Identification of Common Functional Configurations Among Molecules

Doug Barnum,*,† Jonathan Greene,† Andrew Smellie,† and Peter Sprague

Molecular Simulations Incorporated, 555 Oakmead Parkway, Sunnyvale, California 94086

Received July 29, 1995⊗

A new algorithm for identifying three-dimensional configurations of chemical features common to a set of molecules is described. The algorithm scores each configuration based both on the degree to which it is common to the input set and its estimated rarity. The algorithm can be applied to molecules with large (several hundred) conformational models. Results from the application of this algorithm to three data sets are discussed: PAF antagonists, HIV reverse transcriptase inhibitors, and HIV protease inhibitors. Of particular interest is a common configuration identified for a set of HIV reverse transcriptase inhibitors; this configuration is shared by two new, potent inhibitors that were recently described in the literature.

INTRODUCTION

We consider the problem of finding three-dimensional configurations of functional groups common to a set of molecules. Generally, the molecules in question are known to be active on an assay of interest. This problem has several applications:

1. Formulation of a generalized search query that may turn up functionally equivalent but structurally novel molecules in a 3D database.

2. Identification of possible alignments of active conformers for input to comparative molecular field analysis (CoMFA).[1] When molecules are not congeneric, this can be a daunting task requiring some degree of automation. For example, a predictive CoMFA model based on automatically aligned angiotensin converting enzyme inhibitors has recently been reported.[2] Identification of alignments is also a required first step in several other 3D QSAR methods.[3,4]

3. Identification of candidate structure−activity hypotheses as starting points for further optimization.[5]

Finding common configurations requires a search over two large spaces: the conformations that may be assumed by each molecule and the many possible correspondences among features in the various molecules. For molecules that are flexible and have many functional groups that may be relevant, both spaces are individually daunting and the combined search even more so.

The problem has been extensively studied. Mayer[6] et al. applied a systematic search algorithm[7] to simultaneously search the conformational spaces of the molecules. A drawback is that the discrete sampling of torsion space may miss important conformations. Ensemble distance geometry provides an alternative.[8] Either of these approaches assumes knowledge of the correct correspondence of features among the molecules.

The most popular approach in more recent efforts[9−12] is the use of a clique-detection algorithm to find common configurations. In this approach a small number of discrete conformations (10−20) must be provided for each molecule. These are typically determined by distance geometry and energy minimization or by a crystal structure if one is available. Since the clique-detection algorithm works on a pair of conformers at a time, a particular molecule and conformation must be chosen as a reference. Ideally, a reported common configuration need not have any "reference" conformer as its source. In the absence of an efficient solution to this difficult problem, it is preferable to approximate such a solution by allowing many or all conformers of many or all molecules to serve as references, rather than to choose somewhat arbitrarily only one reference conformer. (Let us define a *reference molecule* to be a molecule each of whose conformers is in turn treated as a reference conformer.) The clique-detection algorithm may be run repeatedly with different reference conformers, but in this case the overall run time scales quadratically with the number of conformers per molecule.

In this paper we report an improved search algorithm for identifying common configurations of features among discrete conformations of a set of molecules and a novel way to score them for significance. No reference conformer need be specified, eliminating the requirement of foreknowledge of an active conformer. Allowing every conformer of every molecule to serve as a reference enables us to find more and larger common configurations. All molecules and conformers are treated equivalently.

Despite the elimination of the need for a specific reference, the run time is approximately linear in the number of conformers per molecule. This greatly increases the number of conformers that can be accommodated. Using a few hundred conformers per molecule, coupled with judicious choice of the conformers by the poling method,[13] enables us to effectively handle large, flexible molecules.

Our algorithm also lends itself to relaxing the requirement that all molecules possess all identified features. This is important because it is often the case that a molecule can be highly active despite lacking a feature relevant in the binding of other moleclues. Earlier, alternative approaches to molecular similarity problems have also stressed the value of just this sort of flexibility.[14]

We have found it most effective to operate in a regime with relatively large tolerances and generalized functional definitions of groups that may be involved in binding.[15] The tolerances are determined by the dependence of interaction energies on relative geometry[15] and by the precision of the conformational model, according to the two-set hole-size measure.[13] Adequate selectivity is maintained by identifying

* Author to whom correspondence should be addressed. Email: doug@info.combichem.com.

† Currently with CombiChem, Inc., 1225 Innsbruck Dr., Sunnyvale, CA 94089.

⊗ Abstract published in *Advance ACS Abstracts,* March 1, 1996.

0095-2338/96/1636-0563$12.00/0

larger sets of features and by insisting on superimposability of features (RMS fit) rather than simply similarity of interfeature distances.

Since many common configurations are often present, there is an obvious need to score and rank them. The size of the configuration can sometimes be misleading as a score. For instance a configuration of two charge centers, which are quite rare, may be much more significant than a configuration of four hydrophobes, which are quite common. Allowing configurations to miss one or a few features on one or a few of the molecules further complicates the scoring problem. For instance, a configuration of three features possessed by all molecules may or may not be more significant than a configuration of four features possessed by all but one molecule.

If no missing features are allowed and if a large additional set of inactive molecules can be provided the scoring scheme of Klopman[16] and Golender and Rozenblit[17] can be applied. They suggest computing two statistics (probability and reliability) based on the number of molecules that are or are not active and either do or do not possess the configuration. Because two statistics are used, it is not always possible to obtain a unique rank order on the configurations.

In this paper we propose a new scoring method based on a single maximum likelihood criterion. Unlike previous methods, it can account for partial fits to a configuration. Also, we avoid the need to estimate the rarity of a configuration from its frequency among the set of inactive molecules (which can be unreliable, especially for rare configurations). Instead, we derive a regression equation for this purpose.

We tested our method as follows. The set of eight PAF antagonists previously investigated by Bures[18] et al. was used in a systematic study assessing the impact of various factors on the number of common configurations found. The factors include the following:

- use of a single reference conformation vs all conformations of all molecules,
- generation of conformational models using distance geometry vs the poling method,
- consideration of only ideal hydrogen bond geometries vs including other low energy geometries

We also sought a common configuration for a group of HIV reverse transcriptase inhibitors and compared the result with a docking alignment for one of the compounds proposed by Gussio, Pattabiraman, and co-workers.[19] Finally, we verified that the common feature configuration and alignments found by our method correspond to those indicated by crystal structures of several HIV protease inhibitors.

An earlier version of our algorithm was embedded in the Catalyst/Hypo system[5] to identify candidate structure−activity hypotheses. The improved version reported here has recently been made directly accessible for independent use.[20]

METHODS

The input to the search algorithm is a set of conformers for each molecule and a dictionary of feature definitions.

Conformational Analysis. Conformers were generated using the poling method of Smellie[13] et al. with a 41.8 KJ energy threshold. In brief, poling refers to local modifications of the force field in which successive minimizations occur, which "push" new minimized conformers away from those already discovered. This helps to ensure broad coverage of low energy conformational space.

For comparative purposes, conformers were also generated using the traditional technique of distance geometry and minimization. Specifically, conformers were generated using the basic EMBED algorithm of Crippen and Havel.[21] The metrization algorithm of Havel was used to iteratively refine trial distances during distance selection to guarantee that all triples of randomly selected distances satisfied the triangle inequality. Each trial conformer after embedding was submitted to a standard local energy minimization using the CHARMm force field[22] and a conjugate gradient minimizer.[23] Minimization terminated when (a) the RMS cartesian gradient $\leq$0.1 KJ/A; (b) the energy changes by $<$0.1 KJ over five iterations; or (c) no atom moves by more that 0.01 A over five iterations. A collection of diverse low energy conformers was obtained by embedding and minimizing repeatedly (up to 500 times per molecule), and applying an heuristic[13] to extract a diverse sampling of conformers such that the hole size found by the full conformer set in the extracted set was less than 1.0 Å.

Additionally, in the PAF results, a conformational model composed solely of the crystal structure of antagonist RP-59227 cited by Bures[18] et al. was used instead of a full, generated model of RP-59227.

Feature Definitions. Molecular features considered are hydrogen bond donors and acceptors, negative and positive charge centers, and regions of exposed hydrophobic surface. Both ligand atoms and projected positions of complementary site atoms were considered for hydrogen bonding features. Only projected positions that are outside the ligand surface are allowed. Because interactions often occur even when the atoms are not positioned in the ideal hydrogen bonding geometry, both ideal and nonideal projected positions are considered. On carbonyl oxygens, for instance, hydrogen bonds are considered along the two ideal lone pair positions as well as in the linear position. When the acceptor or donor atom is at the end of a rotatable bond (for example, −OH or −NH2) we consider the three ideal staggered rotations as well as three additional intervening rotations for a total of six samples of the circle swept out by the projected position.

The detailed definitions for the five chemical functions have been reported previously.[15] Features of the various molecules were identified automatically with no manual intervention.

Algorithm. The program begins by identifying configurations of features common to the molecules. More precisely, a *configuration* is a set of relative locations in 3D space, each associated with a type of feature. A molecule matches a configuration if it possesses a set of features and a conformation such that the set of features can be superimposed with the corresponding locations. A set of features is considered superimposed if each feature lies within a specified distance, the *tolerance*, from the corresponding ideal location. [We actually approximate this requirement very closely through a combination of interpoint distance constraints and constraints on the RMS deviations of the entire set and its subsets from the configuration.] Ideally we wish to find configurations that are matched by all of the molecules. One can imagine a common configuration to be the center of a geometric neighborhood within which some conformation of each molecule lies. In practice, each

common configuration we find is precisely matched by one of the reference molecules, that molecule being the source of the configuration itself. Typically, the entire set of molecules would be considered as reference molecules.

Often, the requirement that all molecules match all features in the configuration is relaxed. Certain molecules may be permitted to miss a feature as long as the total number of molecules missing a feature, or possibly even any particular feature (note that this is a different constraint), remains below a specified limit. Furthermore, it is possible to permit up to a specified number of certain appropriately flagged molecules to *completely* miss a "common" configuration: a *complete miss* occurs when more than one feature must be omitted from a configuration, in order that the molecule map to all remaining features. (Note that according to this definition, all complete misses are misses, but the reverse is false: a molecule that can map to all but one feature of a configuration is, in the context of that configuration, a miss, but it is not a complete miss.) The ability to allow a number of complete misses helps account for "active" molecules that may bind at a different site, for without the ability to relax these constraints, such molecules would be likely to unfairly rule out common configurations which explain the activity of the remaining compounds.

To encourage a degree of dispersion amongst the features of a configuration, one may specify a minimum allowed spacing between any two points in a configuration. This parameter is typically set to 2.0 Å.

The configurations are identified by a pruned exhaustive search, starting with small sets of features and extending them until no larger common configuration exists. A modified version of the Ferro−Hermans algorithm[24] is used to verify superposition. We make two more useful definitions before providing an overview of the algorithm. A *partition* is an object associated with all configurations of a particular type, e.g., the DHH partition would be associated with all configurations consisting of one hydrogen-bond **D**onor feature, and two **H**ydrophobic region features. A *subpartition* of a partition P is a partition having all the feature types possessed by P except for one, e.g., both the DH partition and the HH partition are subpartitions of the DHH partition. The core of the algorithm may now be viewed in the following manner:

Working from small partitions to large, for each partition:

Derive associated configurations for each reference molecule, and for each configuration:

Verify that the subconfiguration associated with each subpartition was considered "common", and that the configuration itself passes various tests:

- Surface accessibility of the substituent features.
- Satisfaction of minimum inter-feature spacing.

Enter the configuration in a high-dimensional data structure[25] indexed by inter-feature distances.

Proceed similarly for non-reference molecules, absent the subconfiguration verification.

For each reference configuration, search both the reference and non-reference data structures for the neighbors necessary to deem the configuration "common". This search may be optimized in the following manner:

- We first identify those configurations which are "potentially in common" based on a comparison of interpoint distances,
- Then only for those which pass this filter, we compute the best RMS superposition, weighted by feature tolerances, and if justified, label them as "common".

Proceed to the next partition, until no more common configurations are possible, or parameters limiting configuration size are surpassed.

All satisfactory configurations found are scored and ranked using the formulas below. The final step is to identify and report a small set of high scoring, diverse configurations using a clustering procedure. A simple greedy clustering heuristic is used.

Scoring. Our goal is to compute a score for each configuration that reflects how well it describes the set of active molecules. By putting the problem in the following framework, we can apply standard techniques for assessing statistical hypotheses against experimental data.

Let M denote the number of active molecules supplied as input. Imagine that these M molecules are a subset of a large universe of N arbitrary molecules, $N \gg M$. The available data is that these M molecules are active and the remaining N−M are not. (Although some of the remaining molecules may actually turn out to be physically active, this is not part of the available data. So when we speak of a molecule as being active in this section, we mean only that it is among those M specified.)

Each configuration provides us with an hypothesis of the following form. A molecule that matches the configuration has probability a(Match) of being active. A molecule that does not match the configuration has a probability a(NoMatch) of being active. The constants a(Match) and a(NoMatch) are specified as part of the hypothesis, with a(Match) presumably much greater than a(NoMatch). Thus an hypothesis assigns a probability of activity to a molecule based on whether it matches the configuration.

To account for partial matches, we can generalize this to allow the constants to be a function $a(x)$, where x characterizes the manner in which the molecule matches the configuration. We define x as follows. Observe that a configuration of K features has exactly K subsets of size $K - 1$. We order these subsets based on their selectivity; subset 1 is the least selective (many of the N molecules match it), and subset K is the most selective (very few of the N molecules match it). We can now define $K + 2$ different classes of match as follows:

$x = K + 1$	the molecule matches all K features of the configuration
$0 < x < K$	the molecule matches subset x but does not match any more selective (higher numbered) subset
$x = 0$	the molecule does not match all K features or even any subset of $K - 1$ of the features

The classic maximum likelihood rule rates hypotheses based on the probability of the observed data D under the assumption that the hypothesis H is correct. The hypothesis with the highest value of $P(D|H)$ is preferred. This probability is

$$P(D|H) = \prod_{i=1}^{M} a(x_i) \cdot \prod_{i=M+1}^{N} (1 - a(x_i)) \qquad (1)$$

where x_i is the class of match between molecule i and the configuration, $a(x)$ is the probability of activity for a molecule with a class x match to the configuration, and the molecules are numbered so that $1 \le i \le M$ are the active ones.

Let us now define $q(x)$ to be the fraction of the M active molecules that has a class x match to the configuration, and

$p(x)$ to be the fraction of all N molecules that have a class x match to the configuration. Thus there are $Mq(x)$ active molecules with a class x match and $Np(x)$ molecules with a class x match in all. It can be shown that (1) is maximized by choosing

$$a(x) = \frac{Mq(x)}{Np(x)}$$

In evaluating configurations we choose constants $a(x)$ with these optimal values. Taking the logarithm of (1) for convenience, we have

$$\log_2 (P(D|H)) = \sum_{i=1}^{M} \log_2 \left(\frac{Mq(x_i)}{Np(x_i)}\right) + \sum_{i=M+1}^{N} \log_2 \left(1 - \frac{Mq(x_i)}{Np(x_i)}\right) \quad (2)$$

Since $p(x)$ is a constant independent of N, we expect $Mq(x) \ll Np(x)$. This makes the second summation in (2) negligible. We are left with

$$\log_2 (P(D|H)) \approx \sum_{i=1}^{M} \log_2 \left(\frac{Mq(x_i)}{Np(x_i)}\right) = M \log_2 \left(\frac{M}{N}\right) + \sum_{x} Mq(x) \log_2 \left(\frac{q(x)}{p(x)}\right) \quad (3)$$

Since the first term in (3) is independent of the configuration, we drop it and define just the second term as the score

$$\text{score} = M \sum_{x} q(x) \log_2 \left(\frac{q(x)}{p(x)}\right) \quad (4)$$

The maximum likelihood rule would choose a hypothesis based on a configuration that maximizes this score.

We note that there are other possible interpretations of the scoring function (4). If the factor M is dropped, we have the Kullback−Liebler distance between distributions q and p. If a term penalizing the complexity of the hypothesis is added to the score, we obtain the difference between the minimum description length of the data using the configuration[26] and the length of a naive description.

The one issue remaining is how to determine $p(x)$.

Estimating the Rarity of a Configuration. Conceptually, the value $p(x)$ is the fraction of hits returned from a search on a large, diverse database achieving a class x match. Accordingly we estimate $p(x)$ by means of a regression model for the number of hits found in an actual database when searching with various configurations.

For perfect matches, $x = K + 1$, we use the regression model

$$-\log_2 (p(K+1)) = C_0 + (\sum_i C_i K_i) + C_{\text{disp}} \cdot \text{disp}$$

where K_i is the number of instances of feature definition i among all K features in the configuration, and disp is a measure of the dispersion of the locations in the configuration

$$\text{disp}^2 = \sum_j ||L_j - \bar{L}||^2$$

Here L_j is the position of the jth location in the configuration and $\bar{L}$ is the centroid of the locations.

For partial matches ($0 < x \leq K$), we make the following approximation

$$p(x) = \text{Prob}\langle \text{match subset } x \mid \text{no match to subsets } x+1, ..., K, \text{ nor to all } K \text{ features} \rangle \approx \text{Prob}\langle \text{match subset } x \rangle$$

The latter quantity can in turn be found by applying the above regression model for perfect matches to the subconfiguration including the features in subset x. The approximation is motivated by the fact that the higher numbered subsets are more selective than subset x. In any event, the approximation is a valid upper bound by the correlation inequality,[27] assuming independence of the features present at each location in a conformer.

The C values were empirically determined by multiple least-squares regression on the negative log of the frequency of occurrence of 1500 randomly chosen configurations in the BioByteMasterFile database. The constants are

$$C_0 = -6.6$$

$$C_{\text{acceptor}} = 2.5$$

$$C_{\text{donor}} = 2.7$$

$$C_{\text{hydrophobic}} = 1.9$$

$$C_{\text{negative}} = 3.8$$

$$C_{\text{positive}} = 3.5$$

$$C_{\text{disp}} = 0.54 \text{ per angstrom}$$

The RMS error in the resulting estimates of $\log_2 (p(K+1))$ over the 1500 configurations was 0.7.

RESULTS

1. PAF Antagonists. An initial run, run A_r, was done to approximate the conditions reported in Bures et al. The tolerance was 0.85 Å on locations (equivalent to 1.7 Å on distances). The crystal structure was used for RP-59227, and 20 distance geometry conformers were chosen to represent each of the remaining molecules. The r subscript indicates that the crystal structure was the sole reference conformer; only configurations from the crystal structure which also fully matched some set of features from each other antagonist were reported. Two configurations are found, one with two acceptors and one with an acceptor and two hydrophobes. This is consistent with the finding of Bures et al. that a configuration of two acceptors and a hydrophobe could not be found at this tolerance unless three of the molecules were dropped.

Another run, run A, was similar to A_r except that all conformers of all eight leads are considered as reference conformers. Runs B_r and B were analogous to runs A_r and A except that poled conformers were used instead of distance geometry conformers. These runs are summarized in Table 1.

A comparison of runs A_r and A shows the importance of using multiple reference conformers. Run A finds several

Table 1. PAF Runs[a]

				configurations			
				AAH		AHH	
run	confs	ref confs	H-bonds	no.	best score	no.	best score
A_r	disgeom	one	all	0	N/A	2	38.3
A	disgeom	all	all	69	38.8	286	40.7
A_i	disgeom	all	ideal	13	32.1	4	40.7
B_r	poled	one	all	2	38.0	5	40.9
B	poled	all	all	969	48.7	1580	46.3

[a] Key: A = hydrogen-bond Acceptor, H = Hydrophobic region.

configurations consisting of two acceptors and one hydrophobe, while run A_r finds none of this type. Comparing B_r and *B* we see that the latter finds many more configurations of a given type than the former, and, more significantly, several of the run *B*'s configurations are of considerably higher score than the highest score configuration found by run B_r.

As was mentioned earlier, treating all molecules as reference molecules provides a closer approximation to the solution of the "ideal" common configuration problem, wherein no compound need serve as the source for a configuration. The above results illustrate the benefit of a such a scheme: higher scoring, larger common configurations are identified.

Comparison of runs A_r and *A* with B_r and *B* reveals the importance of using poled conformers instead of distance geometry conformers. Run B_r identifies two acceptor−acceptor−hydrophobe configurations, while run A_r finds none. The configurations identified by run *B* are of significantly higher score than those found by run *A*. This is consistent with the poled conformers' better coverage of low-energy conformational space.

Run A_i was done to demonstrate the loss of configurations if consideration is restricted to ideal hydrogen bond positions only. Comparison or runs *A* and A_i shows that nearly all acceptor−hydrophobe−hydrophobe configurations vanish in A_i, and the number of acceptor−acceptor−hydrophobe configurations is sharply reduced. While consideration of nonideal hydrogen bond positions adds to the computational complexity of the problem, the results may be much improved by the extra effort, as this example illustrates.

In many cases, it is unreasonable to expect that each molecule binds to all relevant features of the receptor; three out of four, for example, may well be enough for the compound to be active. Thus, it is desirable for reported common configurations to allow for this flexibility. In addition, it may well be the case that some molecules bind through a different mechanism, so that requiring configurations to be common among *all* molecules is too strict. The ability to allow some number of the molecules to completely miss a reported "common" configuration is necessary in such cases. Without such freedom, many runs would be required, with the user forced to guess which compounds should be left out in each case. Recall that a *miss* means a molecule could not map to all features of a configuration, and a *complete miss* means a molecule could not even map to any K − 1 features of a *K*-feature configuration.

Three more runs were done, each identical to run B except for noted differences (Table 2). Run C allowed one miss. Run D allowed two misses, one of which was allowed to be a complete miss. Run E allowed three misses, one of which was allowed to be a complete miss. In the run sequence B, C, D, and E, the resulting common configurations of a given run satisfy looser constraints than the configurations of the preceding run, in that more molecules are allowed to miss, or miss completely. As expected, the number of identified configurations increases as the sequence is traversed.

One might be tempted to argue that, of course, loosening the constraints should increase the number of configurations found, but what would be the point? It is important to realize that, in fact, it is a rather common occurrence for a few molecules of an active set to be "missing" some chemical features related to the activity. Thus, not allowing this flexibility in a search for common configurations is likely to *overconstrain* the problem. A chemist might miss an important common configuration, or an identified configuration might be far more common than it appears under a "no misses allowed" search.

2. HIV-1 RT Inhibitors. We considered six nonnucleoside inhibitors for HIV reverse transcriptase, and sought a common configuration of features which they support. Conformers were prepared using the poling method; the number of conformers generated for each compound is shown in Table 3. All conformers of all compounds were considered as reference molecules. One miss, but no complete misses, was allowed. The feature types considered were **H**ydrophobic, hydrogen-bond **D**onor, and hydrogen-bond **A**cceptor. The standard definition of hydrogen-bond acceptor was modified to exclude the sulfur in BI-362528-26 because its electrons are delocalized. A tolerance of 1.4 Å on ligand positions and 1.9 Å on projected hydrogen bonding site positions was specified.

High-scoring configurations were examined visually to evaluate the degree of overlap that occurred when the molecules were aligned to the configuration. A configuration with four hydrophobes and an acceptor appeared to have the greatest overlap. It had the highest score of any configuration in its partition, with the 17th highest score overall. All compounds matched all five features of this configuration with the exception of E-BPU, which fit four.

A few alternate mappings of each compound to the configuration were examined using Catalyst[28] to select those having the largest overlap and structural similarity in the context of the other compounds' alignments. The selected

Table 2. Further PAF Runs[a]

			number of configurations of Type					
run	misses	complete misses	AAH	AHH	AHHH	AAHH	AAAH	overall top score
B	0	0	969	1580	0	0	0	48.7
C	1	0	5975	10362	9	4	0	49.4
D	2	1	11239	20188	533	84	0	52.0
E	3	1	16929	22764	2837	657	27	52.0

[a] Key: A = hydrogen-bond Acceptor, H = Hydrophobic region.

Table 3. HIV-1 RT Compounds

name	conformers
L-697661	90
BI-362528-26	13
Nevirapine	4
Ateverdine	92
R82913s	14
E-BPU	31

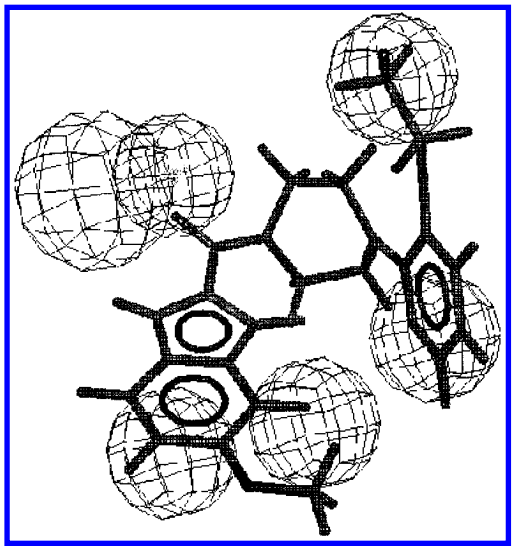

Figure 1. Ateverdine aligned to the common configuration. The two overlapping mesh spheres are the hydrogen bond acceptor (with projected point). All other mesh spheres represent hydrophobic regions.

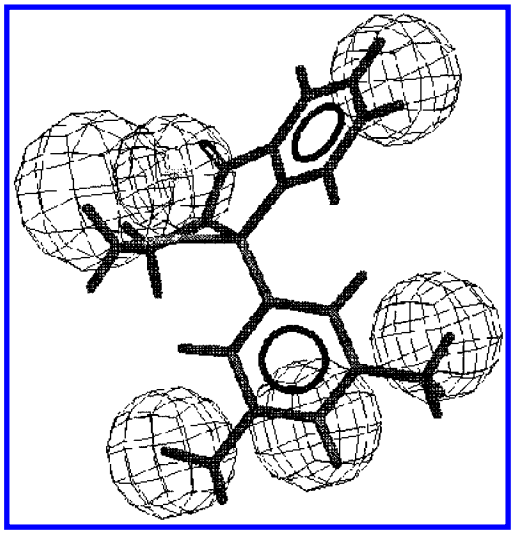

Figure 2. BI-362528-26 aligned.

mappings all fit the configuration within the specified tolerance and were either the first- or second-best fitting mapping. The selected mappings are pictured in Figures 1−6.

The conformer from the nevirapine conformational model that best fit the common configuration was found to superimpose well with a proposed model for the docking of nevirapine into HIV-1 RT.[19] To check the appropriateness of the conformations selected by our algorithm for each of the inhibitors, we measured the enthalpy of binding of each to the Pattabiraman model (1rvo.pdb) using Quanta as follows.

The Pattabiraman model includes an orientation of nevirapine that is a convenient reference point for overlaying

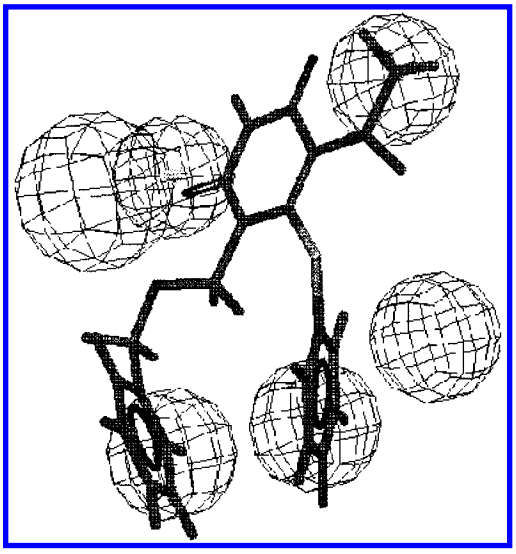

Figure 3. E-BPU aligned.

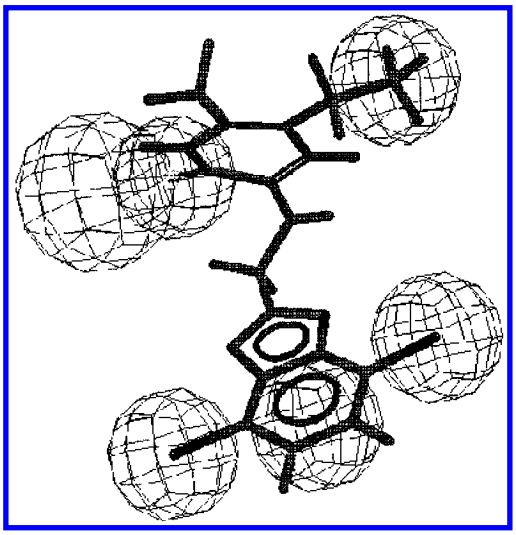

Figure 4. L697661 aligned.

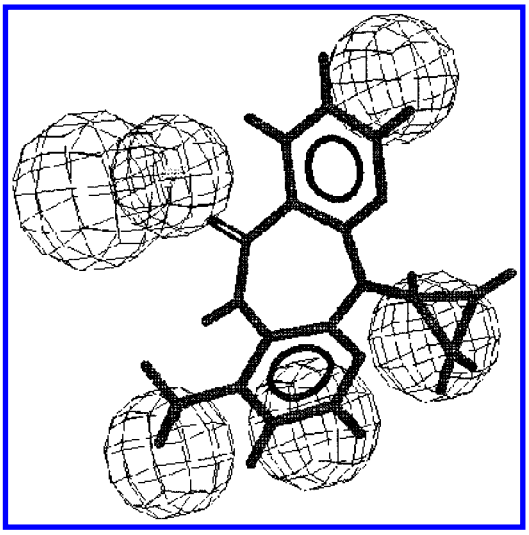

Figure 5. Nevirapine aligned.

the selected conformations. The model of the enzyme in 1rvo.pdb consists of five separate peptide segments. The N and C termini of each segment were converted to amides by capping with acetyl or NH2, respectively, in order to remove inappropriate charged groups. The generated alignments were imported, and the modified enzyme model alignment

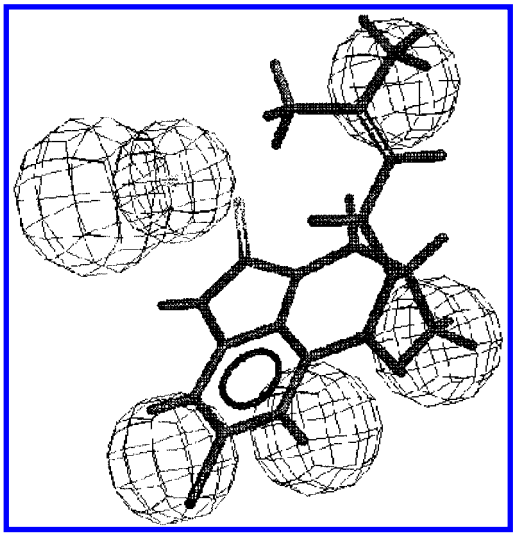

Figure 6. R82931 aligned.

was adjusted manually so that the two nevirapine molecules were superimposed, thus establishing the relative alignment of the model enzyme with the rest of the molecules. Each molecule was then minimized within the pseudo-active site model using 100 steps of CHARMm-based steepest descent minimization followed by the conjugate gradient method until convergence was achieved (typically within 700 steps). During each minimization procedure, the enzyme backbone atoms were fixed, and the enzyme side chains and the ligand were allowed to move. Each measurement began from the same enzyme starting point. After convergence, the energies of the enzyme alone (E), the ligand alone (L), and the enzyme and ligand together (T) were noted. Interaction energy was calculated as (E + L) − T. The results are shown in Table 4.

The results indicate that all of the compounds except L697661 can fit in the enzyme model and have a negative interaction enthalpy. In addition to its positive interaction enthalpy, L697661 appears not to fit when inspected visually in its suggested orientation relative to the enzyme. This suggests that we have not identified the correct orientation and/or conformer for this one compound.

During the minimization step, the conformations of each ligand do change. However, these changes are small as shown in Table 4, row 5 which shows RMS deviations from the beginning selected conformation to the final minimized conformation.

Recently, a paper describing a series of benzophenone HIV-1 RT inhibitors has appeared from Glaxo.[29] In this paper, a figure (Figure 3, p 1662) is presented showing a proposed alignment of a benzophenone and nevirapine. As a further test of our generated hypothesis, we constructed a model of this molecule and calculated the conformation that best fit our common configuration. An overlay of this conformation on nevirapine is shown in Figure 7 and is in excellent agreement with the Glaxo structure.

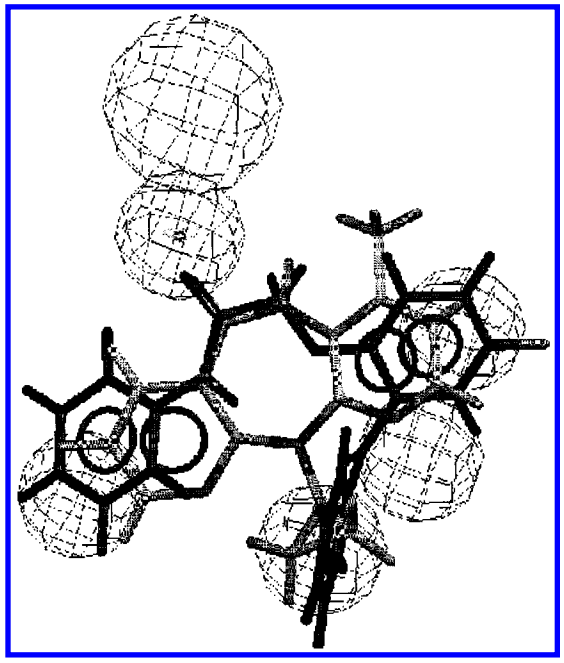

Figure 7. The relative alignments of Nevirapine (grey) and the new Glaxo inhibitor (black) to the configuration.

Just prior to the submission of this paper, a potent new HIV-1 RT inhibitor was reported,[30] referred to as "compound 19". Given the topology of this compound, we generated a conformational model and performed two experiments. First, we fit this compound to the feature-configuration described in section 2 of our Results. Two possible binding modes were observed, both fitting four out of five features. We then added compound 19 to the input set for the algorithm. The highest scoring common feature configuration was slightly perturbed from the configuration of section 2, but the change was enough to allow compound 19 to bind to all features, in both binding modes.

3. HIV-1 PR Inhibitors. To demonstrate the algorithm's ability to handle large conformational models, six HIV-1 PR (protease) inhibitors were chosen. Poled conformers were generated, with average model size of 274 conformers per molecule (Table 5). In each case, the known crystal structure was added to the conformational model to provide an isolated test of the common configuration algorithm independent of the quality of the conformational model. The goal is for the algorithm to identify a configuration common to the crystal conformers in the absence of any special clue as to

Table 4. Generated Alignment Interaction Energies[a] (ΔH)

compound	PatNev[b]	Nev[c]	L[d]	R[e]	Ate[f]	E-BPU[g]	BI[h]
Ligand (L)	70.28	85.58	127.32	46.27	122.62	−47.64	68.80
Enzyme (E)	427.16	427.86	448.49	424.67	445.85	377.27	407.89
Together (T)	431.53	429.25	580.87	435.11	521.92	291.83	472.07
Interaction ΔH	−65.91	−84.19	5.06	−35.83	−46.55	−37.8	−4.62
deviation[i]	0.046	0.52	0.94	0.92	1.2	1.8	1.1

[a] Energy values are in Kcal/mol. [b] Nevirapine [Merluzzi] from the Pattabiraman model (1rvo.pdb). [c] Nevirapine from the generated alignments. [d] L697661 [Hargrave] from the generated alignments. [e] R82931 [Breslin] from the generated alignments. [f] Ateverdine [Romero] from the generated alignments. [g] E-BPU [Miyasaka] from the generated alignments. [h] BI-362528-26 [Saari] from the generated alignments. [i] Overall RMS deviation (in Å) of the minimized ligand conformation from the original, selected conformation.

Table 5. HIV-1 PR Compounds[a]

name	PDB entry	conformers
SKBVa	1aaq	385
SKF108738	1hef	237
U75875	1hiv	254
A74704	9hvp	233
SB203386	1sbg	341
JG365	7hvp	192

[a] Key: A = hydrogen-bond Acceptor, D = Donor, H = Hydrophobic region.

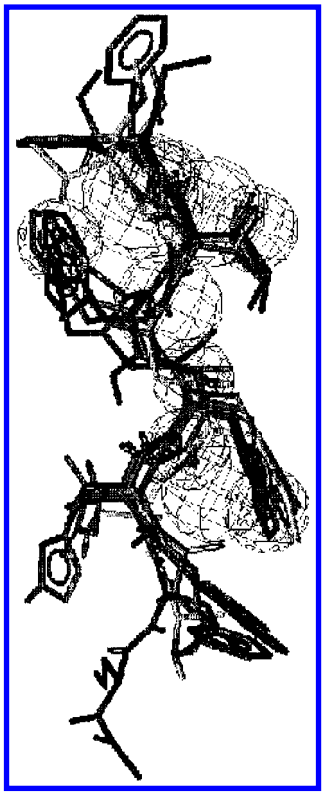

Figure 8. All six protease inhibitors, superimposed, aligned to the highest scoring configuration. (Hydrogens not shown.)

which conformer in the model originated from the crystal structure.

To reduce the computational effort involved in handling these large, feature-rich molecules, we chose the smallest molecule, SB203386, as the sole reference. Note that all 341 conformers of this molecule, not just a single conformer, were treated as references. In addition, only ideal hydrogen bond geometry was allowed. (If a good hypothesis had not been found, additional runs could have been performed, allowing for additional reference molecules or nonideal hydrogen bond geometry.)

Inspection of the compounds revealed the presence of hydrophobic regions located near the ends of the "arms" of each compound. Since there were, in general, many more donors and acceptors than hydrophobes in each compound, we specified that the common configuration should contain at least three hydrophobes. This directed the algorithm first to ensure that at least three hydrophobic regions were in common before seeking additional features. Had no suitable common configuration been discovered under this requirement, our specified constraint could have been relaxed. No misses were allowed. The run completed in roughly 12 h on a 150 MHz MIPS 4400 processor.

Table 6. Highest Scoring Configurations[a]

features	score
AADDHHH	103.013
AADDHHH	98.700
AADDHHH	96.051
ADDDHHH	94.824
ADDHHH	92.658
AADHHH	89.703
ADDHHH	89.635
ADDHHH	89.545
ADDHHH	88.973
DDDHHH	88.812

[a] Key: A = hydrogen-bond Acceptor, D = Donor, H = Hydrophobic region.

The top scoring hypothesis consisted of two acceptors, two donors, and three hydrophobic regions. Each of the six molecules was aligned to the hypothesis based on minimum RMS distance, and the resulting alignment is pictured in Figure 8. In every case the crystal structure conformer fit the hypothesis better than any other conformer in the model. We list the type and score of each of the top ten hypotheses found in this run in Table 6.

SUMMARY

We have presented a new algorithm for determining three-dimensional configurations of chemical features which are common to a set of molecules, along with a novel method for scoring the resulting configurations, based on both the degree to which the configuration is common to all molecules and the estimated rarity of the configuration itself. The algorithm is more precise than existing clique-detection methods which rely only on interfeature distances; although the verification of superposition is computationally expensive, some of this expense is repaid, as the increased precision allows for more aggressive pruning as the search progresses. Furthermore, a configuration which, although satisfying interfeature distance constraints, *cannot* be superposed with the input molecules is of little or no value. The algorithm also accepts relaxed constraints, whereby some of the input set is allowed to fail to match part or all of a reported configuration, obviating the need for multiple runs, in which the operator must prepare many input sets, hoping to discover which compounds are preventing the discovery of common chemical functionality. Finally, the performance of the algorithm under conditions requiring large conformational models is better than clique-detection methods, which have a quadratic time dependence on the number of conformers, assuming (and this is the typical case) that there is no *one* prespecified reference *conformer*.

We have applied the algorithm and scoring method to three data sets, including HIV protease inhibitors and HIV reverse transcriptase inhibitors, obtaining good results, particularly in the latter case, wherein the common configuration discovered by our algorithm matches two newly reported inhibitors. This indicates the power of our approach when used as a tool for finding queries likely to produce new lead compounds.

ACKNOWLEDGMENT

The authors would like to thank Dave Huhta for his assistance with energy calculations and Mark Bures for

providing additional information on the PAF data set. Steven Teig made many useful suggestions. Matt Hahn and Dave Rogers provided valuable feedback on the manuscript.

REFERENCES AND NOTES

(1) Cramer, R.; Patterson, D.; Bunce, R. Comparative molecular field analysis (Comfa). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(2) Norinder, U. The alignment problem in 3d-QSAR: a combined approach using catalyst and 3D QSAR technique. European Symposium on Quantitative Structure−Activity Relationships; Barcelona, Spain, 1994.

(3) Hahn, M.; Rogers, D. Receptor Surface Models. 2. Application to Quantitative Structure−Activity Relationship Studies. *J. Med. Chem.* **1995**, *38*, 2091−2102.

(4) Jain, A.; Dietterich, T.; Lathrop, R.; Chapman, D.; Critchlow, R.; Bauer, B.; Webster, T.; Lozano-Perez, T. Compass: a shape based machine learning tool for drug design. *J. Comp.-aided Mol. Design* **1994**, *8*, 635−652.

(5) Catalyst/Hypo Tutorial, version 2.0; BioCAD Corp.: Moutain View, CA, 1993.

(6) Mayer, D.; Naylor, C. B.; Motoc, I.; Marshall, G. R. A unique geometry of the active site of angiotensin-converting-enzyme consistent with structure-activity studies. *J. Comp.-aided Mol. Design* **1987**, *1*, 3−16.

(7) Dammkoehler, R.; Karasek, S.; Shands, E.; Marshall, G. Constrained Search of Conformational Hyperspace. *J. Comp.-aided Mol. Design* **1989**, *3*, 3−21.

(8) Sheridan, R. P.; Nilakantan, R.; Discon, J. S.; Venkataraghavan, R. The ensemble approach to distance geometry: Application to the nicotinic pharmacophore. *J. Med. Chem.* **1986**, *29*, 899−906.

(9) Brint, A.; Willet, P. Algorithms for the identification of three-dimensional maximal common substructures. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 152−158.

(10) Smellie, A.; Crippen, G.; Richards, W. G. Fast Drug Receptor Mapping by Site-Directed Distances: A novel method of predicting new pharmacological leads. *J. Chem. Inf. Comput. Sci.* **1991**, *31*(*3*), 386−394.

(11) Martin, Y.; Bures, M.; Danaher, E.; DeLazzer, J. New strategies that improve the efficiency of the 3D design of bioactive molecules. Trends in QSAR and Molecular Modelling; Wermuth, C., Ed.; ESCOM: Leiden, 1993; pp 20−26.

(12) Golender, V.; Vorpagel, E. Computer-Assisted Pharmacophore Identification. 3D QSAR in Drug Design; Kubinyi, H., Ed.; ESCOM: Leiden, 1993; pp 137−149.

(13) Smellie, A.; Teig, S. L.; Towbin, P. Poling: Promoting Conformational Variation. *J. Comput. Chem.* **1994**, *16*, 171−187.

(14) Barakat, M. T.; Dean, P. M. Molecular structure matching by simulated annealing. III. The incorporation of null correspondences into the matching problem. *J. Comp.-aided Mol. Design* **1991**, *5*, 107−117.

(15) Greene, J.; Kahn, S.; Savoj, H.; Sprague, P.; Teig, S. Chemical Function Queries For 3D Database Search. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1297−1308.

(16) Klopman, G. Artificial Intelligence Approach to Structure−Activity Studies. Computer Automated Structure Evaluation of Biological Activity of Organic Molecules. *J. Am. Chem. Soc.* **1984**, *106*, 7315−7321.

(17) Golender, V.; Rozenblit, A. Logical and Combinatorial Algorithms for Drug Design; Research Studies Press: Wiley, Letchworth, 1983.

(18) Bures, M.; Danaher, E.; DeLazzer, J.; Martin, Y. New Molecular modeling tools using three-dimensional chemical structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 218−223. See, also: *J. Med. Chem.* **1994**, *37*, 2011.

(19) Entry 1RVO in the Protein Data Bank, described in the following references: (a) Gussio, R.; Pattabiraman, N.; Zaharevitz, D.; Kellogg, G. E.; Rice, B.; Schaeffer, C. A.; Burt, S. K.; Erickson, J. W. Using Nevirapine Analogs as Structural Probes to Model Their Binding Site on HIV-1 Reverse Transcriptase. To be published. (b) Pattabiraman, N.; Gussio, R.; Topol, I.; Burt, S. K.; Erickson, J. W. An Electronic Characterization of a Congeneric Series of Nevirapine Analogs. *Abstr. Pap. Am. Chem. Soc.* **1994**, 208, Abstract number: MEDI 33. (c) Gussio, R.; Pattabiraman, N.; Kellogg, G. E.; Bhat, T. N.; Collins, J.; Burt, S. K.; Erickson, J. W. HIV-1 Reverse Transcriptase Nevirapine Binding Site Model: A 3D Structural QSAR for Ligand Design. *Abstr. Pap. Am. Chem. Soc.* **1994**, 208, Abstract number: MEDI 34. (d) Smerdon, S. J.; Jager, J.; Wang, J.; Kohlstaedt, L. A.; Chirino, A. J.; Friedman, J. M.; Rice, P. A.; Steitz, T. A. Structure of the Binding Site for Nonnucleoside Inhibitors of the Reverse Transcriptase of Human Immunodeficiency Virus Type 1. *Proc. Natl. Acad. Sci.* **1994**, *91*, 3911−3915.

(20) HipHop Tutorial, Version 2.3; Molecular Simulations Inc.: Sunnyvale, CA, 1995.

(21) Crippen, G. M.; Havel, T. F. Distance Geometry and Molecular Conformations; Research Studies Press: Wiley Pub: 1988.

(22) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaninathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4*, 2, 187−217.

(23) Fletcher, R. Practical Methods of Optimization 1: Unconstrained Optimization; Wiley Pub.: 1980.

(24) Ferro, D.; Hermans, J. A different best rigid-body molecular fit routine. *Acta Crystallogr.* **1977**, *A33*, 345−347.

(25) Yao, F. F. Computational Geometry. *Handbook of Theoretical Computer Science. Volume A. Algorithms and Complexity*; van Leeuwen, J., Ed.; The MIT Press/Elsevier: 1990; pp 368−374.

(26) Barron, A.; Cover, T. Minimum Complexity Density Estimation. *IEEE Trans. Inform. Theory* **1991**, *37*(*4*), 1034−1054.

(27) Harris, T. A Lower Bound for the Critical Probability in a Certain Percolation Process. *Proc. Cambridge Phil. Soc.* **1960**, *56*, 13−20.

(28) Catalyst Tutorial, version 2.3; Molecular Simulations Inc.: Sunnyvale, CA, 1995.

(29) Wyatt, P. G. et al. Benzophenone Derivatives: A Novel Series of Potent and Selective Inhibitors of HIV Type 1 Reverse Transcriptase. *J. Med. Chem.* **1995**, *38*, 1657−1665.

(30) Hanasaki, Y. et al. Thiadiazole Derivatives: Highly Potent and Specific HIV-1 Reverse Transcriptase Inhibitors. *J. Med. Chem.* **1995**, *38*, 2038−2040.

(31) Merluzzi, V. J.; Hargrave, K. D.; Labadia, M.; Gronzinger, K.; Skoog, M.; Wu, J. C.; Shih, C.-K.; Eckner, R. J.; Hattox, S.; Adams, J.; Rosenthal, A. S.; Faanes, R.; Eckner, R. J.; Koup, R. A.; Sullivan, J. L. Inhibition of HIV-1 Replication by a Nonnucleoside Reverse Transcriptase Inhibitor. *Science* **1990**, *2150*, 1411−1413.

(32) Hargrave, K. D.; Proudfoot, J. R.; Grozinger, K. G.; Cullen, E.; Kapadia, S. R.; Patel, U. R.; Fuchs, V. U.; Mauldin, S. C.; Vitous, J.; Behnke, M. L.; Klunder, J. M.; Kollol, P.; Skiles, J. W.; McNeil, D. W.; Rose, J. M.; Chow, G. C.; Skoog, M. T.; Wu, J. C.; Schmidt, G.; Engel, W. W.; Eberlein, W. G.; Saboe, T. D.; Campbell, S. J.; Rosenthal, A. S.; and Adams, J. Novel Non-Nucleoside Inhibitors of HIV-1 Reverse Transcriptase. 1. Tricyclic Pyridobenzo- and Dipyridodiazepinones. *J. Med. Chem.* **1991**, *34*, 2231−2241.

(33) Breslin, H. J.; Kukla, M. J.; Ludovici, D. W.; Mohrbacher, R.; Ho, W.; Miranda, M.; Rodgers, J. D.; Hitchens, T. K.; Leo, G.; Gauthier, D. A.; Ho, C. Y.; Scott, M. K.; De Clercq, E.; Pauwels, R.; Andries, K.; Janssen, M. A. C.; Janssen, P. A. J. Synthesis and Anti-HIV-1 Activity of 4,5,6,7-Tetrahydro-5-methylimidazo-[4,5,1-*jk*][1,4]benzodiazepin-2(1*H*)-one (TIBO) Derivatives. 3. *J. Med. Chem.* **1995**, *38*, 771−793.

(34) Romero, D. L.; Morge, R. A.; Biles, C.; Berrios-Pena, N.; May, P. D.; Palmer, J. R.; Johnson, P. D.; Smith, H. W.; Busso, M.; Tan, C.-K.; Voorman, R. L.; Reusser, F.; Althouse, I. W.; Downey, K. M.; So, A. G.; Resnick, L.; Tarpley, W. G.; Aristoff, P. A. Discovery, Synthesis, and Bioactivity of Bis(hetroaryl)piperazines. 1. A Novel Class of Non-Nucleoside HIV-1 Reverse Transcriptase Inhibitors. *J. Med. Chem.* **1994**, *37*, 999−1014.

(35) Miyasaka, T.; Hiromichi, T.; Baba, M.; Hayakawa, H.; Walker, R. T.; Balzarini, J.; De Clercq, E. A Novel Lead for Specific Anti-HIV-1 Agents: 1-[(2-Hydroxyethoxy)methyl]-6-(phenylthio)thymidine. *J. Med. Chem.* **1989**, *32*, 2507−2509.

(36) Saari, W. S.; Hoffman, J. M.; Wai, J. S.; Fisher, T. E.; Rooney, C. S.; Smith, A. M.; Thomas, C. M.; Goldman, M. E.; O'Brien, J. A.; Nunberg, J. H.; Quintero, J. C.; Schleif, W. A.; Emini, E. A.; Stern, A. M.; Anderson, P. S. A New Class of Nonnucleoside, HIV-1-Specific Reverse Transcriptase Inhibitors. *J. Med. Chem.* **1991**, *34*, 2922−2925.

CI950273R