Comparison of Knowledge-Based and Distance Geometry Approaches for Generation of Molecular Conformations

Bradley P. Feuston,*,† Michael D. Miller,†,§ J. Christopher Culberson,† Robert B. Nachbar,‡ and Simon K. Kearsley‡

Molecular Systems Department, Merck Research Laboratories, P.O. Box 4, West Point, Pennsylvania 19486, and P.O. Box 2000, Rahway, New Jersey 07065

Received October 31, 2000

A knowledge-based approach for generating conformations of molecules has been developed. The method described here provides a good sampling of the molecule's conformational space by restricting the generated conformations to those consistent with the reference database. The present approach, internally named *et* for *e*numerate *t*orsions, differs from previous database-mining approaches by employing a library of much larger substructures while treating open chains, rings, and combinations of chains and rings in the same manner. In addition to knowledge in the form of observed torsion angles, some knowledge from the medicinal chemist is captured in the form of which substructures are identified. The knowledge-based approach is compared to Blaney et al.'s distance geometry (DG) algorithm for sampling the conformational space of molecules. The structures of 113 protein-bound molecules, determined by X-ray crystallography, were used to compare the methods. The present knowledge-based approach (i) generates conformations closer to the experimentally determined conformation, (ii) generates them sooner, and (iii) is significantly faster than the DG method.

INTRODUCTION

In the drug discovery process, pharmaceutical databases containing the structures of hundreds of thousands of compounds are routinely searched for both leads and lead optimization. These searches usually involve evaluating database molecules by their similarity to peptide leads, prototype molecules, and/or pharmacophore models. A significant amount of research has been devoted to enhancing the quality of database hits as well increasing the speed of the search, since many of these proprietary databases contain more than 500 000 molecular entities. The number and type of searches performed are far too numerous to discuss or list. Typical searches use two-dimensional (2D) similarity measures. However, some of the latest search algorithms require three-dimensional (3D) structural information for enhancing the quality of database search results.[1−4]

Since 3D structures are playing an increasing role in many computational techniques, a uniform sampling of the energetically accessible conformational space for each molecule of interest is desired. The conformational space sampled by a molecule is largely determined by ring conformations, torsions about rotatable bonds, and chirality. Bond lengths and valence angles, being fairly constant across all conformations, have a lesser effect on differences between conformations of the same molecule. A number of studies comparing various conformation-generating techniques have been previously performed.[5−8] Techniques that rely upon systematic searches of torsion angles typically generate far too many conformations, requiring lengthy search procedures, an additional selection process, or both. Various improvements in the systematic search, including stochastic algorithms and mode following, have successfully reduced both the CPU time and the number of conformations generated.[9−11] However, these previous approaches suffer from the need to perform a final energy minimization procedure for each conformation. Such approaches are inefficient and unsuitable for generating conformations of large molecular databases. Several knowledge-based approaches have been developed to produce physically reasonable and low-energy conformations, in particular MIMUMBA and WIZARD.[12−14] These earlier efforts focused upon identifying substructure fragments and generating conformations by replacing substructures with a number of low-energy fragments selected from a database, the so-called knowledge-based approach. While the smaller number of conformations generated permitted more efficient searches, minimizing the energy of each conformation formed from the combined substructures increased the computational costs significantly.

Klebe et al. were the first to use a knowledge-based approach relying upon torsion angles to generate "biologically relevant conformations."[12] Unlike the expert system embodied in WIZARD, their MIMUMBA algorithm involves separating a molecule into ring and open chain fragments.[14] Conformations of the open chain fragments are obtained through a library of known torsion angles. In this approach neighboring torsion angles are treated as uncorrelated until the final energy minimization procedure. Rings are flexed using the Quantum Chemistry Program Exchange (QCPE) program SCA.[15] Both sets of fragments, i.e., open chain fragments and rings, are brought together and simultaneously

* Corresponding author phone: (215) 652-5048; e-mail: bradley_feuston@merck.com.
† P.O. Box 4, West Point, PA 19486.
‡ P.O. Box 2000, Rahway, NJ 07065.
§ Present address: Pfizer Inc., P.O. Box 8003, Eastern Point Road, Groton, CT 06340.

10.1021/ci000464g CCC: $20.00
Published on Web 04/12/2001

energy minimized using an empirical force field. Though their method yields good results, it requires too much time to be useful in generating a large database of molecular conformations.

The need to generate conformations that uniformly sample the energetically favorable conformational space in a timely fashion provided the motivation for the development of the present method. In the context of the following discussion, a low-energy conformation is one whose torsion angles (rings and chains) are sampled from the distribution of experimentally observed values, while the accessible conformational space is restricted to the Cartesian product of experimentally observed torsion angles and ring conformations.

This paper covers two main topics: (i) the definition and mining of rules and (ii) the conformation generation methodology. Our conformation generator may be used with any of a number of different rule sets, e.g., rules for small molecules, rules for polypeptides, or compound class oriented rules. Since the initial implementation of this approach focused on druglike molecules, the discussion here will be limited to the small molecule rule set.

The definition and development of torsion angle rules will be presented first, followed by a discussion of mining the Cambridge Structural Database (CSD) for the values of the torsion angles and ring conformations.[16] For correlated torsion angles the final conformations generated are sensitive to the algorithm employed. Therefore, a detailed discussion of our conformation generating methodology follows. Finally, 113 molecules, whose protein-bound conformations have been previously determined by X-ray crystallography, are used to compare our knowledge-based method to Blaney et al.'s published distance geometry (DG) algorithm.[17]

DEFINING THE RULES

The present method does not project two-dimensional structures into three dimensions. There are several good algorithms that accomplish this task sufficiently well.[18,19] Since the purpose here is to generate a number of diverse, low-energy conformations, the initial conformation must be a reasonably good 3D structure, with appropriate bond lengths and bond angles and the absence of any steric clashes.

In contrast to MIMUMBA, the present method employs a library of much larger substructures and treats open chains and rings simultaneously. In addition, substructures containing both a ring and open chain fragment are easily handled. While the larger substructures characterize correlation of topologically adjacent torsion angles, the methodology exploits the overlap of substructures to extend this correlation to larger substructures. This both reduces the number of conformations to be considered and enhances the likelihood that the conformations are low energy, i.e., have torsion angles most frequently observed.

Pattern Recognition. The generation of energetically reasonable conformations of a flexible molecule relies upon substructure identification and a database of the observed substructure conformations. Using Merck's proprietary atom typer, a specific molecular substructure may be expressed as an unambiguous pattern and easily identified in a molecule.[20] For example, the moiety in Figure 1 which contains an amide bond may be expressed through the following pattern:

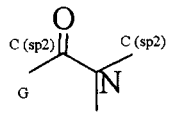

Figure 1. An amide bond moiety, pattern 1 in text.

$$G - C \,\&\, sp2\,(=O) - N(\&X3) - C \,\&\, sp2 \qquad (1)$$

where

G	any non-hydrogen atom
N	nitrogen
C	carbon
O	oxygen
H	hydrogen
—	single bond
=	double bond
&	AND
(•)	branching operator
sp2	sp^2 hybridization property
X3	bonded to three non-hydrogen atoms

The pattern reads as follows: any non-hydrogen atom singly bonded to an sp^2 carbon which is both doubly bonded to an oxygen and singly bonded to a nitrogen which is bonded to three non-hydrogen atoms of which one is an additional sp^2 carbon. The programmable atom typer identifies all such substructures in a given molecule. The pattern and its associated data of observed torsion angles constitute a rule. The rules that are currently in use have evolved over a number of years in Merck Research Laboratories worldwide based upon feedback from users and their application to a large number of medicinal chemistry projects. Therefore, the knowledge embedded in this approach is not restricted to just the observed torsion angles but also includes which patterns are most important. An attempt has been made to capture some of the experience of the medicinal chemist through the specific patterns employed.

Initially, the elementary atom−objects for a pattern were limited to sp^2 and sp^3 C, N, and O. The generic atom property G was used as a match for any non-hydrogen atom. All possible patterns were enumerated for chains containing one to three torsions and rings containing up to eight atoms. Data for each pattern was then mined from the CSD (see below). The torsion angle data for each pattern fell into one of four general categories. The observed torsion angle(s) indicated that the particular pattern was (I) well represented and dominated by a small number of values, (II) well represented but uniformly distributed over a large number of values, (III) poorly represented, or (IV) not represented at all. A pattern is considered well represented if there are many examples in a variety of molecular classes. Category I patterns were deemed to form good rules and were retained while category IV rules could not be used and thus eliminated. The overall objective in the evolution of the rule set was to have exclusively rules from category I. Prototypical category I patterns involve an sp^2 component which severely restricts the torsion angles available, e.g., a pattern containing an amide bond as above, and therefore has a limited number of observed torsion angle values.

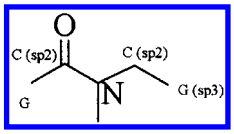

Figure 2. An extension of the moiety in Figure 1 corresponding to pattern 2 in text.

In addition to desiring rules from category I, it is also preferable to have a larger pattern over a smaller pattern. Consider a six-atom pattern covering three contiguous torsion angles versus three four-atom patterns for the same three sequential torsion angles. A well-represented, six-atom pattern dominated by a small number of values will permit a faster and more efficient sampling of the low-energy conformational space. Since the number of possible conformations of the three independent torsion angles is proportional to the product of the number of values for each rule, the number of possible combinations is typically far greater than the six-atom pattern with many to be eliminated by steric clashes, e.g., gauche$^+$ gauche$^-$. Considerable efficiency in conformational sampling is achieved by exploiting the correlation between neighboring torsion angles through the use of the five- and six-atom patterns.

In the example pattern given above, the expected torsion angles about the N−C bond should be split between cis and trans. The database for this simple rule actually contains 12 entries (−46°, −30°, −15°, 1°, 15°, 30°, 45°, 149°, 165°, 179°, 194°, 209°) with the majority of values falling near 0° and 180°. Deviations from the expected values of 0° and 180° are due to steric clashes caused by the substituents on the trisubstituted nitrogen. Though these 12 values are crucial for constructing energetically favorable conformations, they are not particularly interesting. What is interesting, however, is how the amide moiety affects nearby torsion angles. Consider the following pattern, depicted in Figure 2,

$$\text{G} - \text{C \& sp2 (=O)} - \text{N(\&X3)} - \text{C \& sp2} - \text{G \& sp3} \quad (2)$$

where sp3 represents the sp^3 hybridization property.

Since the torsion angle about the bond between the amide N and C(sp^2) has already been found to have 12 potential minima, the total number of conformations for this substructure would expected to be about 144 (12 × 12). However, mining the CSD indicates that only five conformations for this five-atom structure account for over 99% of the observed conformations. The observed torsion angles are (181°, 180°), (23°, 172°), (158°, −14°), (179°, −1°), and (8°, 164°), where the first angle in each pair corresponds to the amide bond. Multiple angle torsion rules can be constructed to reflect correlations between adjacent torsion angles, leading to more efficient sampling of the conformational space.

Thus the order in which patterns are identified becomes very important, with the larger patterns taking precedence over the smaller and patterns containing more constrained moieties, e.g., some sp^2 character, taking precedence over the less constrained. To facilitate pattern recognition, patterns are grouped into classes. A class contains a set of four-, five-, and six-atom patterns with the same principal moiety. For example, all carbonyl carbons bonded to a trisubstituted nitrogen belong to the same class. The pattern recognition algorithm is implemented so that each torsion angle is assigned to the largest and most restrictive pattern first. This limits the number of patterns identified for each torsion angle to those with the most constraints and thereby reducing the number of potential conformations. Using this as a guiding principle, the rules were evolved to rules belonging to category I, i.e., well represented with a relatively small number of observed torsion angle values.

Since category II and III rules do not contain useful information for conformation generation, they require further investigation. Having a category II pattern that is well represented but displaying no preferences in torsion angle values indicates that the pattern needs additional specificity. For example, it was found that a pattern containing an sp^3 C with two hydrogens would be associated with a different distribution of torsion angle values than the same pattern where the sp^3 C had zero or one attached hydrogen. Thus, by increasing the degree of specificity of atoms in a category II pattern, a number of daughter patterns could be defined. Increasing the specificity in this way typically resulted in several new category I rules. Large patterns which fall into category III are typically eliminated, which then allows smaller subpatterns to emerge as category I. For this reason category IV rules are eliminated only if larger inclusive rules have been processed first, i.e., identified as a category I or eliminated as a category IV.

N-member rings (N = 4, 5, 6, 7, 8) were specifically identified and treated as $N - 3$ angle patterns. The values of the $N - 3$ angles for a particular pattern, e.g., fully saturated six-membered ring, were obviously correlated to maintain ring closure. There are a number of patterns for each ring size to account for different bonding configurations and the presence of heteroatoms. Unlike open chain patterns, ring patterns could not be reduced to a combination of smaller patterns in order to avoid ring-closure violations. The seven- and eight-member ring patterns were found to be poorly represented in the CSD, i.e., categories III and IV. An alternative computational approach was taken to complete these rules. A database of all physically reasonable geometries was constructed for each ring size. The sets for seven- and eight-membered rings contained all unsaturated isomers with 0−3 and 0−4 endocyclic double bonds, respectively. For each isomer in turn, 140 (seven-membered ring) or 352 (eight-membered ring) conformations were generated by mapping the double bond pattern onto the optimized conformations (MM2* in BatchMin v5.0[21]) of the fully saturated hydrocarbon ring. For the seven-membered rings, there are five characteristic ring shapes (planar (D_{7h}), chair (C_s), twist-chair (C_2), boat (C_s), and twist-boat (C_2)),[22] 14 permutations of the atom labels under the D_7 symmetry of the planar reference geometry, and 2 mirror image forms. For the eight-membered rings, there are 11 characteristic ring shapes (planar (D_{8h}), crown (D_{4d}), chair-chair (C_{2v}), twist-chair-chair (D_2), boat-boat (D_{2d}), twist-boat (S_4), boat (D_{2d}), twist-chair (C_{2h}), chair (C_{2h}), boat-chair (C_s), and twist-boat-chair (C_2)),[23] 16 permutations of the atom labels under the D_8 symmetry of the planar reference geometry, and 2 mirror image forms. The 2520 structures generated for the seven-membered ring database (10 560 for the eight-membered ring database) are by no means unique. For example, all 28 planar conformations of C_7H_{14} are identical, and of the 32 chair conformations of C_8H_{14} there are only three unique ones. These conformations were used to complete the rules for

each seven- and eight-membered ring pattern. It was decided that rings of greater than eight atoms could not be properly handled with this knowledge-based method due to the lack of data and the large number of possible conformations.

Torsion Angle Values in the Rules. A rule is comprised of a unique pattern, i.e., assembly of connected atoms, which defines one to five torsion angles and a corresponding set of data about the torsion angle(s). The data which were extracted from the database not only included the observed values of the torsion angle but also the relative frequency of occurrence and the observed range about each value. The final rule contains the pattern, a number of torsion angle values with their allowed range, and the frequency of each value. Since the impact of a rule ultimately depends on the allowed values, a great deal of care was taken to ensure the appropriateness of the database for small druglike molecules.

For this reason, a large number of molecules were extracted from the Cambridge Structural Database (CSD), which is known to contain high-quality, small molecule, crystal structural data.[16] Systematic analysis of structures in the CSD has been shown to provide valuable information and insight not available by any other means.[24] Only well-characterized (*R*-factor < 0.120) organic compounds were initially selected. The initial 46 026 molecules were further screened by eliminating all members that contained any atoms, which could not be assigned an atom type according to the MMFF force field,[25] resulting in 40 639 molecules. A constraint of this type is required to ensure that the chemical bonding of each atom can be appropriately identified so that patterns may be correctly assigned.

Though structures passing the early screens are energetically reasonable with well-defined atom types, further manipulation and culling was found to be necessary. Some undesirable observed torsion angle could have significant impact on the conformations generated. For example, approximately 5% of the organic molecules extracted from the CSD that contained an amide bond had cis geometry. Applying this rule uniformly would force the amide bonds *for every set of conformations to contain 5% cis amides*! In practice the result is even much worse: since our methodology also employs a diversity algorithm for conformations returned, it correctly identifies cis amides as contributing to diversity and retains a larger percentage of these undesirable conformations. Therefore, molecules identified as having cis amides, as well as unusual bond lengths and bond angles, were eliminated from consideration.

Finally, a diverse set of the remaining molecules was identified using topological similarities.[26] There are numerous examples in the CSD of molecules of the same series being crystallized. Keeping nearly equivalent molecules would incorrectly bias the database torsion angle distribution. The objective was to identify the torsion angle distributions from a large number of different chemical classes. Not all chemical classes are expected to be equally represented in the CSD. Since the conformations actually characterized are sensitive to the experimental conditions required for crystallization, there is the likelihood that some low-energy minima will be underrepresented. A previous study by F. Allen et al. found that torsion angles associated with high strain energy are rarely found in crystal structures and the distributions of substructure conformations are likely to be a good indicators of the gas-phase potential energy surface.[27] The effect of missing or underrepresented conformations also plays a more significant role in the larger rules. This effect is reduced by considering those values of smaller rules whose patterns are a subset of the larger pattern. In actual applications of the present methodology, this effect has not been found to be serious, and in the several occasions where missing values were identified the database was adjusted accordingly.

The final structures were energy minimized using MMFF starting from the crystal structure. Structures that changed significantly were identified and also eliminated. This last procedure was found to maintain the core structure but helped remove unwanted effects on the outermost atoms ostensibly due to crystal packing. Over 18 000 molecules passed the complete analysis and were ultimately used in building the database of torsion angles.

Binning the Torsion Angles. A number of clustering and binning heuristics were tested to extract the multidimensional torsion angle distributions. However, in the final analysis, a simple discretization approach worked best. For patterns with one to three torsion angles, a bin width of 15° starting at $-7.5°$ was employed while for the four- and five-angle patterns (e.g., seven- and eight-membered rings) a bin width of 30° was used starting at $-10°$. The offsets of $-7.5°$ and $-10°$ were chosen so that the bin boundaries would lie near the minimum of the expected torsion angle distributions. The maximum number of possible bins are 24, 576, and 13 824 for one, two and three-angle patterns, respectively. In the final rule set, the largest number of entries was found to be 76 for a three-angle pattern. For each pattern, all matches in the database were extracted and the torsion angles calculated and assigned to the appropriate bin. For each bin the average angle and the standard deviation were calculated. The average value and the range of allowed values were stored in the database for each bin where the range was set at the average value plus or minus one and a half standard deviations. The relative frequency for each bin was also calculated as the number of values in the bin divided by the total number of matches found for the corresponding pattern. The relative frequency of occurrence of each value is treated as the probability for that value in the conformation generation algorithm. Values which were not well represented, i.e., those containing fewer than 10 instances in the entire database or accounting for less than 1% of the data were not catalogued. These criteria were relaxed for a few rules when some preferred values of torsion angles were found to be absent during application. A single bin for two or more angle patterns actually refers to correlated bins of all the angles in the pattern. Also recall the earlier discussion that the mining of the rules and the definition of patterns were intimately linked and optimized simultaneously. An important contribution to the optimization step was the feedback from a number of computational chemists throughout Merck Research Laboratories worldwide. On occasion an undesirable conformation would be identified through application of the present method in a medicinal chemistry project. The cause of the problem, e.g., lack of specificity in a pattern, unacceptable observed torsion angles, was identified and appropriate modifications of the rules were made.

At present, 797 rules have been developed and included in the knowledge base. The patterns range from a single torsion, e.g., four-atom pattern, to an eight-atom pattern with five torsion angles to describe eight-member rings. The rules'

Table 1. Database for Pattern 2 in Text

	angle 1			angle 2		
frequency	max	avg	min	max	avg	min
0.333 333	184	181	178	183	180	177
0.166 667	34	23	12	183	172	160
0.166 667	169	158	147	357	346	335
0.166 667	190	179	168	10	−1	−13
0.166 667	20	8	−3	175	164	153

patterns have been optimized for generating conformations of druglike molecules utilizing the feedback from a large number of chemists working on medicinal chemistry projects. An example of a database entry is shown in Table 1 for pattern 2 above.

CONFORMATION GENERATION ALGORITHM

The second part of this paper focuses on the conformation generation algorithm. Since this algorithm is independent of the set of rules to be applied, the algorithm is presented below for any rule set. As described above, rules are defined as patterns of connected atoms and the associated torsion angles mined from a database of known structures. Having identified a set of patterns for a particular molecule does not imply that a set of conformations can be easily enumerated. In fact, the molecular conformations generated are highly sensitive to the algorithm employed when correlated torsion angles are to be considered. When identifying patterns for a particular molecule, a large number of patterns will be found to be relevant, with many sharing common substructures. The smallest possible substructure contains four atoms that define a single torsion angle. The task is to find the most appropriate set of patterns and a consistent manner of applying them. A general outline of the algorithm employed is depicted in Figure 3. The algorithm has been given the name *et*, for *e*numerate *t*orsions.

The first step is to identify all potential patterns in the candidate molecule and begin to assign specific rules to each torsion angle. The rule corresponding to the largest and most rigid pattern, and therefore the most restricted substructure, becomes the first rule in the list of rules to be applied. These are typically unsaturated rings or, when present, moieties containing some sp^2 character. Using this substructure as the core or central fragment of the molecule, rules are assigned to each shell of torsion angles moving out from center. Larger rules and those with the most overlap with previously assigned rules take precedence. When a rule is identified that partially overlaps a previously assigned rule, i.e., the patterns contain one or more torsion angles in common, the new rule becomes dependent on the previous rule and only values which are consistent with both rules are applied. Rules are considered consistent when the ranges of allowed values for the same angle overlap. In the case where conflicts arise and no common values are found, the dependency is removed by considering a smaller and less constrained rule for the unassigned torsion. If no overlapping rules are found, then the largest rule containing the targeted torsion is assigned. In the extreme case for chains, one could obtain a set of rules containing only four-atom patterns, i.e., a set of single torsion angles, which are uncorrelated. Though never observed, this particular case is most like the previous knowledge-base approach of Klebe et al.

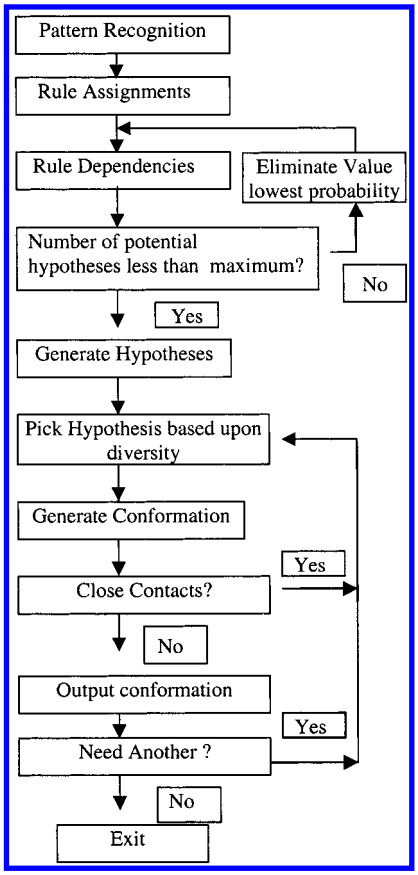

Figure 3. Flow diagram for conformation generation algorithm.

For molecules where two rules of the same size are identified for the same torsion angle, both sets of angles are utilized. This procedure for assigning rules, starting from a central fragment, is repeated for each shell of torsion angles outward from the central core until all torsion angles have been assigned. A unique set of rules and the order in which they are to be applied is ultimately defined for each molecule. The order in which angles are to be rotated creates a dependency relationship between overlapping rules. This order becomes especially important for ring configurations, e.g., single rings, fused rings, and spiro configurations, where closure is required. Cage structures are specifically identified and then held rigid since they have limited flexibility and typically do not add significantly to diversity in conformations.

Additional advantage may also be taken of the overlap between rules. Where small patterns are subsumed by larger rules, the larger rule is usually retained while its values can be restricted to those that are consistent with both rules. By restricting the values of torsion angles in the assigned sets of rules to only those values that are shared by all rules fitting the same substructure, the number of energetically favorable conformations may be further reduced. In practice, this reduction of permitted torsion angles has been found to be too restrictive and is taken advantage of only for very flexible molecules when there are a very large number of potential conformations.

After assigning all the torsion angles in the molecule to the appropriate rules and defining the order of application and the inter-rule dependencies, a few additional issues need to be addressed before generating conformations. In many

cases the number of all possible conformations is still too large even with this knowledge-based approach. To limit the number of possible conformations to be generated, a simple count of the number of potential conformations is performed. The total number of potential conformations may be reduced by judiciously eliminating some torsion angle values from consideration. The torsion angle values that are eliminated depends on the product of the probability of the angle and a weight that has been assigned to each rule. For the purposes of measuring the effect on conformational diversity, weights have been assigned to the central bond of each torsion angle according to its position within the molecule. Torsion angles near the geometrical center of the molecule, where a change in the angle may bring a large change in conformation due to the so-called lever arm effect, are given greater weight than those near the edges. Each rule is then assigned a weight equal to the average weight of each torsion angle in its pattern. This weight is multiplied by the probability of each entry in the rule to arrive at the weighted probabilities. Rings and open chains are treated identically in the determination of the assigned weight. The values with the smallest weighted probabilities are pruned away iteratively until a reasonable number of potential conformations has been achieved. The torsion angle values corresponding to those most frequently observed, and therefore the most likely to occur, survive while values less frequently found are eliminated. Rules on the interior of the molecule keep more of their torsion angle values than those on the periphery. The motivation is to retain the conformations with highest probability of occurring as well as retaining the most diverse.

Though present 3D database searching tools can efficiently search over only a few hundred conformations for each molecule, pruning away torsion angle values to such a low number is not advisable. Such heavy pruning does not guarantee that a sufficiently diverse set of conformations will remain. Since the number of conformations to be generated is a user-input parameter in the present algorithm, much effort has been devoted to returning the best and also the most diverse structures first. The best in the present context are those conformations with the torsion angles corresponding to the most frequently observed, i.e., angles with the greatest probability. Conformational diversity is traditionally evaluated with the root-mean-square deviation (RMSD) between 3D positions of paired atoms in different conformations. However, this requires that the conformations be actually formed first. Since the generation of a large number of conformations is counter to our purposes, an alternative approach has been developed. The objective is to determine the diversity of a set of conformations before they are rendered. The pruning of the torsion angle values is performed until the number of potential conformations is reduced to about 2500, approximately 10 times the number of conformations typically targeted. Though this may be controlled by the user, we have found that 2500 potential conformations is usually sufficiently large for picking a diverse subset. Each potential conformation, which is determined by the order in which specific rules are to be applied and values are to be used, constitute a hypothesis. The hypothesis may be thought of as a recipe for constructing a particular conformation starting from a good 3D structure. The pruning step ensures that the highest probability population has been identified with each member of the population being represented by a hypothesis. The next task is to identify a diverse subset of the required size.

Each hypothesis may be represented by an ordered list with each entry consisting of four atoms in particular order and their torsion angle. The diversity of the final set of conformations may be determined before actual conformation generation by the diversity in hypothesis space. A hypothesis may be treated as an *N*-dimensional vector, where each dimension corresponds to one of the rotatable torsion angles. The components of each hypothesis are the values of those angles for the corresponding conformation. For example, a six-membered ring would be represented by three correlated angles, found sequentially in the hypothesis. Using the hypothesis vector as a conformation descriptor, a dissimilarity algorithm is employed to choose the most diverse hypotheses for generating the final set of conformations.[28] The first hypothesis selected is the one closest to the centroid of all hypotheses. The centroid is just the vector average of all hypotheses. By construction all hypotheses are already in the set of most probable conformations. Subsequent selections are based upon the hypothesis whose distance to the nearest of the previously selected members is a maximum. The selection algorithm is terminated when the targeted number of conformations has been achieved or all the hypotheses have been exhausted. Chirality is not considered for diversity selection but becomes a factor in the final conformation generated.

Conformations are generated from the input structure by simply walking down each component of the selected hypothesis and applying the appropriate rotations. Chiral centers that are not specified by the user nor experimentally determined are randomly assigned a chirality for each conformation.

When changing endocyclic torsion angles, the bond lengths and bond angles should also change slightly to maintain geometric ring closure. To simplify the ring flexing procedure, the bond lengths are kept constant while the bond angles are changed to accommodate changes in the torsion angles to maintain a reasonable ring conformation.

While the rules capture a somewhat extended, or mesoscopic, description of the molecular structures, there is nothing explicit in the procedures described above to prohibit atomic overlap of uncorrelated substructures, e.g., the extreme ends of a long chain. Since the rules prohibit overlap of local topology, only a check of atom pairs whose topological distance is greater than the largest torsion rules needs to be considered. The algorithm only checks the distance between atom pairs that are separated by more than four bonds. A close contact distance of 2.6 Å between any two heavy atoms is permitted. However, if the input molecule has a closer contact distance, that distance will supersede the default value. Note that the input structure is required to be an energetically valid conformation.

EXAMPLES

The crystal structures of 113 molecules bound to receptors were obtained from the Brookhaven Protein Data Bank (PDB) (see Table 2).[29] The molecules chosen were determined to be diverse, sampling each rule class. These structures served as the target conformations for the purposes of testing of the present algorithm and torsion angle database.

Table 2. The 113 PDB Entries Containing Crystal Structures of Bound[29]

1aaq	1abe	1ack	1acm	1aco	1aec	1apt	1ase	1atl	1azm	1baf
1bbp	1blh	1bma	1byb	1cbs	1cbx	1cdg	1cil	1com	1ctr	1dbb
1did	1die	1dr1	1dwd	1eap	1eed	1epb	1eta	1etr	1fen	1fkg
1frp	1ghb	1glp	1glq	1hdc	1hef	1hfc	1hri	1hsl	1hyt	1icn
1ida	1igj	1imb	1IVE	1lah	1lcp	1lic	1lmo	1lna	1lpm	1lst
1mcr	1mdr	1mrk	1mup	1nis	1pha	1phg	1poc	1sts	1rne	1rob
1slt	1snc	1srj	1STP	1tdb	1tka	1tmn	1tng	1tni	1tnl	1tpp
1trk	1tyl	1ukz	1wap	1xid	1xie	2ada	2ak3	2cgr	2cht	2 cmd
2ctc	2dbl	2gbp	2lgs	2mcp	2mth	2PK4	2plv	2r07	2SIM	2yhx
3aah	3cla	3cpa	3GCH	3tpi	4cts	4dfr	4est	4fab	4phv	5p2p
6abp	6rnt	8gch								

Table 3. Average Number of Conformations Generated per Second on an SGI R10000 (Results Averaged over the 113 PDB Test Structures)

no. conformations	et	DG	DG/O
25	19.6	18.8	6.4
50	32.9	19.2	6.4
100	52.3	19.2	6.4
250	85.1	19.9	6.5

Targeting conformations of bound ligands from the Protein Data Bank provides a good test of the present method in typical medicinal chemistry applications. The crystal structures were prepared for this test by first projecting the structures into two dimensions and then projecting each one back into a 3D structure using the heuristic rule-based program CORINA.[18] This had the effect of randomizing the initial configuration of the structures so there would be no sampling bias for the conformation-generating methods. The average non-hydrogen atom root-mean-square deviation (RMSD) per atom between the crystal conformation and the generated conformation for the 113 test molecules was 1.58 Å. In addition to *et*, the knowledge-based methodology presented above, the in-house version of Blaney et al.'s distance geometry (DG) approach was applied for comparison.[17] Since the DG approach does not retain physically reasonable bond lengths and angles, the DG conformations require an additional optimization procedure. An in-house algorithm to optimize the local geometry was employed for this purpose.[30] Though unpublished, this algorithm has been optimized for druglike molecules and considered as efficient as any of the commercially available optimizer, e.g., MSI's CLEAN tool.[31] Below, the DG results are reported with and without optimization, although DG conformations always undergo local optimization before being used in a drug discovery task.

Sets of 25, 50, 100, and 250 conformations were generated for each of the 113 crystal structures starting from the initial 3D structure generated by CORINA. A comparison of execution times of the approaches is given as a function of the number of conformations generated in Table 3. For 25 conformations, *et* generates conformations about 3 times faster than the distance geometry with optimization (DG/O) approach. As expected, the distance geometry approach generates conformations at a fairly constant rate with respect to the number of conformations, whereas *et* becomes increasingly more efficient in generating conformations, and is more than an order of magnitude faster than the DG/O approach at the 250 conformations per molecule level. The initial rule identification and hypothesis generation imposes a fixed overhead on *et*, but once completed allows for the quick enumeration of a large number of conformations.

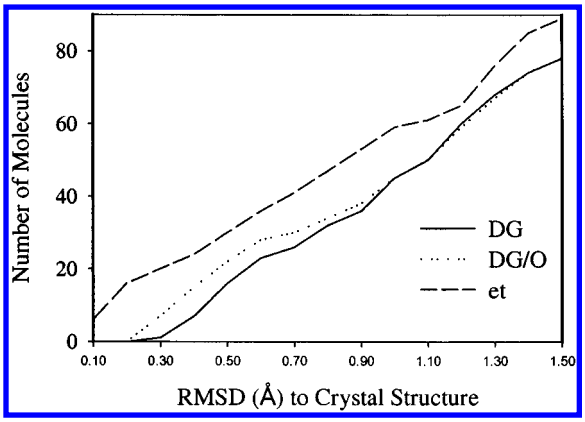

Figure 4. Comparison of present knowledge-based approach (*et*) and published distance geometry algorithm with (DG/O) and without optimization (DG). Number of molecules with a conformation within the given RMSD value of the observed crystal structure. Twenty-five conformations were generated for each molecule.

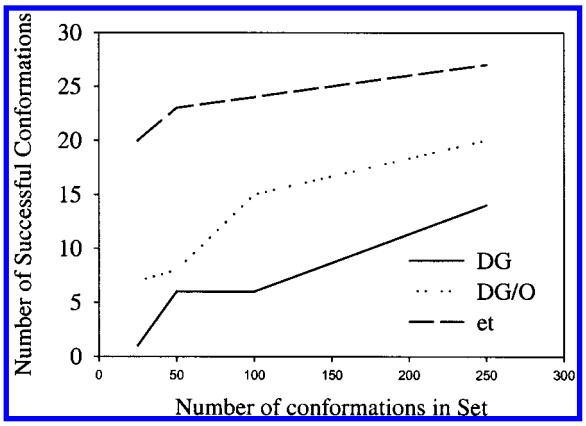

Figure 5. Comparison of the number of molecules which have at least one conformation within 0.3 Å of the crystal structure. There are 113 molecules in the test set.

To evaluate the quality of the conformations generated, the conformation closest to the crystal structure was identified for each molecule in each set of conformations. The number of structures that fall within a range of RMSD values of the corresponding crystal structure for the 25 conformation set is depicted in Figure 4. While *et* generates conformations for 6 (16) of the test molecules within 0.1 Å (0.2 Å) RMSD of the crystal structure, the DG method is seen to fail to generate even one conformation. The DG methods do generate conformations within 0.3 Å, though a significantly lower number than *et*. Note that at this accuracy the DG with optimization is much better than without.

In Figure 5, the number of sets with at least one conformation within 0.3 Å RMSD are plotted for each conformation set. From the figure it can be seen that number

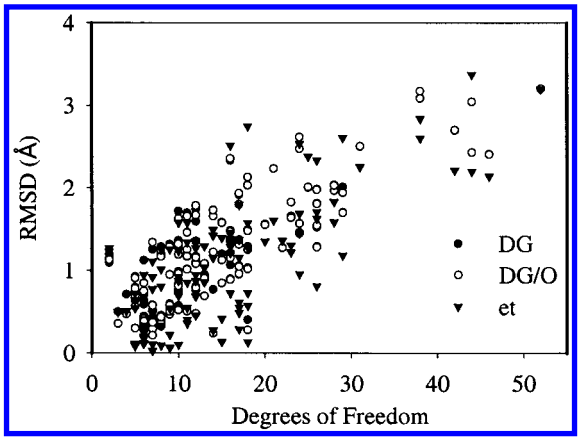

Figure 6. Effect of molecular flexibility on algorithm's ability to generate observed crystal structure. The RMSD is the root-mean-square deviation. Degrees of freedom defined in text.

of sets which had at least one of the 25 conformations within 0.3 Å RMSD of the crystal structure are 20, 1, and 7 for *et*, DG, and DG/O, respectively. Local optimization of the DG conformations can be seen to significantly improve the quality of DG structures. Results are similar for the 50, 100, and 250-conformation sets with *et* always finding more matches than the DG method, though both methods improve with increasing number of conformations. In contrast to *et*, increasing the number of generated conformations significantly improves the results of the distance geometry algorithm. The DG/O 250-conformation set finds nearly 3 times as many conformations within 0.3 Å of the corresponding crystal structure as the 25-conformation set. For *et*, the good conformations are found earlier, requiring fewer conformations to be generated. While the distance geometry performs a more uniform sampling of the conformational space, these results demonstrate that *et* samples conformations more likely to be observed. The 250-conformation DG/O set that is comparable to the 25-conformation *et* set requires 30 times as much CPU. In addition, the time for any modeling task that searches through the conformations would increase 10-fold due to the increased number of conformations that need to be evaluated.

One point that still needs to be addressed is how the methods perform as the number of rotatable bonds increased. Is the better performance of *et* due to the small rigid molecules? In fact, a detailed analysis finds just the opposite: distance geometry does better than *et* on small less flexible molecules. To compare the methods as the intrinsic flexibility of the molecule increases, the RMSD for the conformation closest to the observed structures is plotted against the number of degrees of freedom (DOF) for the 25-conformation sets in Figure 6. For present purposes the DOF is calculated by summing (i) the number of rotatable bonds in chains, (ii) $N - 3$ for N-membered rings, and (iii) number of unspecified chiral centers. The DOF ranges from 2 to 52 for the molecules in the test set. From Figure 6, it is clear that the RMSD of the knowledge-based method is generally lower than the distance geometry method for most degrees of freedom examined. However, at the smallest degrees of freedom, e.g., 2, 3, and 4, DG does better as seen by the lower RMSD at these values. This is not surprising when considering that the torsion angles are coarsely binned in this knowledge-based approach with *et* mining torsion angles in 15° wide bins. This resolution is relatively coarse for the molecules containing a small number of rotatable bonds, resulting in fewer conformations generated and consequently less sampling of the available conformational space. One would therefore expect the DG/O method to sample conformations closer to the crystal structure for smaller and more rigid molecules. Note also that optimization improves the DG performance but the difference in RMSD between DG with and without optimization decreases with increasing DOF. Since the RMSD is reported per heavy atom and the optimization is local, i.e., bond lengths and bond angles, the effect of optimization plays a relatively smaller part in the reported RMSD for larger molecules.

In Figure 7, the best conformations for *et* and DG/O are shown for ligands of the PDB entries 2PK4, 3GCH, 2SIM, and 1IVE. The first three comparisons, 2PK4, 3GCH, and 2SIM, show that *et* found the crystal structure (Figure 7A,C,E) within a set of 25 conformations whereas the best structures for DG/O (Figure 7B,D,F) were found in the 250-conformation sets and still had a relatively large RMSD from the crystal structure. The RMSDs for the 25 conformation DG/O set were 0.90, 0.59, and 1.12 Å for 2PK4, 3GCH, and 2SIM, respectively. However, *et* fails to find the crystal structure for 1IVE as shown in Figure 7G in contrast to DG/O (Figure 7H). In this case the closest *et* conformation is eliminated due to steric constraints. Some flexibility in bond lengths, bond angles, and/or torsion angles could help resolve this problem. Within the context of the present program a better conformation (RMSD = 0.57 Å) is obtained for 1IVE by allowing for closer contacts.

SUMMARY

Motivated by the need to quickly generate biologically relevant low-energy conformations to facilitate the drug discovery process, a new knowledge-based method has been developed. The knowledge-based approach presented above for conformation generation has been shown to be faster than the previously published distance geometry approach while producing conformations more likely to be observed. The present method treats rings and chains simultaneously and in similar fashion. In addition, correlations between torsion angles, up to three open chain and five ring torsion angles, are defined through patterns and the appropriate values mined from a experimentally characterized database. This empirically determined correlation between torsion angles is instrumental in reducing the total conformational space available to a molecule, and it allows for rapid enumeration of the energetically favorable conformations. A pattern and its associated data of observed torsion angles constitute a rule. Current rules have evolved over a number of years in Merck Research Laboratories worldwide. Knowledge embedded in the present method includes not only the torsion angle values mined from a crystal structure database but also contributions from chemists in the form of which patterns play an important role. At present 797 rules have been developed and included in the knowledge base, where a single rule may contain between one and five torsion angles. The database of observed values for the rules has been mined from selected members of the Cambridge Structural Database, representing a diverse group of organic compounds.

Of equal importance to the patterns and database is the conformation generation algorithm that has been imple-

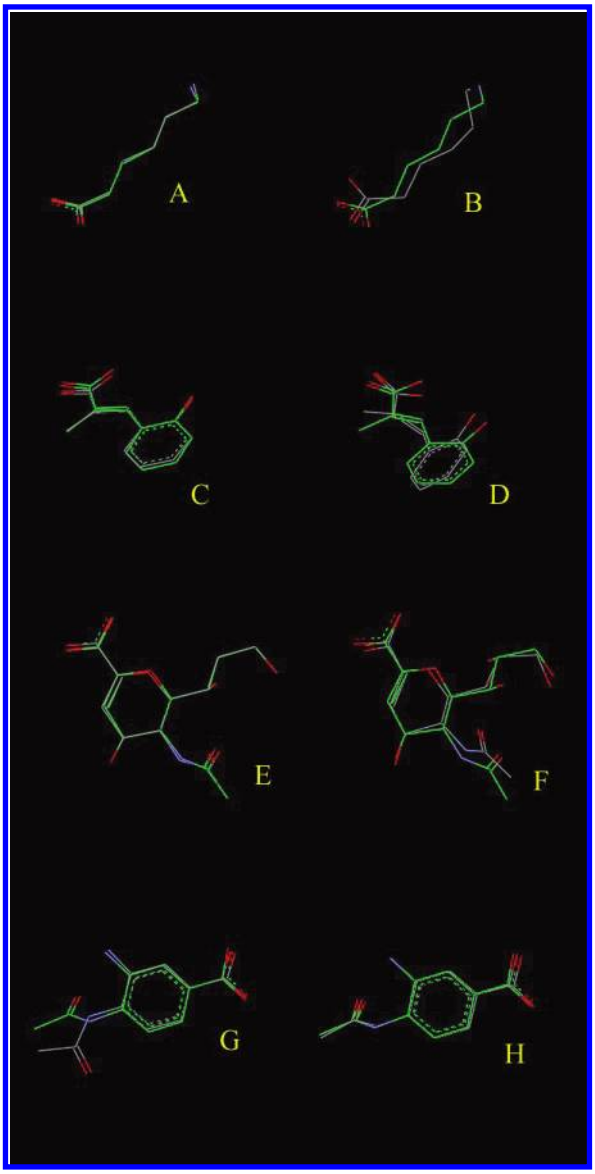

Figure 7. Comparisons of PDB ligand structures with *et* and DG/O conformations. (A) *et* conformation from 25 conformation set with 2PK4. The RMSD is 0.11 Å. (B) DG/O conformation from 250 conformation set with 2PK4. The RMSD is 0.68 Å. (C) *et* conformation from 25 conformation set with 3GCH. The RMSD is 0.11 Å. (D) DG/O conformation from 250 conformation set with 3GCH. The RMSD is 0.55 Å. (E) *et* conformation from 25 conformation set with 2SIM. The RMSD is 0.14 Å. (F) DG/O conformation from 250 conformation set with 2SIM. The RMSD is 0.67 Å. (G) *et* conformation from 250 conformation set with 1IVE. The RMSD is 0.91 Å. (H) DG/O conformation from 250 conformation set with 1IVE. The RMSD is 0.16 Å.

mented. Since many rules share common torsion angles, care has been taken to identify the largest substructure covered by a single rule. Using this fragment as the core structure, the remaining torsions are assigned in the order of appearance as successive shells about the core are examined. In this way the most appropriate rules for the molecule are uniquely identified. The database entries with the lowest probability of occurring and having the least impact on the final conformation, e.g., infrequently observed torsion angles farthest from the center of the molecule, are eliminated until a fixed number of hypotheses is determined. Each hypothesis is a recipe for a molecular conformation. Using diversity in hypotheses to preferentially order conformations before actual construction substantially increases the efficiency of the conformation generation methodology.

The algorithm is sufficiently robust to allow any set of rules and its corresponding database to be utilized. For example, a set of rules derived from proteins may be employed in one effort while rules based on small organic molecule structures could be more appropriate for another. It is clear that the same moiety or substructure may prefer different conformations in different environments, e.g., in proteins versus in small molecules.

Conformations of molecules containing macrocycles, cages, or rings with greater than eight atoms are not sampled well by *et* since only the side chain atoms are allowed to move. In addition, since bond lengths and bond angles are held constant at the values of the input structure, any conformation requiring relaxation of these parameters may be rejected due to steric overlap.

The structures of 113 protein-bound molecules were used to compare the present methodology to the Blaney et al. distance geometry method. When generating over 250 conformations, the present method has proven to be an order of magnitude faster than the distance geometry approach. The new knowledge-based approach has also been found to generate conformations closer to the experimentally determined conformation and to generate them earlier. The superiority of this present methodology extends through a wide range of molecular sizes, performing better for the most flexible molecules tested relative to the published DG method.

ACKNOWLEDGMENT

The authors thank the chemists who helped evolve the rule set through application of *et* to their projects: C. Bayly, H. Broughton, B. Bush, L. Castonguay, Y. Gao, K. Holloway, P. Hunt, M. Mackey, G. McGaughey, R. Mosley, A. Naylor-Olsen, K. Prendergast, R. Sheridan, S. Singh, A. Tebben, D. Underwood, and M. Walker.

REFERENCES AND NOTES

(1) Kearsley, S. K.; Smith, G. M. *Tetrahedron Comput. Methodol*. **1990**, *3*, 615−633.
(2) Miller, M. D.; Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P. *J. Comput.- Aided Mol. Des*. **1994**, *8*, 153−174.
(3) Klebe, G.; Mietzner, T.; Weber, F. *J. Comput.-Aided Mol. Des*. **1994**, *8*, 751−778.
(4) Miller, M. D.; Sheridan, R. P.; Kearsley, S. K. *J. Med. Chem*. **1999**, *42*, 1505−1514.
(5) Howard, A. E.; Kollman, P. A. *J. Med. Chem*. **1988**, *31*, 1669.
(6) Leech, A. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1991; Vol. 2, pp 1−47.
(7) Saunders, M.; Houk, K. N.; Wu, Y.; Still, W. C.; Lipton, M.; Chang, G.; Guida, W. C. *J. Am. Chem. Soc*. **1990**, *112*, 1419−1427.
(8) Treasurywala, A. M.; Jaeger, E. P.; Peterson, M. L. *J. Comput. Chem*. **1996**, *9*, 1171−1182.
(9) Weinberg, N.; Wofe, S. *J. Am. Chem. Soc*. **1994**, *116*, 9860.
(10) Judson, R. S.; Jaeger, E. P.; Treasurywala, A. M.; Peterson, M. L. *J. Comput. Chem*. **1993**, *14*, 1407.
(11) Kolossary, I.; Guida, W. C. *J. Am. Chem. Soc*. **1996**, *118*, 5011−5019.
(12) Klebe, G.; Mietzner, T., *J. Comput.- Aided Mol. Des*. **1994**, *8*, 583−606.
(13) Dolata, D. P.; Carter, R. E. *J. Chem. Inf. Comput. Sci*. **1987**, *27*, 36−47.
(14) Dolata, D. P.; Leach, A. R.; Prout, K. *J. Comput.-Aided Mol. Des.* **1987**, *1*, 73−85.
(15) DeClerq, P. J.; Hoflack, J.; Cauwbergh, S. QCPE Program No. QCMP079; Bloomington, IN.

(16) Cambridge Structural Database, Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, CB2 1EZ, UK.
(17) Blaney, J. M.; Crippen, G. M.; Dearing, A.; Dixon, J. S. DGEOM, #590; Quantum Chemistry Program Exchange; Indiana University: Bloomington, 1990.
(18) Gasteiger, J.; Rudolf, C.; Sadowski, J. *Tetrahedron Comput. Methodol.* **1990**, *3*, 537−547.
(19) Perlman, R. S. Rapid Generation of High Quality Approximate 3-dimension Molecular Structures. *Chem. Des. Auto. News* **1987**, *2*, 1.
(20) Bush, B. L.; Sheridan, R. P. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 756−762.
(21) BatchMin and MacroModel were developed in the laboratories of Professor Clark Still (Columbia University) and are available from Schrodinger, Inc. (Portland, OR). Version 5.0 was released in 1995.
(22) Hendrickson, J. B.; Boeckman, R. K.; Glickson, J. D.; Grunwald, E. *J. Am. Chem. Soc*. **1973**, *95*, 494−505.
(23) Anet, F. A. L.; Krane, J. *Tetrahedron Lett.* **1973**, 5029−5032.
(24) Allen, F. H.; Kennard, O.; Taylor, R. *Acc. Chem. Res*. **1983**, *16*, 146−153.
(25) Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 490−519.
(26) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−127.
(27) Allen, F. H.; Harris, S. E.; Taylor, R. *J. Comput.-Aided Mol. Des*. **1996**, *10*, 247−254.
(28) Willett, P.; Barnard, J. M.; Downs, G. M. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 983−996.
(29) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. E., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **1977**, *112*, 535.
(30) Nachbar, R. B. Merck Research Laboratories, unpublished results.
(31) Clean, Molecular Simulations Inc., Burlington, MA, 1998.

CI000464G