ConfGen: Accurate and Rapid Conformation Generation

Performance in Reproducing Bioactive Ligand Conformations

Conformation generation, a procedure for producing diverse, low-energy 3D structures of compounds, is an essential element of in silico drug design, heavily employed in ligand-based virtual screening. To improve the recovery of bioactive conformations among generated conformers and drastically speed conformer generation, Schrödinger has completely rewritten the ConfGen application. As shown in Table 1, the new algorithm is approximately one order of magnitude faster than the previous version of ConfGen (henceforth referred to as ConfGen Classic), with bioactive recovery equivalent to the most computationally intensive mode of ConfGen previously available. This case study utilized 1,904 ligands, extracted from high quality protein-ligand crystal structures. This dataset includes a broad diversity of functional groups present in drug-like molecules and a range of molecule sizes and flexibility, as shown in Figure 1.

Without minimization of output conformers, ConfGen finds the bioactive conformation with an RMSD of < 1.5 Å in 89% of cases compared to 83% and 87% with ConfGen Classic Intermediate and Comprehensive modes, respectively. The new algorithm processes ligands 25 and 57 times faster than ConfGen Classic Intermediate and Comprehensive modes. When minimization of conformers is performed, ConfGen recovers more bioactive conformations with < 1.0 Å RMSD than without minimization. Because minimization acts as the speed bottleneck, the new algorithm is 5 times faster than both ConfGen Classic Intermediate and Comprehensive modes. It should be emphasized that ConfGen is able to achieve bioactive recovery similar to ConfGen Classic Comprehensive in far fewer conformers.

Table 1. The performance of ConfGen is assessed using 1,904 ligands extracted from high quality protein-ligand crystal complexes. ConfGen Classic/Intermediate and ConfGen Classic/Comprehensive are methods from the older ConfGen application. The RMSD columns show the percent of compounds for which a conformer within a given RMSD against a bio-active configuration was obtained. Calculations were performed using a single thread on a 2.8 GHz Intel Xeon CPU and the Speed includes disk I/O. Not counting disk I/O and without minimization ConfGen now processes about 30 ligands per second.

<table>
<thead>
<tr>
<th>Program</th>
<th>% with RMSD &lt; (Å)</th>
<th>Speed (ligands/s)</th>
<th>Average number of conformers</th>
</tr>
</thead>
<tbody>
<tr>
<td>ConfGen</td>
<td>34% 69% 89% 97%</td>
<td>8.58</td>
<td>52</td>
</tr>
<tr>
<td>ConfGen Classic Intermediate</td>
<td>21% 63% 83% 94%</td>
<td>0.34</td>
<td>47</td>
</tr>
<tr>
<td>ConfGen Classic Comprehensive</td>
<td>33% 71% 87% 96%</td>
<td>0.15</td>
<td>180</td>
</tr>
<tr>
<td>ConfGen + Minimization</td>
<td>39% 73% 90% 97%</td>
<td>0.14</td>
<td>42</td>
</tr>
<tr>
<td>ConfGen Classic Intermediate + Minimization</td>
<td>25% 66% 87% 96%</td>
<td>0.03</td>
<td>59</td>
</tr>
<tr>
<td>ConfGen Classic Comprehensive + Minimization</td>
<td>37% 73% 89% 97%</td>
<td>0.03</td>
<td>150</td>
</tr>
</tbody>
</table>

Figure 1. Distribution of the number of rotatable bonds within the 1,904 molecules used to characterize ConfGen's performance.
Both ConfGen and ConfGen Classic, along with several other commercial conformer generators, have been benchmarked at the Universität Hamburg [1] on a high quality dataset of 2,859 protein-bound ligand conformations extracted from the PDB. In terms of accuracy, the ConfGen program was found to be on par with OMEGA and overall better than MOE (see Table 2). Note that this paper assessed processing time by running conformer generator programs for a single ligand at a time, which is not representative of typical use. In addition, overhead from Schrödinger’s job-control system that enables easy submission to remote queues and hosts was included as part of the ConfGen timing. With these contributions factored out, when applied to the dataset used in that study, ConfGen (Suite 2017-3 release) processes on average 15 ligands per second without optimization, and about 0.4 ligands per second with geometry optimization using OPLS3. These speeds are similar to OMEGA’s and at least an order of magnitude faster than MOE Import, the fastest MOE approach.

### Table 2. Percentage of structures of the Platinum Diverse Dataset successfully reproduced within a specified RMSD threshold (excerpt from Tables 3 and 4 in Ref.[1])

<table>
<thead>
<tr>
<th>Product</th>
<th>% with RMSD &lt; (Å)</th>
<th>Average number of conformers</th>
</tr>
</thead>
<tbody>
<tr>
<td>ConfGen</td>
<td>43% 77% 93% 98%</td>
<td>39</td>
</tr>
<tr>
<td>ConfGen + OPLS3 Minimization</td>
<td>48% 81% 93% 98%</td>
<td>30</td>
</tr>
<tr>
<td>MOE Stochastic</td>
<td>45% 76% 88% 94%</td>
<td>30</td>
</tr>
<tr>
<td>MOE LowModeMD</td>
<td>46% 76% 87% 93%</td>
<td>31</td>
</tr>
<tr>
<td>MOE Import</td>
<td>29% 61% 85% 94%</td>
<td>48</td>
</tr>
<tr>
<td>OMEGA</td>
<td>49% 80% 92% 97%</td>
<td>34</td>
</tr>
</tbody>
</table>

ConfGen Algorithm

The new ConfGen algorithm utilizes a divide-and-conquer strategy to build feasible molecular conformations in three steps: (1) divide input molecules into fragments by breaking exo-cyclic rotatable bonds; (2) obtain conformations for all the fragments; (3) build conformations for the whole molecule by reconnecting the fragments in various ways.

The following types of bonds are considered non-breakable for fragmentation purposes: ring bonds, double or triple bonds, bonds to terminal atoms, and common types of “rigid” bonds such as N-C bonds in amides and thioamides and O-C bonds in esters and thioesters. All other bonds are broken and capped by hydrogens during fragmentation. This process is illustrated in Figure 2.

![Figure 2. Illustration of the ConfGen algorithm. Left: the molecule with individual fragments marked and numbered (there are fifteen fragments of seven distinct types: C, O, N, C(C)CC, Si, O=c1cc[nH]c(=O)[nH]1, and O=S(=O)(O1)C=C(C)1CCOC2). Right: the final conformers with atoms colored according to the fragmentation. The RMSD for the best conformer w.r.t. the crystal structure from 3QO9 is 0.8 Å (PDB resolution is 2.6Å.)](image)

An examination of the Zinc database [2] was used to identify approximately 40,000 template fragments. A MacroModel conformational search was applied to each of these fragments to generate a library of conformations for use in ConfGen searches. Conformations for fragments not found in the library are generated on the fly via a two-stage optimization process starting from random initial atom positions.
The first optimization stage uses a penalty function to efficiently obtain crude coordinates while the second stage uses the OPLS3 force field to fine-tune the conformations. The process is repeated a number of times depending on the number of sp³ centers in the fragment. By studying compounds from several public databases, it was observed that at most roughly $2^{n/2}$ low-energy conformations, where $n$ is the number of sp³ centers, are expected for the fragments generated using the ConfGen fragmentation rules. The algorithm generates this many conformations up to a maximum of 24 by default. Conformations with heavy atom RMSD values less than a configurable threshold (0.25 Å by default) relative to previously generated conformations are classified as duplicates and discarded.

The penalty function used in the first step of the on-the-fly conformer generation for fragments consists of terms designed to produce the desired features of the conformations:

$$E = E_{\text{bonds}} + E_{\text{angles}} + E_{\text{planar}} + E_{\text{tetrahedral}} + E_{\text{repulsion}}$$

The $E_{\text{bonds}}$ contribution is a sum of harmonic terms (one for every bond) that penalize departures of the bond lengths from their ideal values:

$$E_{\text{bonds}} = \sum_{\text{bonds}} (|l_b - l_{\text{ideal}}|^2)$$

Ideal bond lengths, $l_{\text{ideal}}$, are obtained as described in Refs. [3] and [4]. The $E_{\text{angles}}$ term has the same form as $E_{\text{bonds}}$, except that the sum runs over angles. Ideal angles are determined using simple rules based on atom hybridization, number of connections, and ring membership. Instead of using the angles directly, $E_{\text{angles}}$ uses the distance between the first and third atoms in the angle, $l_a$:

$$E_{\text{angles}} = \frac{1}{3} \sum_{\text{angles}} (|l_a - l_{\text{ideal}}|^2)$$

Both of the $E_{\text{planar}}$ and $E_{\text{tetrahedral}}$ terms are given by the sums over “templates”:

$$E_{\text{planar}} = \sum_{\text{planartemplates}} E_t$$

and

$$E_{\text{tetrahedral}} = \sum_{\text{tetrahedraltemplates}} E_t$$

where $E_t$ is proportional to the RMSD of the fragment’s atoms with regard to the corresponding atoms in the template:

$$E_t = \frac{1}{4} \sum_{\text{templateatoms}} (r_{\text{a}} - r_{\text{a,template}})^2$$

Here, template atoms are assumed to be superimposed onto the corresponding fragment’s atoms. ConfGen uses three kinds of templates: one for tetrahedral centers (3 or 4 atoms around a sp³ center) and another two for the planar configurations around sp² atoms and double bonds. In all cases the ideal (template) positions are built using ideal bond lengths and “perfect” angles (120° for the planar templates and 109.5° for the tetrahedral).

Finally, the $E_{\text{repulsion}}$ term is given by a sum of short-ranged soft “bumps” over the non-bonded atom pairs (that is, pairs of atoms that do not contribute to $E_{\text{bonds}}$ or $E_{\text{angles}}$):

$$E_{\text{repulsion}} = \sum_{\text{pairs}} B \left( \frac{r_{ij}}{(R_i^{\text{vdw}} + R_j^{\text{vdw}})^p} \right)$$

where $B(s) = \{(1-s^2)^2, \text{ if } s \leq 1\}$ and $p = \{0.9, \text{ for } 1-4 \text{ interactions}\}$.

The fragments are reconnected along bonds broken in the fragmentation step, one bond at a time. The linking order is chosen to favor joining similarly sized portions of the molecule at each intermediate step. Specifically, at each level, conformations for two fragments (A and B) are used to create the conformations for the combined fragment A+B. First, SMARTS patterns matched against the whole molecule are used to select specific dihedral angles to sample along the A – B bond. When there is only one conformation available for either (or both) of the fragments, the fragment’s symmetry is first identified by rotating it along the new bond by 45, 60, 90, and 180 degrees to see if equivalent structures result.
When symmetry has been identified, only a subset of the allowed dihedral angle values need to be sampled. Given $N_T$ distinct dihedral angles to consider, with $N_A$ and $N_B$ conformers available for each of the A and B fragments, $N_A \times N_T \times N_B$ combinations are possible. By default, the algorithm does not systematically try all the possibilities. Instead, random combinations are tried and checked for gross atom overlaps. The atom overlaps are quantified using van der Waals radii of the atoms:

$$\text{overlap}(i, j) = \begin{cases} 
0 & \text{if } r_{ij} \geq R_i^{vdw} + R_j^{vdw} \\
1 - \frac{r_{ij}}{R_i^{vdw} + R_j^{vdw}} & \text{if } r_{ij} < R_i^{vdw} + R_j^{vdw}
\end{cases}$$

where $r_{ij}$ is the distance between atoms $i$ and $j$, and $R_i^{vdw}$ and $R_j^{vdw}$ are their van der Walls radii from Ref. [5]. If the largest overlap is less than a configurable threshold (0.5 Å by default), the combination is accepted as a viable candidate. This process is repeated until a prescribed number of candidates (96 by default) is collected. If no combination passes the overlap test, all of them are considered to be viable candidates. A prescribed number of the best ranked candidates is then retained (64 by default), while the others are dropped. More conformations are retained for larger molecules: if the number of rotatable bonds exceeds some threshold (9 by default), the values regulating number of conformations (number of conformations for a fragment, number of candidates, and number of conformations to keep after each “join”) are increased by a factor that is a function of the number of the rotatable bonds.

Conformations are ranked using a sum of Lennard-Jones 6-12 non-bonded potentials (for rapid processing the L-J parameters are obtained from a look up table of DREIDING force-field [6] values), ad-hoc tabulated dihedral penalties associated with the inter-fragment bonds, and contributions that facilitate equatorial (versus axial) ring attachment configurations by enhancing specific 1-4 interactions. The process of joining progressively larger fragments together and sampling multiple orientations for the restored bond continues until the entire molecule is reconstructed. Duplicate conformer elimination is only performed for molecules smaller than a certain threshold (having less than 32 heavy or polar hydrogen atoms by default) because this process is slow and rarely impactful for the large molecules.

Key steps of the algorithm are illustrated in Figure 2 and Figure 3 for the ligand from PDB structure with code 3QO9. The structure is split into fifteen fragments as shown in Figure 2. Only the fragment labelled as “2” has more than one low-energy conformation. The algorithm picks the order in which fragments are assembled to minimize the number of the atom pairs that would need to be checked for overlaps. For this structure the order was as follows: 1+3, 2+15, 5+6, 9+13, 12+14, 8+(9+13), 11+(12+14), (1+3)+(2+15), 7+(8+9+13), 10+(11+12+14), 4+(7+8+9+13), (5+6)+(10+11+12+14), (4+7+8+9+13)+(1+3+2+15), (4+7+8+9+13+1+3+2+15)+(5+6+10+11+12+14). Many assembly steps involving simple fragments, e.g., 1+3, 5+6, 12+14, etc., resulting in a merged fragment that has only a single distinct conformation. Illustrative stages of the assembly algorithm are shown in the Figure 3.
Several ConfGen parameters can be used to control tradeoffs in calculation speed and recovery of bioactive conformations. The two most important are: whether the geometry of the conformers is optimized using an OPLS force field and the number of conformers the program is expected to generate. The default dielectric constant of 10 used during minimization was found to generate bioactive conformers better than using 1. A real-world example is shown in Figure 4. Apart from resolving bad atomic contacts, geometry optimization typically leads to mildly improved bioactive conformation recovery (see Table 1). Optimization, however, is about 20 times slower than running without minimization. Optimization also tends to reduce the number of conformers; something that can lead to reduced processing time in subsequent calculations that use these conformations.

Another parameter controlling the speed/quality tradeoff is the number of conformers to be generated per input ligand (-n command line option): larger values of this parameter may often result in better conformer diversity. This is illustrated in Figure 4 for the ligand from 2QP8 PDB entry (1.5Å resolution): using default simulation settings, the conformer closest to the bioactive conformation has an RMSD of 1.3Å (red). The RMSD drops to 0.6Å (green) if the maximum number of conformers is increased from 64 (default) to 256, and geometry optimization using OPLS3 force-field is enabled.

Conclusions

Given its exceptional speed and ability to reproduce low RMSD bioactive conformations, ConfGen is the ideal solution for generating conformations for ligand-based virtual screening. It is tightly integrated within Maestro and Schrödinger's ligand-based drug discovery tools, including Phase, which provides the ability to easily create databases for fast ligand-based virtual screening. As with all Schrödinger products, ConfGen is also constantly being refined and improved, so while the updates described in this paper represent a significant step forward compared to previous iterations, they will also act as the baseline for future enhancements.

Learn More

Interested in learning more about the contents of this paper? Email us at help@schrodinger.com or fill out the online form here. Find additional resources and information on ConfGen online at www.schrodinger.com/confgen.

References