Kothiwale *et al. J Cheminform (2015) 7:47*
DOI 10.1186/s13321-015-0095-1

Journal of Cheminformatics

SOFTWARE

Open Access

BCL::Conf: small molecule conformational sampling using a knowledge based rotamer library

Sandeepkumar Kothiwale[1], Jeffrey L. Mendenhall[1] and Jens Meiler[1,2*]

Abstract

The interaction of a small molecule with a protein target depends on its ability to adopt a three-dimensional structure that is complementary. Therefore, complete and rapid prediction of the conformational space a small molecule can sample is critical for both structure- and ligand-based drug discovery algorithms such as small molecule docking or three-dimensional quantitative structure–activity relationships. Here we have derived a database of small molecule fragments frequently sampled in experimental structures within the Cambridge Structure Database and the Protein Data Bank. Likely conformations of these fragments are stored as 'rotamers' in analogy to amino acid side chain rotamer libraries used for rapid sampling of protein conformational space. Explicit fragments take into account correlations between multiple torsion bonds and effect of substituents on torsional profiles. A conformational ensemble for small molecules can then be generated by recombining fragment rotamers with a Monte Carlo search strategy. BCL::Conf was benchmarked against other conformer generator methods including Confgen, Moe, Omega and RDKit in its ability to recover experimentally determined protein bound conformations of small molecules, diversity of conformational ensembles, and sampling rate. BCL::Conf recovers at least one conformation with a root mean square deviation of 2 Å or better to the experimental structure for 99 % of the small molecules in the Vernalis benchmark dataset. The 'rotamer' approach will allow integration of BCL::Conf into respective computational biology programs such as Rosetta.

Keywords: Conformation sampling, Knowledge-based, Fragment-based, Rotamer-library

Background

The interactions between small molecules and proteins are important for receptors, transporters, or enzymes to recognize their substrates as well as for small molecule therapeutics to bind to their target protein. The molecular interaction and, hence, the biological function of a small molecule is related to its three-dimensional structure when interacting with the protein. In solution, small molecules are often flexible and exist as an ensemble of conformations in equilibrium with one another. The biologically active conformation may be a single conformation or a small subset from the conformations sampled in solution or a new conformation, induced by protein binding. A uniform sampling of all energetically accessible small molecule conformations is essential for the success of protein small molecule docking simulations [1] for example in structure-based computer-aided drug discovery/design (CADD) [1–3]. However, also ligand-based CADD applications such as three-dimensional quantitative structure activity relationships (3D-QSAR) predictions [4] or pharmacophore modeling [5] rely on the use of conformational ensembles of molecules that capture the bioactive conformation as one of a diverse set of energetically accessible conformations [6, 7].

Conformational sampling methods

Table 1 summarizes some of the existing conformational sampling methods. Conformation sampling methods can be characterized in several ways. First, the allowed search

*Correspondence: jens.meiler@vanderbilt.edu
[2] Department of Pharmacology and Biomedical Informatics, Vanderbilt University, Nashville, TN 37212, USA
Full list of author information is available at the end of the article

Table 1 Conformation sampling methods

Method	Search space	Search strategy	Search method	Scoring function
Caesar [39]	Incremental search of torsion angles combined with distance geometry for ring systems	Fragment based	Systematic	CHARMm force field
Catalyst [40]	Incremental search of torsion angles with subsequent energy minimization	Non-fragment based	Simulation (MD)	CHARMm force field
Conan [41]	Incremental search of torsion angles	Fragment based	Systematic	–
Confab [42]	Incremental search of torsion angles	Non-fragment based	Systematic	MMFF94
Confgen [31]	Random walk on energy surface calculated using a truncated version of OPLS_2001	Non-fragment based	Simulation (MC)	MMFFs/OPLs_2001
Corina [43]	Knowledge based rules derived from CSD	Non-fragment based	Systematic	Reduced force field for optimizing only ring systems
Enumerated Torsions (*et*) [18]	Incremental search of rule-based torsion angles	Non-fragment based	Systematic	–
Mimumba [15]	Incremental search of knowledge-based torsion angles from CSD	Non-fragment based	Systematic	Relative frequency of experimentally observed conformations
Moe (low mode MD) [44]	Constant temperature MD	Non-fragment based	Simulation (MD)	MMFF94
Moe (stochastic search) [22]	Random perturbations of rotatable bonds in increments biased around 30°	Non-fragment based	Simulation (MC)	MMFF94
Moe (Confimport) [22]	Pregenerated fragment conformations obtained from stochastic-search	Fragment-based	Simulation	MMFF94
Moe (systematic) [32]	Incremental search of torsion angles	Non-fragment based	Systematic	MMFF94
Omega [33]	Knowledge based torsions from analysis of molecules in PDB and conformations generated by MMFF94	Fragment based	Systematic	MMFF94
RDKit [34]	Distance geometry	Non-fragment based	Simulation (distance geometry)	UFF

space can be analyzed: some methods search the entire conformational space, i.e. bond length, angles and torsions can be altered—for example a molecular dynamics simulation in Cartesian space. Other methods restrict the search space to torsion angles only holding bond length and angles fixed. Another approach involves using pre-existing knowledge of small-molecule conformations to restrict the conformational search space even further to likely torsion angles or combinations thereof. Such knowledge-based methods derive torsion angle preferences from molecular mechanics or quantum chemical simulations of small molecules or structural databases like Cambridge Structure Database [8] (CSD) or Protein Data Bank [9] (PDB).

In addition, it is helpful to single out fragment-based approaches: This search strategy splits a molecule of interest and samples conformations of smaller fragments independently. Candidate conformations of the entire molecule are computed by re-combining constituent fragment conformations. In fragment-based methods, fragments are reused during conformer generation which improves the time-efficiency of sampling. On the other hand, these methods operate on the assumption that all low energy conformations can be created by combinations of low-energy fragments—an assumption that is not always fulfilled.

An alternative classification approach focuses on whether the search space is sampled systematically in its entirety or a search algorithm follows a trajectory that seeks to restrict the search space to low energy conformations. If the conformational space is sufficiently small, systematic approaches can create all possible conformations iteratively and keep all low-energy conformations. An advantage is complete sampling of the entire search

space, one disadvantage is slowness. Trajectory-based methods use random or directed perturbations to alter a starting conformation and the resulting conformation is evaluated energetically. In a feed-back loop this energy and possibly derived forces determine the trajectory of the simulation. Molecular dynamics [10, 11], distance geometry [12], genetic algorithms [5], and Monte Carlo [11] (MC) are commonly used simulation methods for the conformational sampling of small molecules.

Scoring functions

Most methods score conformations using some form of molecular mechanics energy function. Force field based energy calculations use most frequently the Merck molecular force field (MMFF) [13] or the Chemistry at Harvard Molecular Mechanics (CHARMm) force field [14]. Some methods modify the default versions of these force fields by modifying individual scoring terms or using only a subset of the scoring terms. One alternative approach, as used in Mimumba [15], to scoring small molecule conformations can be derived from knowledge-based scoring functions used in protein structure prediction that analyze the frequency of geometric features observed in structural databases such as the PDB or CSD.

Knowledge based conformation sampling

Conformations of small molecules can be restricted in terms of commonly seen conformations of constituent fragments in structure databases like CSD. Brameld et al. [16] have shown that conformations of fragments sampled in the CSD are an accurate representation of conformational space seen in drug-like molecules in complex with protein as observed in the PDB. Fragments occur in these structure databases in different chemical environments, leading to them being observed in different conformations. The central hypothesis of this study is that while not all small molecules have been crystallized in all possible conformations, the conformational space accessible to sufficiently small fragments is adequately sampled.

Existing methods like Confect [17] derive torsion profiles for different dihedral bond types from structure databases. Confect treats dihedral bonds as uncorrelated and does not take into account substituent effects. A rule-based proprietary method, developed by Merck research laboratories for internal use, known as *et* for enumerated torsions uses correlated torsion angles to some extent for conformational sampling [18]. The method overlaps multiple fragments containing topologically adjacent rotatable bonds to extend these fragments until they span the entire small molecule. In *et* a proprietary 'atom typer' is used to express molecular fragments as unambiguous patterns [19]. The pattern along with associated data for observed torsion angles and frequency constitutes a rule. As of 2001, authors reported that 797 rules had been derived over a period of several years. However these patterns consider only the four atoms involved in a dihedral bond and do not take into account effect of substituents on torsional profile of bonds.

The algorithm BCL::Conf described in the present study goes beyond previous work by using torsional profile of multiple consecutive dihedral bonds and capturing effect of substituents on their torsion profiles. All fragment conformations sampled frequently in the CSD and PDB are considered a knowledge-based 'rule' independent of size or number of rotatable bonds. This fragment conformation approach allows BCL::Conf to capture correlations in torsion states for multiple consecutive dihedral bonds in contrast to other methods that treat likely torsion angle states for consecutive bonds in an uncorrelated way. Conformations observed frequently for one fragment are assumed to represent a local energy minimum and are collected in a database. The use of conformations of fragments has also the advantage that these fragment conformations already reside in locally optimal geometries so that only non-local interactions, i.e. clashes, need to be evaluated when fragments are recombined. Lastly, as explicit fragments are used effects of substituents on torsional profiles of rotatable bonds are taken into account. Brameld et al. have shown the effect of substitution on the torsion distribution of common acyclic organic fragments [16].

We expect that the algorithm is therefore particularly tailored for 'drug-like' small molecules which are over-represented in the CSD and PDB databases. BCL::Conf mimics the 'rotamer' libraries created to capture amino acid side chain conformations seen in protein structures within the PDB [20] which, ultimately, will ease its integration with protein modeling packages such as Rosetta [21]. BCL::Conf scoring includes a clash score that avoids atom overlap as well as a knowledge-based scoring function that scores conformations based on probabilities of fragment conformations that it contains.

To benchmark BCL::Conf we use a curated dataset containing drug-like ligands found in complex with proteins in the PDB. The "Vernalis generic compound set" [22] has been used in several studies to evaluate the performance of conformational sampling methods enabling a direct comparison of BCL::Conf to other methods [23, 24]. The benchmark study tests for recovery of protein-bound conformation of the ligand and also the ability of BCL::Conf to produce a diverse set of conformations. To remove any bias during benchmarking, the ligands found in the Vernalis dataset were removed from the PDB ligand library. Additionally, ligands were removed from

the PDB ligand library if bound to proteins or homologues of proteins present in the VERNALIS dataset.

Implementation

BCL::CONF uses fragments generated from decomposing molecules found in CSD and PDB. For this purpose, non-ring bonds of each molecule are broken iteratively to generate all possible fragments. In a second step all occurrences of one fragment within the structure databases are collected and clustered according to discrete dihedral angle bins. A conformer is then defined as a unique conformation represented as a set of integer numbers, one for each dihedral bond, identifying the bin. This procedure is similar to the definition of 'rotamers' that are used to set likely amino acid side chain conformations [20]. A conformer needs to be seen at least four times in the database to be considered a likely conformation of a fragment. It is then added to the rotamer library for sampling. The flowchart for algorithm implemented in BCL::CONF is shown in Fig. 1.

Fragment library

Small organic molecules from the CSD and PDB were used for generating fragments. The PDB ligands were obtained from the refined dataset in the PDBBIND database [25–27]. We removed any molecules for which BCL could not assign correct atom types, molecules with missing 3D coordinates and bad geometries in terms of unrealistic bond-lengths or bond-angles and non-planar aromatic rings or sp^2–sp^2 bonds. This resulted in a database containing 113,339 unique molecules. Molecules were broken iteratively at non-ring bonds which generated 56,818,272 unique fragments.

Rotamer library

The rotamer library was generated for fragments that are seen frequently in same conformations. A unique fragment rotamer/conformation is identified by a set of integers, one for each dihedral bond. The dihedral bonds of a rotamer are represented as a set of integers depending on the angle measure as explained in Fig. 2. The frequency distribution of dihedral angle measures seen in CSD, shown in Fig. 2, suggests that local minima for dihedral angles occur at canonical values of 0°, 60°, 120°, and 180° and so on. In addition, for certain bond types such as *aromatic-chain-aromatic* or *aromatic-chain-any* angles of 90° and 270° are likely (Additional file 1: Figure S1A). Hence, while torsion angles of 90° and 270° are not local maxima when summing over all torsions, they are likely conformations for certain types of torsion angles. Therefore, in order to assign as many likely torsion angles as possible unambiguously and close to a bin center, 12 bins each of which is 30° wide are created centered at 0°, 30°, 60°, 90° and so on. Binning strategies using 30° produces closer to native conformations when 60° binning is used (see "Results and discussion"). All the bonds including the ones that are inside ring systems are described by an integer so that a rotamer can be described as a string of integers. This string is called the bin-signature of a rotamer.

Determining dihedral angles

Since multiple dihedral angles can be measured at each torsion bond, a scheme is required to prioritize which dihedral angle to use and arrive at unambiguous bin-signatures. Therefore a priority dihedral angle is defined. This is accomplished using rules analogous to the Cahn–Ingold–Prelog (CIP) system [28]. For example, as shown in Fig. 3a, 2-butanol has one torsion bond but two dihedral bonds about the single rotatable bond. According to CIP rules, the O–C–C–C dihedral angle will have a higher priority over the C–C–C–C dihedral angle. If out of three possible dihedral angles, two dihedral angles of equally high priority exist, then the third dihedral angle with lowest priority is used. If ambiguity still exists in assigning unique dihedral bonds, for example in the case where all dihedral angles have the same priority, the one with the smallest angle measure is chosen. Priority dihedral bonds in rings are defined in a special way in that all atoms constituting a priority bond are contained in the ring, as shown in Fig. 3b for cyclohexanol. This ensures that for the same ring conformation, a substituted ring system has the same dihedral-signature as an un-substituted ring system. If a fused ring system is present, then priority dihedrals are determined using atom priorities and the assumption that all atoms of the ring system are part of one ring (Fig. 3c). BCL::CONF can identify different ring conformations and use these in conformational sampling. Since dihedral angles are assigned in a unique way for a molecule of interest, a unique rotamer of the molecule has a unique dihedral bin signature. Table 2 shows different rotamers for a fragment from the rotamer library and their bin signatures.

Searching rotamers

In building the rotamer library, all instances of every fragment are collected in the molecular database using a graph isomorphism search [29]. For each fragment, all unique rotamers are identified using dihedral bin signatures. Then statistics is gathered for each rotamer including rotamer counts, i.e. the number of times a rotamer is seen in the database, and dihedral angle statistics, i.e. the average angle measure and standard deviation of dihedral bonds within each bin. A representative structure for a fragment is obtained by clustering all instances of the most frequently observed rotamer in the structure database on the basis of root mean square

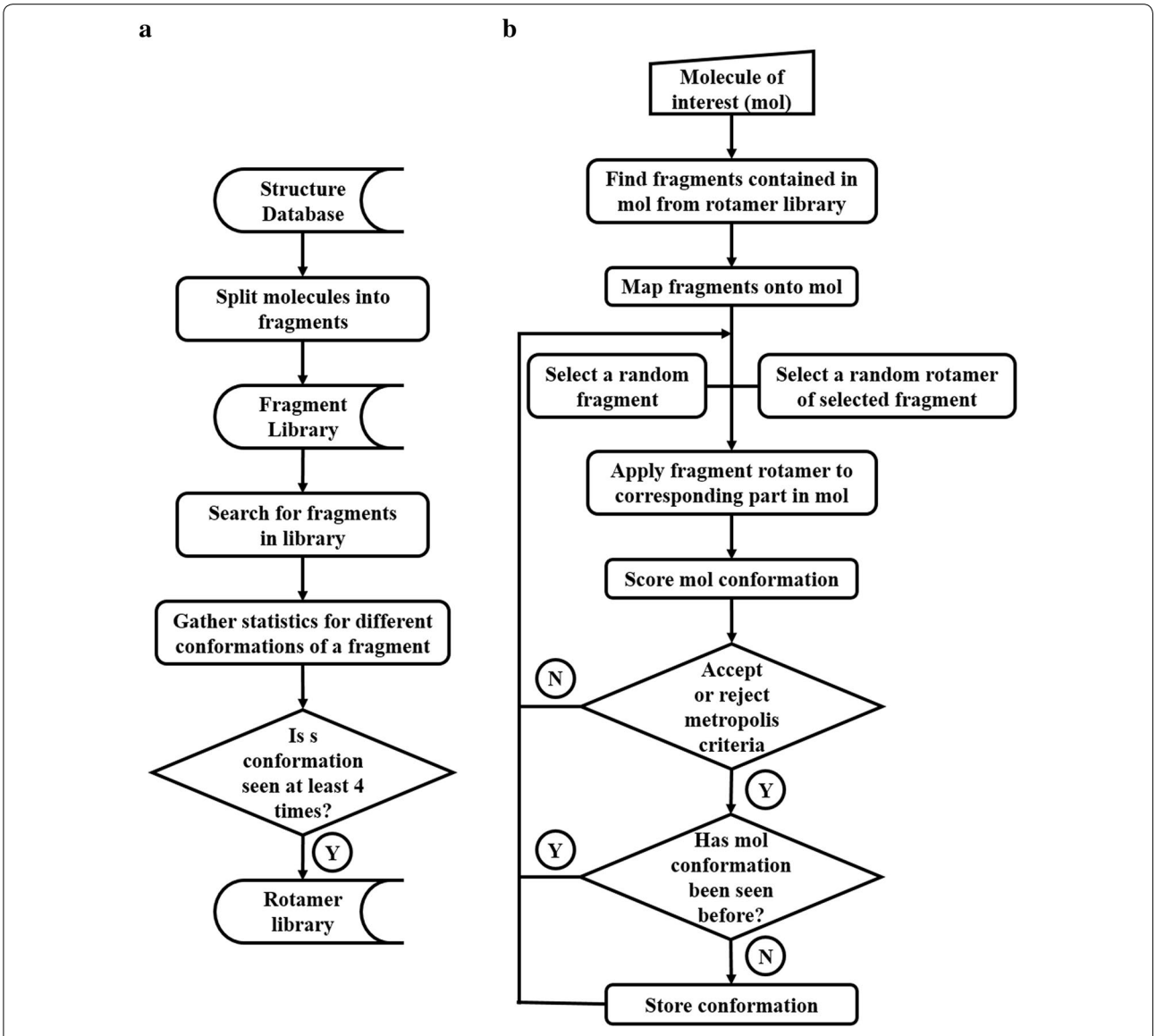


Fig. 1 General scheme for BCL::Conf conformation generator. **a** Scheme for generating the rotamer library. **b** Flowchart depicting conformation sampling process. See text for a detailed description

deviation (RMSD) after superposition. In addition, if a fragment contains a ring in different conformations, explicit coordinates are stored for each rotamer. A conformer is added to the rotamer library of a fragment if it is seen at least four times in the structure databases (combined CSD and PDB), i.e. it can be considered a likely conformation for that fragment. A total of 231,049 fragments are observed that have at least one conformer which is seen at least four times in the molecular database and hence these fragments are retained in the rotamer library. Table 3 shows the rotatable bond distribution and rotamer distribution of fragments in the rotamer library.

Search fragments from the rotamer library that are contained in the molecule of interest

Conformational sampling begins with searching fragments contained in a molecule of interest. This involves substructure searches to identify all suitable fragments in the rotamer library. A hierarchical search has been implemented to minimize the number of substructure searches. The rotamer library is represented as multiple

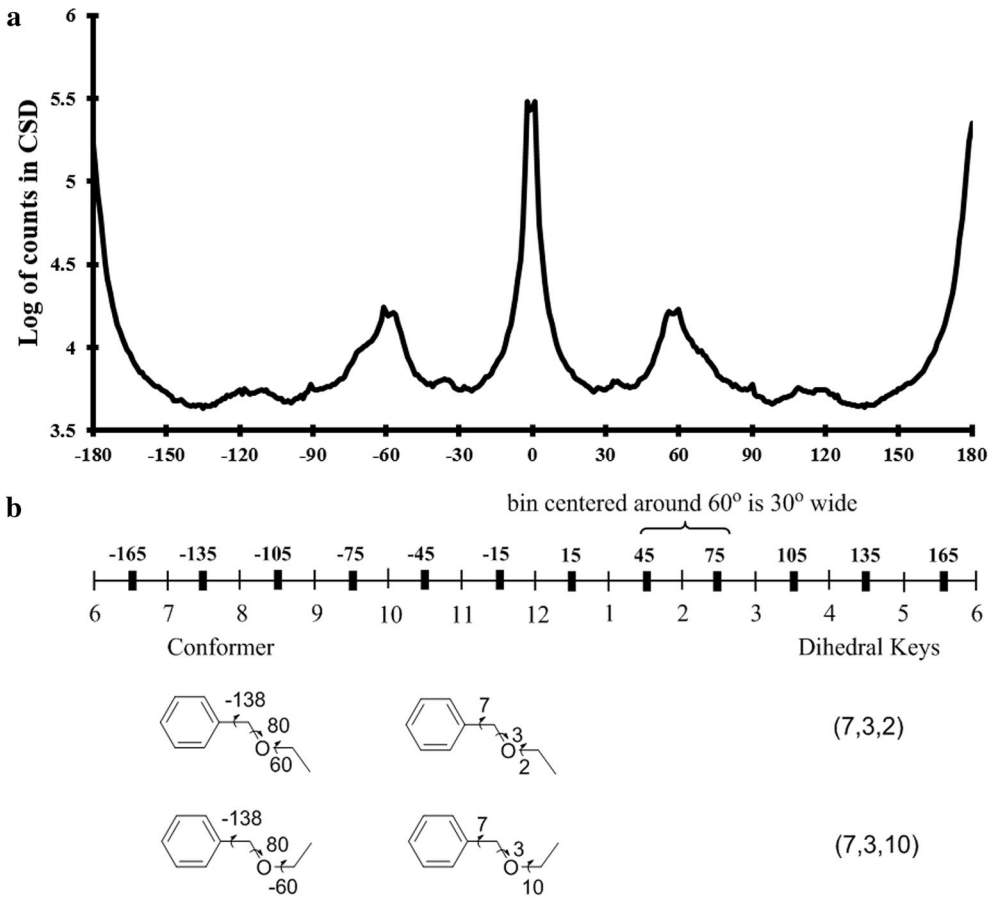


Fig. 2 Scheme for torsion angle binning. **a** The *line graph* shows the distribution of dihedral angle measurements of all dihedral bonds over all the molecules in the CSD. **b** Torsional angles are binned into 12 uniform parts with each bin represented as an integer. For example −135 to −165 belongs to bin number 7. Rotamers can thus be represented by a unique key of integers representing each dihedral angle bin

rooted graphs where each node is a unique constitution. The root nodes are not contained in any other fragments. Child nodes are such that the parent node is an immediate substructure. Figure 4 illustrates a rooted graph with benzene as root. Benzene is an immediate substructure of its child nodes i.e. toluene-like fragment which is an immediate substructure of cyclohexylbenzene-like fragment.

The fragment searching begins at the root node of graphs. If the root node is contained in the query molecule, all its immediate child nodes will be searched to determine if they are contained within the molecule of interest. For all child nodes contained, their immediate child notes are considered and so on. In Fig. 4 fragments that are part of molecule are colored in blue—i.e. a successful substructure search. Fragments colored red indicate that a substructure search was performed but unsuccessful. This terminates further searches in this branch of the tree. Fragments colored in black are not considered for a substructure search, because their parent fragments were not contained within the molecule of interest (colored red). The edges in the graph are directed from parent to child nodes and represent search paths that can be taken to find all constituent fragments in a query molecule. Paths in blue color are actual paths that are taken to identify all the fragments contained in the molecule interest while the paths in red or black are never explored. Search paths in black originate from fragments that are not contained with the molecule. Red paths represent redundant searches in the tree. This hierarchical tree structure of the data enables fast and efficient searching of all the fragments contained within a molecule of interest.

Generation of initial 3D structure from minimum set of fragments with most likely conformation

An initial 3D conformation is necessary for using the conformer sampler implemented in BCL::Conf. The BCL software suite accepts molecules in the MDL [30] format. A 3D structure generator has been implemented to generate

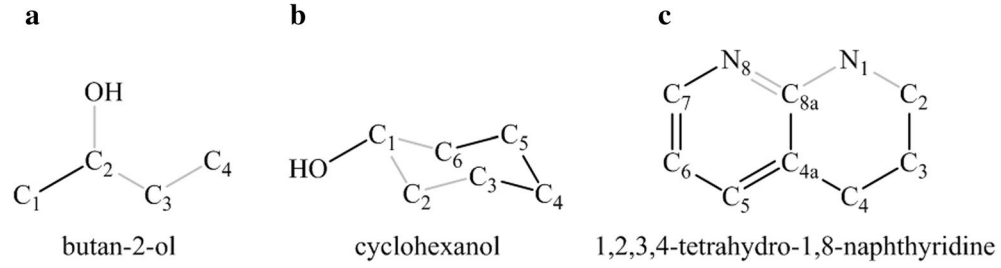


Fig. 3 Determination of priority dihedral bonds in molecules. Bond priorities are determined using rules analogous to Cahn–Ingold–Prelog (CIP) rules. In the figure priority dihedral bonds are colored in *grey*. **a** The priority dihedral angle of 2-butanol is c4–c3–c2–O. **b** Priority dihedral bonds in cyclohexanol are defined such that all atoms that define priority dihedral angles are in the ring. Thus for bond C_1–C_2, C_3–C_2–C_1–C_6 is the priority dihedral angle instead of C_3–C_2–C_1–O. **c** For multiple ring systems like 1,2,3,4-tetrahydro-1,8-naphthyridine, priority angles are determined by atom priority using the assumption that all atoms in the multiple ring system are part of one ring. Thus C_2–N_1–C_{8a}–N_8 is the priority dihedral angle instead of C_2–N_1–C_{8a}–C_{4a} as N_8 is counted to be in the same ring system as the N_1–C_2 bond of interest

an initial 3D structure if coordinates are not provided. BCL::Conf can generate starting coordinates from connectivity information provided in the MDL format. When coordinates or 3D structure is not available, BCL::Conf first searches for all fragments from the rotamer library that are contained in a molecule of interest. The algorithm identifies the minimum number of fragments that can be connected to generate molecule of interest. The most likely conformers of fragments are then connected to assemble the molecules of interest and generate an initial 3D structure which may or may not have clashes between atoms. As this conformation only serves as a starting point with the objective to place all torsion angles into a locally reasonable conformation and is not necessarily part of the output ensemble of conformations, atom clashes are not a problem.

Monte-Carlo Metropolis sampling for efficient search of conformational space for likely non-clashing conformations

Conformational sampling begins by identifying fragments from the rotamer library that are contained in the molecule of interest whose conformations need to be sampled. From the fragments contained in the molecule of interest, a random one is selected and one of its rotamers is applied to change the conformation of the molecule. The rotamer is selected based on probability of its occurence in the structure database (Fig. 1b). If the chosen fragment rotamer contains a different ring conformation, then the whole molecule is reassembled by using the chosen conformer as the starting fragment. By default only a subset of rotamers that are observed most frequently are used in sampling. The cutoff value is specified at half of the probability of the most likely rotamer. If more sampling is desired, an option to use the full rotamer set can be specified at the command line.

Starting with the input structure of the molecule of interest, new conformations are created in a continuous MC trajectory. A MC step is accepted or rejected based on the Metropolis criterion. The energy or score used is a combination of atom clashes and propensity of observing constituent fragment rotamers in structure database. The atom clash score is calculated by evaluating non-bonded atom pairs for clashes using Eq. 1.

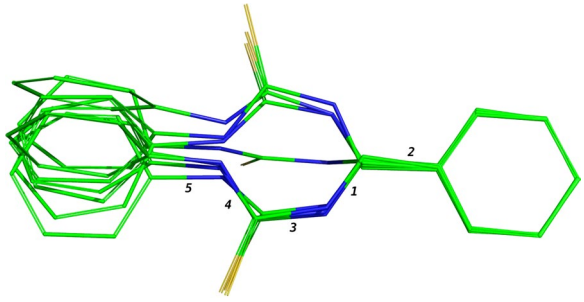
$$Atom\ Clash\ Score = \frac{\sum_{i>j} 2 * score_{atom_j} \begin{cases} 0, & dist \geq cov \\ 1, & dist \leq cov \end{cases}}{Number\ of\ atoms\ in\ the\ molecule} \quad (1)$$

where $dist \stackrel{\text{def}}{=}$ distance between non-bonded atoms i and j, $cov \stackrel{\text{def}}{=}$ sum of covalent radii of atoms i and j.

Rotamer propensity score (Eq. 2) leverages the statistics on the rotamer of a particular fragment to estimate the likelihood of a particular conformation. The hypothesis is that there is a correlation between frequency of occurrence and free energy of a fragment conformation. For a given molecular conformation, the observed rotamer of each of the constituent fragments is determined. The observed rotamer propensity for a fragment is calculated by dividing observed rotamer count by average rotamer counts. The overall conformation score is obtained by summing up observed rotamer propensities of all the constituent fragments. If, for a fragment none of the rotamers are seen in a given conformation, then a pseudo rotamer count equal to half of the least common rotamer count is used instead. The propensity score is normalized by dividing it by absolute value of maximum possible propensity score for the molecule of interest.

$$Propensity\ Score = \sum_{i=0}^{N} \left(-ln \frac{R_i \times F_i R_j}{\sum_j F_i R_j} \right) \Bigg/ \sum_{i=0}^{N} \left(ln \frac{R_i \times F_i R_{max}}{\sum_j F_i R_j} \right) \quad (2)$$

Table 2 The rotamers of a fragment from the rotamer library



Rotamer #	Bond1	Bond2	Bond3	Bond3	Bond4	Bond5
1	6	5	6	12	2	6
2	6	3	6	12	5	6
3	6	1	6	12	2	6
4	6	5	6	12	4	6
5	6	1	6	12	5	6
6	6	5	6	12	5	6
7	6	4	6	12	2	6
8	6	1	6	12	1	6
9	6	5	6	12	1	6

Five rotatable dihedral bonds are labeled in the figure and for each rotamer, the dihedral bins are shown

Table 3 (a) Rotatable bond distribution in the rotamer library, (b) conformation statistics in the rotamer library

Number of rotatable bonds	Number of fragments
0	47,205
1	38,616
2	31,225
3	20,500
4	15,221
5	13,665
6	14,014
7	14,693
8	14,435
9	13,492
≥10	10,064
Number of rotamers	**Number of fragments**
1–5	219,684
6–10	10,840
11–15	1768
16–20	488
21–25	209
26–30	82
31–35	47
36–40	18
41–45	8
46–50	1
>50	3

where $N \stackrel{\text{def}}{=}$ number of fragments that are part of the molecule of interest, $F_i \stackrel{\text{def}}{=} i$th fragment of molecule, $R_i \stackrel{\text{def}}{=}$ number of rotamers of the ith fragment, $R_{max} \stackrel{\text{def}}{=}$ counts of the most common rotamer, $F_iR_j \stackrel{\text{def}}{=}$ counts of jth rotamer of the ith fragment.

Results and discussion

We assess the performance of BCL::Conf (BCL) with curated generic ligand dataset known as the Vernalis dataset [22], in comparison with Confgen [31], Moe (Confimport) [32], Omega [33] and RDKit [24, 34]. The first metric defined as the completeness criteria is the fraction of molecules for which any conformation was generated. The second comparison is the ability of the method to produce ligand conformations within a specified RMSD value to the native conformation of ligands in protein–ligand complexes. This analysis is reported as the percentage of molecules whose conformations are recovered within a given threshold RMSD value. The third criteria for comparison is diversity, that is how similar or different are the generated conformations. Finally a comparison of the methods on computational speed is provided. We also report results for different flavors of BCL that use different schemes for rotamer library generation—(a) using a 60° torsion binning (BCL_60), (b) rotamer library derived from only the CSD (BCL_CSD), (c) rotamer library containing only single dihedral bond torsion profiles (BCL_D).

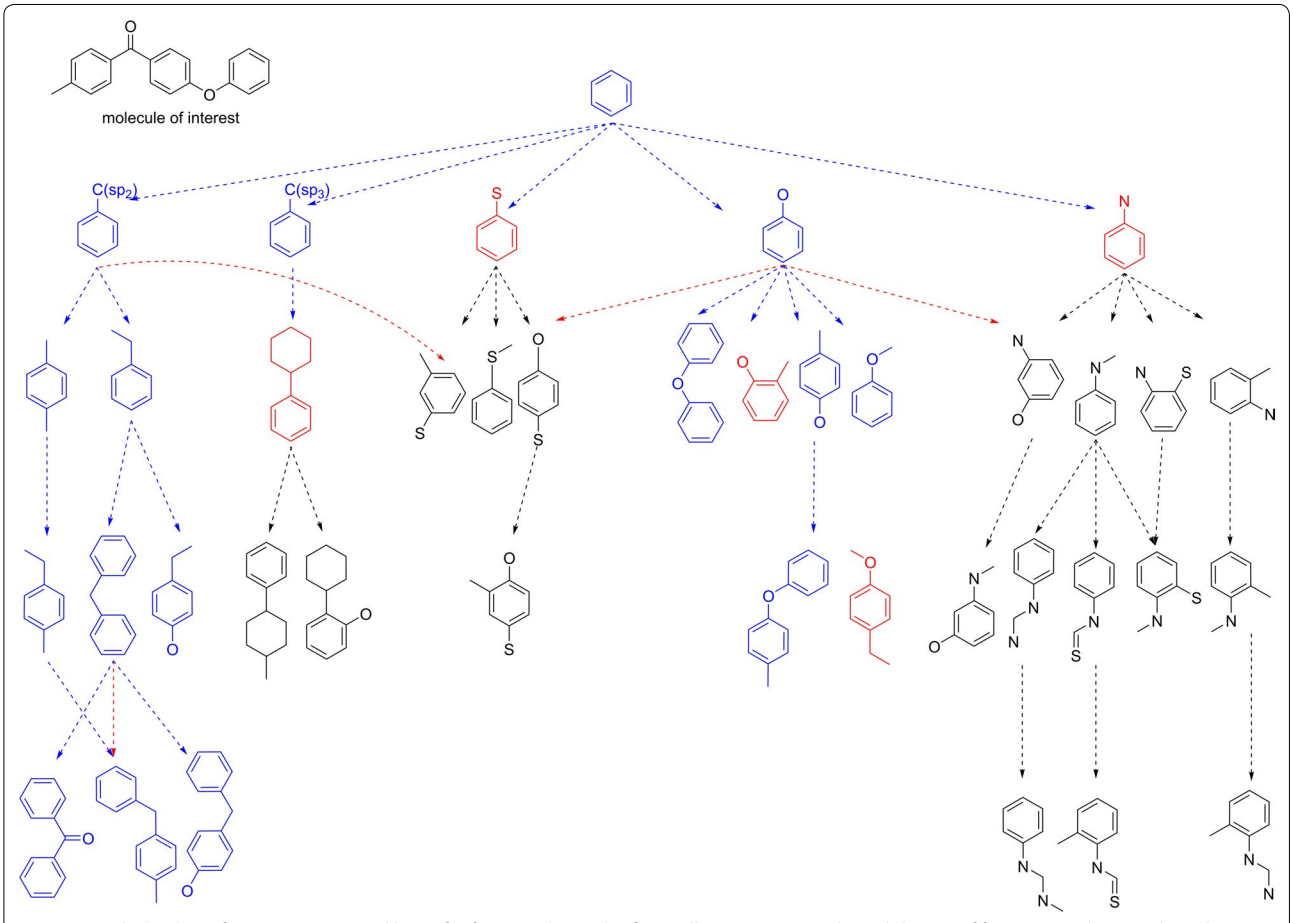


Fig. 4 Graph database for storing rotamer library for fast searching. The figure illustrates a rooted graph layout of fragments where each node is a unique constitution. The child nodes originating from the root are such that the root (in this case, benzene fragment) is their immediate substructure among all the fragments shown in the graph. Fragments contained in the molecule of interest are colored in *blue* while those that are not are in *red* or *black*. For fragments in *black* no substructure search is performed because their parent fragments were not found in the molecule of interest. The edges represent all possible search paths for finding fragments contained in the molecule of interest. Paths in *blue* are the actual searches that were performed for finding fragments for the query molecule. Paths in *red* and *black* are never taken during the search. *Red* colored paths are redundant search paths that have already been covered in a previous search

Conformational sampling with different methods was performed to yield a symmetry corrected RMSD diversity of 0.25 Å—i.e. no two conformations have a RMSD smaller than 0.25 Å—and a maximum of 100 conformers per molecule.

Ligand dataset

Vernalis dataset is used here to compare BCL::Conf to other existing methods in the field. The Vernalis Dataset (Additional file 2), compound set introduced by Chen and Foloppe [22–24], contains 253 ligands derived from high-resolution protein–ligand complexes found in the PDB and includes the Bostrom [35, 36] ligand set and Perola [37] ligand set. The Vernalis Dataset has been used in previous benchmark studies to compare Moe, Catalyst and Confgen methods for conformation sampling [22–24].

Conformer generation methods

BCL::Conf (BCL): Conformation sampling was carried out by providing ligands in the MDL format with all atom coordinates set to zero to remove any initial conformation bias. The rotamer library uses the 30° torsion binning scheme to determine dihedral keys. It is derived from the CSD and the refined set of PDBbind database minus the Vernalis dataset ligands to remove any bias. Conformers were generated in 200 iterations of MC fragment sampling at a temperature of 3.0 such that they were at least 0.25 Å away from each other. Table S2 (see Additional file 1, Additional file 3) shows parameter optimization for native conformer recovery in terms of RMSD with different temperature and iteration values. The row shaded in gray corresponds to parameters used for comparing to other methods.

BCL_60: Conformations were sampled using the same settings as described for BCL::Conf except that 60° torsion binning was used instead of 30°. This experiment tests the effect of 60° binning on conformation sampling.

BCL_CSD: Same parameters as used for BCL:Conf with the only difference being that the rotamer library was sourced from only the CSD. This experiment shows the effect of adding PDB fragment conformations.

BCL_D: Conformation sampling was performed by using torsion angle statistics for single dihedral bonds derived from molecules in the CSD and PDBbind databases. Fragments containing only four atoms and a single dihedral bond from the rotamer library were used for this experiment—i.e. the smallest possible fragments. This experiment tests the impact of the addition of larger fragments that sample the correlation between multiple torsion angles. Initial conformation bias in benchmark dataset molecules was removed by perturbing all dihedral angles to random values. The conformers were generated using the same set of parameters as that for BCL.

Confgen: Confgen systematically samples rotatable bonds, ring conformations, nitrogen atom inversions and amide bond conformations. Force field OPLS_2001 is used for calculating potential for rotating about each rotatable bond [31]. In the present study, conformer generation was done starting from SMILES string of ligands in the Vernalis dataset. SMILES string were generated using Maestro from the dataset ligands in MDL format. Confgen has been reported to reproduce 93 % of molecules within 1.5 Å in the comprehensive mode [31]. 250 conformers were generated with Confgen in the comprehensive mode by keeping RMSD cutoff at 0.25 Å, energy cutoff at 104.6 kJ/mol (default value). 100 conformations were saved per ligand for comparison.

Moe-conformation_import (Moe): Conformational import is a high-throughput conformer generation method in Molecular Operating Environment (Moe). Molecule of interest is divided into overlapping fragments and these are searched in a pregenerated library of fragment conformations. If a fragment is not found, conformations are generated using a stochastic conformation search algorithm available in Moe. For this study, the Vernalis dataset was provided such that all atom coordinates were set to zero. The default parameters specified with Moe have been determined to perform best in previously reported benchmark studies [22–24]. The MMFF94x force field and Generalized Born solvation model was during ligand conformation generation. Fragment conformation energy cutoff was kept at a default of 4 kcal/mol. The program was constrained to maintain stereochemistry of the input structures but allowed to sample ring conformations. The stochastic search protocol that conformation import uses for creating conformations of fragments missing in database was modified to generate fragment conformers that were 0.25 Å apart in RMSD. Fragment conformations that were within 15 kcal/mol window of the lowest energy conformer were retained for the stochastic search.

Omega: Omega is a systematic knowledge based conformer generator developed by Openeye Scientific Software. It exhaustively enumerates all rotatable torsions using a knowledge-based list of angles which are then sampled by geometric and energy criteria [33]. The torsion library is derived from analysis of a set of experimental crystal structures from the PDB and from energy scans of torsions against MMFF94. Default parameter values were used except RMSD and MaxConfs which was set to 0.25 and 100 respectively to specify custom conformation diversity level and limit the number of output conformations.

RDKit: RDKit uses distance geometry algorithm described by Blaney et al. for sampling ligand conformations [38]. A distance bound matrix is calculated for a molecule of interest based on connection table and a set of rules. The matrix is smoothed using a triangle-bounds smoothing algorithm. Random distance matrices that satisfy the bounds matrix are generated followed by embedding in 3D dimension to generate conformations. In a final step, embedded coordinates are cleaned up using a crude force field and the bound matrix [23]. In this study, ligand conformations generated using RDKit were minimized using the Universal Force Field 'uff' as suggested by Ebejer et al. [24]. 100 conformations were generated followed by minimization and pruning to remove conformations that measure less than 0.25 Å away from each other in RMSD.

BCL::Conf generates conformations for all drug-like small molecules

While BCL, Confgen, Moe and RDKit are able to generate conformations for all the molecules of the Vernalis dataset, Omega could not for 16 molecules due to missing fragments in its library.

Recovery of experimentally observed conformations

The native conformation recovery by BCL, Confgen, Moe, Omega and RDKit is plotted in Fig. 5a. Figure 5a shows the percent recovery of native conformation of ligands at different RMSD cutoff values. BCL recovers native conformer for 11 % of ligands within 0.25 Å, 79 % within 1.0 Å and 99 % within 2.0 Å. Figure 5c shows the effect of rotamer library source (CSD; single dihedral

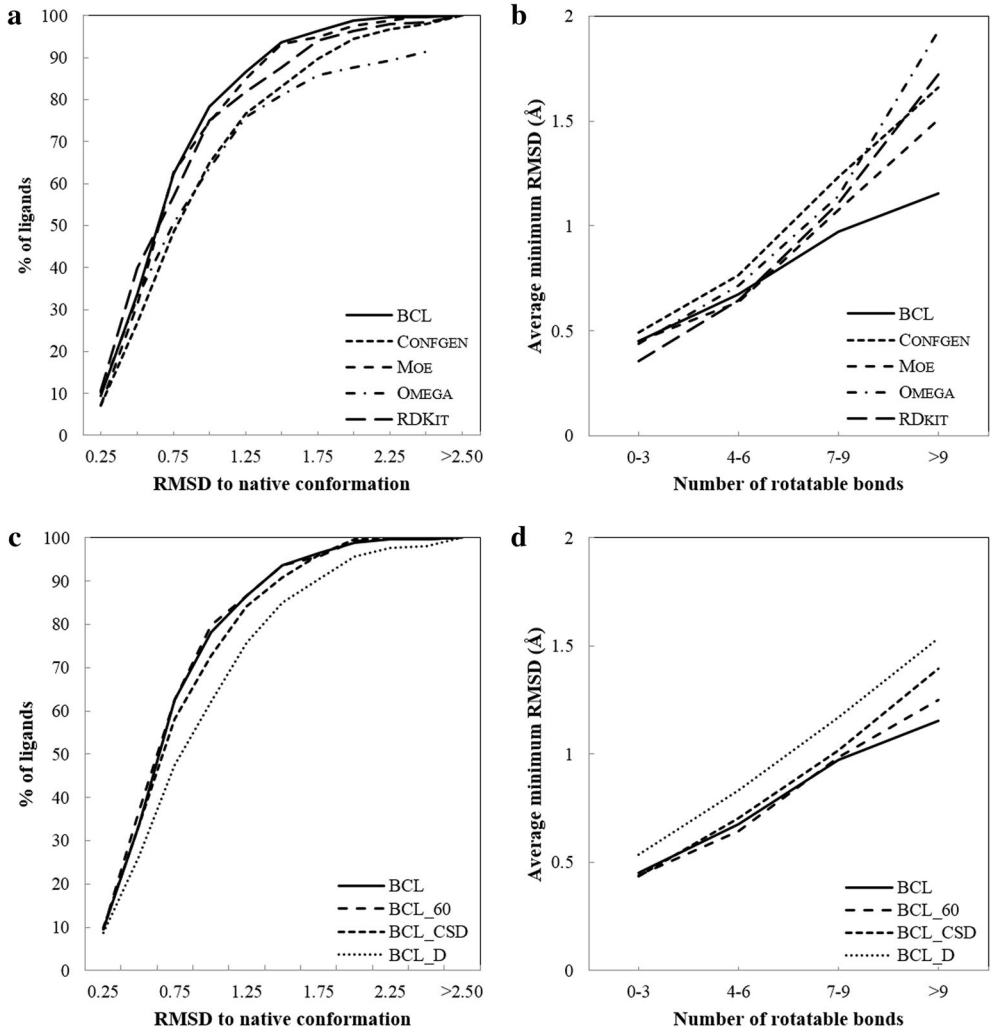


Fig. 5 Benchmarking results for the VERNALIS dataset. **a** The plot represents percentage of ligands (y-axis) for which different methods produce at least one conformer within an RMSD value less or equal to the RMSD value on the x-axis. **b** Quality of conformations sampled as the number of rotatable bonds increases. The average RMSD of conformers closest to native structure is plotted on the y-axis as the number of rotatable bonds increases (x-axis). **c** Same as **a** for different rotamer libraries used. "BCL" refers to recovery using 30° dihedral bins with rotamers derived from both the CSD and PDB. "BCL_CSD" leverages conformations from the CSD only. "BCL_D" refers to experiments in which instead of fragments containing multiple torsion angles, statistics on single dihedral angles were used for sampling conformations. "BCL_60" refers to a rotamer library that uses 60° dihedral bins are. **d** Quality of conformations sampled as the number of rotatable bonds increases for BCL, BCL_CSD, BCL_D, and BCL_60

torsion profiles; and CSD + PDB) and binning strategy (30° or 60°) on conformation recovery. Conformation recovery is slightly lower when fragment rotamers observed in only the CSD are used suggesting unique rotamers or significant deviation from canonical values that are observed in ligands bound to proteins. Recovery is not effected significantly when 60° bins are used.

Figure 6 shows pairwise comparison of CONFGEN, MOE, OMEGA, RDKIT, BCL_60, BCL_CSD and BCL_D to BCL in generating conformer closest to native. Each point corresponds to a molecule in a test set. The coordinates of a point corresponds to the RMSD of closest to native conformer generated by BCL (x-axis) and the method being compared (y-axis). Molecules for which closest to native conformation generated by the pair of methods is within 0.25 Å RMSD of each other are plotted in shaded gray area. For points above the shaded region, BCL recovers lower RMSD conformer compared to the other method referenced. The molecules for which OMEGA could not generate conformations are omitted from the graph and statistical analysis when comparing to BCL.

Figures 5 and 6 suggest that BCL is better than other methods and other flavors of BCL being compared.

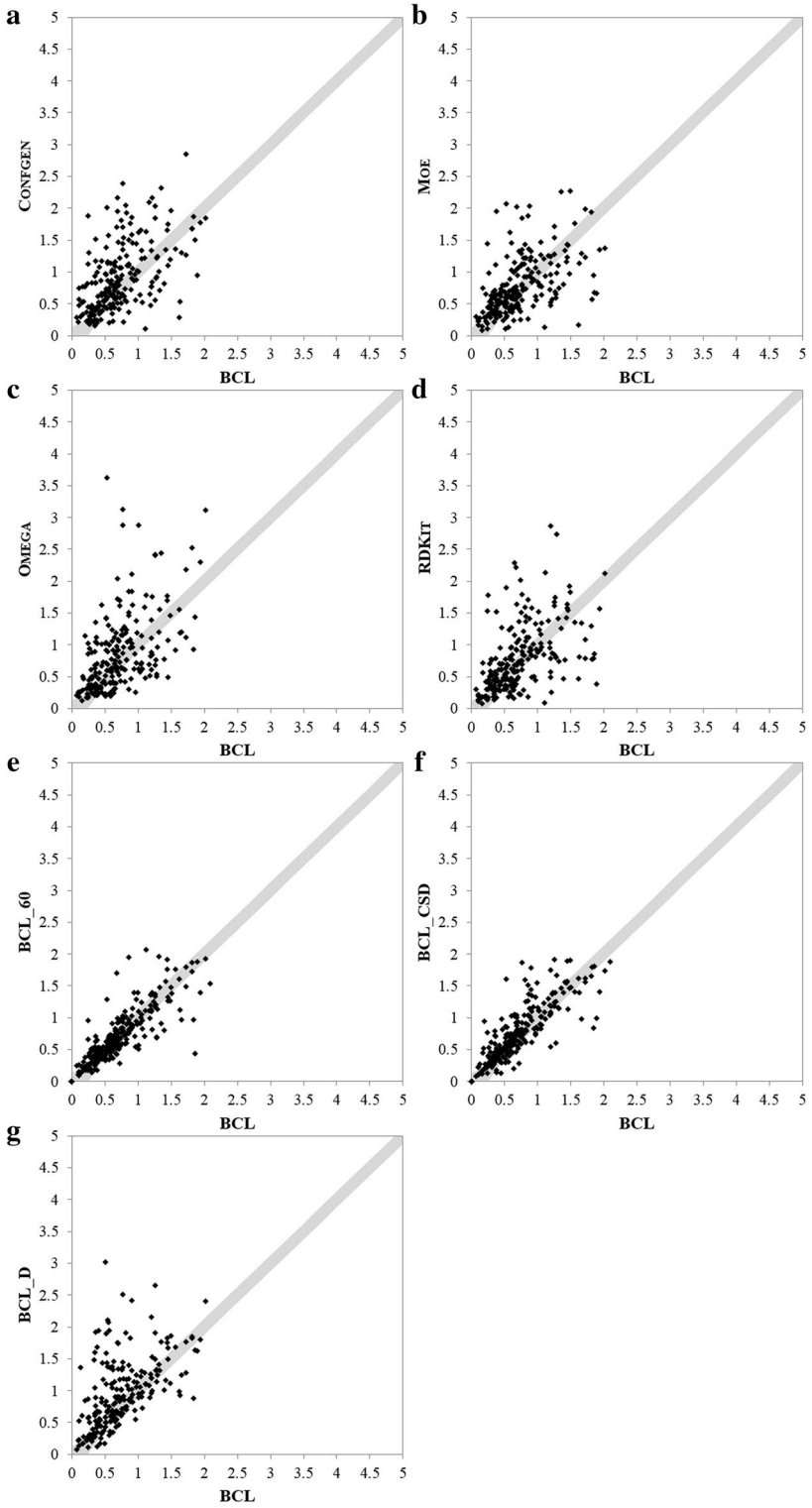


Fig. 6 Pair-wise comparison of BCL::Conf to other methods. **a**–**g** Plot the RMSD to native for the BCL on the x-axis, for other methods or flavors of the BCL on the y-axis. BCL::Conf samples closer to native conformations for points that lie above the diagonal. Conformations plotted within the *shaded region* differ by less than 0.25 Å

Wilcoxon Matched-Pairs Signed-Ranks statistical test was performed to compare conformations generated by BCL to those produced by other methods for each molecule in the Vernalis dataset. The statistics test was performed using R software package. BCL generated closer to native conformations compared to Confgen, Moe, Omega and BCL_D at p value <0.01 over all the molecules. When compared to BCL_CSD, BCL generates more native like conformations at p value <0.05. Statistically there is no significant difference in native recovery between BCL, BCL_60 and RDKit. However, 30° binning allows recapitulation of frequently observed 90° or 270° rotamers of dihedral bonds containing *aromatic-single-aromatic* or *aromatic-single-any* (Additional file 1: Figure S1B).

Effect of the number of rotatable bonds on native conformation recovery

Figure 5b, d show the average RMSD of closest to native conformation of molecules plotted against number of rotatable bonds. Figure S2 (Additional file 1) plots the average number of conformations generated by different methods for molecules of different rotatable bonds. BCL is better than other methods at producing closer to native conformers for molecules with greater than six rotatable bonds as suggested by Wilcoxon Paired test at p value <0.05. For molecules containing four to six rotatable bonds, BCL performs better than Confgen and Omega respectively at p value <0.01. There is no significant difference between quality of conformations generated between BCL, Moe and RDKit for molecules with up to six rotatable bonds. For different flavors of BCL, there is no significant difference between BCL and BCL_60 in native conformation recovery based on rotatable bonds. However, statistical analysis clearly shows that using extended fragments improves native conformation recovery compared to using single dihedral bond statistics (BCL_D) for molecules greater than three rotatable bonds at p value <0.01. BCL produces closer to native conformations compared to BCL_CSD for molecules with greater than 10 rotatable bonds.

Diversity of conformational space sampled

Diversity of ligand conformations is an important consideration for ligand docking studies. A representative sample that covers ligand's sample space is therefore desired. Figure 7a, b show the distribution of RMSDs of conformers against the number of rotatable bonds. Box plots show the distribution of conformer RMSD with respect to native structure. The upper and lower edges of box correspond to the first and third quartiles. The whiskers extend from edge to highest/lowest value that is within 1.5 × Inter-Quartile Range (IQR) of the box, where IQR is the distance between the first and third quartile.

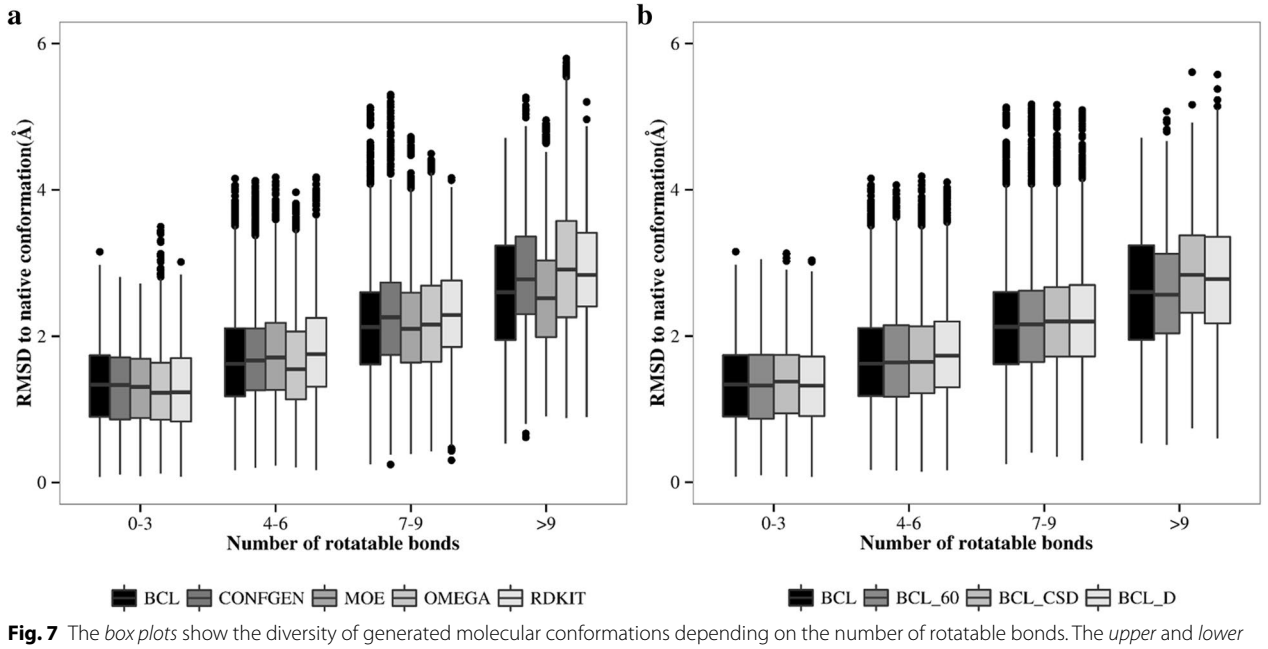


Fig. 7 The *box plots* show the diversity of generated molecular conformations depending on the number of rotatable bonds. The *upper* and *lower edges* of *box* correspond to the first and third quartiles. The *horizontal dash* in the *box* represents the median value. The *whiskers* extend from edge to highest/lowest value that is within 1.5 × Inter-Quartile (IQR) of the *box*, where IQR is the distance between the first and third quartile. **a** Conformation diversity produced by different methods, **b** conformation diversity obtained by using different flavors of BCL

The data beyond whiskers are plotted as outliers. The horizontal dash in the box represents the median value. Diversity of conformations generated by all the methods is comparable. Confgen, Moe and RDKit sample conformations more efficiently compared to BCL for molecules with up to three rotatable bonds (see Additional file 1: Figure S2). The reason is that smaller fragments have large number of rotamers with similar energy profiles. Larger fragments on the other hand have fewer local minima allowing sampling of relevant conformations in fewer steps.

Comparison of CPU time requirements

The computational run time for the different methods except Omega was compared on Intel Xenon model 26 running at 3.2 GHz with 24 Gb of RAM. All the methods take less than 2 Gb of RAM. BCL generated conformations for a single molecule in 1.6 s compared to 1.9 s taken by Confgen, 5.1 s for Moe, 0.5 s for Omega and 10.2 s for RDKit. Computation time of when using only dihedral torsion profiles i.e. BCL_D is 0.7 s/molecule.

Conclusions

We have developed a conformational search method called the BCL::Conf and validated it against other methods in the field like Confgen, Moe, Omega and RDKit. The method utilizes the conformational space seen in the structure databases, CSD and PDB, to sample conformations of small-molecules. BCL::Conf is compared to other methods in three measures which are critical in computational drug discovery process: (a) the ability to generate conformation close to experimentally observed structure, (b) diversity of conformations indication coverage of sample space of molecules, (c) performance in terms of speed. The benchmark study was performed using a curated dataset of high resolution X-ray crystal structures from the PDB, Vernalis datasets, containing 253 molecules.

BCL::Conf is capable of reproducing bioactive conformations generating conformers that are structurally close to experimentally determined structures. Analysis of coverage space shows that BCL::Conf generates a diverse set of conformers performing as well as Moe and RDKit, however in much shorter time. BCL:Conf is better and more efficient in sampling molecules with greater than three rotatable bonds as indicated in Fig. 5b and Figure S4 (Additional file 1). Using extended fragments gives BCL:Conf a distinct advantage over other methods in sampling more flexible molecules efficiently. The study shows utility of using explicit fragment conformations to recapitulate protein-bound ligand conformations. A slightly reduced performance is seen when using rotamers derived from only the CSD (Fig. 5c). The somewhat reduced accuracy could result from biases in the fragment sets between CSB and PDB or biases in dihedral angles between ligands bound to proteins and ligands residing in a crystal. Nonetheless results reported in this paper suggest that fragment conformations obtained from the CSD seen in structure databases can be used to adequately model small molecule conformations bound to proteins.

BCL::Conf extends the idea of protein side-chain conformer sampling to fragments of small molecules. The method is novel as it takes into account torsion correlations and substituents effects on fragment torsion profiles. It has been designed and developed to be integrated with RosettaLigand which is part of the macromolecular modeling suite Rosetta.

Availability and requirements

- Project name: BioChemicalLibrary
- Project home page: http://meilerlab.org/bclcommons
- Operating system(s): Supported on Linux, Apple and Windows
- Programming language: C++
- Other requirements: Access to current CSD license (individual or institutional)
- License: Open source with restrictions, See http://meilerlab.org/bclcommons/license
- Any restrictions to use by non-academics: commercial license needed

Additional files

Additional file 1. Supplementary data and protocol capture describing steps to reproduce data.

Additional file 2. The vernalis dataset molecules without 3D coordinates.

Additional file 3. protocol capture files mentioned in Supplement.docx.

Abbreviations

CADD: computer aided drug discovery/design; CHARMm: Chemistry at HARvard molecular mechanics; Moe: molecular operating environment; CSD: Cambridge structure database; MC: Monte Carlo; MMFF: Merck molecular mechanics force field; PDB: protein databank; RMSD: root mean square deviation; 3D-QSAR: three-dimensional quantitative structure activity relationships.

Authors' contributions

The manuscript was written through contributions of all authors. All authors read and approved the final manuscript.

Author details

[1] Department of Chemistry, Center for Structural Biology, Vanderbilt University, Nashville, TN 37232, USA. [2] Department of Pharmacology and Biomedical Informatics, Vanderbilt University, Nashville, TN 37212, USA.

Acknowledgements

We acknowledge Dr. I Chen and Dr. N Foloppe for the transfer of the Vernalis dataset used in benchmark studies in the current work. We thank Dr. Steven Combs, post-doctoral researcher at Eli Lilly for access to Omega conformation sampling tool.

Compliance with ethical guidelines

Competing interests
The authors declare that they have no competing interests.

Received: 13 May 2015 Accepted: 3 September 2015
Published online: 30 September 2015

References

1. Taylor RD, Jewsbury PJ, Essex JW (2002) A review of protein-small molecule docking methods. J Comput Aided Mol Des 16(3):151–166
2. Sliwoski G, Kothiwale S, Meiler J, Lowe EW Jr (2014) Computational methods in drug discovery. Pharmacol Rev 66(1):334–395
3. Song CM, Lim SJ, Tong JC (2009) Recent advances in computer-aided drug design. Brief Bioinform 10(5):579–591
4. Shim J, Mackerell AD Jr (2011) Computational ligand-based rational design: role of conformational sampling and force fields in model development. Medchemcomm 2(5):356–370
5. Jones G, Willett P, Glen RC (1995) A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. J Comput Aided Mol Des 9(6):532–549
6. Ekins S, Mestres J, Testa B (2007) In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. Br J Pharmacol 152(1):9–20
7. Leach AR, Gillet VJ, Lewis RA, Taylor R (2010) Three-dimensional pharmacophore methods in drug discovery. J Med Chem 53(2):539–558
8. Allen FH (2002) The Cambridge Structural Database: a quarter of a million crystal structures and rising. Acta Crystallogr B 58(Pt 3 Pt 1):380–388
9. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28(1):235–242
10. Vangunsteren WF, Berendsen HJC (1990) Computer-simulation of molecular-dynamics—methodology, applications, and perspectives in chemistry. Angew Chem Int Ed 29(9):992–1023
11. Jorgensen WL, TiradoRives J (1996) Monte Carlo vs molecular dynamics for conformational sampling. J Phys Chem USA 100(34):14508–14513
12. Lagorce D, Pencheva T, Villoutreix BO, Miteva MA (2009) DG-AMMOS: a new tool to generate 3D conformation of small molecules using distance geometry and automated molecular mechanics optimization for in silico screening. BMC Chem Biol 9:6
13. Halgren TA, Bush BL (1996) The Merck molecular force field (MMFF94). Extension and application. Abstr Pap Am Chem Soc 212:2-Comp
14. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M (1983) Charmm—a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 4(2):187–217
15. Klebe G, Mietzner T (1994) A fast and efficient method to generate biologically relevant conformations. J Comput Aided Mol Des 8(5):583–606
16. Brameld KA, Kuhn B, Reuter DC, Stahl M (2008) Small molecule conformational preferences derived from crystal structure data. A medicinal chemistry focused analysis. J Chem Inf Model 48(1):1–24
17. Scharfer C, Schulz-Gasch T, Hert J, Heinzerling L, Schulz B, Inhester T, Stahl M, Rarey M (2013) CONFECT: conformations from an expert collection of torsion patterns. ChemMedChem 8(10):1690–1700
18. Feuston BP, Miller MD, Culberson JC, Nachbar RB, Kearsley SK (2001) Comparison of knowledge-based and distance geometry approaches for generation of molecular conformations. J Chem Inf Comput Sci 41(3):754–763
19. Bush BL, Sheridan RP (1993) Patty—a programmable atom typer and language for automatic classification of atoms in molecular databases. J Chem Inf Comput Sci 33(5):756–762
20. Dunbrack RL (2002) Rotamer libraries in the 21(st) century. Curr Opin Struct Biol 12(4):431–440
21. Davis IW, Baker D (2009) RosettaLigand docking with full ligand and receptor flexibility. J Mol Biol 385(2):381–392
22. Chen IJ, Foloppe N (2008) Conformational sampling of druglike molecules with MOE and catalyst: implications for pharmacophore modeling and virtual screening. J Chem Inf Model 48(9):1773–1791
23. Chen IJ, Foloppe N (2010) Drug-like bioactive structures and conformational coverage with the LigPrep/ConfGen suite: comparison to programs MOE and catalyst. J Chem Inf Model 50(5):822–839
24. Ebejer JP, Morris GM, Deane CM (2012) Freely available conformer generation methods: how good are They? J Chem Inf Model 52(5):1146–1158
25. Wang R, Fang X, Lu Y, Wang S (2004) The PDBbind database: collection of binding affinities for protein–ligand complexes with known three-dimensional structures. J Med Chem 47(12):2977–2980
26. Wang R, Fang X, Lu Y, Yang CY, Wang S (2005) The PDBbind database: methodologies and updates. J Med Chem 48(12):4111–4119
27. Wang R, Lu Y, Fang X, Wang S (2004) An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein–ligand complexes. J Chem Inf Comput Sci 44(6):2114–2125
28. Cahn RS, Ingold C, Prelog V (1966) Specification of molecular chirality. Angewandte Chemie International Edition 5(4):385
29. Krissinel EB, Henrick K (2004) Common subgraph isomorphism detection by backtracking search. Softw Pract Expert 34(6):591–607
30. Dalby A, Nourse JG, Hounshell WD, Gushurst AKI, Grier DL, Leland BA, Laufer J (1992) Description of several chemical-structure file formats used by computer-programs developed at molecular design limited. J Chem Inf Comput Sci 32(3):244–255
31. Watts KS, Dalal P, Murphy RB, Sherman W, Friesner RA, Shelley JC (2010) ConfGen: a conformational search method for efficient generation of bioactive conformers. J Chem Inf Model 50(4):534–546
32. MOE (Molecular Operating Environment) (2015) 2013.08; Chemical Computing Group Inc., Montreal, Canada
33. Hawkins PCD, Skillman AG, Warren GL, Ellingson BA, Stahl MT (2010) Conformer generation with OMEGA: algorithm and validation using high quality structures from the protein databank and Cambridge structural database. J Chem Inf Model 50(4):572–584
34. RDKit documentation (2015) RDKit, Open-Source Cheminformatics. http://www.rdkit.org
35. Bostrom J (2001) Reproducing the conformations of protein-bound ligands: a critical evaluation of several popular conformational searching tools. J Comput Aided Mol Des 15(12):1137–1152
36. Bostrom J, Greenwood JR, Gottfries J (2003) Assessing the performance of OMEGA with respect to retrieving bioactive conformations. J Mol Graph Model 21(5):449–462
37. Perola E, Charifson PS (2004) Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. J Med Chem 47(10):2499–2510
38. Blaney JM, Dixon JS (2007) Distance geometry in molecular modeling. In: Reviews in Computational Chemistry, vol 5, pp 299–335
39. Li J, Ehlers T, Sutter J, Varma-O'Brien S, Kirchmair J (2007) CAESAR: a new conformer generation algorithm based on recursive buildup and local rotational symmetry consideration. J Chem Inf Model 47(5):1923–1932
40. Kirchmair J, Laggner C, Wolber G, Langer T (2005) Comparative analysis of protein-bound ligand conformations with respect to catalyst's conformational space subsampling algorithms. J Chem Inf Model 45(2):422–430
41. Smellie A, Stanton R, Henne R, Teig S (2003) Conformational analysis by intersection: CONAN. J Comput Chem 24(1):10–20
42. O'Boyle NM, Vandermeersch T, Flynn CJ, Maguire AR, Hutchison GR (2011) Confab—systematic generation of diverse low-energy conformers. J Cheminform 3. doi:10.1186/1758-2946-3-8
43. Gasteiger J, Sadowski CRJ (1990) Automatic generation of 3D-atomic coordinates for organic molecules. Tetrahedron Comput Methodol 3(6):537–547
44. Labute P (2010) LowModeMD—implicit low-mode velocity filtering applied to conformational search of macrocycles and protein loops. J Chem Inf Model 50(5):792–800