ConfGen: A Conformational Search Method for Efficient Generation of Bioactive Conformers

K. Shawn Watts,[†] Pranav Dalal,[‡] Robert B. Murphy,[§] Woody Sherman,[§] Rich A. Friesner,[||] and John C. Shelley*,[†]

Schrödinger, LLC, 101 SW Main Street, Suite 1300, Portland, Oregon 97204 and 120 West 45th Street, 17th Floor, New York, New York 10036, D. E. Shaw India Software Private Limited, Sanali Infopark, 8-2-120/113, Road No. 2, Banjara Hills, Hyderabad 500 034, Andhra Pradesh, India, and Department of Chemistry, Columbia University, 3000 Broadway, New York, New York 10027

Received January 9, 2010

We describe the methodology, parametrization, and application of a conformational search method, called ConfGen, designed to efficiently generate bioactive conformers. We define efficiency as the ability to generate a bioactive conformation within a small total number of conformations using a reasonable amount of computer time. The method combines physics-based force field calculations with empirically derived heuristics designed to achieve efficient searching and prioritization of the ligand's conformational space. While many parameter settings are supported, four modes spanning a range of speed and quality trades-offs are defined and characterized. The validation set used to test the method is composed of ligands from 667 crystal structures covering a broad array of target and ligand classes. With the fastest mode, ConfGen uses an average of 0.5 s per ligand and generates only 14.3 conformers per ligand, at least one of which lies within 2.0 Å root-mean-squared deviation of the crystal structure for 96% of the ligands. The most computationally intensive mode raises this recovery rate to 99%, while taking 8 s per ligand. Combining multiple search modes to "fill-in" holes in the conformation space or energy minimizing using an all-atom force field each lead to improvements in the recovery rates at higher resolutions. Overall, ConfGen is at least as good as competing programs at high resolution and demonstrates higher efficiency at resolutions sufficient for many downstream applications, such as pharmacophore modeling.

INTRODUCTION

The ability to accurately and robustly generate bioactive conformations of small molecules is a key step in both ligand- and structure-based drug design. Many widely used techniques, such as docking,[1−4] pharmacophore searching,[5] and shape-based screening,[6,7] depend heavily on the ability to generate conformers that are close to the structure that the ligand assumes in the protein−ligand complex of interest. In benchmarking studies, closeness is usually measured by calculating the root-mean-squared displacement (rmsd) of the heavy atoms of the ligand relative to a conformation of the ligand from a crystal structure of a protein−ligand complex. Existing conformational search algorithms sample the ligands using a variety of techniques, including random torsional angle changes,[8−10] random coordinate changes,[11] distance geometry,[12,13] and rule-based methods,[14,15] or locate minima using normal modes.[16,17] A number of recent studies have been published that attempt to assess and/or optimize the performance of various tools for predicting bioactive conformations.[18−27]

One challenge in conformational searching comes from the fact that bound-state ligand conformations are often not in the global minimum energy state.[22] Ligand conformations in protein−ligand complexes do not exactly correspond to conformations at local potential energy minima for unbound ligands because the binding process induces some degree of strain in the ligand. Therefore, one would not expect that the lowest energy conformation for the unbound state, even with a perfect energy function and complete sampling, to always be a bioactive conformer. However, recent results suggest that bioactive conformers are often closer to energy minima than previously reported.[28] Another important question to consider when analyzing conformational search methods is whether there can be multiple bioactive conformations for a given ligand, and if there is something special about a bioactive conformation relative to other energetically accessible unbound-state conformations. It has been shown that very similar ligands can adopt different conformations upon binding[29] or even that the same ligand can adopt multiple binding modes with different conformational states.[30−33] Clearly, to reliably reproduce bioactive conformations for ligands, a conformational search algorithm must produce multiple conformations.

Regarding whether there is something special about bioactive conformers relative to other low-energy conformations, thermodynamics tells us that any state can be a bioactive conformer as long as the energy cost associated with adopting that conformation can be compensated by favorable interactions in the receptor−ligand complex. While this is true in theory, general observations as well as previous work have shown that ligands typically bind in relatively

* Corresponding author. E-mail: John.Shelley@schrodinger.com.
[†] Schrödinger, LLC, Portland, Oregon.
[‡] D. E. Shaw India Software, Andhra Pradesh, India.
[§] Schrödinger, LLC, New York, New York.
[||] Columbia University, New York, New York.

10.1021/ci100015j
Published on Web 04/07/2010

extended conformations because it maximizes the surface area for favorable interactions with the receptor as well as the displacement of water molecules during the binding process.[22]

Ideally, a conformational search would generate all of the potential bioactive conformations and would not generate any unrealistic or energetically inaccessible conformations. For practical drug discovery applications, it is important to strive toward this ideal scenario. While it is clear that one needs the bioactive conformations, each inappropriate conformation generated in the conformational ensemble requires disk space to store and requires time to process in downstream applications (docking, pharmacophore searching, shape-based screening, etc.). More importantly, each inappropriate conformation increases the chances that the downstream applications will produce inaccurate results. For example, if a docking pose or pharmacophore alignment is obtained using an unrealistic high-energy conformer, this can lead either to invalid insights or to the wrong answer. So there are significant benefits to be gained by having a fast conformational search application that generates a relatively small number of high-quality conformations that include those conformations that have a reasonable likelihood of being bioactive.

Program accuracy is often measured relative to known bioactive conformations for the ligands (e.g., from crystal structures of protein−ligand complexes) using rmsd values of nonhydrogen atoms of optimally overlaid ligand structures. For some applications, such as pharmacophore modeling, rmsd values of 2.0 Å may be adequate, as can be inferred in the work by Dixon et al.[5] where conformations aligned using accurate pharmacophore hypotheses have in most cases rmsd values larger than 2.0 Å.[34] However, for other applications, such as docking, smaller values may be needed. While a high-resolution search would cover the needs of both target applications, it would require more conformations than needed for the former, and the additional conformations may require more time and computer resources than those available for the project. Conformation generation applications should be versatile enough to meet the diverse requirements of different projects.

Producing bioactive conformers is, in principle, an easily tractable task if one does not impose limits on the computational time or the total number of conformations generated, since a full enumeration of all internal degrees of freedom is almost guaranteed to yield conformations that are very close to the bioactive one. An exhaustive search of internal torsional degrees of freedom, ignoring bond length and angle variations, should in most cases generate accurate bioactive conformations given that bond lengths and angles in bound-state conformations are in general reasonably close to their equilibrium solution phase values. However, this approach will produce an impractically large number of conformations for most drug-like molecules.

A number of studies have been performed to analyze bioactive conformations and to assess the ability of various conformational search methods to find such conformations. Typically, the ligand geometries from crystal structures of protein−ligand complexes are taken as the bioactive conformations. In a study by Boström et al., published in 2003 that employed a small set of molecules (36 molecules), search parameters for the program Omega v1.0[35] were varied to improve the probability of generating bioactive conformations.[23] A later study by Perola and Charifson[22] utilized a larger set of ligands (100 from the public domain and 50 in-house) to explore the energetics of conformations associated with binding. They also characterized how well three search programs that were fast enough to be run on large numbers of ligands, namely Catalyst v4.6,[36] ICM v3.0,[37,38] and Omega v1.2,[39] generated bioactive conformations. The searches were run with mostly default parameters and returned on average 100−200 conformations per ligand. All three methods reproduced the bioactive conformations within 1.0 Å rmsd for 60−70% of the ligands. For an rmsd value of 2.0 Å, the recovery rate rose to 94−98%. Kirchmair et al. compared Omega v2.0[40] and Catalyst v4.11[41] using 778 structures from the Protein Data Bank (PDB) that contained drug and pharmacologically relevant molecules that both programs could process without issues and found that Catalyst Fast performs better in high-throughput screenings, while Omega could more accurately reproduce experimental conformations.[26] In a more recent study by Loferer et al., a set of 604 molecules were studied with 5 methods (DGeom,[42] QXP,[43] ROTATE v1.15,[44] LMOD,[45] and Omega v1.8[46]).[45] The most computationally expensive and physics-based method LMOD performed the best according to energetic criteria, whereas other methods were faster and generated more diverse conformational ensembles. Li et al.[27] describe a new method CAESAR and using a test set of 918 ligand structures from the PDB find that it is 5−20 times faster and slightly better at reproducing the ligand conformations than Catalyst FAST. Chen and Floppe[21] performed conformational searches using Catalyst v4.10[47] and MOE v2006.08[48] for a variety of parameter values in the contexts of detailed conformational analysis and high-throughput three-dimensional (3D) library conformational searching. They found that MOE performed at least as well as Catalyst for both categories of conformation generation tasks.

Overall, most studies generally focused on accuracy as opposed to limiting the number of conformers and thus did not address efficiency. One notable exception is a recent study in which a new program TCG[20] was compared with Catalyst v4.10[47] and Omega v2.0[40] considering the speed, accuracy, and ensemble size trade-offs needed to screen large databases of ligands. These three programs were found to be capable of generating conformers with similar accuracy. TCG was judged to be able to reproduce bioactive conformations with an average rmsd of roughly 1.0 Å for ligands with fewer than 9 rotatable bonds using less than 20 conformers on average. The Chen study also showed that Catalyst v4.10[47] and MOE v2006.08[48] could be used to generate relatively few conformers (as low as 33) with nondefault settlings, but the authors generally preferred settings that produced a larger number of conformers.

In this work, we describe the methodology and application of ConfGen v2.0,[49] a conformational search program to efficiently generate bioactive conformations. First, we describe the general ConfGen methodology, including both the physics-based and heuristic rules. We then present the results of applying ConfGen to a set of 667 ligands from cocrystal PDB structures that include compounds from previous publications (Boström[23] and Perola)[22] as well as additional structures added in this work. Four standard ConfGen search strategies with progressive qualitative/resource consumption

trade-offs are employed with and without energy minimization. As well, an additional mode is examined in which structures from the fastest and slowest modes are combined to fill in gaps from each individual method to provide an approach with the best overall reproduction of bioactive conformations. Finally, comparisons with other programs will be made using previously published results from the Perola[22] and Chen[21] studies.

METHODS

The core technology for rapidly generating diverse conformers in ConfGen was originally developed for the docking program Glide[2,3] and has been modified for the task of reproducing bioactive conformations in a relatively small set of total conformations. ConfGen uses the infrastructure from the general molecular modeling program MacroModel,[8] which allows for access to multiple all-atom force fields, redundant conformer elimination, and multiple processor computing. Technology development is an ongoing process, and the description presented here is for ConfGen version 2.1211.

In this section, we first provide a detailed explanation of how ConfGen generates conformers. We then describe the ligand data sets used in this work. Finally, we discuss how the parameters for the ConfGen modes were selected. ConfGen processes a ligand in three main phases: (i) Variable feature identification; (ii) Conformation generation; and (iii) Conformer selection and refinement.

Variable Feature Identification. Variable feature identification focuses on identifying rotatable bonds, flexible ring systems, and invertible nitrogen atoms. A bond is generally considered rotatable if rotations about it, that are thermally accessible on an experimentally relevant time-scale, can lead to significant structural variation. ConfGen identifies a bond as potentially rotatable if it meets the following conditions: (i) It is a single bond; (ii) It does not lie within a ring; (iii) Neither of the atoms connected by the bond is terminal (i.e., has no other atoms bonded to it); (iv) Neither end of the bond is a CH_3, NH_2, or NH_3^+ group; and (v) Neither atom in the bond is bonded to two or three atoms that are all equivalent and are arranged with two- or three-fold rotational symmetry. An exception to this is made if there are two equivalent nitrogen atoms (e.g., amidinium groups).

While ConfGen can optionally consider N−C bonds from amides and O−C bonds from carboxylic acids or esters as nonrotatable (i.e., fixed in the input geometry), this study treats these bonds as rotatable.

ConfGen generates ring conformations using the same template-based facility available in LigPrep,[50] Glide,[2,3,51] MacroModel,[8] and Phase.[5,52] This facility is designed to produce a complete set of low-energy ring conformations with high-quality geometries and accurate relative energies. Individual rings are identified using the smallest set of smallest rings (SSSR) approach.[53] Ring fusion (rings sharing at least two atoms) complicates ring templating since it alters the relative energies of the conformations of the individual rings, sometimes to the extent that some individual ring conformations do not occur in the fused ring system. Ring systems consist of single or fused flexible rings and rigid rings that share at least two atoms with a flexible ring in the collection. When a ring system is identified in a molecule, a matching template for the ring system as a whole is sought from a pre-existing collection of templates. This stage of template matching is quite strict requiring that all corresponding atoms have the same element and the same hybridization, with the exception that templates involving neutral sp^3 nitrogen atoms are used to generate ring conformations for the corresponding ring systems with that atom protonated. In addition, for fused flexible rings, the atoms joining the rings can have distinct topographies and are often chiral. Since these distinct geometries correspond to ring systems with fundamentally different ring conformations and energies, different templates are needed for the different topographies of the atoms common to different individual rings. The increase in the number of templates required for this stereochemical criterion is reduced somewhat by automatically reflecting the template geometries to generate enantiomers, as needed to obtain a match. If an exact match with a template cannot be made, the matching criterion is relaxed to allow: (i) a rigid ring attached to a flexible ring to be matched by a double bond; and (ii) substitution by atoms with similar geometries, e.g., a sp^3 C in a template can stand in for a sp^3 O in the molecule.

Table 1. Ligand Data Set Properties[a]

	Boström	Perola	full	Chen
total ligands	36	100	667	256
average no. of atoms	43.0	50.4	46.0	46.8
average no. of rotatable bonds	5.7	8.1	6.8	6.7
average no. of aliphatic rings	0.72	0.69	0.67	0.67

[a] The full data set includes the ligands used in the Boström[23] and Perola[22] studies plus an additional 538 ligands from in-house prepared structures. The Chen[21] data set is from a recent study and has 130 ligands that are not in the full data set.

High-quality templating requires many ring templates. At present, there are 1252 templates in the collection. The conformations for each template were originally generated using a MacroModel conformational search.[8] By default, ring systems are searched using 1000 Monte Carlo torsional sampling search steps (MCMM) with molecular mechanics minimization using either the MMFFs[54] or the OPLS_2005 force field[55] and the GB/SA solvation model.[56] For these searches, 500 steps of truncated Newton conjugate gradient (TNCG) minimization were used for input and output structures. Non-mirror image conformers with rms values less than 0.25 Å were eliminated as well as those conformers with energies more than 11.94 kcal/mol (50 kJ/mol) higher than the lowest energy conformer. The results were inspected for completeness, and if necessary, the search is continued by seeding it with the current set of conformers and in some cases with additional manually generated conformers. A maximum of 50 conformations based on the lowest energy structures were stored for each template. The energy of each conformation relative to the energy of the lowest energy conformation was saved in the template.

Once a ring template is found, the relative energies of the ring conformations within the molecule, E_{rrel}, are calculated using the stored relative energy for that ring system with corrections for the axial vs equatorial positioning of attachments to the ring (see, for example, Table 1 conformational analysis by N. L. Eliel et al.).[57] ConfGen has adjustable controls to limit the number of ring system conformations

sampled, including an upper limit on E_{rrel}, a maximum number of the lowest energy ring conformations per ring system to use, and a maximum overall number of ring conformations for cases where multiple ring systems are present. Not all ring systems have templates, and if a template cannot be found for any ring system in a ligand, then no ring conformation sampling is performed, and the original geometries for the ring systems are retained.

The sp^3 nitrogen atoms that are bonded to three other atoms are often pyramidal, and many such nitrogen atoms can rapidly invert under thermal conditions in solution. If the nitrogen atom in one of these invertamers can be regarded as chiral when the lone pair is taken into account, the invertamers are distinct but should be sampled as a conformational variation. The ring templates automatically include conformations with such inversions for nitrogen atoms in rings. ConfGen also explicitly identifies such nitrogen atoms in linear portions of the molecule for sampling during conformation generation.

Conformation Generation. There are N_{ri} combinations of ring conformations and invertible nitrogen atoms:

$$N_{\mathrm{ri}} = 2^{N_{\mathrm{i}}} \prod_{r} N_{\mathrm{cr}}$$

where N_{i} is the number of invertible nitrogen atoms, r runs over all flexible ring systems, and N_{cr} is the number of conformers selected for use for a particular ring system. Each of these combinations are processed in three stages: (i) Generate tabulated potentials for all rotatable bonds; (ii) Identify minima in the potentials for each of the rotatable bonds; and (iii) Sample core conformations.

The potential for rotating about each rotatable bond are calculated using a truncated version of OPLS_2001[58,59] from a combination of dihedral angle potentials and Lennard-Jones non-bonded interactions for a subset of the atoms on either side of the bond that are deemed key to avoiding topologically local van der Waals clashes. These subsets include:

I. All atoms immediately attached to the atoms at each end of the bond.
II. For cases when either or both atoms in the bond are in rings:
    A. All atoms in the ring and atoms bonded to atoms in the ring. If the ring is aromatic, also include all atoms within two bonds of the ring.
    B. If both atoms are in aromatic rings, include all atoms in any aromatic rings ortho to the bond.
    C. If one atom is in an aromatic ring and the other atom is bonded to another aromatic ring and is one of: an oxygen, sulfur, sp^2 nitrogen, sp^2 carbon or two-coordinate phosphorus atom, include all of the atoms in the other aromatic ring as well as all atoms within two bonds of that aromatic ring.

The potentials are generated and stored on a grid of 1000 equally spaced angles spanning a complete rotation.

Topologically symmetric portions of a molecule can be distorted in well-minimized structures. While the distortions are small, they can have dramatic effects on the torsional potentials. An example of this is trifluoroacetate in which the F atoms adopt a slightly asymmetric conformation upon energy minimization and have distinctly nonequivalent minima for rotations around the carbon−carbon bond. ConfGen has a procedure to adjust the potentials to reproduce the correct periodicity. The ideal periodicity, r_s, of the tabulated potentials is recognized from the topological and geometric molecular symmetries. The code locates the lowest potential minimum within the range of angles $\pm \pi/r_s$ and averages the potential for $\pm \pi/r_s$ on either side of that minimum with its reflection about that minimum. This averaged potential is then repeated r_s times to generate a new, symmetric periodic potential for a complete rotation.

Weak dihedral angle potentials with a small number of potential minima can present challenges for generating bioactive conformations, since the binding site can force the molecule to adopt conformations in which this dihedral angle is far away from any of the minima inherent in the ligand's potential energy surface. ConfGen can optionally replace the tabulated potential with a cosine potential having a specified periodicity (typically 6) and an amplitude, if the extremes in the original force field based potential differ by less than that amplitude. The cosine potential is shifted so that one of its minima line up with the global minimum in the original tabulated potential.

Local potential minima are identified in the tabulated potentials for use in conformation generation. These minima are manipulated and filtered as follows: (i) When the atoms at both ends of a rotatable bond lie within an aromatic ring, ensure that all potential minima have a corresponding minimum for an angle of opposite sign; (ii) Remove minima that map onto equivalent minima for symmetric groups (e.g., minima separated by π radians for 1,4-substituted phenyl rings); and (iii) Optionally set esters, carboxylic acids, and amides to "trans" geometries by selecting the potential minimum closest to trans geometry and eliminate all other potential minima for these bonds.

For each rotatable bond, the energy of the lowest energy minimum is subtracted from all of the minima for that bond to convert them into relative energies.

The molecular core is defined as the portion of the molecule remaining when each terminus of the ligand is figuratively severed at the last rotatable bond. The core plus these terminal rotamer groups comprise the entire ligand. For each combination of ring system conformation and invertible nitrogen atom geometry, all combinations of minima for each rotatable bond in the core are examined with the peripheral groups in their lowest energy orientation. If the sum of the relative potential energies for all the rotatable bonds and the ring conformations exceeds a preset maximum, the conformation is eliminated from consideration.

A core conformation that passes the energetic filter is then checked for steric clashes. If there are one or more such clashes, a repulsive Gaussian function between the clashing atoms is defined and added to the torsional energy, and the resulting total energy function is minimized to relieve the steric clashes. The success of this procedure in producing an acceptable conformation without unphysical overlaps is checked in subsequent steps described below.

As noted above, ligands in protein−ligand complexes tend to adopt extended conformations. A heuristic scoring function provides an assessment of the degree of extendedness for each candidate conformation. This extension score (ES) consists of a sum over pairs of heavy atoms in the ligand, where the contribution for each atom pair is computed via a Gaussian function of the interatomic distance, with the parameters of the function dependent upon the number of

rotatable bonds separating the atoms and the atom types involved in the pair. Generally speaking, the Gaussian widths are narrow, so the main effect of this heuristic is to penalize close contacts between topologically distant atoms. The minimum number of rotatable bonds for any of these terms is two, so atoms separated by three or fewer bonds are not included in the sum (i.e., the prefactor multiplying the Gaussian is set to zero for these cases). As an example, pentane gives only one term between the first and last carbon atoms in the chain using this heuristic. The function is truncated (set to 0) beyond 2.9 Å and has a σ value of 1.29 Å and a height of 0.1 kcal/mol. Conformers are ordered by the ES, and the total number of top-ranked core conformations retained for a molecule with N_{rotc} core rotatable bonds, and N_{ri} conformational variations due to invertible nitrogens and ring conformations are limited to at most:

$$1000 \text{ if } N_{\mathrm{rotc}} \leq 5$$

$$64 \times (N_{\mathrm{rotc}} + N_{\mathrm{ri}} - 1) - 234 \text{ otherwise}$$

This relation was determined empirically to include enough core conformations to yield low rmsd values for ligand conformations as compared to the original PDB structures for protein−ligand complexes. As mentioned above, all core conformations are generated with the peripheral groups in their lowest energy orientations. Rotamers for the peripheral groups are sampled when the sampling of core conformations is complete for all combinations of ring conformations and invertible nitrogen atoms. In the rapid mode, peripheral groups are sampled one at a time, while maintaining all others in their lowest energy orientation. For clarity, in this mode, all conformers derived from a core conformation will differ from that core conformation in the orientation of only one peripheral group. All conformations generated by varying rotamers are also subjected to the relative energy cutoff filter and then minimized using the tabulated potentials and scored. In addition to this rapid mode of sampling, ConfGen also supports a thorough mode in which all combinations of potential minima for terminal group orientations are explored. In our trials, the thorough mode did not significantly affect the quality of the results, so we will not explore this further here.

In addition to the internal energy cutoff described above, conformations with close contacts, defined as atoms that lie within 60% of the sum of their van der Waals radii, are eliminated. This small distance is used because the structures have not been energy minimized and can be considered as representing structures that might relax into low-energy conformations by alleviating the van der Waals clashes. In unusual cases where all conformations would be eliminated, the overlap tolerance is increased by 15% (that is to 60% × 0.85 = 51% of the van der Waals radii). In very rare instances, that too eliminates all conformations, and then the tolerance is further increased by 15% (to 43%). Reducing tolerance below 60% is a pragmatic approach driven by the need to produce some conformations rather than none. We plan to investigate and, if possible, to eliminate these problematic cases, which are likely due to general molecular features in which torsional minima determined by small sets of atoms consistently generate topologically distant atom−atom clashes. This close contact filtering is applied to the cores before variations of the peripheral group orientations are sampled and to each new conformation generated by sampling a peripheral group. One limitation in this filtering scheme is that filtering the cores before sampling peripheral groups may potentially prevent the generation of conformations where sampling the peripheral group would eliminate atom−atom clashes found in the core conformation. This is something that we plan to investigate in future studies. All remaining conformers are saved in memory in order of increasing score (energy using the truncated potentials).

Conformer Selection and Refinement. There are optional additional levels of filtering that may be used to further eliminate undesirable conformations or to limit the number of conformations. Similar filtering approaches are used to eliminate conformations with high-energy electrostatic interactions or conformations with local concentrations of heavy atoms. In all cases, the overall penalty, P_k, for conformer k, is calculated from the individual penalty contributions, p_{ij}, based upon the appropriate atom−atom distances:

$$P_k = \sum_{i,j} p_{ij}$$

The smallest penalty for any conformer is subtracted from the penalties for each conformer to give P'_k. Conformers with values of P'_k larger than a specified cutoff value, $P'^{\max}_{\mathrm{filter\ type}}$, are eliminated. For filtering out conformers in which atoms with the same net formal charge approach each other closely, p_{ij} is given by

$$p_{ij} = A_q q_i q_j e^{-r_{ij}^2/2\sigma_q^2}$$

where A_q and σ_q are adjustable parameters corresponding to the amplitude and standard deviation for the Gaussian formal charge weighting function. The formal charges on atoms i and j are defined by q_i and q_j.

For eliminating conformations with polar hydrogens pointing toward functional groups with a net positive formal charge, p_{ij} is calculated as the difference of two Gaussian weighting functions, one for the polar hydrogen, H, and the other for the donor atom, D, as given by

$$p_{ij} = p_{i\mathrm{H}} - p_{i\mathrm{D}}$$

$$p_{i\mathrm{X}} = A_{\mathrm{hb}} q_i e^{-r_{i\mathrm{X}}^2/2\sigma_{\mathrm{hb}}^2}, \quad \mathrm{X} = \mathrm{H}, \mathrm{D}$$

where A_{hb} and σ_{hb} are adjustable parameters corresponding to the amplitude and the standard deviation for the Gaussian hydrogen bonding weighting function. D must correspond to an oxygen atom or a nitrogen atom, and atom i must carry a positive formal charge (q_i).

For eliminating conformations with high local concentrations of heavy atoms, such as occurs when rings within the ligand stack:

$$p_{ij} = A_{\mathrm{ha}} e^{-r_{ij}^2/2\sigma_{\mathrm{ha}}^2}$$

where A_{ha} and σ_{ha} are adjustable parameters corresponding to the amplitude and standard deviation for the Gaussian heavy atom weighting function. In this case, i and j run over all pairs of nonhydrogen atoms in the molecule.

A typical drug-like ligand will generate an intractable number of conformers for storage and future use. ConfGen selects a subset of N_{target} conformers, where $N_{target} = N_{mult}N_{rot}$. N_{mult} is user-specified, and N_{rot} is the total number of rotatable bonds in the molecule. If N_{target} is larger than the total number of conformers produced within ConfGen, then all of those conformers are used. Otherwise N_{target} conformers are selected uniformly from those stored within memory using the following procedure: (i) Keep the first conformation (the one with the lowest score); (ii) Keep the last conformation (the one with the highest score); (iii) Select conformations midway between previously selected conformations in the ordered list of conformations; and (iv) Repeat stage iii until N_{target} conformers have been selected.

The rationale for this protocol is that a distribution of conformers produced by a heuristic scoring approach roughly matches that found in crystal structures (see the Results Section for further information), and this method for generating a subset should maintain this behavior.

All-atom energy and energy minimization calculations can be carried out as part of the ConfGen search. These calculations use the OPLS_2005 force field,[55] with nonbonded electrostatic and van der Waals interactions truncated at 12 and 7 Å, respectively. A distance-dependent dielectric constant of 4 is used. The truncated Newton conjugate gradient (TNCG) minimizations proceed until the rms force drops below a specified threshold (0.012 kcal/mol Å) for a maximum of 500 steps.

Elimination of redundant conformers after processing is performed by comparing all pairs of conformers and eliminating one conformer (the one with the higher energy or score, if the energy has not been calculated) from those that are found to be too similar. If the smallest rmsd value calculated between all combinations of symmetrically equivalent heavy atoms is larger than a specified threshold, then the conformers are considered different, and both are retained. In addition, if the difference in dihedral angles involving polar hydrogen atoms is greater than a specified threshold, then the conformers are considered different, and both are retained. An additional energy filter is employed based upon the all-atom energy relative to that for the lowest energy conformer. If this relative energy for a particular conformer is larger than a specified value, it is eliminated.

Overall, while ConfGen was derived from and still uses the same basic conformer generation approach as used in Glide, many changes have been introduced in both programs at all levels. For instance, the recognition and enforcement of symmetry in torsional potentials, filtering out conformations (electrostatic or concentrated heavy atoms), the selection of a uniform subset of conformations, and the coupling with MacroModel's infrastructure for energy evaluation, minimization, and redundant conformer elimination are specific to ConfGen.

Small Molecule Data Set. Ligands were obtained by combining the Boström (36 molecules)[23] and Perola (100 molecules)[22] sets with additional PDB structures from our internally prepared structures. The Boström and Perola sets have seven structures in common (PDB codes 1fcz, 1fkg, 1frb, 1ppc, 1pph, 3std, and 5std). An additional 538 structures selected from the PDB[60−62] were added from in-house prepared structures, resulting in a total of 667 structures used in this study. This collection of ligands will be referred to as the full data set. The complete list of PDB codes used in this work is available in the Supporting Information. A representative example of the molecules used in this study is shown in Figure 1. Table 1 shows the average properties of the ligands from the full data set as well as for the Boström and Perola subsets. In addition, the Chen set of ligands consisting of 256 ligands (130 of which are also in the full data set) will be used in separate calculations to characterize how well ConfGen performs relative to competing programs.

Protocol for Parameter Tuning. In order to optimize the performance of ConfGen, multiple iterations of calculations were performed varying the search, energy, and filtering parameters. While there has been no explicit optimization with a defined objective function in this work, our goal has been to produce collective settings or modes that could be classified as very fast, fast, intermediate, and comprehensive. Roughly speaking, these modes are a pragmatic progression with subsequent levels of increasing accuracy and demands on computer resources (CPU time and disk space). The primary variables adjusted were the search steps per rotatable bond, total number of conformers, energy window, criteria for duplicate conformer elimination, ring sampling, and minimization steps. Since all combinations of all search parameters could not be varied simultaneously, we modified a subset of search parameters in a ± direction to see the direction resulting in the best performance. The best parameters from each perturbation established the starting point for a subsequent iteration of variable perturbations. In this work, we present the end results from this parameter tuning strategy. The results for each of these modes are presented below.

All modes employ redundant conformer elimination based upon rmsd thresholds between topologically equivalent mappings of atoms in the molecules. The very fast mode involves torsional searching with only five steps per rotatable bond and heuristic scoring without any all-atom minimization or energy filtering. The fast method is very similar to the very fast one but uses a smaller rmsd (1 vs 1.25 Å) threshold and includes all-atom energy filtering. The intermediate search mode increases the sampling to 75 steps per rotatable bond and increases the threshold for weak torsional potentials to 5.74 kcal/mol (vs 2.87 kcal/mol), to allow more conformations. Finally, the comprehensive mode uses an even greater energy window with a smaller rmsd threshold for duplicate pose elimination and additional ring sampling (more conformers per ring, more ring combinations, and higher energy limit for rings), resulting in the largest total number of conformations and the highest potential for retrieval of bioactive conformers. Table 2 contains a detailed compilation of the parameters for these four search modes.

Modeling and Calculations. Modeling and calculations were carried out using Maestro[63] and the accompanying software in Schrödinger Suite2009. Coordinate biases initially present for all ligands were removed before the conformational search study by converting the ligand conformation from the crystal structure of the complex to a SMILES string and then back to 3D coordinates using LigPrep.[50] Conformational searching was done with ConfGen.[49] The four standard ConfGen modes with or without all-atom energy minimization can be started from a simple GUI panel in Maestro. The conformational analysis was performed with in-house Schrödinger Python scripts, available upon request.

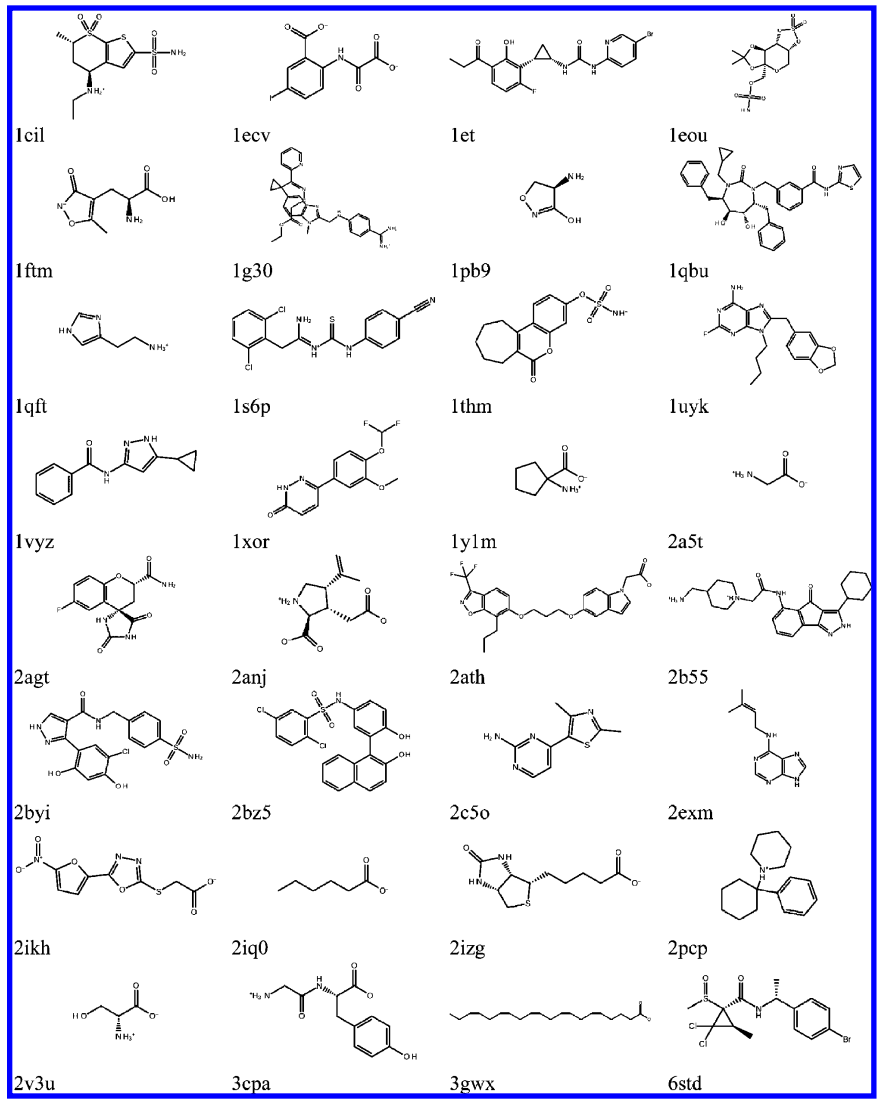

Figure 1. Thirty-two diverse, representative ligand structures from the full data set.

Table 2. Optimized Search Parameters for the Four Confgen Modes: Very Fast, Fast, Intermediate, and Comprehensive

parameter	very fast	fast	intermediate	comprehensive
maximum number of search steps	1000	1000	1000	1000
search steps per rotatable bond	5	5	75	75
minimum heavy atom rmsd (Å) for distinct conformers	1.25	1	1	0.5
minimum dihedral angle difference for polar hydrogens (°)	60	60	60	60
maximum relative energy for flexible rings (kcal/mol)	2.39	2.39	2.39	23.9
maximum number of ring conformations per ligand	16	16	16	128
maximum number of ring conformations for a ring	8	8	8	64
maximum relative ConfGen energy (kcal/mol)	25	25	25	119.5
energy threshold for periodic torsions (kcal/mol)	2.87	2.87	5.74	5.74
σ_q (Å)	5	5	5	5
A_q	2.5	2.5	2.5	2.5
$P'^{\max}_q$	0.75	0.75	0.75	0.75
σ_{hb} (Å)	5	5	5	5
A_{hb}	1	1	1.5	1.5
$P'^{\max}_{hb}$	0.25	0.25	0.25	0.25
σ_{hb} (Å)	2.5	2.5	2.5	2.5
A_{ha}	0.1	0.1	0.1	0.1
$P'^{\max}_{ha}$	1	1	1	1
restraint potentials for weak torsions in MacroModel (kcal/mol)	N/A	239	239	239
restraint potential half width (°)	N/A	10	10	10
suppress hydrogen-bond electrostatics in MacroModel	N/A	Yes	Yes	Yes
maximum relative energy all-atom energy in MacroModel (kcal/mol)	N/A	25	25	119.5

The degree of extension/folding was computed, as described by Perola and Charifson,[22] using a Python script (downloadable from the Schrödinger Script Center) as the ratio between the maximum pairwise atom distance for each

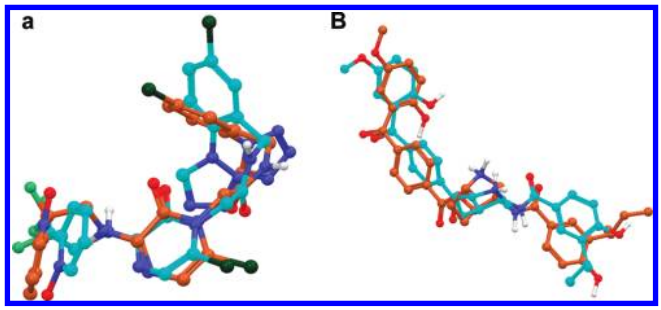

Figure 2. Examples of ConfGen-generated ligand conformations (orange) with rmsd values greater than 2.0 Å to the crystal structure (cyan), yet resemble the crystal conformation for (a) 1sl3 (rmsd: 2.01 Å) and (b) 1erk (rmsd: 2.22 Å).

conformer over the maximum pairwise atom distance across the entire set of conformers for a given molecule.

All calculations were run using a single core on Intel(R) Core(TM)2 Quad CPU Q6600 at 2.40 GHz with 1.3 GB memory under the CentOS 5.3 operating system. Each calculation mode was run for all ligands in the full set or for the Chen data set in a single run. CPU times per molecule are calculated by dividing the total compute time for this collective run by the number of molecules and averaging this quantity over three separate runs.

RESULTS AND DISCUSSION

Studies were conducted on either the full set of 667 ligands (the small molecule data set described earlier) or the Chen[21] data set as indicated. It is rare for a conformation generation program to exactly reproduce the bioactive conformation. As a result, the performance of a conformation generation program is usually expressed statistically, typically as the percent of ligands for which at least one conformer with a heavy atom rmsd (relative to that from the protein−ligand complex) less than a specified value. We will refer to this as percent reproduced for a given threshold rmsd value. While standards vary, as in some other studies,[20,21,25] we regard finding at least one conformation for a given ligand with a rmsd less than 2.0 Å from the ligand conformation in the crystal structure as a success. Although this limit may be too generous for smaller ligands, an rmsd cutoff of 2.0 Å typically results in conformations that look reasonably close to the bioactive conformation of interest (for example, see Figure 2) and provides adequate resolution for downstream methods, like pharmacophore[5] or shape-based searching. Results for lower rmsd cutoff values are reported to provide insights into the accuracies of the methods when higher precision is needed.

The percent recovered for the entire ligand data set as a function of rmsd for each of the four ConfGen search modes is plotted in Figure 3. Table 3 also lists these results for selected rmsd values along with two performance measures: (i) the average number of conformations per ligand and (ii) the average CPU time per ligand. Over the full data set, the fastest search mode ("very fast") yielded 96, 84, 52, and 16% with rmsd values less than 2.0, 1.5, 1.0, and 0.5 Å, respectively. The most computationally expensive mode ("comprehensive") without pre- and postminimization of conformers resulted in 99, 90, 71, and 35% at these resolutions. The average number of conformers generated for the very fast and comprehensive methods were 14.3 and

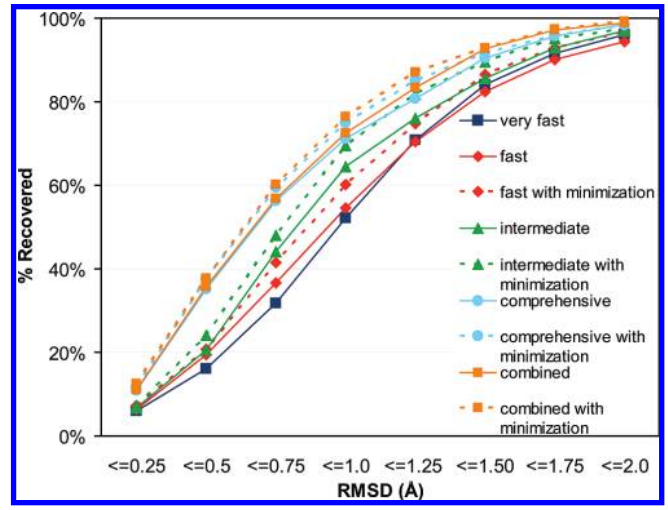

Figure 3. Percent recovered for all ConfGen modes without and with all-atom energy minimization. Percent recovered refers to the percent of ligands for which ConfGen produces at least one conformation with an rmsd value less than or equal to the rmsd value on the horizontal axis.

146.4, respectively. In general, the percent recovered, the number of conformers generated, and the CPU times increase with mode in the order of very fast, fast, intermediate, and comprehensive. One exception to this rule is that the number of conformers and percent recovered at 2.0 Å decrease somewhat in fast as compared to very fast. Filtering by relative energy in the fast mode reduces the number of conformers; something not performed in the very fast mode. Such filtering can lead to worse results when assessed by the ability to generate a bioactive conformer within the set of all conformations, but it should also produce a more energetically reasonable set of conformations, which may be desirable for some applications. At 0.5 s per ligand, the very fast mode is fast enough to support conformational searching of large databases of ligands, particularly if the jobs are distributed over many processors. Its high percent recovered value (96%) at 2.0 Å with just 14.3 conformers may seem surprising but is consistent with a recent study.[64] Depending on the project and the computational resources available, comprehensive searches, at 8 s per ligand, may also be fast enough for such applications in practice.

Table 4 gives the ConfGen results for the Boström and Perola ligand subsets. The trends as a function of search mode and rmsd threshold value are the same as those of the full set of ligands. From Tables 3 and 4, it is apparent that the percent recovered decreases in the progression from the Boström set, through the full set to the Perola set. This trend can be rationalized by the increasing size of the conformational space for these data sets in the same order as measured by the average number of rotatable bonds for these data sets (see Table 1) and evidenced by the increasing number of conformations found for each of them.

ConfGen produces a profile of conformational extendedness[22] similar to that observed in crystal structures, as shown in Figure 4 for each of the four modes. The fast method most closely reproduces the profile from the crystal structures. The intermediate and comprehensive modes generate significantly more conformations, and, given that a limited number of conformations can be fully extended, it is expected that the relative amount of highly extended structures with these methods is decreased. The fast method eliminates

Table 3. Results for Each ConfGen Mode for the Full Data Set of Ligands

ConfGen mode	bioactives recovered (%) ≤0.5 Å	≤1.0 Å	≤1.5 Å	≤2.0 Å	average no. of conformers per ligand	average time per ligand (s)
very fast	16	52	84	96	14.3	0.49
fast	20	55	82	94	13.2	1.09
intermediate	21	65	85	97	37.9	3.33
comprehensive	35	71	90	99	146.4	8.00

Table 4. Results for Each ConfGen Mode for the Boström and Perola Ligand Sets[a]

ConfGen mode	Bioactives recovered (%) ≤0.5 Å	≤1.0 Å	≤1.5 Å	≤2.0 Å	Average # of conformers per ligand
		Boström			
very fast	22	56	92	100	10.2
fast	14	61	94	97	9.7
intermediate	17	69	92	97	18.8
comprehensive	42	72	94	100	97.2
		Perola			
very fast	6	46	86	100	19.3
fast	10	46	82	95	15.0
intermediate	8	58	81	97	40.0
comprehensive	17	65	88	98	162.2

[a] See refs 22 and 23 for the Perola and Boström ligand sets, respectively.

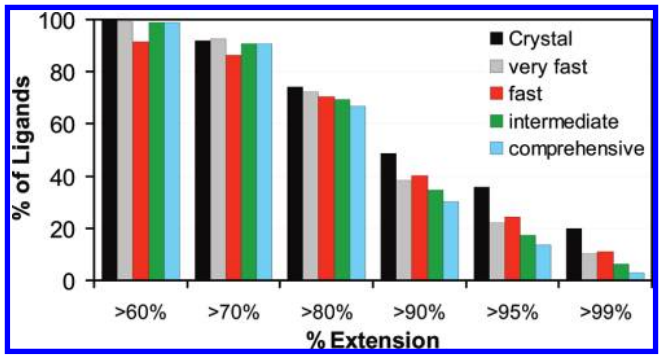

Figure 4. Degree of extendedness[22] for the crystal structure conformations and the four ConfGen modes (without minimization). These distributions for the ConfGen modes resemble that of the crystal structures except for extensions greater than 90%, indicating that the collections of ConfGen conformers include more compact structures. The fast mode is closest to reproducing the most extended portion of the crystal structure distribution.

energetically unfavorable conformations that are retained by the very fast method. While conformational extendedness is not explicitly taken into account in energy filtering, compact structures are more likely to have atom clashes and higher net torsional energies and are, therefore, preferentially eliminated. As a result, the profile with energy filtering shifts to more extended conformations and is closer to the experimentally observed distribution. The overall similarity of the distributions from the crystal structure and those generated by ConfGen supports the uniform selection of conformers from the full set of conformers.

For some ligands ConfGen had trouble finding conformers within 2.0 Å of the bioactive one. An examination of these ligands revealed two general classes of problems. The first class is due to specific shortcomings in the search methods, such as a missing ring template, too strongly favoring trans conformations in linear segments of the ligands (e.g., C−C−O−C), a difficulty in finding appropriate orientations of aniline N atoms that have large substituents (see Figure S1 in the Supporting Information), and a need for additional parameters for axial vs equatorial energies for substituents on flexible rings. The second class of ligands is made up of very flexible molecules, typically those with more than 15 rotatable bonds, suggesting that efficient sampling of the conformational space is a very difficult task.

Table 5. Results for the Combined Search for the Full Data Set of Ligands and the Boström and Perola Subsets[a]

data set	bioactives recovered (%) ≤0.5 Å	≤1.0 Å	≤1.5 Å	≤2.0 Å	average no. of conformers per ligand
full set	36	73	93	99	160.8
Boström	42	75	97	100	107.4
Perola	17	70	93	100	181.4

[a] See refs 22 and 23 for the Perola and Boström ligand sets, respectively.

Since the search modes select conformers differently, we explored combining the conformers generated from different modes to see if these conformations would complement each other and thus give even better results. In general, this did improve the percent recovered, and the combination of the most dissimilar search modes, very fast and comprehensive, seemed to offer significant gains in accuracy without significantly increasing the computer resources needed over just the comprehensive method. The results for this combination are given in Table 5. There was a general improvement in the results over the comprehensive results for the full set and the Boström and Perola sets particularly at the 1.0 and 1.5 Å rmsd level. The results did not improve much at the 2.0 Å level because the comprehensive results were already quite good. At the 0.5 Å level, too few conformers are contributed by the very fast mode to significantly improve the comprehensive results.

For the fast, intermediate, and comprehensive methods, we also performed a full minimization (using OPLS_2005 as described in Methods Sections) of each molecule before the search, after the search, or both. The very fast method gains much of its speed over the fast mode by not assigning potential parameters, and thus it does not make sense to use it in combination with minimizations. Minimizing the input structures had no detectable effect on the quality of the search methods and added minimal CPU time. This is most likely due to the fact that the input structures were already prepared with LigPrep and, therefore, started with reasonable internal geometry. The results of the searches with all-atom minimization of all input structures and conformers generated are in Table 6. Minimization slowed the search by factors of 3.8, 12.9, and 5.2 for the fast, intermediate, and comprehensive methods, respectively. Minimization in-

Table 6. Results for Search Modes on Full Data Set with an All-Atom Energy Minimization of All Conformers for the Full Set of Ligands[a]

ConfGen mode	bioactives recovered (%) ≤0.5 Å	≤1.0 Å	≤1.5 Å	≤2.0 Å	average no. of Conformers per ligand	average time per ligand (s)
fast	21	60	87	97	16.3	4.2
intermediate	24	70	90	98	43.0	34.3
comprehensive	38	75	90	98	111.9	41.6
combined	38	76	92	99	128.2	45.8

[a] The average time for the combined mode is the sum of the times for the fast and comprehensive modes.

Table 7. Search Results for the Public Subset of Ligands (100 Ligands) from the Perola Study for Catalyst, ICM, Omega, and ConfGen[a]

Program/mode	bioactives recovered (%) ≤0.5 Å	≤1.0 Å	≤1.5 Å	≤2.0 Å	average no. of conformers per ligand	average time per ligand (s)
Catalyst/FAST	23	65	93	97	190	4.9
ICM	18	61	85	97	141	151.8
Omega	24	67	87	94	179	1.7
ConfGen/						
very fast	6	46	86	100	19	0.71
comprehensive	17	65	88	98	162	9.1
comprehensive + minimization	20	68	92	98	128	51.0
combined	17	70	93	100	181	9.8
combined + minimization	21	72	94	99	150	56.4

[a] The study that produced the results for Catalyst, ICM and Omega is described in ref 22, which reported results for 150 ligands including 50 proprietary ones. The values for just the 100 public structures were provided by the authors of that study in a private communication. The average time per ligand for Catalyst, ICM, and Omega are only roughly comparable to those from ConfGen as they are for the entire 150 ligand data set and have been scaled for the same processor (CPU times scaled down by factors of 0.0613, 0.857, and 0.857, respectively) used in the current study based upon SPEC CFP2000[65] and SPECfp95[66] performance figures.

creases the number of conformers generated for the fast and intermediate methods because more structures have relative energies small enough to pass the energy filter. While this same effect is present in the comprehensive mode, there is a net reduction in the number of conformations produced when minimizing because this mode generates enough conformers to effectively saturate the local minima, resulting in a significant number of redundant structures upon minimization. Overall, minimizations improve the percent recovered, particularly at the 1.0 Å rmsd level with the exception on the comprehensive mode, where this value decreases slightly at 2.0 Å rmsd likely due to the smaller number of conformations. The "combined with minimization" mode uses the conformers generated by the "fast with minimization" and "comprehensive with minimization" modes. The results for this mode are the best overall produced by ConfGen for the full set.

Two recent studies afford us the opportunity to compare ConfGen's performance with that of other commonly used programs. The Perola[22] study included results for Catalyst[47] (v4.6) FAST, ICM v3.0,[38] and Omega v1.2[39] run in a manner consistent with rapidly processing a large number of structures (using largely default parameters) for 150 structures of which 100 are in the public domain. Privately, the authors of that study provided the corresponding results for the public structures. These results along with those for ConfGen's very fast, comprehensive, comprehensive with minimization, combined, and combined with minimization modes are presented in Table 7. The second study, from Chen,[21] involved conformational searches of 256 ligands that are in the public domain (characterized in Table 1) using MOE v2006.08[48] IMPORT and Stochastic modes as well as Catalyst's v4.10[47] BEST and FAST modes. While the full set and the Chen set overlap, 130 of the ligands from the latter are not in the full test set. While a number of parameter variations were explored in the Chen study, we will focus on the results using the default parameters for these programs. These results along with those for ConfGen's very fast, comprehensive, comprehensive with minimization, combined, and combined with minimization modes are presented in Table 8. We have selected these ConfGen modes rather than all of the standard modes in order to focus the comparisons on two very different yet common tasks; rapid searches of many ligands (very fast) and thorough searches (comprehensive, comprehensive with minimization, combined, and combined with minization) of relatively few ligands.

At the highest resolution (0.5 Å rmsd), Catalyst and Omega performed slightly better than that of ConfGen for the Perola set. However, ConfGen's thorough searches performed as well as or better than MOE and Catalyst for the Chen set at this resolution, while the very fast mode performed poorer. At the other resolutions, ConfGen's comprehensive mode performed at a comparable level as the competing programs, while ConfGen's combined mode produced either the best or second best results as compared to the competing programs in all cases. Minimization improved ConfGen's comprehensive and combined mode results. The combined + minimization mode matched or outperformed the competing programs at resolutions lower than 0.5 Å rmsd except at 1.0 and 1.5 Å for the Chen set, where the Catalyst BEST mode was slightly better.

Except for the very fast mode, ConfGen modes generated roughly the same number of conformers as those of the other programs and ran faster than ICM, Catalyst/BEST, and MOE/Stochastic but slower than MOE/Import, Catalyst FAST, and Omega. ConfGen's very fast mode performed poorer than

Table 8. Search Results for the Chen Data Set of Ligands (256 ligands) for MOE, Catalyst, and ConfGen[a]

program/mode	bioactives recovered (%) <0.5 Å	<1.0 Å	<1.5 Å	<2.0 Å	average no. of conformers per ligand	average time per ligand(s)
MOE/Import	32	75	94	98	110	4.5
MOE/Stochastic	34	71	84	92	67	64
Catalyst/FAST	24	70	95	100	106	3
Catalyst/BEST	29	81	98	100	111	94
ConfGen/						
very fast	13	53	89	99	14	0.45
comprehensive	34	74	93	99	127	6.4
comprehensive + minimization	34	78	95	99	97	36.6
combined	34	77	95	100	141	6.9
combined + minimization	35	80	97	100	113	43.2

[a] Values for MOE and Catalyst are as reported for the default settings in ref 21. CPU times for MOE and Catalyst are scaled by 0.76 to approximately reflect the speed differences between the CPU used in the Chen study and the processor used for the current study based upon the SPEC CFP2006[67] benchmarks.

the other programs for resolutions of 0.5 and 1.0 Å. For resolutions of 1.5 (86% Perola, 89% Chen) and 2.0 Å (100% Perola, 99% Chen), it performed comparably to the other programs. This is significant given that this mode runs at a rate of 0.45 s/ligand and 0.71 s/ligand for the Perola and Chen sets, respectively. We estimate this to be more than 2.4, 6, 10, 140, 200, and 200 times faster than Omega, Catalyst/FAST, MOE/Import, MOE/Stocastic, Catalyst/BEST, and ICM, respectively. In addition, it produces on average only 19 and 14 conformations for the Perola and Chen sets, respectively, which is roughly 7 and 4.8 times lower than the values for any of the other programs. Thus ConfGen's very fast mode, due to its relatively small computational demands, is very efficient at reproducing bioactive conformations at resolutions of 1.5 and 2.0 Å.

CONCLUSIONS

In this work, we presented the methodology and validation of a conformational search method called ConfGen designed to efficiently generate bioactive conformations of drug-like molecules. This application has four standard modes and a new command-line-based combined mode that can be used with or without all-atom energy minimization. These modes span a wide range of accuracy/resource consumption tradeoffs. Calculations across a large number of diverse molecules from the Protein Data Bank (PDB) showed that ConfGen is able to accurately reproduce bioactive conformations with all four modes. Combining the fastest and most comprehensive modes improved the overall accuracy, while minimally increasing the computational overhead relative to the comprehensive mode. Similarly, minimization improves the results for all modes, including the combined mode, at high resolutions but requires significantly more CPU time.

At resolutions of 1.5 Å or better, appropriate for applications where accurate reproduction of bioactive conformations is desirable, ConfGen's comprehensive and combined modes, particularly with all-atom energy minimization, perform at least as well as competing programs in terms of accuracy and speed. At a resolution of 2.0 Å, the fastest mode produced results as good as or better than competing programs in terms of accuracy, and it was more efficient since it ran significantly faster and produced considerably fewer conformers. The latter is noteworthy, since downstream applications will run correspondingly faster with fewer input conformations, and the results will likely be better because fewer low-quality conformations are produced. Thus, the very fast mode is appropriate for performing conformational searches on large collections of ligands for applications where generating conformations within 2.0 Å of the bioactive conformation is acceptable.

Improvements will be sought to address specific problems in ligands that did not perform well, such as missing ring templates, occasional poor dihedral angle preferences when heteroatoms are present in linear segments of the molecule, and more specific parameters for axial vs equatorial energies of side chains attached to flexible rings. A challenge of growing importance in drug discovery is the treatment of large ring systems. Specifically, macrocycles derived from natural products are becoming more prevalent as early leads in drug discovery projects. In the present study, we did not include any macrocycles in the validation set. ConfGen currently does not search the torsions within macrocyclic rings and, therefore, would likely fail to reproduce bioactive conformations for these molecules. In future work, we plan to improve performance for macrocycles by including more templates for them and perhaps by permitting further relaxations of the strict matching criteria for ring systems.

Another challenge in conformational searching is the dependence of the results on input geometry. While the aggregate summary of results will remain largely unaffected by variations in the input geometry, it is likely that individual cases will produce slightly different results. This sensitivity to the input geometry can be easily resolved through a canonicalization of the input structures, for example using unique SMILES, to produce the exact same structure for the ConfGen algorithm independent of the initial input structure. However, this is avoiding rather than confronting the actual problem, which is complicated due to the large conformational space available on a relatively rough energy landscape. A better solution is to continue improving the program to ensure that, independent of the input geometry, a consistent and representative ensemble of conformations is generated.

ACKNOWLEDGMENT

We are grateful to Emanuele Perola for providing the results for the 100 public structures from the Perola study.[22] We thank the reviewers for asking questions that suggest our approach for eliminating conformations with atom−atom clashes can be improved.

Supporting Information Available: The complete list of PDB codes used in this work is available. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Rarey, M.; Kramer, B.; Lengauer, T. Time-efficient docking of flexible ligands into active sites of proteins. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1995**, *3*, 300–8.

(2) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47* (7), 1739–49.

(3) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **2004**, *47* (7), 1750–9.

(4) DesJarlais, R. L.; Sheridan, R. P.; Seibel, G. L.; Dixon, J. S.; Kuntz, I. D.; Venkataraghavan, R. Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. *J. Med. Chem.* **1988**, *31* (4), 722–9.

(5) Dixon, S.; Smondyrev, A.; Knoll, E.; Rao, S.; Shaw, D.; Friesner, R. 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* (10), 647–671.

(6) Rush, T. S.; Grant, J. A.; Mosyak, L.; Nicholls, A. A Shape-Based 3-D Scaffold Hopping Method and Its Application to a Bacterial Protein-Protein Interaction. *J. Med. Chem.* **2005**, *48* (5), 1489–1495.

(7) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A fast method of molecular shape comparison: A simple application of a Gaussian description of molecular shape. *J. Comput. Chem.* **1996**, *17* (14), 1653–1666.

(8) *MacroModel*, v9.6; Schrödinger, Inc.: New York, NY, 2008.

(9) Mohamadi, F.; Richards, N. G. J.; Guida, W. C.; Liskamp, R.; Lipton, M.; Caufield, C.; Chang, G.; Hendickson, T.; Still, W. C. MacroModel - An integrated software system for modeling organic and bioorganic molecules using molecular mechanics. *J. Comput. Chem.* **1990**, *11* (4), 440–467.

(10) Chang, G.; Guida, W. C.; Still, W. C. An Internal Coordinate Monte Carlo Method for Searching Conformation Space. *J. Am. Chem. Soc.* **1989**, *111*, 4379–4386.

(11) Ferguson, D. M.; Raber, D. J. A new approach to probing conformational space with molecular mechanics: random incremental pulse search. *J. Am. Chem. Soc.* **1989**, *111*, 4371–4378.

(12) Crippen, G. M. Rapid calculation of coordinates from distance matricies. *J. Comput. Phys.* **1978**, *26*, 449–452.

(13) Izrailev, S.; Zhu, F.; Agrafiotis, D. K. A distance geometry heuristic for expanding the range of geometries sampled during conformational search. *J. Comput. Chem.* **2006**, *27* (16), 1962–9.

(14) Rusinko, A. I.; Sheridan, R. P.; Nilakantan, R.; Haraki, K. S.; Bauman, N. Using CONCORD to construct a large database of three-dimensional coordinates from connection tables. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 251–255.

(15) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Comput. methodol.* **1990**, *3*, 537–547.

(16) Kolossváry, I.; Guida, W. C. Low mode search. An efficient, automated computational method for conformational analysis: application to cyclic and acyclic alkanes and cyclic peptides. *J. Am. Chem. Soc.* **1996**, *118*, 5011–5019.

(17) Kolossváry, I.; Guida, W. C. Low-mode conformational search elucidated: Application to $C_{39}H_{80}$ and flexible docking of 9-deaza-guanine inhibitors into PNP. *J. Comput. Chem.* **1999**, *20* (15), 1671–1684.

(18) Agrafiotis, D. K.; Gibbs, A. C.; Zhu, F.; Izrailev, S.; Martin, E. Conformational sampling of bioactive molecules: a comparative study. *J. Chem. Inf. Model.* **2007**, *47* (3), 1067–86.

(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* (12), 1137–1152.

(20) Griewel, A.; Kayser, O.; Schlosser, J.; Rarey, M. Conformational sampling for large-scale virtual screening: accuracy versus ensemble size. *J. Chem. Inf. Model.* **2009**, *49* (10), 2303–11.

(21) Chen, I. J.; Foloppe, N. Conformational sampling of druglike molecules with MOE and catalyst: implications for pharmacophore modeling and virtual screening. *J. Chem. Inf. Model.* **2008**, *48* (9), 1773–91.

(22) 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* (10), 2499–2510.

(23) Boström, J.; Greenwood, J. R.; Gottfries, J. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J. Mol. Graph. Model.* **2003**, *21*, 449–462.

(24) Liu, X.; Bai, F.; Ouyang, S.; Wang, X.; Li, H.; Jiang, H. Cyndi: a multi-objective evolution algorithm based method for bioactive molecular conformational generation. *BMC Bioinformatics* **2009**, *10*, 101.

(25) 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* (2), 422–30.

(26) 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* (4), 1848–61.

(27) Li, J.; Ehlers, T.; Sutter, J.; Varma-O'brien, S.; Kirchmair, J. CAESAR: a new conformer generation algorithm based on recursive buildup and local rotational symmetry consideration. *J. Chem. Inf. Model.* **2007**, *47* (5), 1923–32.

(28) Keith, T.; Butler, F. J. L. Xavier Barril,, Toward accurate relative energy predictions of the bioactive conformation of drugs. *J. Comput. Chem.* **2009**, *30* (4), 601–610.

(29) Bostrom, J.; Hogner, A.; Schmitt, S. Do Structurally Similar Ligands Bind in a Similar Fashion. *J. Med. Chem.* **2006**, *49* (23), 6716–6725.

(30) Lewis, P. J.; de Jonge, M.; Daeyaert, F.; Koymans, L.; Vinkers, M.; Heeres, J.; Janssen, P. A. J.; Arnold, E.; Das, K.; Clark, A. D.; Hughes, S. H.; Boyer, P. L.; de Béthune, M.-P.; Pauwels, R.; Andries, K.; Kukla, M.; Ludovici, D.; De Corte, B.; Kavash, R.; Ho, C. On the detection of multiple-binding modes of ligands to proteins, from biological, structural, and modeling data. *J. Comput.-Aided Mol. Des.* **2003**, *17* (2), 129–134.

(31) Wu, N.; Pai, E. F. Crystal Structures of Inhibitor Complexes Reveal an Alternate Binding Mode in Orotidine-5′-monophosphate Decarboxylase. *J. Biol. Chem.* **2002**, *277* (31), 28080–28087.

(32) Erika De Moliner, N. R. B.; Louise, N. Johnson,. Alternative binding modes of an inhibitor to two different kinases. *Eur. J. Biochem.* **2003**, *270* (15), 3174–3181.

(33) Gunther, S.; Senger, C.; Michalsky, E.; Goede, A.; Preissner, R. Representation of target-bound drugs by computed conformers: implications for conformational libraries. *BMC Bioinformatics* **2006**, *7*, 293.

(34) Dixon, S. L., Private communication; 2010.

(35) *Omega*, v1.0; Openeye Scientific Software: Santa Fe, NM, 2002.

(36) *Catalyst*, 4.6; Accelrys: San Diego, CA, 2003.

(37) Abagyan, R.; Totrov, M. Biased Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.* **1994**, *235*, 983–1002.

(38) *ICM*, v3.0; Molsoft L.L.C.: La Jolla, CA, 2004.

(39) *Omega*, v1.2; Openeye Scientific Software: Santa Fe, NM, 2004.

(40) *Omega*, v2.0; Openeye Scientific Software: Santa Fe, NM, 2008.

(41) *Catalyst*, 4.11; Accelrys: San Diego, CA, 2005.

(42) Blaney, J.; Crippen, G. M.; Dearing, A.; Dixon, J. S. *DGEOM*, program no. 590; Quantum Chemistry Program Exchange: Indiana University, Bloomington, IN, 1995.

(43) McMartin, C.; Bohacek, R. S. QXP: powerful, rapid computer algorithms for structure-based drug design. *J. Comput.-Aided. Mol. Des.* **1997**, *11* (4), 333–44.

(44) *ROTATE*, v1.15; Molecular Networks GmbH Computerchemie: Erlangen, Germany, 2006.

(45) Loferer, M.; Kolossváry, I.; Aszódi, A. Analyzing the performance of conformational search programs on compound databases. *J. Mol. Graph. Model.* **2007**, *25* (5), 700–710.

(46) *Omega*, v1.8; Openeye Scientific Software: Santa Fe, NM, 2007.

(47) *Catalyst*, Accelrys: San Diego, CA, 2004.

(48) *MOE*, v2006.08; Chemical Computing Group: Montreal, Canada, 2006.

(49) *ConfGen*, v2.0; Schrödinger, Inc.: New York, 2008.

(50) *LigPrep*, v2.1; Schrödinger, Inc.: New York, 2008.

(51) *Glide*, v5.5; Schrödinger, Inc.: New York, 2008.

(52) *Phase*, v3.0; Schrödinger, Inc.: New York, 2008.

(53) Zamora, A. An Algorithm for Finding the Smallest Set of Smallest Rings. *J. Chem. Inf. Comput. Sci.* **1976**, *16* (1), 40–43.

(54) Halgren, T. A. MMFF VI. MMFF94s Option for Energy Minimization Studies. *J. Comput. Chem.* **1999**, *20*, 720–729.

(55) Banks, J. L.; Beard, H. S.; Cao, Y.; Cho, A. E.; Damm, W.; Farid, R.; Felts, A. K.; Halgren, T. A.; Mainz, D. T.; Maple, J. R.; Murphy, R.; Philipp, D. M.; Repasky, M. P.; Zhang, L. Y.; Berne, B. J.; Friesner, R. A.; Gallicchio, E.; Levy, R. M. Integrated Modeling Program, Applied Chemical Theory (IMPACT). *J. Comput. Chem.* **2005**, *26* (16), 1752–80.

(56) Still, W. C.; Tempczyk, A.; Hawlely, R. C.; Hendrickson, T. A General Treatment of Solvation for Molecular Mechanics. *J. Am. Chem. Soc.* **1990**, *112*, 6127.

(57) Eliel, E.; Allinger, N.; Angyal, S.; Morrison, G. *Conformational Analysis*; Wiley: New York, 1965; p 200.

(58) Jorgensen, W. L.; Tirado-Rives, J. T. The OPLS Potential Functions for Proteins. Energy Minimization for Crystals of Cyclic Peptides and Crambin. *J. Am. Chem. Soc.* **1988**, *110*, 165.

(59) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11235.

(60) The RCSB Protein Data Bank.

(61) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242.

(62) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. F., 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* (3), 535–42.

(63) *Maestro*, v8.5; Schrödinger, Inc.: New York, 2008.

(64) Borodina, Y. V.; Bolton, E.; Fontaine, F.; Bryant, S. H. Assessment of conformational ensemble sizes necessary for specific resolutions of coverage of conformational space. *J. Chem. Inf. Model.* **2007**, *47* (4), 1428–37.

(65) SPEC CFP2000; Standard Performance Evaluation Corporation: Warrenton, VA; http://www.spec.org/cpu2000/results/cfp2000.html. Accessed January 6, 2010.

(66) SPECfp95; Standard Performance Evaluation Corporation: Warrenton, VA; http://www.spec.org/cpu95/results/cfp95.html. Accessed January 6, 2010.

(67) SPEC CFP2006; Standard Performance Evaluation Corporation: Warrenton, VA; http://www.spec.org/cpu2006/results/cfp2006.html. Accessed January 6, 2010.

CI100015J