Conformational Sampling for Large-Scale Virtual Screening: Accuracy versus Ensemble Size

Axel Griewel,[†] Ole Kayser,[†] Jochen Schlosser,[†] and Matthias Rarey*

Center for Bioinformatics, University of Hamburg, Bundesstrasse 43, 20146 Hamburg, Germany

Received July 9, 2009

We introduce the TrixX Conformer Generator (TCG), a novel tool for generating conformational ensembles. The tool addresses especially the requirements of large-scale computer-aided drug design applications using conformer databases. For these, the trade-off between accuracy, i.e. rmsd to biologically active conformers, and database size, i.e. the number of conformers in an ensemble, is of central interest. Based on a tree data structure representing the molecule, conformations are generated incrementally in a best-first-search build-up process employing an internal rmsd clustering. This way TCG builds conformational ensembles of low energy conformers utilizing conformational energy as a scoring function. A crucial parameter is the amount of search space to be covered in the build-up process. This parameter is determined according to an exponential function employing a user-specified quality level as base and an exponent which depends on the molecule's flexibility. The quality level allows the user to set the aforementioned trade-off while taking into account the exponentially growing number of combinations of torsion angles. Tested on a set of 778 molecules, we show that on average 20 conformers per ensemble suffice to achieve an average accuracy of 1.13 Å. We observed that an improvement in accuracy goes along with an exponential rise of the number of conformations per ensemble (e.g., 100 conformations per ensemble yield an accuracy of 0.99 Å). Furthermore, we show that for molecules with less than nine rotatable bonds, ensembles with an average accuracy better than 1 Å can be generated with an average ensemble size of 20 conformers. However, this value deteriorates for more flexible molecules. A comparison to CATALYST and OMEGA shows that TCG achieves a comparable performance in terms of accuracy. Furthermore, it performs well with respect to the trade-off between accuracy and ensemble size.

INTRODUCTION

Conformational flexibility of small, pharmaceutically relevant molecules in computer-aided drug design (CADD) applications can either be handled in an online process during the application or by employing precomputed conformational ensembles containing probable bioactive conformations of the ligand. These ensembles bear the advantage that flexibility has been taken care of and no further online conformational sampling is necessary. In current research, many applications relying on precomputed ensembles can be found that show the relevance of this paradigm.[1,2]

Bioactive conformations often comprise a certain amount of conformational strain.[3−5] Nevertheless, conformations found in experiment are close to local minimal energy conformations of the molecule. An optimal computer-generated ensemble should, therefore, be populated by conformations covering a broad range of low-energy conformational space.[6−8]

First approaches using ensembles describing conformational flexibility of small molecules were undertaken in the 1970s. Here a pharmacophore-like descriptor was used to query a database of crystallographically determined conformations.[9,10] Later, in the 1980s, the computation of a single minimal energy conformer of small molecules was addressed by knowledge-based tools like CONCORD[11] and CORINA.[12] Parallel to the development of these programs first approaches to virtual screening tools were published. These included methods for ligand similarity searching like SEAL,[13] docking tools such as DOCK,[14] and tools querying databases of molecules according to pharmacophore-like constraints like ALADDIN.[15] These tools used molecular conformations determined by X-ray crystallography and lacked the automatic treatment of molecular flexibility. To address this question, programs generating ensembles of low-energy ligand conformations were developed. These ensembles were then used to carry out virtual screening tasks. One of the first programs to follow this paradigm was FLOG, which docks computer-generated conformers from a database to an active site.[16,17] The development of such tools showed the applicability of conformational ensembles in practical applications and led to the implementation of stand-alone tools like ROTATE,[18] OMEGA,[19] and CATALYST,[20] which are able to compute conformational ensembles of molecules given user-specified parameters.

OMEGA utilizes fragment template libraries and histograms to represent torsion-angle-energies. An exhaustive sampling of molecular fragments comprising up to five rotatable bonds is performed, and the molecules are then reassembled by successively choosing fragments with the lowest energy.[21] The algorithm underlying CATALYST makes use of a coarse conformational sampling of the molecule[22,23] and then increases diversity in the set of generated conformers by adding a term to the utilized force

* Corresponding author e-mail: rarey@zbh.uni-hamburg.de.

[†] These authors contributed equally to this work.

10.1021/ci9002415 CCC: $40.75 © 2009 American Chemical Society
Published on Web 09/29/2009

field.[24] In order to speed up calculations, a fragment library can be applied.

A number of structure-based applications relying on conformational ensembles have been presented. Besides the already mentioned DOCK and FLOG approaches, FRED[25] utilizes conformational ensembles to dock ligands. Furthermore, MS-DOCK[26] incorporates both conformer sampling as well as virtual screening. In terms of ligand-based applications, SEAL and ROCS[27] rely on conformational ensembles. All these tools exploit the preprocessing of molecular flexibility by utilizing fast rigid-fitting algorithms. To speed up calculations even further, searching can be supported by index structures. A CADD tool exploiting this fact is TrixX BMI,[28] which allows very fast access to molecules with simultaneous pharmacophore and shape fit.

In this paper we present the TRIXX Conformer Generator (TCG), which is geared to the generation of conformational ensembles for large-scale CADD applications. This includes mainly applications like virtual screening, chemical similarity searching, and pharmacophore matching. These applications have certain requirements to the ensembles in common. On the one hand, the ensembles need to be highly accurate. Here, *accuracy* is measured as rmsd between the conformer in the ensemble which best resembles a bioactive conformation and the bioactive conformations itself. On the other hand, the number of conformers generated in an ensemble has to be as low as possible. This is desirable, since the *number of conformations* (NOC) equals also the number of entities that have to be handled in downstream applications, thus affecting the overall run time. These two goals are contradictory, and an optimal trade-off between accuracy and NOC has to be found.

We analyzed TCG with respect to accuracy using the rmsd to a biologically active structure. For this, we utilized a publicly available test set consisting of 778 druglike molecules bound to their receptors from the PDB.[29] The performance of OMEGA 2.0 and CATALYST 4.11 has been assessed on this test set facilitating a comparison to the results of TCG. Furthermore, we carried out principle investigations on the question of how to determine the trade-off between accuracy and NOC. We also assessed the quality of the generated ensembles with respect to the flexibility of the ligand. To further discuss the aspect of accuracy, we present case studies of ligands which can be found in different conformations in the PDB.

METHODS

FlexX Model. TCG treats molecules according to the FLEXX model:[30] Each atom of a molecule is assigned to a specific component. The set of components is generated by splitting the molecule at each acyclic, nonterminal rotatable bond. Therefore, a component consists of either rigidly connected atoms or atoms within a ring system. Eventually, a tree structure, the so-called *component tree,* is generated: Components are represented as nodes, the connecting bonds as edges.

The conformational search is based on the component tree as well as the MIMUMBA[31] model for torsion angles. Within this model, each rotatable bond outside a ring is assigned to a specific torsion profile. This profile is calculated according to the neighborhood of the considered bond. The pattern of each combination of atoms which determine the dihedral angle of the current bond is used as a query to a local torsion library based on crystallographic data. This query also takes symmetry and periodicity into account. Subsequently, the matching torsion-histograms are retrieved and transformed into potentials. Eventually, the postoptimization routines return the preferred dihedral angles of the given bond in conformational space.

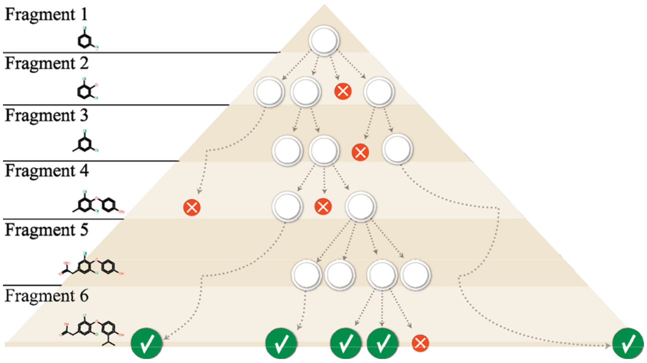

Figure 1. The conformation tree and its traversal for an example molecule during the course of the TCG build-up algorithm.

The connectivity of the component tree, which represents the fragmentation of the molecule, and the torsion angles assigned to each bond yield the basis for the TCG *conformation tree* (see Figure 1). From this data-structure conformations of the molecule are constructed successively by traversing the tree starting with the molecule's central component as root. One component after another is added using the previously assigned torsion angles. Ring components with a cycle length less than eight are expanded using the respective torsion angle plus precomputed CORINA[12] ring conformations. Each valid torsion angle is represented as an edge in the conformation tree. If component C_i is connected via bond b with n different torsion angles to component C_j, this results in n edges from node C_i to n nodes C_{jk} $(1 \leq k \leq n)$. Thus, each C_{jk} represents a different conformation of the partial molecule in 3D space. Leaf nodes represent conformations of the complete molecule and thereby valid solutions. Inner nodes correspond to partially built-up molecules in certain conformations. The actual conformation of a [partial] molecule represented by a node can be determined by traversing the path from that node to the root-node in the conformation tree using the specific torsion angles assigned to the edges on the path.

Sampling. Within the TCG conformational sampling algorithm, the most central component of the molecule is chosen as root of the conformation tree. This choice is based on the strong influence central, rotatable bonds have on the overall molecular shape. TCG performs a best-first-search on the conformation tree. In each step, the partial solution performing best according to the scoring function is selected for expansion, similar to the strategy of the A*-based sampling algorithm.[32] During expansion, the next component is placed using the precomputed torsion angles and, in case of ring systems, also the precomputed ring conformations. The conformational energy calculated using the Tripos force field[33] is employed as scoring function to select the next node for expansion and to guide the search toward local energy minima. Since TCG needs to evaluate scores for partial molecule structures, only already built-up parts of the

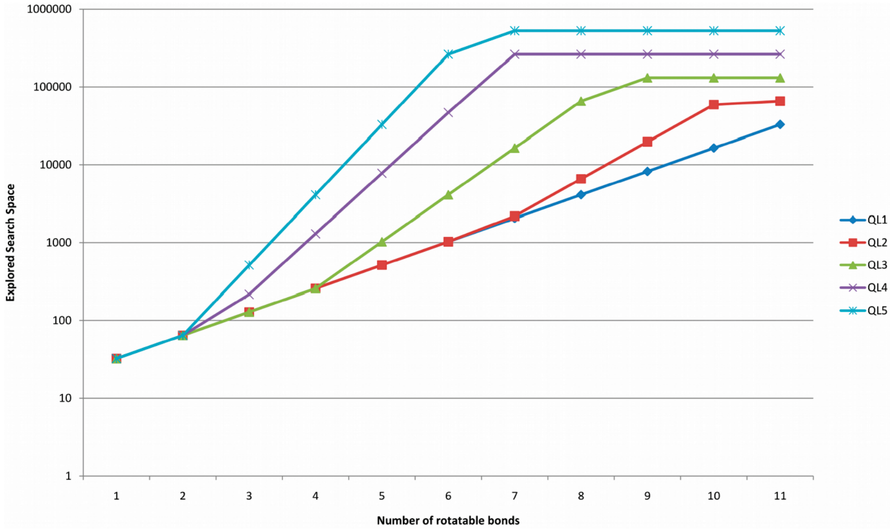

Figure 2. *Explored search space* as a function of a molecule's flexibility and quality level (QL). Quality level 1 (dark blue) is used as a lower bound.

Table 1. Specification of the Base Exponent for the Different Quality Levels[a]

Quality level (q)	1	2	3	4	5
b_q	2	3	4	6	8

[a] In combination with the function f_{ESS} (k,q) this determines the maximum amount of search spaced to be explored by TCG.

molecule are used as input to the regular force field. The score contribution of the remaining, nonplaced components is evaluated using a heuristic energy function: In case of ring systems, TCG chooses the energy of the lowest energy conformation. Rigid components account with their internal energy as calculated by the force field. The final TCG score of a [partial] molecule consists of the sum of these terms. Since some ligands also bind in compact conformations[6,34] and TCG aims at covering all bioactive conformations, we decided against a special scoring term which treats the radius of gyration of the molecule.[7,35]

Stereochemistry. In general, the TCG methodology allows the generation of different isomeric forms of a molecule: Within ring systems CORINA is able to supply different starting points for the conformational search. In all other cases the FlexX model allows to enumerate stereoisomers. Nevertheless, the default setting is to generate just one isomeric form, either preserving the user-supplied 3D structure or, if none is available, the CORINA configuration that is generated using the no-stereo flag.

Size Thresholds and Depth Probes. Since the search space grows exponentially with the number of components, TCG supplies different quality levels to control the granularity of the sampling. Depending on the downstream modeling task, TCG can be adapted to generate high-quality ensembles with many conformations per ensemble as well as compact representations which cover the molecule's flexibility reasonably well.

Quality levels determine the amount of conformations produced and the overall run time by constraining the amount of *explored search space* (*ESS*) (see Figure 2). The idea behind the *ESS* constraint is based on the exponential growth of the total search space depending on the number of rotatable bonds k of the molecule. In combination with a user-defined quality level q, TCG deploys the following formula

$$f_{ESS}(k,q) = \min\{2^{q+14}, \max\{b_q^k, 2^{k+5}\}\}$$

At most f_{ESS} (k,q) nodes of a molecule's conformation tree are expanded during the search, with b_q defining the base of the exponential function according to the desired quality level (see Table 1). A larger quality level results in a larger b_q value leading to higher accuracy and more conformations in the ensemble. Correspondingly, if b_q is smaller, fewer conformations are generated by trading off accuracy. Furthermore, a certain minimal ESS is guaranteed using 2^{k+5} as lower bound. This mechanism allows a fine-grained control and enables the user to specify the quality/quantity trade-off concerning accuracy and NOC.

In analogy to the *ESS* constraint, which limits the search space, TCG also employs a constraint guaranteeing a *minimum number of conformers* (*MNC*) produced before clustering (see Figure 3) as described later. Again, this number depends on the molecule's number of rotatable bonds k and the desired quality level q. It is calculated as follows

$$f_{MNC}(k,q) = \min\{2^{k+2}, 2^{q+4}\}$$

The *MNC*-limit outweighs the *ESS*-constraint. This constraint hierarchy ensures that a certain minimum number of solutions is found before the *ESS* constraint is enforced and the calculation concludes. This only holds for cases in which the conformation tree is not fully traversed. If the traversal concludes, no more conformations can be generated, and the TCG result corresponds to all conformations that have been found so far.

While expanding the conformation tree, TCG uses so-called depth-probes to generate fully built-up molecules. With a depth probe frequency (*DPF*) depending on the molecule's flexibility, ranging from every 100th to every 30th expansion, the best-first-search is converted to a depth-first search of the currently best scored partial molecule. This partial solution is then expanded in a depth-first-manner until either a clash occurs or a complete molecule is generated. Subsequently, TCG reverts to best-first-search. As a result of this

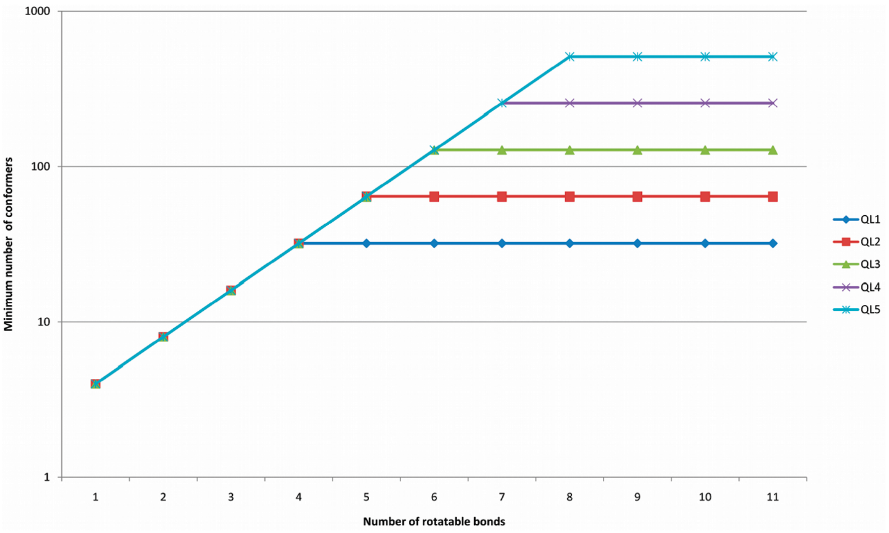

Figure 3. *Minimum number of conformer* as a function of a molecule's flexibility and quality level (QL) before clustering.

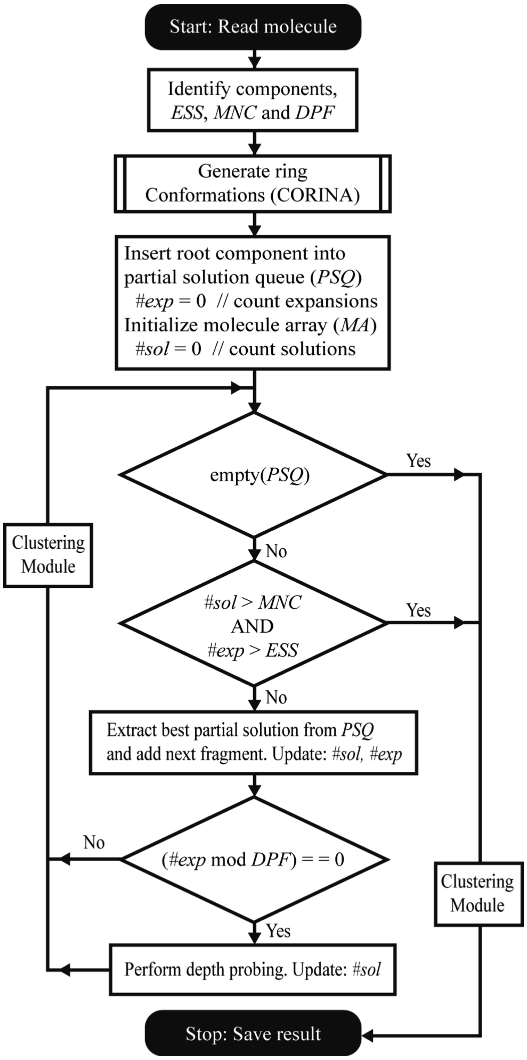
Start: Read molecule
Identify components, ESS, MNC and DPF
Generate ring Conformations (CORINA)
Insert root component into partial solution queue (PSQ)
#exp = 0 // count expansions
Initialize molecule array (MA)
#sol = 0 // count solutions
empty(PSQ)
Yes
No
#sol > MNC AND #exp > ESS
Yes
No
Extract best partial solution from PSQ and add next fragment. Update: #sol, #exp
(#exp mod DPF) = = 0
No
Yes
Perform depth probing. Update: #sol
Clustering Module
Clustering Module
Stop: Save result

Figure 4. Flowchart of the TCG sampling algorithm.

heuristic, TCG generates valid solutions from various starting points during the course of the algorithm. A summary of the TCG sampling strategy is illustrated in Figure 4.

Clustering. Large search spaces lead to large conformational sets. Therefore, TCG performs clustering already during the search process using a window of fixed size to reduce the number of conformers. Instead of guaranteeing diversity within the whole conformational set, only sets of 3000 conformations are subject to clustering. Besides a reduction of memory and run time requirements, this intermediate online clustering procedure ensures a reasonably sized conformer set which is clustered once the algorithm concludes. Notably, the *MNC* constraint is enforced independently of the clustering procedure: It is based on the number of generated conformations prior to clustering. In addition to the internal clustering procedure, another interactive clustering can be performed. This algorithm allows for clustering using an rmsd-diversity criterion or an upper NOC bound.

The default value used for the online clustering procedure is 0.8 Å rmsd. [All rmsd-values throughout this paper are symmetry-corrected and rely solely on heavy atom distances.] This value can be adapted in order to control the granularity of the results: Low clustering thresholds yield larger conformational sets and lower accuracy, while higher values decrease the size of the set by compromising on accuracy.

RESULTS

In a first step we analyze TCG's capabilities to reproduce conformations found in the Cambridge Structural Database (CSD).[36] For this purpose we utilize a previously published subset[37] of the CSD consisting of approximately 71,200 high quality structures of molecules containing only H, C, N, O, S, and halogens. Furthermore, we filtered this set by applying the Oprea leadlike criteria[38] and discarding molecules containing crowded ring systems yielding a final set-size of 43,047 structures. Conformational ensembles were then generated for each molecule using quality level 1 with a clustering threshold of 1.2 Å. An average accuracy of less than 1.0 Å was achieved in this test showing, that TCG is able to reproduce conformations found in the CSD within small error bound.

TCG is then evaluated using a publicly available test set[29] consisting of 778 druglike ligands from protein–ligand complexes retrieved from the PDB in a resolution mostly better than 2.5 Å. Differing from the described protocol,[29] we downloaded the ligands in SDF format from the ligand expo database.[39] The downloaded SDF-files were converted

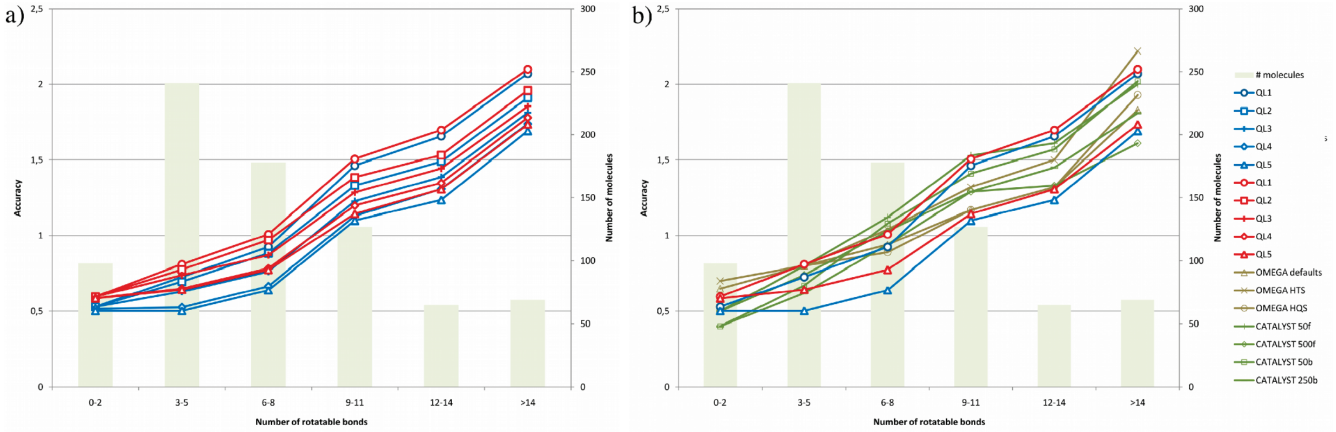

Figure 5. Relationship between flexibility of the molecule and accuracy. The diagrams are separated for clarity: a) shows only results for TCG (green bars: number of molecules with a specific number of rotatable bonds; blue lines: cluster threshold 0.8 Å; red lines: cluster threshold 1.2 Å; quality levels (QL) increasing from 1 to 5: circle, square, cross, diamond, triangle). Diagram b) displays a comparison to the tools OMEGA (parameter settings "defaults", "high throughput screening", and "high quality screening") as well as CATALYST. For the latter the number indicates the final number of conformers, while the letter indicates if the BEST or FAST algorithm was used. For a better overview, this diagram comprises only the results of TCG in the highest and lowest quality level for each clustering threshold. Data for OMEGA and CATALYST are taken from a comparative study.[29]

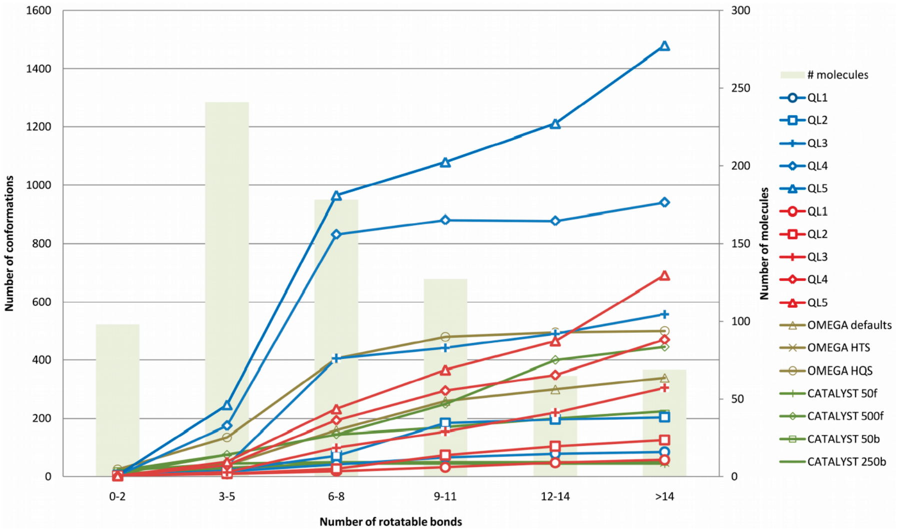

Figure 6. Relationship between flexibility of the molecule and NOC; green bars: number of molecules with a specific number of rotatable bonds; blue: cluster threshold 0.8 Å; red: cluster threshold 1.2 Å; quality levels (QL) increasing from 1 to 5: circle, square, cross, diamond, triangle. Data for OMEGA and CATALYST taken from a comparative study.[29]

to MOL2-fomat using CORINA. The input structures for TCG were CORINA-generated structures, while reference conformations were produced by using CORINA as a file-format-converter, not changing the spatial arrangement of the atoms in the complex. The molecules in the test set show an average flexibility of seven rotatable bonds. Furthermore, 605 of the molecules comprise less than 11 rotatable bonds and, therefore, fulfill the Oprea flexibility criterion[38] for lead structures. The distribution of molecular flexibility within the test data set assessed by the number of rotatable bonds is shown in Figure 5.

In a first step, we analyzed the essential properties accuracy and NOC separated by the number of rotatable bonds as displayed in Figures 5 and 6. As expected, both average NOC and accuracy rise with the molecule's number of rotatable bonds. For molecules with up to 8 rotatable bonds on average, high quality ensembles with an accuracy of less than 1.0 Å were generated; however, for more flexible molecules this value deteriorates (see Figure 5). The NOC rises linearly with the number of rotatable bonds of the molecule when using a clustering threshold of 1.2 Å (see Figure 6). For molecules with a relatively small number of rotatable bonds, this yields ensembles covering the conformational space very well. For more flexible molecules, more conformers are needed to adequately represent the accessible conformational space. When analyzing the high quality setting with a clustering threshold of 0.8 Å, it can be seen that the average number of conformations per molecule rises strongly for molecules comprising 6−8 rotatable bonds. For molecules with more than eight rotatable bonds the developed ESS bounds prevent the so-called exponential explosion and, therefore, creates the plateau found in the diagram for quality levels four and five.

Figure 6 displays the percentage of ensembles with an accuracy below a certain threshold for TCG as well as for OMEGA and CATALYST.[29] For each of these tools, two lines are shown. Upper and lower lines of equal color correspond to one of the aforementioned programs. The upper line displays the parameter set yielding the best performance with respect to accuracy, while the lower line

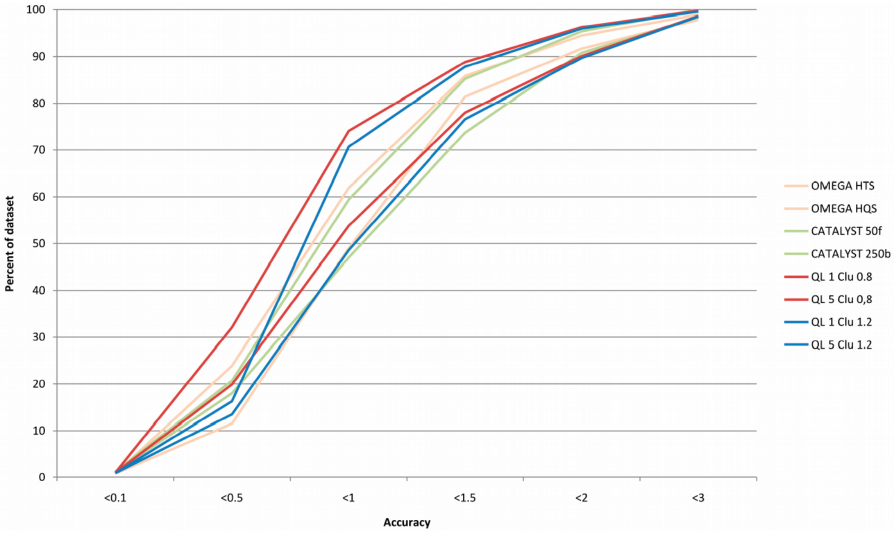

Figure 7. Accuracy-plot. Abscissa: rmsd to the biologically active structure. Ordinate: Percent of test set sampled with an accuracy below the value indicated on the abscissa.

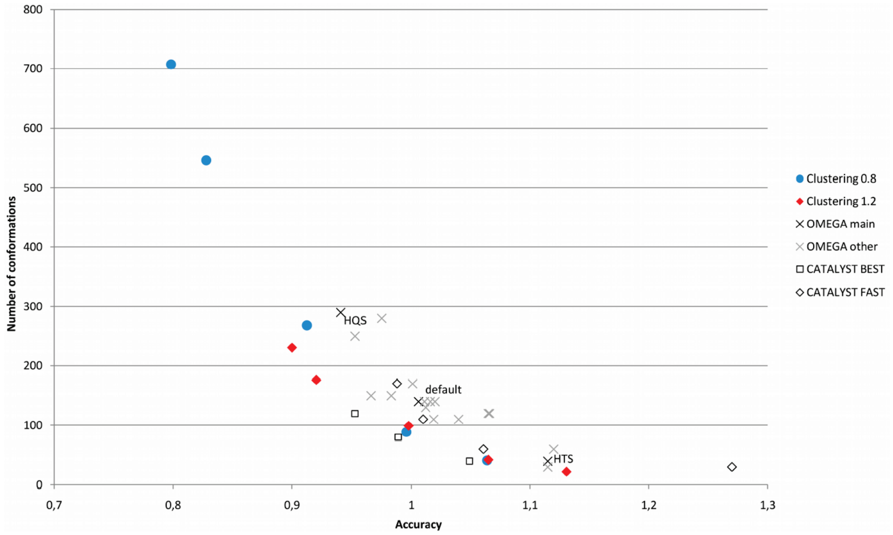

Figure 8. Relation between average accuracy and the average number of conformations per ensemble for the complete data set. Red diamonds and blue circles indicate the performance of the TRIXX Conformer Generator in the five quality levels with respective clustering threshold of 1.2 Å and 0.8 Å. Data for OMEGA are displayed as crosses: previously mentioned settings are shown in the darker color, labeled explicitly, and referred to as "OMEGA main" in the legend; data points of "OMEGA other" display results for additional settings. CATALYST results are displayed distinguishing between FAST and BEST.[29]

indicates the parameter set yielding the least favorable performance.

Success and failure of the conformer sampling are discriminated by the resulting accuracy. Cases in which the rmsd is larger than 2 Å are generally considered a failure. For molecules with an rmsd between 1.5 Å and 2.0 Å, the overall structure of the conformer is usually close to the bioactive conformation, while structural details may differ significantly. An rmsd below 1.5 Å indicates an acceptable reproduction of the conformer, while rmsds below 1 Å are considered good fits between generated and biologically active conformer.

As can be seen in Figure 7, all conformer generators yield good conformational ensembles for more than 45% of the considered molecules in all settings. The most accurate generator at 1 Å is TCG in its highest quality setting yielding 74% conformational ensembles which contain a conformer closer than 1 Å to the bioactive conformation. This, however, bears the cost of an increased amount of conformations. All sampling modes generating small NOC numbers show similar behavior. Here, roughly 50% of the ligands are sampled with an accuracy better than 1 Å rmsd. Furthermore, the area between the upper and lower curves in quality level one and five is large for TCG, which shows that the detail of sampling can be specified in fine granularity by the quality levels.

The relation between accuracy and the number of generated conformers is assessed in Figure 8. On the abscissa of the diagram the average accuracy is plotted, while the ordinate shows the average number of conformations for the complete data set. For each run of a program with a specific parameter set, one dot was created in the diagram. The exponential relationship between accuracy and NOC can easily be observed from the diagram. For most applications ensembles with high accuracy, i.e. low rmsds, and few conformations are desired. Thus, an optimal trade-off can be determined by comparing the gain in accuracy to the number of additionally generated conformations. For more

Table 2. rmsds between Conformers in the Case Study Test Set and the Generated Ensembles in All Quality Levels Using a Clustering Threshold of 0.8 Å[a]

OHT	2bj4	2gpu	3ert
2bj4	0	1.17	1.02
2gpu		0	1.33
3ert			0
TCG ensemble QL1	0.76	0.61	0.75/38
TCG ensemble QL2	0.38	0.47	0.64/100
TCG ensemble QL3	0.38	0.47	0.46/841
TCG ensemble QL4	0.38	0.47	0.56/1124
TCG ensemble QL5	0.38	0.47	0.55/1200

IMN	1s2a	2dm6	2zb8	3fo7
1s2a	0	2.28	0.90	1.12
2dm6		0	2.23	2.35
2zb8			0	1.13
3fo7				0
TCG ensemble QL1	0.46	1.51	0.89	1.21/9
TCG ensemble QL2	0.46	1.14	0.89	1.06/12
TCG ensemble QL3	0.46	1.14	0.89	1.06/12
TCG ensemble QL4	0.45	1.07	0.86	1.02/22
TCG ensemble QL5	0.46	0.52	0.69	0.65/114

STI	1t46	1xbb	2pl0	3fw1
1t46	0	2.69	0.89	2.56
1xbb		0	2.62	1.34
2pl0			0	2.58
3fwd				0
TCG ensemble QL1	1.90	1.15	1.81	1.04/30
TCG ensemble QL2	1.22	1.13	1.57	0.89/94
TCG ensemble QL3	1.22	1.15	1.39	0.89/202
TCG ensemble QL4	0.97	0.99	1.08	0 89/495
TCG ensemble QL5	0.50	0.98	0.69	0.82/963

[a] The number of conformations for the generated ensembles is given in the last column behind the slash.

than about 100 conformations in an ensemble, the accuracy rises more slowly than for smaller NOCs. Therefore, those parametrizations resulting on average in less than 100 conformations per ensemble achieve already optimal results with respect to the trade-off between NOC and accuracy. With regard to CADD applications utilizing databases of conformational ensembles, however, one might want to lower the NOC even further to minimize the amount of entities to be handled downstream. For the setup resulting in the lowest NOC (quality level 1 with a clustering threshold of 1.2 Å), the averaged accuracy is already in the acceptable range of 1 Å. For comparison, we added the values for OMEGA and CATALYST to Figure 8. In most cases TCG reaches a very good ratio of accuracy to NOC. Only CATALYST BEST yields conformational ensembles which are comparable to those generated with TCG with respect to this number.

We studied the quality of the generated conformers in detail using three ligands and the corresponding crystal structures from the PDB as case studies. The first ligand is 4-hydroxy-tamoxifen (Ligand-Expo ID OHT), which comprises 8 rotatable bonds as well as three ring systems and is present in nine PDB structures. We restricted this set further by selecting only structures with a resolution better than 2 Å and discarded all those structures having an rmsd below 0.5 Å to any of the other included conformers. This guarantees high quality structures while maximizing conformational diversity resulting in three conformers in total. The contained PDB-IDs as well as the intrinsic diversity can be found in Table 2. Furthermore, Table 2 contains accuracy and NOC for conformational ensembles generated in quality level one and five, using a cluster threshold of 0.8 Å. For all examples in the test set, a conformation with accuracy below 0.8 Å was found, making the ensemble a good approximation of the experimentally determined conformers. In Figure 9a) a superposition of the crystal structure as found in PDB entry 2gpu,[40] and the best match of the highest and lowest quality ensemble is displayed. The conformation of the three rings on the right-hand side of the figure in quality level one is already very similar to the characteristic binding mode. The alignment of the aliphatic part of the molecule is less precise but improves largely in quality level five, yielding an rmsd of 0.47 Å. This accuracy is already reached in quality level 2.

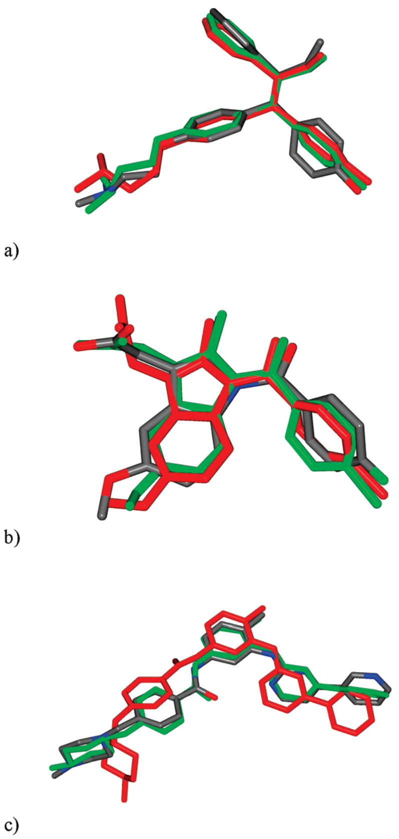

Figure 9. Superposition of the closest generated conformer and the crystal structure for (a) a tamoxifen derivate (OHT) as found in PDB-structure 2gpu, (b) indomethacin (IMN) from 2zb8, and (c) Imatinib (STI) as found in 1t46. CPK colored structure: crystal structure; red structure: quality level one clustering 0.8 Å; green structure: quality level 5, clustering 0.8 Å.

In a second case we investigated the binding modes of indomethacin (IMN). This molecule comprises four rotatable bonds and is contained in 11 PDB structures. After applying the aforementioned quality- and diversity threshold, 4 conformers are left, which show pairwise rmsds between 0.9 Å to 2.35 Å. Ensembles generated in quality levels one and five contain conformers resembling the crystal structure by an average rmsd of 0.95 Å, respectively 0.55 Å (Table 2). Figure 9b) shows the superposition analogous to the previous example for PDB entry 2zb8.[41] In this example, the para-substituted phenyl-ring is well aligned to the crystal structure even in the low quality level. In the high quality level (green), also the methoxy group as well as the carboxylate align well to the crystal structure. The analysis of the results in different

Table 3. Run Times of TCG with Different Parameters

quality level	cluster threshold	time (s)	cluster threshold	time (s)
1	0.8	5.6	1.2	5.2
2	0.8	15.8	1.2	12.8
3	0.8	70.2	1.2	42.8
4	0.8	188.0	1.2	113.6
5	0.8	293.3	1.2	201.6

quality levels shows that the conformations found in 1s2a and 1zb8 are resembled very well by conformers generated in quality level one. For the other two ligands the conformers generated in quality level one resemble the conformation only coarsely, but the accuracy improves here significantly in higher quality levels.

In the third case we studied imatinib (STI), which comprises nine rotatable bonds and is present in 8 PDB structures. Four of these structures comply with our requirements yielding a test set with intrinsic rmsds between 0.89 Å and 1.34 Å (Table 2). Ensembles generated in quality level 1 reflect only the coarse shape of the bound ligand with rmsd values between 1.15 Å to 1.90 Å, while the ensembles generated in quality level 5 resemble the structures quite well with rmsds between 0.5 Å and 0.98 Å. Similar to the increase in accuracy, the NOC in this case also rises from 30 to 963 conformers in the ensemble. In Figure 9c) a superposition of the generated structure closest to the conformation as found in 1t46[42] is shown. This conformer binds to the c-Kit Tyrosine Kinase forming three hydrogen bonds at the pyridine-nitrogen, the aminopyrimidine nitrogen, and the carbonyl oxygen. Besides these interactions, the binding site forms a hydrophobic tunnel. The superposition of the conformers displays that the best structure found in the conformational ensemble in quality level one differs significantly not only in the orientations of the terminal rings but also in the orientation of the peptide bond. In the high quality ensemble, however, a structure with an rmsd of 0.5 Å is found, which resembles the spatial alignment of the atoms very well and has also all hydrogen-bond partners directed correctly. The results for different quality levels show improvement in accuracy in all four cases. This is due to the high flexibility of the molecule, which is not exhaustively explored in low quality levels but better covered in higher quality levels.

The performance of the presented algorithm with respect to time was assessed on single 2.4 GHz Xeon CPUs with 4 Gbyte of main memory. In Table 3 the average run time of the test for the computation of the conformational ensembles can be found. It ranged from 5.2 s in the lowest quality setting to 293.3 s in the highest quality setting. The major reason for the increase in computing time using higher quality levels is the enlarged search space. The internal clustering threshold has also an impact on the run time, since it influences the number of generated conformers. The performance of CATALYST and OMEGA on the test set has been assessed on Intel Pentium IV 2.8 GHz workstations with 1GB RAM.[29] The average run times of OMEGA ranged from 6.0 to 12.9 s, while those of CATALYST were reported to be in the range of 1.5 s for the FAST algorithm to 155.0 s for the BEST algorithm. Taking the similar setups into account, the run times for low quality settings of TCG lie in the same range as those of OMEGA and CATALYST FAST, while the high quality configurations require more time.

CONCLUSION

We presented the TRIXX Conformer Generator for conformational sampling of druglike molecules. The underlying algorithm utilizes a best-first-search on a tree representation of the molecule. In addition, complete conformations are generated in a regular manner leading to a reasonable coverage of the conformational space. Five different quality levels limiting the exploration of the entire search space and an online clustering threshold are available. These enable the user to determine the trade-off between accuracy and the size of the ensembles.

TCG was analyzed on a publicly available test set containing 778 ligands. On this test set an average of 100 conformations per ensemble suffices to produce ensembles with an average accuracy of 1 Å. When restricting the test set to molecules which comprise less than nine rotatable bonds, even less than 20 conformers are necessary to achieve similar results. A comparison to the tools OMEGA and CATALYST revealed that TCG produces ensembles with comparable accuracy. Furthermore, TCG performs very well with respect to the trade-off between the number of conformers per ensemble and accuracy; i.e. a low amount of conformations suffices to generate accurate ensembles. This property is essential for downstream high-throughput CADD applications intending to utilize conformational ensembles as representations of a flexible molecule.

Our future research in the field of conformational sampling will include improvements in the rule-based chemical model as well as in the implementation of different force fields to guide the best-first-search. The resulting conformational ensembles are then going to be utilized in applications like small molecule docking, in which ensembles of TCG have already been used productively.[28] Furthermore, we are about to investigate the applicability in ligand-based virtual screening and 3D-pharmacophore search.

REFERENCES AND NOTES

(1) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49* (20), 5912–5931.

(2) Moitessier, N.; Englebienne, P.; Lee, D.; Lawandi, J.; Corbeil, C. R. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br. J. Pharmacol.* **2008**, *153* (1), S7–26.

(3) Mobley, D. L.; Dill, K. A. Binding of small-molecule ligands to proteins: "what you see" is not always "what you get". *Structure* **2009**, *17* (4), 489–498.

(4) Bostrom, J.; Norrby, P. O.; Liljefors, T. Conformational energy penalties of protein-bound ligands. *J. Comput.-Aided Mol. Des.* **1998**, *12* (4), 383–396.

(5) Nicklaus, M. C.; Wang, S.; Driscoll, J. S.; Milne, G. W. Conformational changes of small molecules binding to proteins. *Bioorg. Med. Chem.* **1995**, *3* (4), 411–428.

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

(7) 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.

(8) Wang, Q.; Pang, Y. P. Preference of small molecules for local minimum conformations when binding to proteins. *PLoS One* **2007**, *2* (9), e820.

(9) Gund, P. Pharmacophoric Pattern Searching and Receptor Matching. *Annu. Rep. Med. Chem.* **1979**, *14*, 299–308.

(10) Gund, P.; Wipke, W. T.; Langridge, R. In *Computer searching of a molecular structure file for pharmacophoric patterns*, International Conference on Computers in Chemistry Research and Education, Ljubljana, 1974, 1973; Hadzi, D., Zupan, J., Eds.; Elsevier: Amsterdam, Ljubljana, 1973; pp 5/33−5/38.

(11) Pearlman, R. S. Rapid generation of high quality approximate 3D molecular structures. *Chem. Des. Autom. News* **1987**, *2* (1), 5–7.

(12) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Comput. Methodol.* **1990**, *3* (6, Part 3), 537–547.

(13) Kearsley, S. K.; Smith, G. M. An alternative method for the alignment of molecular structures:Maximizing electrostatic and steric overlap. *Tetrahedron Comput. Methodol.* **1990**, *3* (6), 615–633.

(14) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161* (2), 269–288.

(15) Van Drie, J. H.; Weininger, D.; Martin, Y. C. ALADDIN: an integrated tool for computer-assisted molecular design and pharmacophore recognition from geometric, steric, and substructure searching of three-dimensional molecular structures. *J. Comput.-Aided Mol. Des.* **1989**, *3* (3), 225–251.

(16) Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P.; Miller, M. D. Flexibases: a way to enhance the use of molecular docking methods. *J. Comput.-Aided Mol. Des.* **1994**, *8* (5), 565–582.

(17) Miller, M. D.; Kearsley, S. K.; Underwood, D. J.; Sheridan, R. P. FLOG: a system to select ‘quasi-flexible’ ligands complementary to a receptor of known three-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8* (2), 153–174.

(18) *ROTATE*; Molecular Networks: Erlangen, Germany, 2009.

(19) *OMEGA*; OpenEye Scientific Software, Inc.: Santa Fe, NM, U.S.A., 2009.

(20) *Catalyst*; Accelrys: San Diego, CA, U.S.A, 2009.

(21) Leach, A. R.; Smellie, A. S. A combined model-building and distance-geometry approach to automated conformational analysis and search. *J. Chem. Inf. Model.* **1992**, *32* (4), 379–385.

(22) Smellie, A.; Kahn, S. D.; Teig, S. L. Analysis of Conformational Coverage. 2. Applications of Conformational Models. *J. Chem. Inf. Model.* **1995**, *35* (2), 295–304.

(23) Smellie, A.; Kahn, S. D.; Teig, S. L. Analysis of Conformational Coverage. 1. Validation and Estimation of Coverage. *J. Chem. Inf. Model.* **1995**, *35* (2), 285–294.

(24) Smellie, A.; Teig, L. T.; Towbin, P. Poling: Promoting Conformational Variation. *J. Comput. Chem.* **1995**, *16* (2), 171–187.

(25) *FRED*; OpenEye Scientific Software, Inc.: Santa Fe, NM, U.S.A., 2009.

(26) Sauton, N.; Lagorce, D.; Villoutreix, B. O.; Miteva, M. A. MS-DOCK: accurate multiple conformation generator and rigid docking protocol for multi-step virtual ligand screening. *BMC Bioinf.* **2008**, *9*, 184.

(27) *ROCS*; OpenEye Scientific Software, Inc.: Santa Fe, NM, U.S.A., 2009.

(28) Schlosser, J.; Rarey, M. Beyond the Virtual Screening Paradigm: Structure-Based Searching for New Lead Compounds. *J. Chem. Inf. Model.* **2009**, *49* (4), 800–809.

(29) 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), 1848–1861.

(30) Rarey, M.; Wefing, S.; Lengauer, T. Placement of medium-sized molecular fragments into active sites of proteins. *J. Comput.-Aided Mol. Des.* **1996**, *10* (1), 41–54.

(31) Klebe, G.; Mietzner, T. A fast and efficient method to generate biologically relevant conformations. *J. Comput.-Aided Mol. Des.* **1994**, *8* (5), 583–606.

(32) Leach, A. R.; Prout, K. Automated conformational analysis: Directed conformational search using the A* algorithm. *J. Comput. Chem.* **1990**, *11* (10), 1193–1205.

(33) Clark, M.; Cramer III, R. D.; van Opdenbosch, N. Validation of the General Purpose Tripos 5.2 Force Field. *J. Comput. Chem.* **1989**, *10* (8), 982–1012.

(34) Stockwell, G. R.; Thornton, J. M. Conformational diversity of ligands bound to proteins. *J. Mol. Biol.* **2006**, *356* (4), 928–944.

(35) Diller, D. J.; Merz, K. M. J. Can we separate active from inactive conformations. *J. Comput.-Aided Mol. Des.* **2002**, *16* (2), 105–112.

(36) Allen, F. H.; Davies, J. E.; Galloy, J. J.; Johnson, O.; Kennard, O.; Macrae, C. F.; Mitchell, E. M.; Mitchell, G. F.; Smith, J. M.; Watson, D. G. The development of versions 3 and 4 of the Cambridge Structural Database System. *J. Chem. Inf. Comput. Sci.* **2002**, *31* (2), 187–204.

(37) Maass, P.; Schulz-Gasch, T.; Stahl, M.; Rarey, M. Recore: a fast and versatile method for scaffold hopping based on small molecule crystal structure conformations. *J. Chem. Inf. Model.* **2007**, *47* (2), 390–399.

(38) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is there a difference between leads and drugs? A historical perspective. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (5), 1308–1315.

(39) Feng, Z.; Chen, L.; Maddula, H.; Akcan, O.; Oughtred, R.; Berman, H. M.; Westbrook, J. Ligand Depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics* **2004**, *20* (13), 2153–2155.

(40) Wang, L.; Zuercher, W. J.; Consler, T. G.; Lambert, M. H.; Miller, A. B.; Orband-Miller, L. A.; McKee, D. D.; Willson, T. M.; Nolte, R. T. X-ray crystal structures of the estrogen-related receptor-gamma ligand binding domain in three functional states reveal the molecular basis of small molecule regulation. *J. Biol. Chem.* **2006**, *281* (49), 37773–37781.

(41) Wu, Y. H.; Ko, T. P.; Guo, R. T.; Hu, S. M.; Chuang, L. M.; Wang, A. H. Structural basis for catalytic and inhibitory mechanisms of human prostaglandin reductase PTGR2. *Structure* **2008**, *16* (11), 1714–1723.

(42) Mol, C. D.; Dougan, D. R.; Schneider, T. R.; Skene, R. J.; Kraus, M. L.; Scheibe, D. N.; Snell, G. P.; Zou, H.; Sang, B. C.; Wilson, K. P. Structural basis for the autoinhibition and STI-571 inhibition of c-Kit tyrosine kinase. *J. Biol. Chem.* **2004**, *279* (30), 31655–31663.

CI9002415