CAESAR: A New Conformer Generation Algorithm Based on Recursive Buildup and Local Rotational Symmetry Consideration

Jiabo Li,*,† Tedman Ehlers,† Jon Sutter,† Shikha Varma-O'Brien,† and Johannes Kirchmair‡

Accelrys Inc., 10188 Telesis Court, San Diego, California 92121, and Computer Aided Molecular Design Group, Department of Pharmaceutical Chemistry, Institute of Pharmacy, University of Innsbruck, Innrain 52, A-6020 Innsbruck, Austria

Received April 16, 2007

A highly efficient conformer search algorithm based on a divide-and-conquer and recursive conformer build-up approach is presented in this paper. This approach is combined with consideration of local rotational symmetry so that conformer duplicates due to topological symmetry in the systematic search can be efficiently eliminated. This new algorithm, termed CAESAR (**C**onformer **A**lgorithm based on **E**nergy **S**creening **a**nd **R**ecursive Buildup), has been implemented in Discovery Studio 1.7 as part of the Catalyst Component Collection. CAESAR has been validated by comparing the conformer models generated by the new method and Catalyst/FAST. CAESAR is consistently 5−20 times faster than Catalyst/FAST for all data sets investigated. The speedup is even more dramatic for molecules with high topological symmetry or for molecules that require a large number of conformers to be sampled. The quality of the conformer models generated by CAESAR has been validated by assessing the ability to reproduce the receptor-bound X-ray conformation of ligands extracted for the Protein Data Bank (PDB) and assessing the ability to adequately cover the pharmacophore space. It is shown that CAESAR is able to reproduce the receptor-bound conformation slightly better than the Catalyst/FAST method for a data set of 918 ligands retrieved from the PDB. In addition, it is shown that CEASAR covers the pharmacophore space as well or better than Catalyst/FAST.

INTRODUCTION

Conformation generation plays an integral role in molecular modeling.[1−6] Many applications require an accurate sampling of the conformational space in order to ensure success. Examples include, but are not limited to, 3D pharmacophore modeling,[7−11] database building, and searching. As motivated by high throughput virtual screening, there is always a strong demand for even faster and more reliable conformer search algorithms[12−20] in order to speed up computer-aided drug discovery.[7] From a methodology perspective, the existing conformer generation algorithms can be classified into two major categories: (1) a deterministic search[12,13,21−26] and (2) a stochastic search.[27−43] While from the view point of applications, the conformer search methods can be classified into the following two categories: (1) a search for a specific conformation or generating a single representative conformation[24,25] or structure determination using distance restraints from NMR experiments[22] and (2) a diverse conformer sampling of conformational space (e.g., generating a collection of conformers for 3D database searching and pharmacophore generation).[11,14−20,44]

Stochastic methods such as random search,[27,28] molecular dynamics,[29,30] distance geometry,[31−34] genetic algorithm,[35−39] and simulated annealing[40−43] are used mostly for searching specific conformations, such as GEM. These methods rely on certain random perturbations and are usually followed by a geometric optimization. Stochastic methods are usually not the most efficient general conformer sampling methods. Moreover, the geometric optimization tends to lead to uneven conformer sampling as two initially different structures can approach the same local minima resulting in conformer duplicates. One remedy for such a problem is the poling algorithm introduced in Catalyst.[20] The poling method is a technique for promoting conformational variation. The function that is typically minimized to generate conformers is modified to force similar conformers away from each other. The algorithm has been shown to be extremely effective at creating a diverse sampling of the conformational space.[20] A most recent study shows that Catalyst conformer generation with poling can cover most of the pharmacophore space with significantly fewer conformers than other methods.[46] Although the poling algorithm is effective, the use of poling may slow down the sampling process since the geometry optimization with poling penalty is involved, and the conformer pairwise contribution to the poling penalty needs to be computed for each cycle of optimization. In the CAESAR algorithm, the geometry optimization is replaced by energy pruning on fine torsion grids.

The deterministic methods, including systematic search (or torsion grid search)[12,13,21,22] and rule-based approaches,[24,25] are potentially the most efficient ones. However, the actual performance depends greatly on the implementation details. For instance, most systematic search methods require an exhaustive search of predefined torsion grids even if only a small subset of conformers is requested. Such an approach becomes prohibitive for large molecules due to a combina-

* Corresponding author e-mail: jli@accelrys.com.
† Accelrys, Inc.
‡ University of Innsbruck.

10.1021/ci700136x CCC: $37.00
Published on Web 08/11/2007

torial explosion. For efficient implementation of a systematic search method, a divide-and-conquer strategy is necessary. Such an approach has been used in some fashion within recent developments in programs such as OMEGA[12] and CONAN.[13] In OMEGA,[12] a molecule is divided into small pieces with each having at most five rotatable bonds. The exhaustive search is then performed for each piece, and the conformations of the entire molecule are then constructed from all pieces at once. The intersection method in CONAN[13] also has a similar strategy: a molecule is divided into several overlapping fragments, an exhaustive search is performed for each fragment, and finally the conformations of the whole molecule are constructed by intersecting all the overlapping pieces. The efficiency of such an approach comes from the reuse of fragment conformations and a fast check for intersection matching. One can expect that such an algorithm is particularly efficient for combinatorial libraries. However, such an approach will suffer from combinatorial overflow if fine torsion grids are used. If torsion grids are not fine enough, then matching intersections may not be found. A similar strategy was also proposed for NMR structure determination of polypeptides,[22] in which an exhaustive search for a specific 3D conformation which satisfied experimental NMR restraints was sought.

Before we discuss the development of a new conformer search algorithm we must answer the following questions: (1) What is the purpose of the conformer search? (2) How should we evaluate the quality of the conformer model? In this paper, we discuss a new conformer generation algorithm with the purpose of efficiently and effectively sampling both conformational and pharmacophore space for small, druglike molecules. The pharmacophore space for a druglike molecule can be considered as a collection of its chemical features and their all possible 3D locations due to the conformational variation. The new algorithm can be used to quickly build conformational databases that can be searched with pharmacophore queries. Two different validation methods are used to measure the quality of our new conformer models. We attempt to measure how effective CAESAR is in finding the X-ray conformation (bound to a receptor) and how effective it is in covering the pharmacophore space. The new algorithm is based on a more efficient divide-and-conquer approach: recursive decomposition of a molecule into the smallest units followed by recursive buildup of molecular conformations from the smallest units. We also put careful consideration in the implementation for achieving maximum efficiency. While here we are focusing on the diverse sampling of low-energy conformers, conformer energy screening is performed at each level of the conformer buildup. This allows for the conformers with high-energy structures or bad VDW clashes to be pruned at a very early stage, even before the conformers are fully constructed. This can be thought of as a tree pruning exercise that allows us to focus on energy reasonable portions of the tree without wasting time on energy unreasonable portions. The algorithm also considers local rotational symmetry so that duplicate conformers due to topological symmetry are effectively eliminated during conformer buildup. The new algorithm, CAESAR, has been implemented in Discovery Studio Version 1.7 as part of the Catalyst Component Collection. This paper discusses the algorithm in detail and presents validation results.

METHODS

Four data sets have been used in the current validation work. The data sets include 918 ligands extracted from the PDB using LigandScout,[45] 168 randomly selected molecules from the World Drug Index database, and the entire Maybridge database (v2004). Furthermore, we added 10 sulfonamide containing compounds to account for adequate chemical space coverage for this important class of drugs.

Software and Hardware. Discovery Studio (DS) v1.7 and Pipeline Pilot v6.0 (Accelrys, Inc.) were used to generate all the conformations as well as for building and searching databases. We developed a standard validation protocol using Pipeline Pilot v6.0 for fitting conformer models to receptor-bound ligand conformations and calculating rmsd values. These protocols provide a consistent framework to standardize validation of any conformation generation method. The conformation generation algorithms are Windows OS compatible; however, in the current study CAESAR and Catalyst/FAST conformations were generated on a Pentium IV/3.4GHz processor with 1GB RAM, running Red Hat Enterprise Linux WS 4.0. The validation protocol was developed and processed using Pipeline Pilot running on Windows XP.

Database Building. Maybridge v2004 was used to build Catalyst formatted 3D conformer databases using both CAESAR and Catalyst/FAST. The CAESAR option for database build is currently accessible via command line mode in DS 1.7. The default settings are used for both CAESAR and Catalyst/FAST, the maximum number of conformers is 100, and the energy threshold is 20 kcal/mol.

Conformational Model Generation with CAESAR and Catalyst/FAST. Three settings for the maximum number of generated conformers are used: 100, 250, and 500. The conformer energy threshold is 20 kcal/mol. For speed test, all conformations were generated using command line mode with pregenerated ring fragment data file for both CAESAR and Catalyst/FAST. To avoid the influences on the generation of conformers from existing crystal structures, starting crystal structure conformations were discarded by converting the raw data into a random 3D conformation using only the topology information of the structures (catConf with −*scratch* option). The resulting file was used as input for both CAESAR and Catalyst/FAST for comparison.

THE CAESAR ALGORITHM

The CAESAR algorithm can be represented schematically in Figure 1. A molecule can be represented as a tree, as shown in Figure 1A. Each tree node is either a ring or a rigid structure, and each tree edge represents a rotatable bond. The algorithm involves several steps: the first step is recursive partitioning of a molecule tree into the smallest units as (Figure 1A). At the top level, a molecule tree is divided into two subtrees of approximately equal complexity. The same step is applied to the subtrees recursively until no further partitioning can be performed. Ring structures (such a cyclohexane) or rigid structures (such as $-CH_2-$ and $>C=C<$) are considered as the smallest units. The partitioning step can be done very quickly because only topological analysis is involved, and it needs to be done only once. The next step is the tree-node conformer initialization and ring sampling, as shown Figure 1B. For simple tree nodes, a node

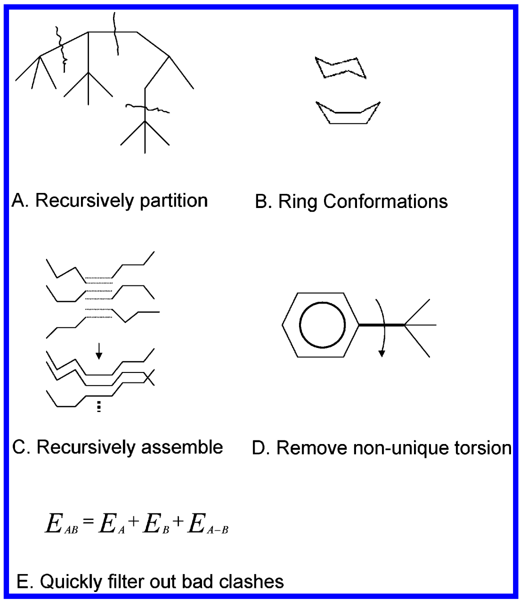

A. Recursively partition
B. Ring Conformations
C. Recursively assemble
D. Remove non-unique torsion
$E_{AB} = E_A + E_B + E_{A-B}$
E. Quickly filter out bad clashes

Figure 1. Schematic presentation of the new algorithm.

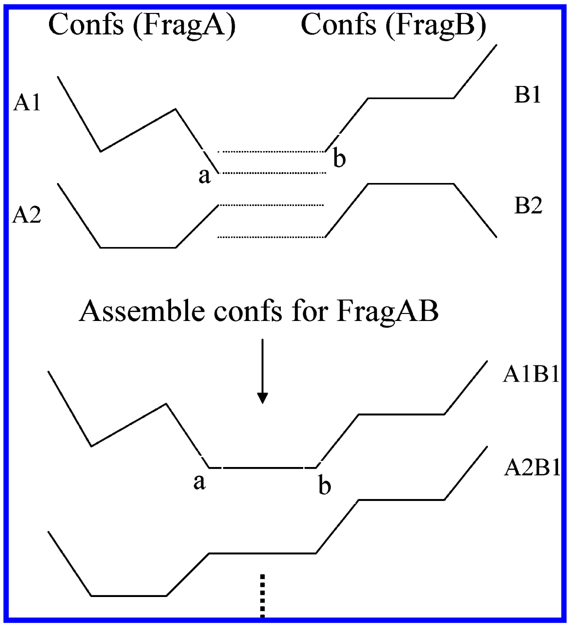


Figure 2. Conformer assembling from smaller fragments.

conformer consists of only one atom center and its bonds. The orientation of those bonds contains the stereo information for a stereo center. For unknown chiral centers, different chiral configurations will be enumerated. For tree nodes of rings, the ring-fragment library technology found in the Catalyst product is used where the ring conformers are generated using distance geometry methods, and the number of ring conformers is adjusted according to the flexibility of the ring. The ring conformers are cached in memory and reused for other molecules. The ring conformers can also be exported to a library file and can be reloaded into memory for a different conformer generation task. Step C, as shown in Figure 1C, is the recursive conformer assembling which starts from the smallest fragments. We have put careful consideration on efficiency in step C. At each level of recursive assembling, the local rotational symmetry and quick energy pruning are performed as shown in Figure 1D,E. As an illustration, let us consider the assembling of two small fragments FragA and FragB into a larger fragment FragAB through a connection bond a-b. At this stage, the conformations of fragment FragA and FragB are available, either from the tree-node conformers or from assembly at a lower level. At this stage, all small fragment conformers have been cleaned for bad clashes and also satisfy the energy threshold requirement as specified by the user. The assembling includes the following substeps as shown in Figure 2. The first step is to reset the orientation for all conformers in both FragA and FragB so that the connection bond a-b is in the same standard orientation, say bond a-b is on the *x*-axis with atom a at the origin and atom b at a distance of bond length of a-b. The next step is to select a fragment conformer from FragA and another one from FragB and join them together with a selected torsion angle. Assuming that there are N_A conformers in FragA, N_B in FragB, and the torsion grid contains N_R possible torsions, then the total number of conformer candidates for FragAB can be expressed as follows:

$$N_{\text{candidate}} = N_A \times N_B \times N_R \tag{1}$$

At this stage, we have the following restrictions of conformer sampling for the subfragment FragAB: (1) a fast energy screening, (2) a torsion angle check for nonunique torsions due to local rotational symmetry, and (3) a control for the maximum number of intermediate conformers for FragAB. For example, the conformer energy of the joint fragment FragAB can always be expressed as follows:

$$E_{\text{Conf(FragAB)}} = E_{\text{Conf(FragA)}} + E_{\text{Conf(FragB)}} + E_{\text{Conf(FragA)}-\text{Conf(FragB)}} \tag{2}$$

The conformer energy of FragAB consists of three terms as shown in eq 2. The first two values are subconformer energies which are already available at this level of assembling. Generally speaking, the energy of a subfragment conformer is calculated with a force field using the same atom typing for all fragment atoms as in their parent molecule. In the current implementation of CAESAR, the Catalyst force field parameters are used for energy calculation. Since CAESAR is a torsion search algorithm, the bond lengths and bond angles are all fixed, and their contribution to the total energy is a constant and is therefore skipped from calculation. The third term, the interaction energy between Conf(FragA) and Conf(FragB), includes both torsion energy and VDW interaction energy from the Catalyst force field. Both need to be computed for each combination of subfragment conformers. Potentially, the torsion energy could be precomputed for each grid point, however, we are not exploring this advantage as the interfragment VDW interaction energy computation is the dominant cost. If the total energy is larger than a threshold, then the joint conformer is rejected. We use the following method to set the energy threshold for subfragment conformers. Assume the user specified energy threshold for the complete conformers of the entire molecule is E_T, then at the top level of conformer assembly the energy threshold is also set to E_T. At the previous level of assembly, the threshold is set to $^2/_3E_T$. The threshold is set to be $^1/_2E_T$ for any level lower than the top two. In our experience, this simple method efficiently and effectively samples the low-energy conformers and prunes the high-energy substructures at an early stage. A very interesting feature of the CAESAR algorithm is that the energy of a conformer is accumulated from its building fragments during the recursive buildup instead of calculating the energy after the conformer is fully created.

We also employ a very simple method to control the number of intermediate subconformers of fragments in order to achieve maximum efficiency. It is not necessary to exhaustively sample all conformers in the conformer space when only a subset of conformers is requested as long as the subset sampling is evenly distributed. Let us assume that the required number of conformers of FragAB is N_i, and its subfragments FragA and FragB have similar flexibility, then the number of intermediate conformers, N_{i+1}, for each of the subfragments can be estimated as the square root of N_i. According to eq 1, the total combination for FragA, FragB, and torsion grids is larger than N_i. This is necessary since some of the combinations give high-energy conformers and need to removed. For small N_i, say less than 50, we find that setting $N_{i+1} = N_i$ gives a better chance of sampling lower energy conformers. Based on the above consideration, we proposed the following recursive formula to calculate how many conformers for a subfragment is needed

$$N_{i+1} = \begin{cases} N_i, \text{ (if } N_i \le 50) \\ 50 + \sqrt{N_i - 50}, \text{ (if } N_i > 50) \end{cases} \qquad (3)$$

where N_i is the required number of conformers for the fragments at level i, and N_{i+1} is the number of conformers needed for level i+1 fragments. At the top level (the whole molecule), the number of conformers is equal to the number of conformers specified by the user. Our experience shows that the above formula works quite well. With this simple formula, we avoid the combinatorial explosion of the conformer space while always keeping enough intermediate conformer candidates of substructures to ensure adequate sampling.

Another important element in the new algorithm is a general method which is used to eliminate the generation of duplicate conformers for molecules with topological symmetry. This is discussed in detail in the following section.

Topological Symmetry Consideration. Previously within Catalyst a unique technology for duplicate conformer checking using rmsd between conformer pairs was used. For example, if the molecule has topological symmetry, then all symmetry mappings up to a maximum value can be considered. This is normally a quick postconformer generation process. The computational cost of this procedure is proportional to $M \times N^2$, where M is the number of topological symmetry operations, and N is the number of conformers. For a molecule with high topological symmetry, M can be prohibitively large making the duplicate check quite expensive, especially when the number of conformers is large. Since the postconformer generation check can be expensive, the best way to avoid this is to eliminate duplicate conformations before they are generated. Since we construct conformers from small units in a recursive way, it is possible to avoid generating duplicate conformers during the assembly process. In the CAESAR algorithm, a simple and general approach is employed to achieve this. The idea is to consider local rotational symmetry of fragment conformations. Theoretically speaking, a systematic search algorithm does not create strict duplicate conformers if the molecule does not have any topological symmetry. For molecules with topological symmetry, a systematic search algorithm may lead to some duplicates. If a fragment has local rotational symmetry, such as a 1,4-disubstituted benzene, phenyl group, *tert*-butyl group, etc., then some of the torsion angles are redundant when we try to rotate the fragments to assemble new conformers of their parent fragment. Some ad-hoc treatments for special cases, such as 1,4 disubstituted benzene, have been previously reported.[12] Here we generalize such kind of treatment. The most common cases are 2-fold symmetry such as phenyl groups and 3-fold symmetry such as *tert*-butyl groups. In the new algorithm, only 2-fold and 3-fold rotation symmetries are considered. To have a better understanding of this issue, let us consider a simple example shown in Figure 3. Assuming the default torsion grid for a sp^2-sp^3 rotatable bond is 6 and if the two fragments have no rotational symmetry, then one can get 6 conformations by rotating the bond angle by 60° each time. For a phenyl-*tert*-butyl molecule, as shown in Figure 3, the phenyl group has a 2-fold rotational symmetry, and the *tert*-butyl has a 3-fold rotational symmetry along the connection bond. In such a case, a 60 degree rotation does not lead to a new conformation due to topological symmetry and local rotational symmetry. If we look at this molecule from the top view (a), then a 60 degree rotation of the *tert*-butyl group leads to conformation (b). If we rotate the whole molecule by 180°, then the same conformation (b) becomes (c). Apparently, (a) and (c) can overlap perfectly on each other since all carbon atoms are equal, and the atom mapping from (a) to (c) for a perfect alignment is (1 → 3, 2 → 1, 3 → 2). This method can be generally formulated in the following way. Let us assume that the two fragments are A and B, and fragment A has a n_a-fold rotational symmetry along the A-B connection bond, and fragment B has a n_b-fold symmetry. Let us also further assume that the default rotational grid number is N_{ab} and the symmetry unique rotations are (0, $2\pi/N_{ab}$, ... $2\pi(N_x$-1)/N_{ab}), where N_x is reduced from N_{ab} by removing the primitive integer factors of n_a and n_b from N_{ab}. This can be illustrated by the phenyl-*tert*-butyl example as

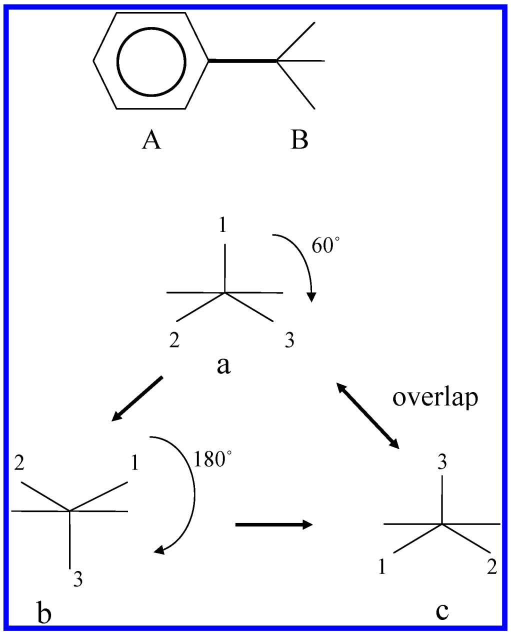
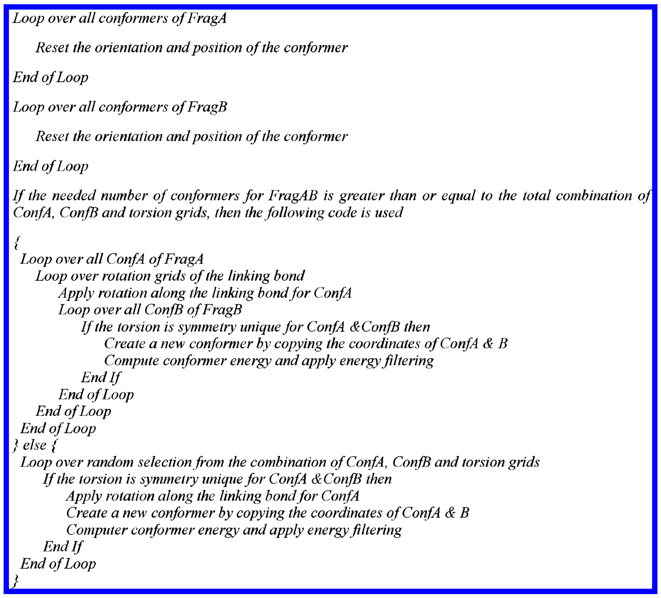


Figure 3. Symmetry unique torsion.

shown in Figure 4. In this example, fragment A (phenyl) has a 2-fold symmetry, thus $n_a = 2$, and fragment *B* (*tert*-butyl) has a 3-fold symmetry, and $n_b = 3$. The number of torsion grids of 6 has two primitive factors, 2 and 3. Reducing the two factors 2 and 3 from 6, one obtains $N_x = 1$. Table 1 shows the most common cases. There is a simple way to check the rotational symmetry for a fragment conformation. For instance, if we test the symmetry of a 3-fold rotation of a fragment conformation, then one can rotate this conformation by 120° and check whether it overlaps the original conformation or not. If the overlap is perfect, i.e., each atom from one fragment conformation matches another atom from another fragment conformation with the same atom type with a zero or very small distance, then the fragment conformation has a 3-fold rotational symmetry along the rotatable bond connecting the two fragments. In the above discussion, we only show the cases where the molecules have only two basic units in each. Actually, the same formula can be applied at any level of conformer assembling except one additional criterion is added for detecting the rotational symmetry of a fragment conformation: the fragment conformation does not have any dangling bond that is not collinear with the current rotational bond. Even though this is a more restrictive condition than necessary, the tradeoff is that we have a very efficient way to eliminate most of the potential duplicates with the exception of some very rare cases in real scenarios. The flow diagram of CAESAR is shown in Figure 5

Pseudocode. The pseudocode of conformer assembling from small fragments conformer is shown as follows.



```
Loop over all conformers of FragA

    Reset the orientation and position of the conformer

End of Loop

Loop over all conformers of FragB

    Reset the orientation and position of the conformer

End of Loop

If the needed number of conformers for FragAB is greater than or equal to the total combination of
ConfA, ConfB and torsion grids, then the following code is used

{
  Loop over all ConfA of FragA
    Loop over rotation grids of the linking bond
        Apply rotation along the linking bond for ConfA
        Loop over all ConfB of FragB
            If the torsion is symmetry unique for ConfA &ConfB then
                Create a new conformer by copying the coordinates of ConfA & B
                Compute conformer energy and apply energy filtering
            End If
        End of Loop
    End of Loop
  End of Loop
} else {
  Loop over random selection from the combination of ConfA, ConfB and torsion grids
     If the torsion is symmetry unique for ConfA &ConfB then
         Apply rotation along the linking bond for ConfA
         Create a new conformer by copying the coordinates of ConfA & B
         Computer conformer energy and apply energy filtering
     End If
  End of Loop
}
```

The first ***if*** block says, if the required number of conformers is greater than or equal to $N_A \times N_B \times N_R$, then only up to this number of conformer candidates will be created. In the ***else*** block, a subset of conformers is randomly selected from the pool of $N_A \times N_B \times N_R$ candidates. The above design of the conformer assembling has an obvious advantage. The transformation of atom coordinates can be reduced to a minimum. Once the conformers of two child fragments have been reoriented and repositioned, there is no need for the matrix transformation for any combination of the child fragment conformers. One can directly copy the coordinates from the child fragment conformers. Although there are some recent discussions about the efficiency of coordinate transformation for conformation construction,[47,48] we believe that the reorientation and repositioning for subfragment conformations in the recursive assembling achieves the maximum efficiency by reducing matrix transformations. For even better efficiency, the reorientation transformation matrix for each connection of each tree-edge atom is updated during the recursive assembling using simple quaternion multiplication instead of recomputing from the coordinates. The latter is a more expensive operation.

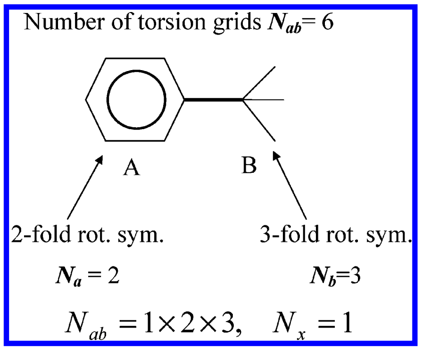

Number of torsion grids N_{ab}= 6
A
B
2-fold rot. sym.
3-fold rot. sym.
N_a = 2
N_b=3
$N_{ab} = 1 \times 2 \times 3, \quad N_x = 1$

Figure 4. Number of symmetry unique rotations.

Table 1. Symmetry Unique Grid Points for Rotatable Bonds with Local Symmetry

example	default no. of grids	n_a	n_b	unique rotation angles	N_x
Ph−CF_3	6[a]	2	3	0.0	1
CH_2═CH−CF_3	6[b]	1	3	0.0, π	2
CH_2═CH−CF_2Cl	6	1	1	0.0, $\pi/3$, ..., $5\pi/3$	6
Ph−C(F)(Cl)Br	6	2	1	0.0, $2\pi/3$, $4\pi/3$	3
R−CF_3	3	1	3	0.0	1
CF_3−CCl_3	3[c]	3	3	0.0	1

[a] The default number 6 is a product of 2 and 3, and the phenyl group has a 2-fold rotation symmetry, and the −CF3 group has a 3-fold symmetry, thus N_x is reduced to 1. [b] Only the −CF3 group has a 3-fold symmetry. Thus $N_x = 2$. [c] Even though both groups have 3-fold symmetry, the default number of grids has only one factor of 3, thus $N_x = 1$.

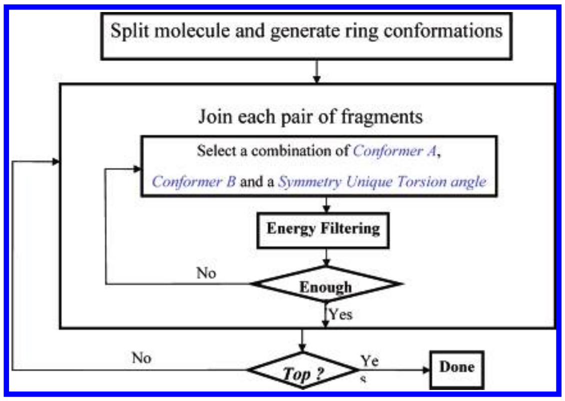

Split molecule and generate ring conformations
Join each pair of fragments
Select a combination of Conformer A, Conformer B and a Symmetry Unique Torsion angle
Energy Filtering
No
Enough
Yes
No
Top ?
Yes
Done

Figure 5. Flow diagram of the new algorithm.

Convergence and Systematic Improvability. One frequently asked question is do we sample the conformer space enough. Traditionally, some convergence methods are used in conformer generation methods. For instance, in a random search algorithm, one can use a number of successive failures of conformer generation as the convergence criteria. If the conformer sampling fails to generate new conformers for a consecutive number of tries, then the conformer sampling is terminated. Our experience shows that such a control is

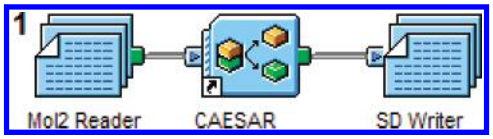


Figure 6. A simple protocol of conformer generation using CAESAR component.

not perfect. The "convergence" can be sensitive to the sampling condition. If we restart conformer sampling for a converged conformer model, then often more new conformers can be generated. This can be repeated quite a few times before the conformer model truly converges with a certain criteria. Therefore, the convergence in stochastic methods has a degree of randomness associated with it. This can be easily explained when you consider that the failure of a conformer sampling in a stochastic method can occur due to a high ratio of VDW clashes or due to poor and uneven sampling. It is not necessarily an indication that the conformer sampling is truly converged. In CAESAR, we adopt a different control on conformer space sampling. The refinement of the conformer sampling is totally controlled by the requested number of conformers and the torsion grids. Therefore, the conformer sampling can be systematically refined by increasing the number of conformers and the torsion grids. In the current implementation, CAESAR uses very fine torsion grids as indicated below:

sp3-sp3: 6

sp3-sp2: 12

sp2-sp2: 8

Implementation and Deployment. The CAESAR algorithm is a standalone code which has been integrated in Pipeline Pilot as a component. This component can be exposed in the DS 1.7 environment by a user created protocol. The Pipeline Pilot componentization of CAESAR also allows one to build customized workflows independent of DS1.7. For instance, Figure 6 shows a very simple protocol which takes a mol2 input file and streams the data through the CAESAR component for 3D conformer generation and writes an output SD file. This CAESAR component can be dropped into any Pipeline Pilot protocol to develop a desired workflow.

VALIDATION AND PERFORMANCE

For a new conformer search algorithm, an obvious question is how to evaluate the quality of the conformer model and the speed. In this work, the purpose of conformer search is to cover the pharmacophore space of druglike molecules; therefore, our focus of validation studies is concentrated in this area. As mentioned earlier, we selected four data sets from varied sources to validate the quality and performance of CAESAR. For quality validation, instead of using a hole-size metric,[49] two kinds of tests which are directly related to the pharmacophore space coverage were used. One is the similarity comparison of hit lists of a database search using various 3D pharmacophore queries. The other is to determine how closely a conformation generated by our algorithms matches (rmsd) an experimental conformation of a bound ligand, as observed in X-ray structures.

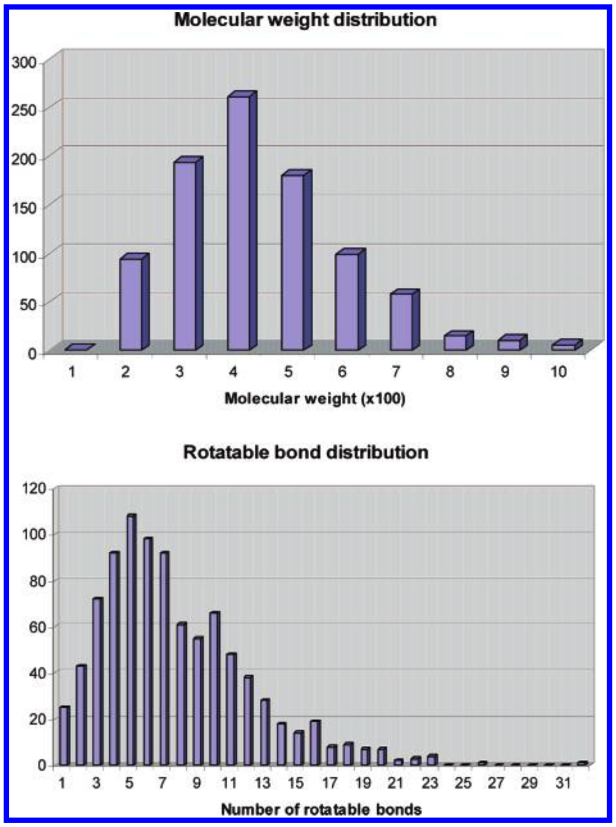


Figure 7. Molecular weight and the number of rotatable bonds distribution of 918 PDB ligands.

Data Set 1: 918 PDB Ligands. This data set is originally from the same source used in a recent study by Kirchmair et al.[18] for validation studies of Catalyst/FAST and OMEGA. Kirchmair originally used LigandScout to assign bond characteristics and to extract the small organic molecules from 918 PDB compounds. Kirchmair found that approximately 1% of the extracted ligands could not be used due to parsing issues, and an additional 7% would require manual intervention to compute the rmsd calculation. In our study, we visually checked all 918 compounds and made careful corrections for tautomerization, hybridization, and bond type for each problematic structure according to the 3D geometries (bond lengths, bond angles, and dihedral angles of heavy atoms) of the compounds. Changes were made for 193 structures of the entire data set. With this cleaned data, our Pipeline Pilot validation protocol was able to process all 918 compounds without error. The distributions of both molecular weight and the number of rotatable bonds are shown in Figure 7 indicating that the molecular weight ranges from 100 to 1000, where most compounds have rotatable bonds in the range of 3−9 and the most flexible compound containing 31 rotatable bonds. This highlights the diversity of the chemical space and properties covered by this data set.

Data Set 2: **168 Molecules from WDI Database**. This data set was randomly selected from the Derwent World Drug Index database. With respect to molecular weight and the number of rotatable bonds, it is also quite diverse as shown in Figure 8**Data Set 3: 10 Sulfonamide Molecules.** On the basis of previously reported findings,[14,18] Catalyst force field parameters were enhanced and updated to account for exhaustive conformational coverage of sulfonamide

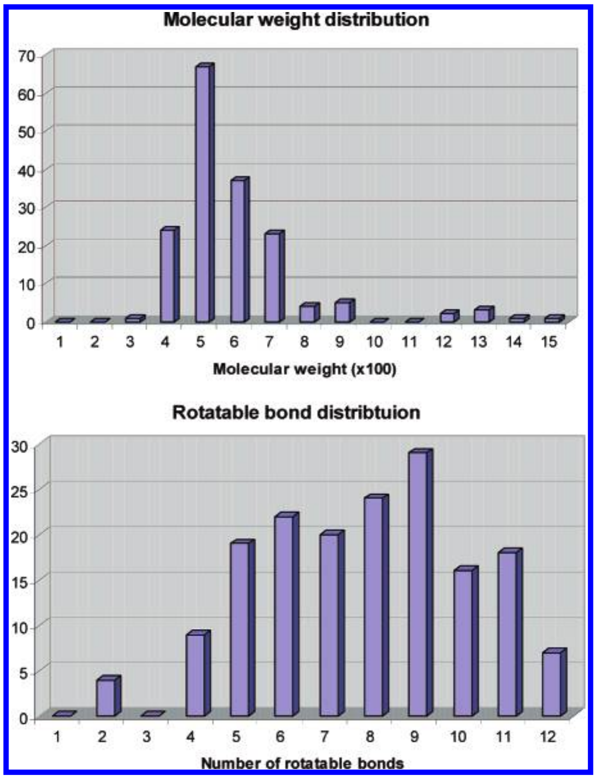


Figure 8. Molecular weight and rotatable bond distribution of 168 WDI compounds.

Table 2. Speed Test with 3 Data Sets and Different Numbers of Maximum Conformers[a]

		CPU time (s)		
data set	MaxConfs	Catalyst/ FAST	CAESAR	speedup
WDI168 (168 molecules)	100	304	63	4.9
	250	694	77	9.0
	500	1700	96	17.7
sulfonamide10 (10 sulfonamides)	100	12	2	6.0
	250	29	2.5	11.6
	500	60	3.8	15.8
PDB918 (918 ligands from PDB)	100	941	144	6.5
	250	2601	191	13.6
	500	5600	265	21.1

[a] A prebuilt ring fragment library file is used for both Catalyst/FAST and CAESAR.

containing compounds. These 10 molecules were retrieved from the PDB and also used as part of this validation

Data Set 4: Maybridge Database (v2004). The entire Maybridge (∼60 000 compounds) database was used to construct Catalyst 3D conformer databases using both Catalyst/FAST and CAESAR algorithms.

Speed Test. The first three data sets were used for conformation generation speed tests. For comparison, both CAESAR and Catalyst/FAST were used. All calculations are performed on a Linux machine with a single Intel Pentium IV/3.40GHz processor running RHEL WS 4.0. The results are summarized in Table 2For three quite different data sets, the performance improvement compared to Catalyst/FAST is in the range of 5−20-fold, depending on the maximum number of conformers requested in conformer generation. The larger the maximum requested conformers, the greater the speedup. Another interesting observation is that the CPU time for Catalyst/FAST increases slightly faster than the linear increment with the maximum number of conformers, while CAESAR increases slower than the linear increment. A detailed analysis on the breakdown of CPU time for CAESAR algorithm shows that, for typical druglike molecules, about 70% of CPU time was spent on the initialization of the conformer search and the exporting of the generated conformers when the maximum number of conformers is set to 250. Similar to Catalyst/FAST, the initialization stage includes parsing and validating input molecules from a SD file and the initialization of tree-node conformations, which can be either generated on the fly or retrieved from a fragment library file. The CPU cost for initialization is independent of the number of conformers requested and therefore is a constant. While this part is only a small fraction of the total cost of conformer search for Catalyst/FAST, it becomes a major part of CPU cost in CAESAR since the actual conformation search part as described in this paper is so efficient. This explains why the speedup for CAESAR vs Catalyst/FAST is much higher for larger number of conformers. This also suggests that the performance of CAESAR can be further increased if the initialization and conformer exporting can be done more efficiently.

Reproducing Receptor-Bound Conformations. One of the measures for the quality of conformer sampling is to examine how close a conformer model is to the bound conformation of a ligand from, for example, an X-ray structure. For this study, the 918 compounds that were extracted from the PDB were used. The conformer generation is done by taking an input SD file without any information of the crystal structure. Since Catalyst/FAST has been extensively validated with respect to its high performance and its ability to retrieve receptor-bound conformations,[14−19] the CAESAR algorithm was directly compared to Catalyst/FAST. The maximum number of conformers was set to 100, and the conformer energy threshold was set to 20 kcal/mol. The average rmsd of the best fitting conformers to the crystal structures were then calculated using a Pipeline Pilot protocol as shown in Figure 9. Pipeline Pilot provides a convenient tool for automating processes[50] such as the validation studies in this work. The statistics for best fitting rmsd for both CAESAR and Catalyst/FAST are given in Table 3. As we can see, CAESAR and Catalyst/FAST both show excellent ability to reproduce receptor-bound ligand conformations with high accuracy. The conformer models of over 60% of the compounds can be fitted to the receptor-bound ligand conformations within 1 Å and over 90% within 2 Å.

It is also interesting to check the diversity of conformation sampling by analyzing the rmsd distribution of all generated conformers against the X-ray structures. Table 4 shows the percentage of all generated conformers within different rmsd ranges from the X-ray structures. The statistics are divided into three categories: ligands with less flexibility, i.e., the number of rotatable bonds (NRB) is less or equal to 3, ligands with moderate flexibility (NRB in the range of 4−6), and the ligands with more flexibility (NRB > 6). The results for both CAESAR and FAST are given in the table. For ligands with no more than 3 rotatable bonds, CAESAR has nearly half of all generated conformers within 1 Å rmsd of the X-ray structures. FAST has 39% of its generated conformers within 1 Å. The rmsd distributions for moderately flexible molecules are very similar for CAESAR and FAST. 57% of conformers are within 1−2 Å range. For more flexible molecules with

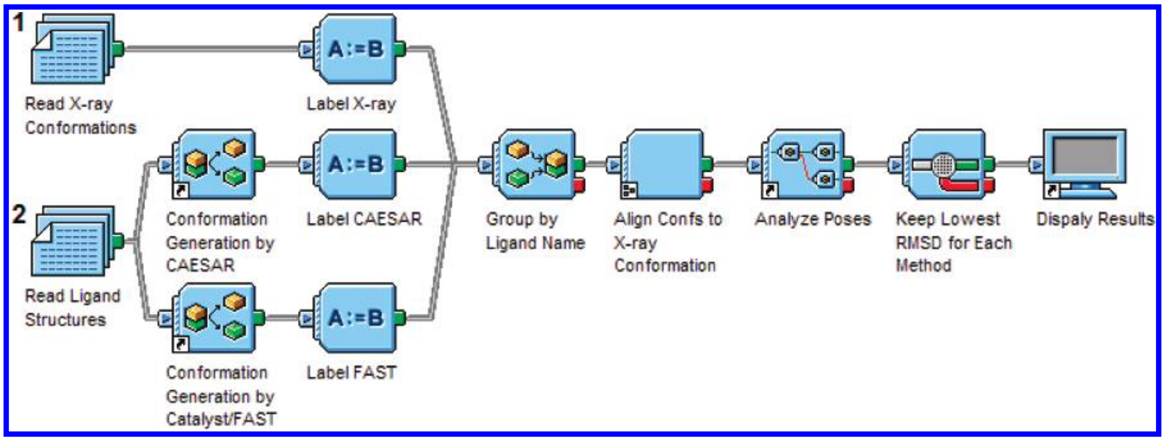


Figure 9. Validation protocol created with Pipeline Pilot. Conformations are generated from SD file without crystal structure coordinates and aligned to the corresponding X-ray pose. The lowest rmsd found for each method is kept. The protocol can easily be extended and customized to provide additional analysis.

Table 3. Average RMSD of the Best Fitting Conformers to the Receptor-Bound Conformers

PDB data	CAESAR	Catalyst/FAST
average rms	0.95	0.97
rms < 0.5	29%	26%
rms < 1.0	61%	60%
rms < 2.0	93%	93%

Table 4. Percentage of All Generated Conformers within Different Ranges of RMSD from the X-ray Structures

	rmsd range (Å)					
	CAESAR			Catalyst/FAST		
number of rotatable bonds	0.0−1.0	1.0−2.0	>2.0	0.0−1.0	1.0−2.0	>2.0
0−3	49	44	7	39	52	9
4−6	18	57	25	17	57	26
>6	2	30	68	1	26	73

more than 6 rotatable bonds, the percentage of conformers within 1 Å is very small.

Database Search. 3D conformer database search with various pharmacophore queries provides a more direct test for the 3D pharmacophore space coverage. Consider a comparison between a 3D database built with a new conformation generation algorithm (for example, CEASAR) and a reference database (for example, Catalyst/FAST). Once searched with a 3D pharmacophore query, if the CAESAR database returns more hits, it would be considered to cover a more complete pharmacophore space. Therefore, it is assumed that the conformation generation method used for constructing this database is superior. This can be further supported if different queries also retrieved more hits from this database. In our current study, four different pharmacophore queries were used for database searching, as shown in Figure 10. Two simple 3D queries containing 3−5 chemical features were selected from Catalyst tutorial, and the feature and shape combined query was created from a database search hit (combine the features and the shape of the hit conformer). The pure shape query was created from a randomly selected conformer of a database molecule. Since we are only doing statistics of hits from a database search, the details of the queries are not essential. Two Catalyst databases were built from Maybridge database (2004) using both CAESAR and Catalyst/FAST. All default settings were used (e.g., maximum number of conformers was set to 100).

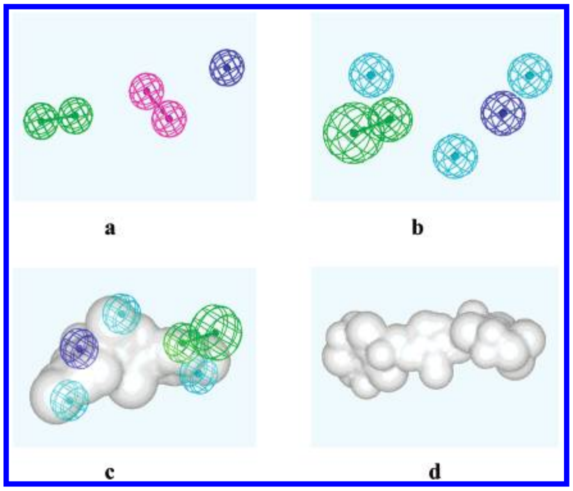


Figure 10. Four 3D queries for database search: (a) a three-feature pharmacophore model, (b) a 5-feature pharmacophore model, (c) a 5-feature pharmacophore model with a shape constraint, and (d) a shape query.

Table 5. Database Search Similarity for 4 Different 3D Queries[a]

	number of hits			
query	Catalyst/ FAST	CAESAR	number of common hits	similarity (%)
Pharm_3F	106	117	93	83
Pharm_5F	51	50	41	81
Shape	68	98	58	70
Pharm_5F+Shape	10	13	6	52
total hits	235	278	198	77

[a] The similarity of two sets A and B is calculated in the following way: similarity = A∩B/[(A+B)/2], i.e. the number of common hits divided by the half of total number of hits in the two hit lists.

The search results with the 4 different queries are shown in Table 5. For simple 3D queries containing 3−5 chemical features, the two databases have more than 80% common hits. For the pure shape query, CAESAR database returns 40% more hits than the Catalyst/FAST database does, while the overlap between the hits indicates that out of 68 hits 58 are common between the two hit lists. The fourth query is a more sophisticated one, which has 5 features and a shape constraint. Even for this extremely restrictive query, 6 out of 10 hits in Catalyst/FAST database are retrieved from the CAESAR database. In addition, the CAESAR database has three more hits than the Catalyst/FAST database. On average, CAESAR database has 77% similarity with Catalyst/FAST database and returns 18% more hits. We conclude that the CAESAR conformer model has 18% more pharmacophore space coverage than Catalyst/FAST conformer model.

CONCLUSIONS AND DISCUSSIONS

We present a new, highly efficient conformer search algorithm, CAESAR, and report the results of validation and performance studies as compared with Catalyst/FAST. The study shows that the new algorithm is 5−20 times faster than Catalyst/FAST for three sets of very diverse compounds with a prebuilt ring fragment library. The ability of the new conformer search algorithm to reproduce the receptor-bound X-ray conformations has also been tested and compared with Catalyst/FAST. The current study shows that CAESAR performs slightly better than the Catalyst/FAST method both in terms of quality and throughput. The pharmacophore space coverage of CAESAR conformer search has been evaluated by building a 3D Maybridge database and searching it with different 3D queries. The results suggest that CAESAR conformer algorithm has better pharmacophore space coverage than Catalyst/FAST. The significant speedup of CAESAR for large conformer model suggests that the new algorithm has more advantage for extensive conformer sampling. The scaling of computational cost with the number of conformers in CAESAR also suggests that there is room for further performance improvement if the initialization of conformer search and the ring conformer generation can be more efficient.

ACKNOWLEDGMENT

We express our sincere thanks to Professor Thierry Langer and Dr. Wolber Gerhard for inspiring discussions. We also thank Dr. Paul Flook, Dr. Remy Hoffman, and Dr. C. M. Venkatachalam for their strong support for this research project. The authors thank Dr. Marvin Waldman for many interesting discussions on this project. Many Discovery Studio team members are also acknowledged for their work of integrating CAESAR into Discovery Studio 1.7. In particular, we thank Mr. Roman Kuchkuda, Dr. Deqiang Zhang, Mr. Thomas Helvey, and Dr. Al Maynard. We also thank Dr. Xuan Hong for the assistance of an early validation study. We would like to gratefully acknowledge Dr. Luke Fisher, Ms. Robin Truelove, and Dr. Derek Stonich for their technical support on Pipeline Pilot protocol development.

REFERENCES AND NOTES

(1) Leach, A. R.; Prout, K. Automated conformational analysis: directed conformational search using the A* algorithm. *J. Comput. Chem.* **1990**, *11*, 1193−1205.

(2) Leach, A. R. A survey of methods for searching the conformational space of small and medium-sized molecules. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1991; Vol. 2, pp 1−55.

(3) Leach, A. R. *Molecular Modeling: Principles and Applications*, 2nd ed.; Pearson Education: England, 2001.

(4) Saunders, M.; Houk, K. N.; Wu, Y. D.; Still, W. C.; Lipton, M.; Chang, G.; Guida, W. C. Conformations of cycloheptadecane. A comparison of methods for conformational searching. *J. Am. Chem. Soc*. **1990**, *112*, 1419−1427.

(5) Vasquez, M.; Nemethy, G.; Scheraga, H. A. Conformational energy calculations on polypeptides and protein. *Chem. Rev*. **1994**, *94*, 2183−2239.

(6) Vengadesan, K.; Gautham, N. A new conformational search technique and its applications. *Curr. Sci.* **2005**, *88*, 1759−1770.

(7) Güner, O. *Pharmacophore Perception, Development, and Use in Drug Design*; International University Line: La Jolla, CA, 2000.

(8) Dixon, S. L.; Smondyrev, A. M.; Rao, S. N. PHASE: A novel approach to pharmacophore modeling and 3D database searching. *Chem. Biol. Drug Des*. **2006**, *67*, 370−372.

(9) Dixon, S. L.; Smondyrev, A. M.; Knoll, E. H.; Rao, S. N.; Shaw, D. E.; Friesner, R. A. PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. *J. Comput.-Aided Mol. Des*. **2006**, *20*, 647−671.

(10) Barnum, D.; Greene, J.; Smellie, A.; Sprague, P. Identification of Common Functional Configurations Among Molecules. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 563−571.

(11) Kristam, R.; Gillet, V. J.; Lewis, R. A.; Thorner, D. Comparison of conformational analysis techniques to generate pharmacophore hypothesis using Catalyst. *J. Chem. Inf. Model.* **2005**, *45*, 461−476.

(12) *Omeag, version 2.0*; OpenEye Scientific Software: Santa Fe, NM, 2005.

(13) Smellie, A.; Stanton, R.; Henne, R.; Teig, S. Conformational analysis by intersection: CONAN. *J. Comput. Chem*. **2003**, *24*, 10−19.

(14) Kirchmair, J.; Laggner, C.; Wolber, G.; Langer, T. Comparative analysis of protein-bound ligand conformations with respect to Catalyst's conformational space subsampling algorithms. *J. Chem. Inf. Model.* **2005**, *45*, 422−430.

(15) Krovat, E. E.; Frühwirth, K. H.; Langer, T. Pharmacophore identification, in silico screening, and virtual library design for inhibitors of the human factor Xa. *J. Chem. Inf. Model.* **2005**, *45*, 146−159.

(16) Perola, E.; Charifson, P. S. Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *J. Med. Chem*. **2004**, *47*, 2499−2510.

(17) Wolber, G.; Langer, T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model.* **2005**, *45*, 160−169.

(18) Kirchmair, J.; Wolber, G.; Laggner, C.; Langer, T. Comparative performance assessment of the conformational model generators Omega and Catalyst: A large-scale survey on the retrieval of protein-bound ligand conformations. *J. Chem. Inf. Model.* **2006**, *46*, 1848−1862.

(19) Boström, J. Reproducing the conformations of protein-bound ligands: a critical evaluation of several popular conformational searching tools. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 1137−1152.

(20) Smellie, A.; Teig, S. L.; Towbin, P. Poling: promoting conformational variation. *J. Comput. Chem*. **1995**, *16*, 171-.187.

(21) Bruccoleri, R. E.; Karplus, M. Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers* **1987**, *26*, 137−168.

(22) Gippert, G. P.; Wright, P. E.; Case, D. A. Distributed torsion angle grid search in high dimensions: a systematic approach to NMR structure determination. *J. Biomol. NMR* **1998**, *11*, 241−263.

(23) Smellie, A.; Kahn, S. D.; Teig, S. L. Analysis of conformational coverage. 2. Applications of conformational models. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 295−304.

(24) Sadowski, J.; Gasteiger, J. From atoms and bonds to three-dimensional atomic coordinates: automatic model build. *Chem. Rev*. **1993**, *93*, 2567−2581.

(25) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparision of automatic three-dimensional model builders using 639 x-ray structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000−1008.

(26) Sadowski, J.; Boström, J. MIMUMBA revisited: torsion angle rules for conformer generation derived from X-ray structures. *J. Chem. Inf. Model.* **2006**, *46*, 2305−2309.

(27) Chandrasekhar, J.; Sanders, M.; Jorgensen, W. Efficient exploration of conformational space using the stochastic search method: application to β-peptide oligomers. *J. Comput. Chem.* **2001**, *22*, 1646−1654.

(28) Saunders, M. Stochastic search for the conformations of bicyclic hydrocarbon. *J. Comput. Chem*. **1989**, *10*, 203−208.

(29) Chen, J.; Im, W.; Brooks, C. L., III. Application of torsion angle molecular dynamics for efficient sampling of protein conformations. *J. Comput. Chem.* **2005**, *26*, 1565−1578.

(30) Sun, Y.; Kollman, P. A. Conformational sampling and ensemble generation by molecular dynamics simulations: 18-Crown-6 as a test case. *J. Comput. Chem.* **1992**, *13*, 33−40.

(31) Leach, A.; Smellie, A. combined model-building and distance-geometry approach to automated conformational analysis and search. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 379−385.

(32) Crippen, G. M. Exploring the conformation space of cycloalkanes by linearized embedding. *J. Comput. Chem*. **1992**, *13*, 351−361.

(33) Peishoff, C. E.; Dixon, J. S. Improvements to the distance geometry algorithm for conformational sampling of cyclic structures. *J. Comput. Chem*. **1992**, *13*, 565−569.

(34) Grippen, G. M.; Havel, T. F. *Distance Geometry and Molecular Conformation*; John Wiley: New York, 1988.

(35) McGarrah, D. B.; Judson, R. S. Analysis of the genetic algorithm method of molecular conformation determination. *J. Comput. Chem.* **1993**, *14*, 1385−1395.

(36) Judson, R. S.; Jaeger, E. P.; Treasurywala, A. M.; Peterson, M. L. Conformational searching methods for small molecules. II. Genetic algorithm approach. *J. Comput. Chem*. **1993**, *14*, 1407−1414.

(37) Glen, R. C.; Payne, A. W. R. A genetic algorithm for the automated generation of molecules within constraints. *J. Comput.-Aided Mol. Des*. **1995**, *9*, 181−202.

(38) Schulze-Kremer, S. Genetic algorithm and protein folding. In *The Protein Structure Prediction − Methods and Protocols*; Webster, D. M., Ed.; Humana Press: NJ, 2000; pp 175−222.

(39) Le Grand, S. M.; Merz, K. M. The genetic algorithm and protein structure prediction. In *The Protein Folding Problem and Tertiary Structure Prediction*; Merz, K. M., Le Grand, S. M., Eds.; Birkhauser: Boston, MA, 1994; pp 109−124.

(40) Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Optimization by simulated annealing. *Science* **1983**, *220*, 671−680.

(41) Wilson, S. R.; Cui, W. L. Applications of simulated annealing to peptides. *Biopolymers* **1990**, *29*, 225−235.

(42) Scheraga, H. A.; Lee, J.; Pillardy, J.; Ye, Y. J.; Liwo, A.; Ripoll, D. Surmounting the multiple-minima problem in protein folding. *J. Global Optim*. **1999**, *15*, 235−260.

(43) Li, Z.; Scheraga, H. A. Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 6611−6615.

(44) 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.

(45) Wolber, G.; Langer, T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model.* **2005**, *45*, 160−169.

(46) Agrafiotis, D. K.; Gibbs, A. C.; Zhu, F. Q.; Izrailev, S.; Martin, E. Conformational sampling of bioactive molecules: a comparative study. *J. Chem. Inf. Model.* **2007**, in press.

(47) Zhang, M.; Kavraki, L. E. A new method for fast and accurate derivation of molecular conformations. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 64−70.

(48) Choi, V. On updating torsion angels of molecular conformations. *J. Chem. Inf. Model.* **2006**, *46*, 438−444.

(49) Smellie, A.; Kahn, S. D.; Teig, S. L. Analysis of conformational coverage. I. Validation and estimation of coverage. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 285−294.

(50) Brown, R. D.; Varma-O'Brien, S.; Rogers, D. Data pipelines and virtual screening: automating the process. *QSAR Comb. Sci*. **2007**, in press.

CI700136X