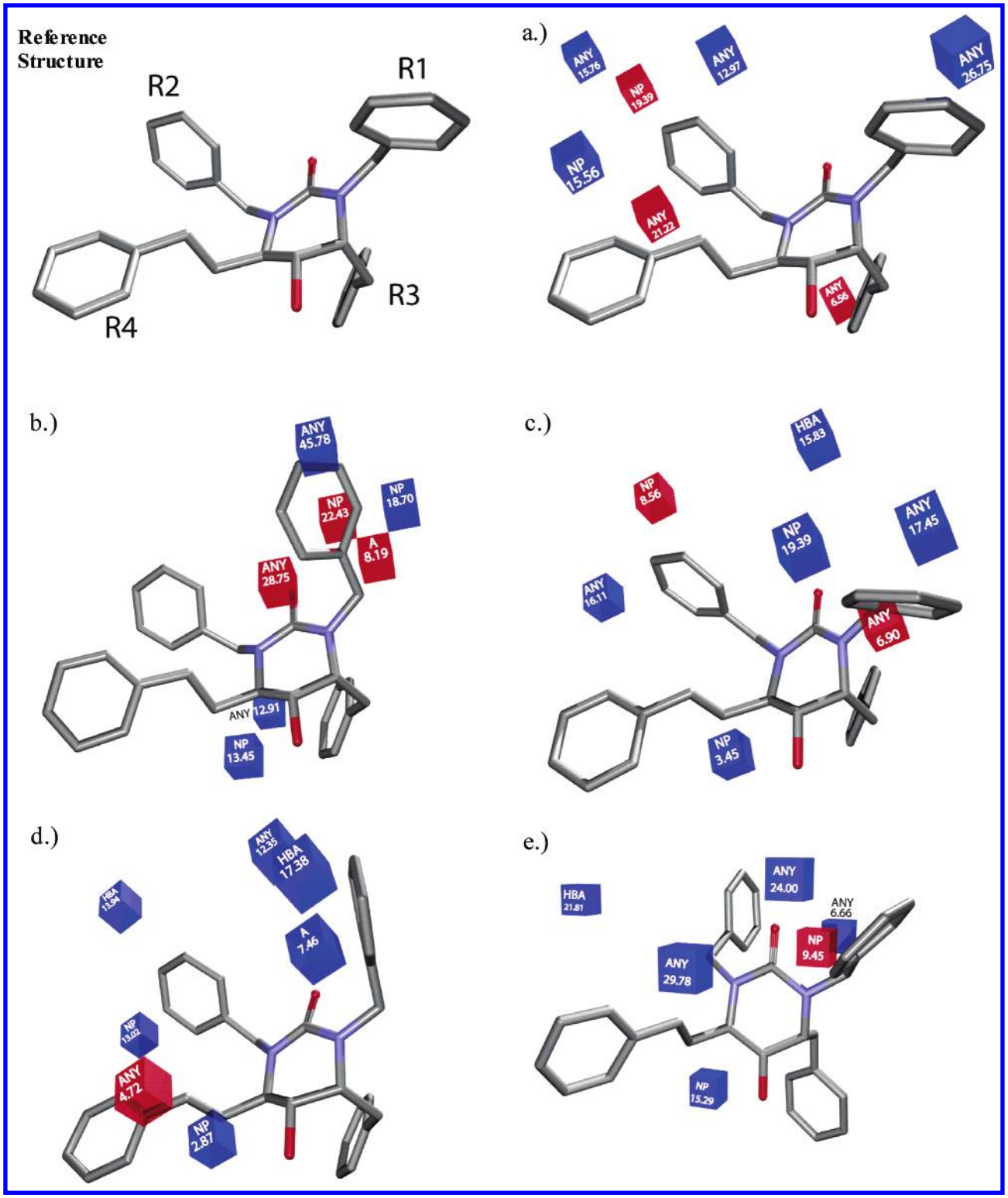


Figure 2. Predicted active conformations for compound 2 using each of the five unique 4D-QSAR models, eq 1 (a), eq 2 (b), eq 3 (c), eq 4 (d), and eq 5 (e). The grid cells from a model are the colored one angstrom cubes. Blue grid cells are those that contribute positively to inhibition potency, while the red grid cells reduce potency. The grid cells shown are labeled with their corresponding IPE type as well as the magnitude of the regression coefficient. Also shown is a reference structure where the four regions discussed in the text are defined.

that relate to activity differentiation among potent inhibitors. Compound 28 contains a 3-methylindazole moiety in the R_1 and R_2 region, while compound 29 contains a 3-aminoindazole moiety in this same region. The methyl of the indazole ring on compound 28 can occupy the activity reducing grid cell of type "np" in the R_2 region. Compound 29 contains the same indazole ring, but with an amino substituent, thereby eliminating nonfavorable interactions with the seemingly electrostatic (polar) region of the receptor present in the R_2 region.

Test Set Compound Prediction and Cluster Analysis. Table 6 contains the observed and predicted $-\log K_i$ values for a set of eleven test compounds. To evaluate the predictive ability of the models given by eqs 1−5, the r^2 of test set

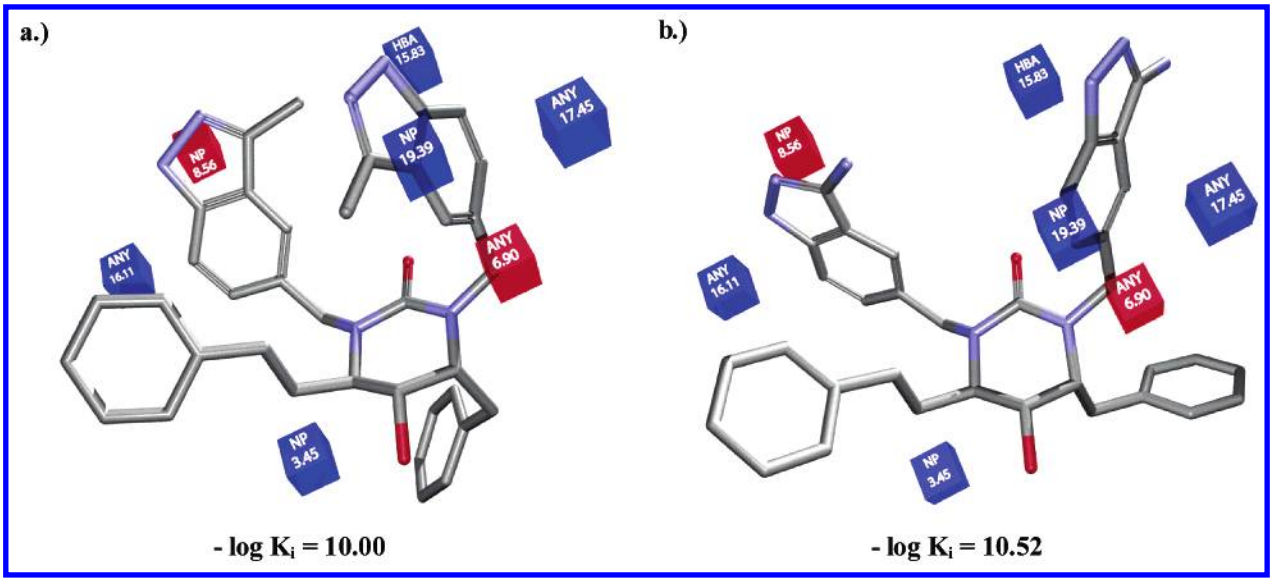


Figure 3. Predicted active conformations of compounds 28 (a) and 29 (b) by the 4D-QSAR model, eq 3. The IPE type and magnitude of the regression coefficient is given for each grid cell. Blue grid cells contribute positively to inhibition potency, while red grid cells reduce potency.

Table 6. Predicted and Observed $-\log K_i$ Values and the r^2 Values of Prediction of the Test Set Compounds for the Five Unique 4D-QSAR Models

	observed	predicted −log Ki					residual −log Ki				
compound	−log Ki	eq 1	eq 2	eq 3	eq 4	eq 5	eq 1	eq 2	eq 3	eq 4	eq 5
t1	10.70	9.05	8.22	9.23	10.79	9.80	−1.65	−2.48	−1.47	0.09	−0.90
t2	6.34	12.16	7.77	6.43	6.30	6.50	5.82	1.43	0.09	−0.04	0.16
t3	7.00	6.11	6.95	6.50	6.68	7.68	−0.89	−0.05	−0.50	−0.32	0.68
t4	7.85	9.21	9.83	8.87	8.12	8.05	1.36	1.98	1.02	0.27	0.20
t5	9.60	9.63	9.88	11.45	8.00	8.86	0.03	0.28	1.85	−1.60	−0.74
t6	10.10	11.78	10.22	8.82	8.39	10.91	1.68	0.12	−1.28	−1.71	0.81
t7	10.52	11.21	11.71	10.32	8.14	9.59	0.69	1.19	−0.20	−2.38	−0.93
t8	7.36	6.92	7.06	6.21	6.03	6.23	−0.44	−0.30	−1.15	−1.33	−1.13
t9	10.10	8.12	9.79	7.44	8.09	7.61	−1.98	−0.31	−2.66	−2.01	−2.49
t10	10.22	8.85	9.29	6.91	7.51	7.59	−1.37	−0.93	−3.31	−2.71	−2.63
t11	10.70	7.73	9.32	6.75	7.30	7.73	−2.97	−1.38	−3.95	−3.40	−2.97
		r-squared of prediction					−1.10	0.39	−0.60	−0.33	0.02

prediction was calculated as

$$r^2 = 1 - \sum(y_{i(pred)} - y_{i(obs)})^2 / \sum(y_{i(obs)} - k)^2 \quad (6)$$

where k is equal to the average of the observed test set values and $y_{i(pred)}$ and $y_{i(obs)}$ are the predicted and observed activity values for the ith test set compound, respectively. The *rank-difference correlation coefficient*, defined by eq 6, was chosen for its ease in interpretability. A rank-difference r^2 of greater than zero indicates that the model predicts better than if the average activity value over the test set compounds was chosen for each of the *predicted* activity values. Conversely, a value less than zero indicates that using the average activity value of the test set compounds gives a better overall prediction for the test set compounds than the model being used to make the prediction. Overall, eq 2 is the only model that produces significant predictions for the test set, *as a whole*, with an r^2 of prediction of 0.39, and with no outlier predictions (predicted values more than twice the standard deviation of the observed test set compound values). An inspection Table 6 reveals that specific compounds are predicted better by certain models. For example, eq 1 predicts the activity of test set compound t2 nearly 6 orders of magnitude larger than the observed activity measure. Conversely, eq 4 predicts the $-\log K_i$ of compound t2 nearly identical to that observed, with a residual of prediction of only −0.04 log units. Test set compound t2 is very similar to training set compound 1, the only difference being a double bond in the R_4 region of test set compound t2, a chemical structure feature not represented in the training set. Equation 1 contains no descriptors in the R_4 region, while eq 4 contains two descriptors in this region, possibly explaining why eq 4 predicts well for this compound, while eq 1 does not. The importance of representing the appropriate region of ligand−receptor interaction space in the 4D-QSAR model used to make a test set prediction is well illustrated in this example but may not be obvious in other cases. Therefore, the clustering method described above has been applied to improve test set predictions.

The five unique 4D-QSAR models contain a total of 34 unique descriptors, 16 of which are of IPE type “any”, and 12 are of IPE type “np”. This suggests that clustering the compounds based on properties related to the IPE “any” or “np” grid cell types may prove successful. Recently, we have developed a molecular similarity method based on the 4D-QSAR paradigm, called 4D-QSAR Molecular Similarity or *4D-QSARMS*.[22] Briefly, a set of measures of pairwise molecular similarity, based on the eight 4D-QSAR IPE types, can be obtained by comparing eigenvectors generated from

Table 7. Sum-Squared Error of Prediction Using the Five Unique 4D-QSAR Models for Training Set Compounds 1, 4, 12, 32, 42, 43, and 48

model	sum-squared error of prediction
eq 1	3.606
eq 2	0.726
eq 3	1.375
eq 4	0.516
eq 5	2.576

Table 8. r^2 of Test Set Prediction for the Three Clustering Levels of the Two IPE Types

IPE basis for clustering	cluster level		
	0.0025	0.005	0.01
any	0.61	0.68	−0.11
Np	−1.09	0.08	−0.18

principle component analysis of a distance-average matrix or DAM. The DAM is created by using distances between atoms of specified types from the MDS trajectories of a pair of molecules. In other words, to arrive at a measure for molecular similarity based on the "any" IPE type, one needs only to compare the eigenvectors from the "any" distance matrices of the two compounds of interest. The eigenvectors provide a convenient way in which to arrive at molecular similarity values based on the 4D-QSAR IPE types. Therefore, clustering compounds based on the "any" or "np" 4D-QSARMS eigenvectors will yield clustered groups that are similar with respect to these IPE types.

Figure 4 shows the results of clustering the training and test set compounds based on their "any" and "np" molecular similarity eigenvectors. Due to the large number of eigenvalues composing a 4D-QSARMS eigenvector of each compound, data reduction, by means of principle component analysis, was first performed. The cluster trees given in Figure 4 are based on the first three (largest) principle components of the independent variable matrix. An inspection of the two dendrograms reveals that test set compounds 6 and 7 do not cluster well with the other compounds in the remainder of the data set. This finding suggests that these two compounds are unique with respect to their "any" and "np" properties. The uniqueness of these two compounds most likely arises because they each contain relatively extended chains in the R_1 and R_2 region making them larger than the other compounds in the data set. Also apparent from the cluster trees is the unique nature of training set compound 3 with respect to its nonpolar and steric properties. Only protons are substituents in the R_1 and R_2 region of compound 3, making this compound unique in that it is much smaller than the other compounds in the data set.

The cluster trees were evaluated at three different levels, corresponding to cluster distances of 0.0025, 0.005, and 0.01, since this is the range where the majority of the separation takes place. The three levels were evaluated based on the resulting r^2 of the test set predictions. For example, in the "any" cluster tree at a level of 0.005, test set compounds 2 and 4 belong to a cluster containing training set compounds 1, 4, 12, 32, 42, 43, and 48. The sum-squared error of predictions by the five unique 4D-QSAR models for these seven training set compounds is given in Table 7. It is evident from Table 7 that the model having eq 4 predicts the best for this group of training set compounds. Therefore, eq 4 is used to predict test set compounds 2 and 4. Table 6 indicates that using eq 4 to predict these test set compounds, as opposed to the equation that represents the "overall" best model for prediction, namely eq 2, improves substantially the accuracy of prediction for these two test set compounds.

If, however, a test compound does not belong to a cluster at a given level, as is the case with test set compounds 6 and 7, the branches in the cluster tree of these singleton compounds are followed up the levels until they join a cluster, at which point the cluster is evaluated in the same manner as described for test set compounds 2 and 4. This procedure was carried out for each test set compound at each clustering distance level to arrive at an r^2 of test set prediction. The results are given in Table 8. The eq 2 of the best single unique 4D-QSAR model yielded an r^2 of test set prediction of 0.39. Clustering based on the "np" IPE type does not improve upon the predictions of eq 2. However, clustering based on the "any" IPE type at levels of 0.0025 and 0.005, to determine which equation to use to predict which test set compound, results in substantial improvements in test set prediction ($r^2 = 0.68$ for 0.005).

The purpose of the cluster analysis in improving test set prediction is to determine a priori which of the five unique models to use to predict each of the test set compounds. Since there are five unique models to choose from for each of the 11 test set compounds, there are $5,^{11}$ or 4.9×10^7 possible combinations of models available for the test set predictions, with each combination producing a unique r^2 of test set prediction value. Thus, there is the very real possibility that the improved predictive ability achieved by the clustering method developed in this study may be due to a random occurrence. A distribution plot of the r^2 value of prediction for the possible combinations of models to make the test set predictions was generated and is given as Figure 5. The table indicates that nearly half of the random combination models generated (46.6%) yield an r^2 of test set prediction within the range of −0.3 to 0.1. With an r^2 of test set prediction value of 0.39, eq 2, is within the top predictive 5% of all the combination of models (4.9% of all the combination of models have an r^2 of prediction greater than 0.35). Note that eq 2 would appear in the random selection process if every test compound was determined to be best predicted by eq 2. Finally, the r^2 of test set prediction (0.68) arrived at by clustering based on the "any" IPE type falls within the top 0.23% of all combination of models generated, i.e., 0.23% of all models generated have an r^2 of test set prediction greater than 0.65. This simulated prediction modeling experiment supports the idea that the improved test set prediction is not based on a random occurrence.

CONCLUSION

The work presented in this paper utilizes the 4D-QSAR paradigm to provide further analysis of the binding characteristics of the THP analogues to the aspartyl protease of the human immunodeficiency virus type-I. Four important binding regions of the THP ligands to the HIVPR receptor are identified, described here as R_1 through R_4. The five statistically significant models generated ($q^2 > 0.8$) are in consensus concerning the importance of steric and nonpolar effects on potency, and each model appears to represent a

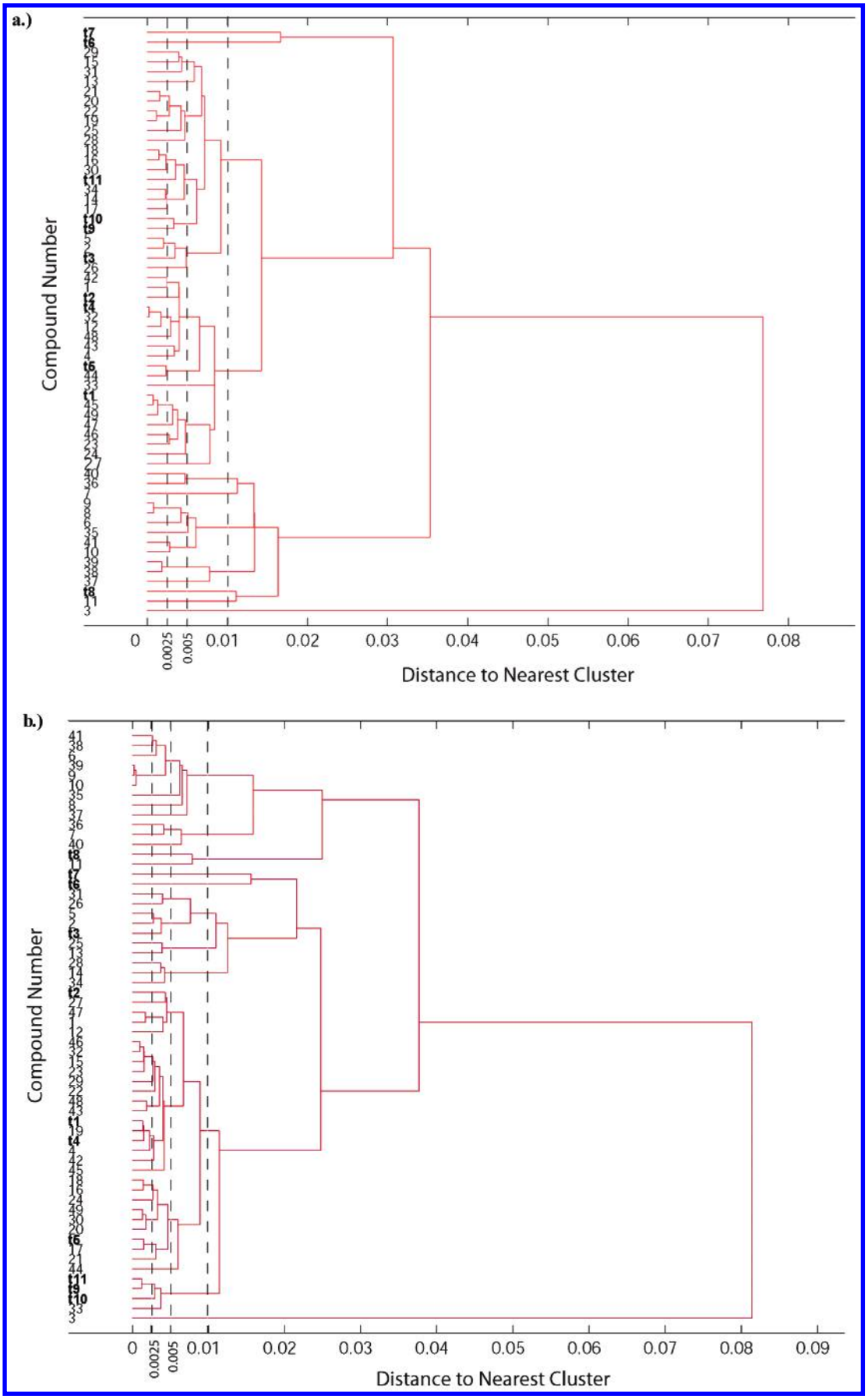


Figure 4. Dendrograms resulting from the clustering of the first three principle components using the (a) "any" similarity eigenvectors and (b) the "np" similarity eigenvectors. The compound numbers in bold on the left correspond to the test set compounds, while the other numbers correspond to training set compounds.

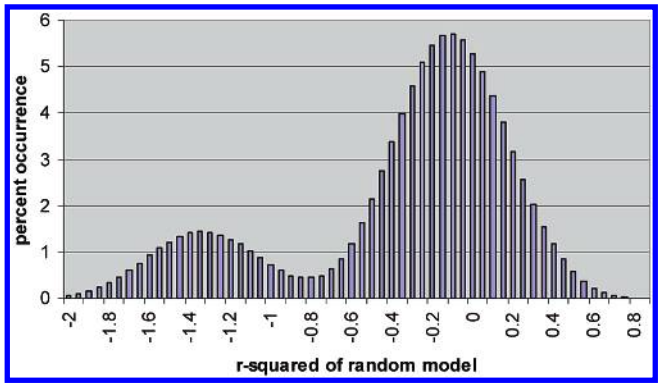


Figure 5. Histogram plot of percent occurrence of varying r^2 values for the 10^8 random models generated.

particular region of the receptor environment. To the best of our knowledge, the current work also represents the first application of a simple clustering algorithm to aid in the identification of the components of a manifold model as well as to assist in the prediction of test set, or unknown, compounds.

The need for a discrete algorithm to identify the components of a manifold model becomes apparent when it is realized that there cannot be a single solution, or model, that can fully represent the multidimensional nature inherent to ligand−receptor binding. Employing multiple alignments, for example, in a 4D-QSAR study is one way in which to approach this problem. The multiple alignment criteria offers a possible solution to a system that involves more than one binding mode or a dependence on different regions of the ligand molecule. The alignment analysis therefore provides a clever way to probe the local environment of the ligands in the receptor. By holding constant a portion of the ligand, by means of the alignment, the region, or feature that is not held constant is therefore analyzed. If a certain alignment does not produce significant models, the region being probed with that alignment is not of interest for the data set. When the situation occurs that several alignments provide significant models, it becomes important to choose those models that are unique in their representation of the data.

The cluster analysis described in this work provides a quantitative and convenient method to identify independent solutions resulting from a single data set. The eight alignments utilized in the 4D-QSAR analysis of the THP data set produced 24 five-, six-, and seven-term models that were all statistically significant. The result of the cluster analysis method was five models that were unique in their representation of the training set data. It is important to note that the five models from the cluster analysis are independent in their activity prediction as well as their *pharmacophore representation*, as shown in Figure 2. A possible explanation for this is that a single model can only represent a limited region of space as determined by the particular alignment of the training set compounds for that model. Therefore, it is increasingly important to systematically identify the complete manifold model to fully evaluate the training set data.

By acknowledging the possibility that a single model may only represent a limited region of space, and, therefore, may only produce reliable results for compounds that apply to that region of space, one must realize that the model is *real* only for correspondingly appropriate compounds. In all other cases use of the model would be inappropriate. This becomes clear upon inspection of Table 6. The r^2 value for each of the five models would increase significantly if one or two of the test set compounds that were not predicted well were removed. For example, the r^2 value for eq 1 increases from −1.10 to 0.14 when test compounds 2 and 11 are removed. This is evidence that these two test compounds interact with the receptor in a way not fully, or appropriately, described by eq 1. Therefore, there must also be present in the QSAR method a way to identify discretely which models produce the *most real* predictions for each of the test, or unknown, compounds.

The second clustering method described in this work shows that the accuracy of test set predictions can be significantly improved by identifying in a practical manner which of the models produce *real* predictions for the corresponding test set compounds. The r^2 of prediction of the combination model is nearly 2-fold higher than the best single model, an increase that is achieved by simply acknowledging that the predictions of each of the models are only *real* for specific subsets of the test set compounds. The caveat of this procedure is that there must be a systematic means to identify which model is appropriate for which compound. This is achieved in the current work by a simple clustering algorithm.

Overall, the methods employed in this paper illustrate the success of simple clustering in a 4D-QSAR analysis and provide general guidelines as to how this method can be utilized to help achieve more complete data extraction in a QSAR paradigm.

ACKNOWLEDGMENT

This work was supported, in part, with gift funds from The Chem21 Group, Incorporated. Resources from the Laboratory of Molecular Modeling and Design were used.

REFERENCES AND NOTES

(1) Roberts, N. A.; Craig, J. C.; Duncan, I. B. HIV proteinase inhibitors. *Bio. Soc. Trans.* **1992**, *20*, 513−516.
(2) Sakurai, M.; Sugano, M.; Handa, H.; Komai, T.; Yagi, R.; Nishigaki, T.; Yabe, Y. Studies of HIV-1 protease inhibitors. I. Incorporation of a reduced peptide, simple amino alcohol, and statine analogue at the scissile site of substrate sequences. *Chem. Pharm. Bull.* **1993**, *41*, 1369−1377.
(3) Kaldor, S. W.; Kalish, V. J.; Davies, J. F.; Shetty, B. V.; Fritz, J. E.; Appelt, K.; Burgess, J. A.; Campanale, K. M.; Chirgadze, N. Y.; Clawson, D. K.; Dressman, B. A.; Hatch, S. D.; Khalil, D. A.; Kosa, M. B.; Lubbehusen, P. P.; Muesing, M. A.; Patrick, A. K.; Reich, S. H.; Su, K. S.; Tatlock, J. H. Viracept (nelfinavir mesylate, AG1343): a potent, orally bioavailable inhibitor of HIV-1 protease. *J. Med. Chem.* **1997**, *40*, 3979−3985.
(4) De Lucca, G. V.; Liang, J.; Aldrich, P. E.; Calabrese, J.; Cordova, B.; Klabe, R. M.; Rayner, M. M.; Chang, C.-H. Design, synthesis and evaluation of tetrahydropyrimidines as an example of a general approach to nonpeptide HIV protease inhibititors. *J. Med. Chem.* **1997**, *40*, 1707−1719.
(5) Gupta, S. P.; Babu, M. S. Quantitative structure−activity relationship studies on cyclic cyanoguanidines acting as HIV protease inhibitors. *Bioorg. Med. Chem.* **1999**, *7*, 2549−2553.
(6) Hagen, S. E.; Domagala, J.; Gajda, C.; Lovdahl, M.; Tait, B. D.; Wise, E.; Holler, T.; Hupe, D.; Nouhan, C.; Urumov, A.; Zeikus, G.; Zeikus, E.; Lunney, E. A.; Pavlovsky, A.; Gracheck, S. J.; Saunders, J.; VanderRoest, S.; Brodfuehrer, J. 4-Hydroxy-5,6-dihydropyrones as inhibitors of HIV protease: the effect of heterocyclic substituents at C-6 on antiviral potency and pharmacokinetic parameters. *J. Med. Chem.* **2001**, *44*, 2319−2332.
(7) Kohl, N. E.; Emini, E. A.; Schleif, W. A.; Davis, L. J.; Meimbach, J. C.; Dixon, R. A. F.; Scolnick, E. M.; Sigal, I. S. Active human immunodeficiency virus is required for viral infectivity. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *85*, 4686−4690.

(8) Ashorn, P.; McQuade, T. J.; Thaisrivongs, S.; Tomasselli, A. G.; Tarpley, W. G.; Moss, B. An inhibitor of the protease blocks maturation of human and simian immunodeficiency viruses and spread of infection. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 7472−7476.
(9) McQuade, T. J.; Tomasselli, A. G.; Liu, L.; Karacostas, V.; Moss, B.; Sawyer, T. K.; Heinrikson, R. L.; Ratpley, W. G. A synthetic HIV-1 protease inhibitor with antiviral, activity arrests HIV-like particle maturation. *Science* **1990**, *247*, 454−456.
(10) Sakurai, M.; Higashida, S.; Sugano, M.; Komai, T.; Yagi, R.; Ozawa, Y.; Handa, H.; Nishigaki, T.; Yabe, Y. Structure−activity relationships of HIV-1 PR inhibitors containing AHPBA. *Bioorg. Med. Chem.* **1994**, *2*, 807−825.
(11) Huang, X.; Xu, L.; Luo, X.; Fan, K.; Ji, R.; Pei, G.; Chen, K.; Jiang, H. Elucidating the inhibiting mode of AHPBA derivatives against HIV-1 protease and building predictive 3D-QSAR models. *J. Med. Chem.* **2002**, *45*, 333−343.
(12) Senese, C. L.; Hopfinger, A. J. Receptor independent 4D-QSAR analysis of a set of norstatine derived inhibitors of HIV-1 protease. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1297−1307.
(13) Barnard, J. M.; Downs, G. M. Clustering of chemical structures on the basis of two-dimensional similarity measures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 644−649.
(14) Fan, Y.; Leming, M. S.; Kohn, K. W.; Pommier, Y.; Weinstein, J. N. Quantitative structure-antitumor activity relationships of camptothecin analogues: cluster analysis and genetic algorithm based studies. *J. Med. Chem.* **2001**, *44*, 3254−3263.
(15) Brown, R. D.; Martin, Y. C. Use of structure−activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.
(16) McFarland, J. W.; Gans, D. J. Cluster significance analysis contrasted with three other quantitative structure−activity relationship methods. *J. Med. Chem.* **1987**, *30*, 46−49.
(17) Downs, G. M.; Barnard, J. M. *Rev. Comput. Chem*. **2002**, *18*, 1−40.
(18) Nair, A. C.; Jayatilleke, P.; Wang, X.; Miertus, S.; Welsh, W. J. Computational studies of tetrahydropyrimidine-2-one HIV-1 protease inhibitors: improving three-dimensional quantitative structure−activity relationship comparative molecular field analysis models by inclusion of calculated inhibitor and receptor based properties. *J. Med. Chem.* **2002**, *45*, 973−983.
(19) DeLucca, G. V.; Liang, J.; DeLucca, I. Stereospecific synthesis, structure−activity relationship and oral bioavailability of tetrahydropyrimidine-2-one HIV protease inhibitors. *J. Med. Chem.* **1999**, *42*, 135−152.
(20) DeLucca, G. V.; Liang, J.; Aldrich, P. E.; Calabrese, J.; Cordova, B.; Klabe, R. M.; Rayner, M. M.; Chang, C.-H. Design, synthesis and evaluation of tetrahydropyirimidinones as an example of a general approach to nonpeptide HIV protease inhibitors. *J. Med. Chem.* **1997**, *40*, 1707−1719.
(21) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, G. Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509−10524.
(22) 4D-QSAR, Version 3.0; The Chem21 Group, Inc., 1780 Wilson Drive, Lake Forest, IL 60045, 2002.
(23) Pan, D.; Tseng, Y.; Hopfinger, A. J. Quantitative Structure-Based Design: Formalism and Application of Receptor-Dependent RD-4D-QSAR Analysis to a Set of Glucose Analogue Inhibitors of Glycogen Phosphorylase. *J. Chem. Inf. Comput. Sci.* **2003**, publication in progress.
(24) *HyperChem Program Release 6.01 for Windows;* Hypercube, Inc.; 2000.
(25) *MOLSIM V3.0*; D. C. Doherty and The Chem21 Group, Inc., 1780 Wilson Drive, Lake Forest IL 60045, 1998.
(26) Rogers, D.; Hopfinger, A. J. Application of the genetic function approximation to quantitative structure−activity relationships and quantitative structure−property relationships. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 854−866.
(27) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

CI034168Q