Assessment of Conformational Ensemble Sizes Necessary for Specific Resolutions of Coverage of Conformational Space

Yulia V. Borodina, Evan Bolton, Fabien Fontaine, and Stephen H. Bryant*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland 20894

Received March 12, 2007

The size of conformational ensembles required for regular coverage of the conformational space of druglike molecules was examined. Using the conformer generation program Omega, the number of regularly distributed conformers (NRC) of flexible compounds was determined as a function of the root-mean-square deviation (RMSD) resolution of coverage. A regression equation was developed predicting the NRC of a molecule as a function of RMSD. The model yielded R^2 of 0.91 for both training and test sets, which consisted of 3414 and 3352 compounds, respectively. Utilizing 14 504 ligands from the Protein Data Bank with experimentally determined 3-D conformations, the regression equation was applied to the estimation of the NRC and the success rate of reproduction of experimental conformations from a theoretical conformation ensemble as a function of RMSD and flexibility was explored.

1. INTRODUCTION

Conformer generation software[1−6] is widely used as a theoretical construct in protein docking, 3-D pharmacophore analysis, and shape searching.[7] This is not surprising given that the biologically active ("bioactive") conformation(s) of a compound cannot, generally speaking, be represented by one, simple-to-obtain, "privileged" conformer. Rather, a compound's bioactive conformation(s) is found on variable locations of its conformational hyperspace. To detect and locate a bioactive conformation(s), one must select and analyze potentially large representative subsets of all possible conformers, with the concomitant demands on generation, evaluation, and storage.

For drug design purposes, theoretical models of conformational flexibility would be most ideal if the conformer ensemble were relatively small in size and contained only relevant conformations that are within a knowable degree of accuracy to all possible bioactive conformations. Correspondingly, the quality of conformational ensembles generated by specific theoretical methods or software is often gauged by their ability to reproduce, for example, a molecule's protein-bound conformation responsible for biological activity.

Two major approaches (not mutually exclusive) exist for meeting the challenge of balancing the size of the ensemble against accuracy of reproduction of protein-bound conformations: use of energy thresholds, which consider only conformations within an energy range relative to the conformation of least energy; and providing regular coverage, where a single conformation is chosen to represent a particular region of possible conformational space. The first approach excludes those high energy conformations that are considered to be biological irrelevant but assumes that the theoretical force-field employed can accurately represent actual biological environment energetics. The second approach excludes conformations that are too similar to each other and assumes that the conformer generation procedure will properly sample all possible conformational space.

A number of studies[8−12] have been carried out to understand the energetics of biologically active conformers. Some authors[8−10] have shown that bioactive conformations can have relatively high energy, up to 9−20 kcal/mol. Others[11,12] have argued that the high energy estimates are a direct result of deficiencies in the molecular mechanics force-fields and that, as theoretical methods improve, the estimates of strain energy will decrease to 3−4 kcal/mol. However, no consensus has emerged on what energy threshold and force-field combinations are best able to partition the conformational hyperspace of druglike molecules into "bioactive" and "bioinactive" parts.

One problem in this context may be that the majority of these energetic studies were performed in isolation from the geometric conformational concept of a molecule. By "geometric concept" we mean the intrinsic conformational characteristics of a molecule as a function of its size, flexibility, and possibly other structural descriptors, regardless of energetic considerations. This study attempts to provide some quantitative characteristics of the geometric space of drug-sized molecules to enable a deeper understanding of the possible size of geometrical space of an arbitrary molecule and, perhaps, provide the basis for a better understanding of the role of energy thresholds relevant to answering the following fundamental question: What is the bioactive relevant conformational space.

The characteristic of the geometric conformational space of a molecule, as analyzed here, is the number of regularly distributed conformers (NRC) in a conformational ensemble at a given resolution of coverage. The resolution of coverage is the minimum RMSD (root-mean-square deviation of atomic coordinates of non-hydrogen atoms) between conformers in an ensemble. We denote this RMSD of resolution

* Corresponding author phone: (301)435-7792; e-mail: bryant@ncbi.nlm.nih.gov.

10.1021/ci7000956
Published on Web 06/15/2007

of conformational coverage in the remainder of the article as simply "RMSD" unless otherwise noted.

The concept of the NRC embodies a number of useful aspects. The potential number of conformers of a flexible molecule, according to the RMSD definition above, is infinite at the limit of RMSD approaching zero. At nonzero RMSD values, however, the NRC provides a *quantitative* description of a molecule's conformational space. As a structural characteristic, NRC is directly related to a molecule's size and flexibility and should, therefore, be *predictable* from the molecule's structure, ideally from just a few structural descriptors. Since NRC is also a characteristic of a complete and regular ensemble, which must, by definition, contain the bioactive conformation within the ensemble's RMSD resolution of coverage, NRC defines *the size of the ensemble* that must be generated to provide the desired accuracy of reproduction of any bioactive conformation. Application of energy thresholds and other more sophisticated rules of conformer selection should then, ideally, be able to decrease the size of the ensemble required for the same accuracy of reproduction of the bioactive conformation. Therefore, the NRC can be used as a reference point for evaluation of the effectiveness of energy thresholds and of other rules used in the exploration of bioactive conformational space.

Although the NRC can be thought of as a general characteristic of a molecule, one has to determine it with a specific method and application. In this study, we exemplify our approach with the conformer generation program Omega (OpenEye), version 1.8.1.[13] Modern conformer generation software typically uses an approach that systematically omits some parts of conformational space by using a predefined knowledge base of allowed torsion angles and ring conformations, by limitations on the maximum number of conformers, by using energy thresholds, and by other algorithmic limitations.[14] As a result, the produced ensembles are "quasi-regular". Omega falls in this class of software, and we discuss its effects on the results of our study. The specific feature of Omega that made this study possible is the ability to use the RMSD-based filtering of the preliminary pool of conformers, which provides RMSD-spaced ensembles as output.

Using Omega, we analyze the average size of ensembles produced for RMSD resolutions ranging from 0.6 Å to 2.4 Å for compounds with different numbers of effective rotors. We develop a regression equation predicting the dynamics of NRC as a function of RMSD. Finally, we discuss the coverage of the conformational space by Omega by comparing generated conformational ensembles with experimentally determined conformations of protein-bound ligands in the MMDB[15] data source as deposited in PubChem.[16]

2. COMPUTATIONAL METHODS

2.1. Conformer Generation Using Omega 1.8.1. Omega[13] generates ensembles of conformers in three stages. The first stage involves the generation of initial 3-D conformations from the connection table of the input chemical structure. The second stage generates conformers from the resulting first stage 3-D conformers using torsion driving. The last stage performs optional postprocessing, e.g., energy minimization, and final filtering by RMSD of the produced conformers.

Table 1. Values of Omega 1.8.1 Parameters Used in This Study[a]

parameter	default settings	settings used in the study
MMFF94S	"false"	"true"
NUMCONFS	1	25
KEEPCONFS	1	25
RMS	1.0 Å	variation between 0.6 and 2.4 Å
EWINDOW	14.0 kcal/mol	15.0 kcal/mol
MAXCONFS	400	10 000
MAXROT	12	15
FINALRMS	0.0	same as RMS parameter

[a] MMFF94S − this parameter changes the force-field such that anilinic nitrogens are considered planar. NUMCONFS − maximum number of conformers allowed to be generated by the first stage of ensemble generation. KEEPCONFS − actual number of conformers generated by the first stage of conformer generation that are passed on to the second stage of conformer generation for independent torsion driving. If KEEPCONFS is less than NUMCONFS, then the first KEEPCONF conformers lowest in energy are selected from the initial NUMCONFS conformers. RMS − this parameter directly controls the resolution of coverage of the conformational space by specifying the minimum RMSD difference between conformers. When filtering by RMSD, Omega considers the lowest energy conformations first, resulting in higher energy conformations being eliminated. EWINDOW − the maximum energy difference between the conformer of lowest and highest energy as determined by the force-field used. Conformers beyond this energy range are eliminated from consideration. MAXCONFS − maximum number of conformers produced during torsion driving, the second stage of conformer generation, from a conformer produced in the first stage of conformer generation, after filtering by EWINDOW and RMS. MAXROT − maximum number of rotatable bonds allowed in a structure for processing by Omega. FINALRMS − minimum RMSD between any two conformations output in the final stage of conformer generation process.

The first stage employs a random-coordinate distance-geometry method[17] to generate initial conformers that are refined using the Merck Molecular Force-Field[18,19] (MMFF) to create the initial conformers used in the torsion driving stage. In the second stage, Omega generates multiple conformers from each first-stage conformer by using a depth-first, divide-and-conquer algorithm, where the input structure from the first stage is transformed into small fragments that are then, using a knowledge base of allowed torsions and ring conformations, joined together to build the molecule's conformations. A selectable energy window for conformations, using an Omega-specific variant of the Dreiding force-field[20] favoring macromolecule-bound conformations, optionally limits which conformers are generated.

In the final stage, the resultant set of conformations produced by the earlier stages is filtered using the RMSD distance between conformations. This step achieves the actual RMSD "grid" resolution for the ensemble. Considering each conformation produced in the first stage is treated separately in the second stage of conformer generation, this final filtering can dramatically reduce the size of the ensemble by guaranteeing a minimum RMSD distance between all conformers. No additional postprocessing, beyond this RMSD filtering, was performed in this study.

2.2. Parametrization of Omega. Omega 1.8.1 has 26 parameters that the user can modify to fine-tune both construction and optimization of the generated conformational ensembles. Default settings were used except for the eight parameters in Table 1. The choice of parameter values used in this study resulted from experimentation and experience. The first stage of conformer generation used non-default parameter values for "MMFF94S", "NUMCONFS",

and "KEEPCONFS". The second stage of conformer generation used nondefault parameter values for "MAXCONFS", "RMS", "EWINDOW", and "MAXROT". The last stage of conformer generation used a nondefault parameter value for "FINALRMS".

One of the central entities in this study is the number of conformers necessary to achieve full conformational space coverage as provided by Omega. For this reason, we chose a relatively high value of 10 000 for MAXCONFS to enable conformational ensembles to be as large as possible using Omega.

We chose both NUMCONFS and KEEPCONFS to be 25, which is the recommended value, for example, when building custom Omega ring system template libraries. In our experience, the choice of this value makes Omega's first stage random-coordinate distance-geometry method nearly deterministic for druglike structures, likely due to better sampling of initial conformers.

Our choice of parameter values for NUMCONFS, KEEPCONFS, and MAXCONFS, being a tradeoff between unlimited ensemble size and calculation expense, allows generation of up to a maximum of 250 000 conformers per molecule in the second stage of model building. This large number enabled us to explore how variation in the parameter RMS affects the total number of conformations produced.

2.3. Modeling of the Relationship between NRC and RMSD Resolution. *2.3.1. Premise.* We assume the general dependency

$$\mathrm{NRC} = f(\mathrm{RMSD}, \mathrm{DF}, \mathrm{SIZE}) \quad (1)$$

where NRC is the number of conformers in the output ensemble; RMSD is the resolution of coverage; DF is the number of the molecule's degrees of freedom; and SIZE is a descriptor of the molecule's size.

In theory, DF and SIZE are purely structure-dependent parameters and thus knowable prior to conformer generation. This would suggest that only the RMSD parameter embodies the variability of the conformer generation software.

In order to explicitly take into account the peculiarities of the Omega algorithm and knowledge base used for generation of the conformers, we introduce an additional, calibrating, parameter: the number of conformers generated by Omega for each small molecule at a fixed $RMSD_0$, which we denote by NRC_0. Besides accounting for the software idiosyncrasies, NRC_0 should reflect, for example, such properties of molecules as distribution of chemical groups that may cause steric clashes. Calculation of NRC_0 requires the expense of an additional run of the Omega application. We set $RMSD_0$ to a relatively high value of 2.0 Å to make this step relatively cheap in terms of computational resources. The general form of this relationship describes the dynamics of NRC relative to NRC_0 for RMSD values relative to $RMSD_0$, where $RMSD_0$ is 2.0 Å, and is expressed by eq 2:

$$\mathrm{NRC} = f(\mathrm{NRC}_0, \mathrm{RMSD}/\mathrm{RMSD}_0, \mathrm{DF}, \mathrm{SIZE}) \quad (2)$$

The value of RMSD in eq 2 was taken to be identical to the FINALRMS parameter used to generate the Omega conformer ensembles. The ensembles' conformer count was taken as the value of NRC. To account for a molecule's degrees of freedom (DF), we counted the number of rotors (*nr*), as determined by the OEChem software.[21] In addition, we attempted to account for flexible ring systems by using a count of non-hydrogen atoms in nonaromatic rings (*nnara*), where the description of aromaticity was selected to be the OpenEye aromaticity model implemented within the OEChem software. We also removed bridgehead atoms and non-sp^3 hybridized ring atoms from the "*nnara*" value. For the value of SIZE, we selected the count of the number of non-hydrogen atoms (nha).

In this study, logarithmic forms of NRC and NRC_0 were used as variables in eq 2, and, since the equation describes the dynamic of NRC relative to NRC_0, we likewise used $\ln(\mathrm{RMSD}/\mathrm{RMSD}_0)$ instead of RMSD. The relationship explored in this study is expressed in eq 3:

$$\ln(\mathrm{NRC}) = f(\ln(\mathrm{NRC}_0), \ln(\mathrm{RMSD}/\mathrm{RMSD}_0), nr, nnara, nha) \quad (3)$$

2.3.2. Data Sets. From the PubChem[16] Compound database, 10 000 random chemical structures were selected, each alternating between being placed in either a training set or a test set. Both structure sets were then filtered to remove entries that possess the following: "nonorganic" elements, being those other than H, C, N, O, F, P, S, Cl, Br, and I; more than a single covalently bonded unit; a molecular weight greater than 1000; more than 15 rotatable bonds; and undefined tetrahedral sp^3 or planar sp^2 stereo centers.

The resulting training and test sets consisted of 3414 and 3352 compounds, respectively. Omega was utilized to create a 3-D conformer ensemble for each compound in the training and test sets using the parameter values described above. For every compound, Omega was run twice: the first time to generate the calibration parameter NRC_0 of the model at RMSD = 2.0 Å ($RMSD_0$) and the second time with an RMSD picked randomly from the set of {0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0, 2.2, 2.4} Å (RMSD).

2.3.3. Method Used for Model Development. To explore and determine the relationship in eq 3, the combinatorial algorithm of the Group Method of Data Handling[22] (GMDH) was selected. GMDH is a method that iteratively selects, among a series of tested models, the model with both highest prediction accuracy and lowest complexity. The general functional form of the models tested was a polynomial of second order, thus allowing for terms that are the product of any two independent variables.

The first iteration of GMDH produced models of the simplest form $y = a_0 + a_1x_i$, $i = 1,2,...,M$, and the five best of these models ($F = 5$) were selected. A second series of more complex models, $y = a_0 + a_1x_i + a_2x_j$, $i = 1,2,...,F$, $j = 1,2,...,M$, was generated using the first series of models. The third series of models produced, $y = a_0 + a_1x_i + a_2x_j + a_3x_k$, $i = 1,2,...,F$; $j = 1,2,...,F$; $k = 1,2,...,M$, is even more complex, and so on. The iterative procedure was carried on as long as the convergence criterion decreased in value. The external criterion for GMDH algorithm convergence is the forecast error variation criteria, RR, whose functional form used in this study is shown in eq 4

$$RR_{B/A} = \frac{\sum_{i=1}^{N_B} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N_B} (y_i - \bar{y})^2} \rightarrow \min \quad (4)$$

where y_i is the actual variable value, $\hat{y}_i$ is the value calculated according to the model, $\bar{y}$ is the mean value, B is the test set, and A is the training set. Minimization of RR consists essentially in selecting the model developed on training set A that gives the best prediction accuracy on the test set B.

2.4. Assessment of NRC Needed for Given RMSD for Flexible Compounds. *2.4.1. Data Set.* A total of 17 200 small-molecule, experimental 3-D crystal structures were downloaded from the PubChem Substance database from the MMDB data source. The MMDB database[23] is a subset of three-dimensional ligand structures obtained from the Protein Data Bank,[24] excluding theoretical models. This data set represents the bioactive conformations considered to be the "conformations of interest".

If ligands occurred in multiple protein−ligand complexes in the MMDB, all instances were included in our data set. Only compounds passing the filtering step described in section 2.3.2 were considered, resulting in a final set of 14 504 experimentally determined 3-D structures. The stereochemistry present in a given 3-D structure was enforced for all conformers generated in an ensemble.

2.4.2. Computing and Predicting NRC. For every structure, Omega was run using the method described in section 2.3.2. Starting conformations were generated independent of the input 3-D coordinates, i.e., from connection tables. The resulting NRC_0 values were used to predict NRC for the entire range of investigated RMSD values from 0.6 to 2.4 Å with step size 0.2 Å, using the model developed in section 2.3.

2.4.3. Calculation of $RMSD_{X\text{-}ray}$. The RMS deviation between the X-ray structure and each conformer in the Omega produced ensemble was calculated using the OEChem function "OERMSD".[21] The least RMS deviation to the X-ray structure is referred to in this study as $RMSD_{X\text{-}ray}$.

2.4.4. Determination of "Holes" in the Conformational Coverage. For an ideal conformational ensemble perfectly spaced by a particular RMSD value, one should be guaranteed that the RMSD is an upper bound of the $RMSD_{X\text{-}ray}$. However, the $RMSD_{X\text{-}ray}$ can be greater than the RMSD for the ensemble in cases where the bioactive conformation is located in a region lacking conformational coverage. Using $RMSD_{X\text{-}ray} \leq$ RMSD as an indicator, we looked for "holes" in the conformational coverage generated by Omega. Obtaining a quantitative estimate of this effect is important because in cases where such "holes" exist, the NRC of ensembles produced by Omega underestimates the total number of conformers needed for regular coverage of the conformational space. The relative number of cases in which $RMSD_{X\text{-}ray} \leq$ RMSD is referred to from now on as the success rate of reproduction.

3. RESULTS AND DISCUSSION

3.1. Model of the Relationship between NRC and RMSD. The convergence of the GMDH RR criterion was achieved at the complexity level of three. The best regression equation obtained for eq 3 is depicted in eq 5

$$\ln(NRC) = 0.02 + 1.01 \cdot \ln(NRC_0) - (0.503 \cdot nr + 0.119 \cdot nnara) \cdot \ln\left(\frac{RMSD}{RMSD_0}\right) \quad (5)$$

where *nr* is the number of rotatable bonds; *nnara* is the number of atoms in nonaromatic rings; NRC_0 is the actual number of conformers generated at $RMSD_0 = 2.0$ Å; and NRC is the total number of conformers produced at RMSD.

For eq 5, the training set yielded an R^2 of 0.92 and an RSE of 0.56 with *N* being 3414, and the test set yielded an R^2 of 0.91 and an RSE of 0.57 with *N* being 3352. The variable *nha* was not included in the equation by the GMDH procedure, presumably because the NRC_0 variable already encodes information about the size of the molecule. The variable nr is primary in determining the influence of the molecule's degrees of freedom on the NRC value. The measure of pseudorotors, *nnara*, adds a significant contribution due to the flexibility of ring systems.

After the form of the model was determined, coefficients were estimated more precisely using the combined training and test sets, yielding eq 6 with an R^2 of 0.92 and an RSE of 0.57, where *N* is 6766.

$$\ln(NRC) = 0.02 + 1.007 \cdot \ln(NRC_0) - (0.513 \cdot nr + 0.108 \cdot nnara) \cdot \ln\left(\frac{RMSD}{RMSD_0}\right) \quad (6)$$

Examination of eq 6 suggests an "effective number of rotors" relationship, in that five nonaromatic ring atoms are roughly equivalent to a single rotatable bond. Rerunning the regression using $nr_{effective}$, where $nr_{effective} = nr + nnara/5$, in place of the parameters *nr* and *nnara* in eq 6 yields eq 7 with an R^2 of 0.92, an RSE of 0.56, and *N* being 6766.

$$\ln(NRC) = 0.02 + 1.006 \cdot \ln(NRC_0) - 0.515 \cdot nr_{effective} \cdot \ln\left(\frac{RMSD}{RMSD_0}\right) \quad (7)$$

Figure 1 displays the predicted versus actual values of ln(NRC) for eq 7 for the combined training and test sets. The distributions of number of heavy atoms and of effective rotors for the combined set (Figure 2) demonstrate the diversity of size and flexibility of the compounds used for model development.

To demonstrate how eq 7 performs, Figures 3−5 provide predicted and actual numbers of conformations, depicted as ln(NRC), as a function of RMSD for typical chemical structures.

As expected, the NRC value grows faster or slower depending on the molecule's flexibility as a function of RMSD, with the rate of change of NRC depending mostly on the number of effective rotors. One can see, for example, that for molecules with four effective rotors, the ln(NRC) value increases much slower than for molecules with 5.4 or 10.2 effective rotors. The molecule with the PubChem Compound identifier (CID) 1212923 (33 heavy atoms) is nearly the same size as CID 2941020 (32 heavy atoms). These two molecules have three ($\ln(NRC_0)$=1.1) and six ($\ln(NRC_0)$=1.79) conformers at RMSD = 2.0 Å, respec-

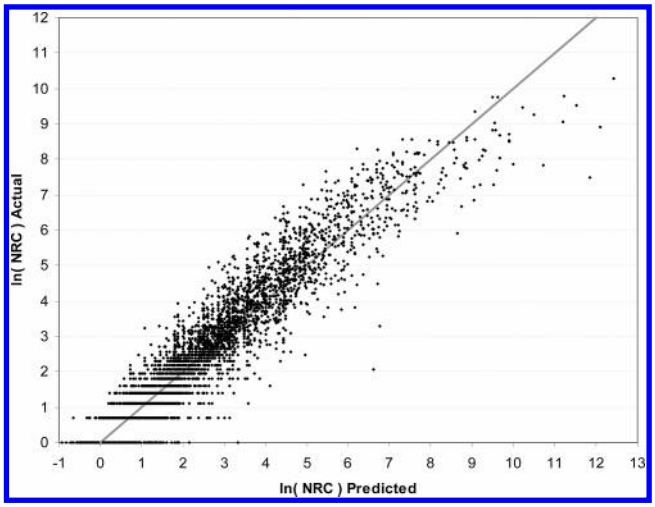

Figure 1. Plot of predicted versus actual ln(NRC) using the model in eq 7. Each dot represents one of the 6766 structures in the combined training and test sets. The solid line depicts an ideal fit of predicted to actual values.

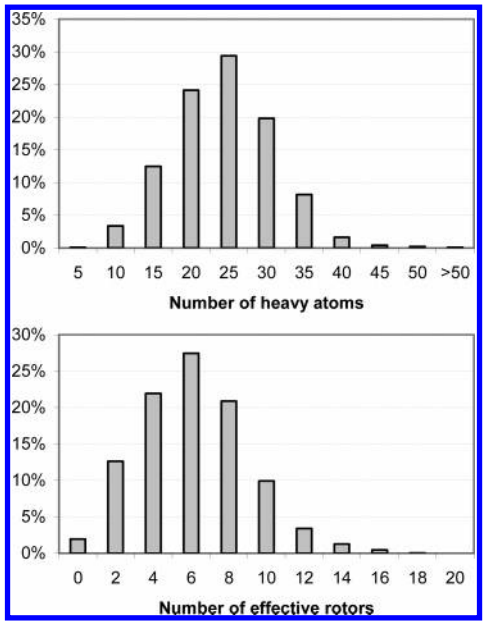

Figure 2. Distributions of the number of heavy atoms and effective rotors for the combined training and test sets.

tively. However, due to their different number of effective rotors, 5.4 for CID 1212923 versus 10.2 for CID 2941020, already at RMSD = 0.6 Å the respective NRC values are 121 (ln(NRC)=4.8) and 1536 (ln(NRC)=7.33) conformers.

Applying the model in eq 7 to the MMDB data set, we use prediction intervals instead of single point predictions, with the prediction intervals for NRC being calculated according to eq 8

$$\ln(\mathrm{NRC})_{\mathrm{calc}} - \mathrm{RSE}\cdot t_{\alpha,n} \le \ln(\mathrm{NRC}) \le \ln(\mathrm{NRC})_{\mathrm{calc}} + \mathrm{RSE}\cdot t_{\alpha,n} \quad (8)$$

where $\ln(\mathrm{NRC})_{\mathrm{calc}}$ is a value calculated by eq 7; RSE is the residual standard error of the regression; n is the number of degrees of freedom equal to that used for the RSE calculation; t is the critical value for the Student's distribution; and α is the significance level. In this study, we use $\alpha = 0.1$ to

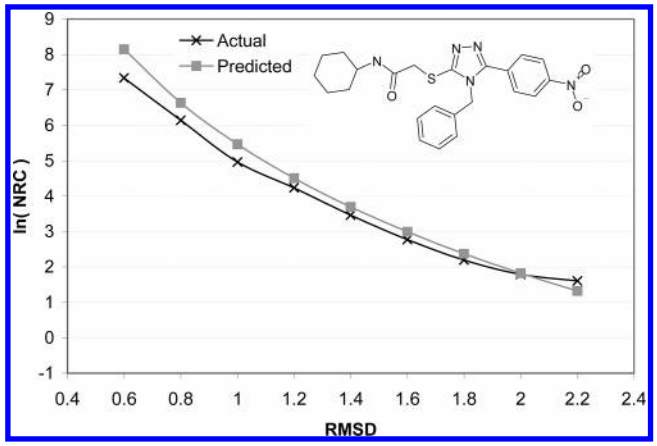

Figure 3. Comparison of ln(NRC), actual and predicted, using eq 7, as a function of RMSD for PubChem CID 2941020, depicted above, which contains 32 non-hydrogen atoms, 9 rotors, and 6 "nonaromatic" ring atoms. Effective rotors: 10.2.

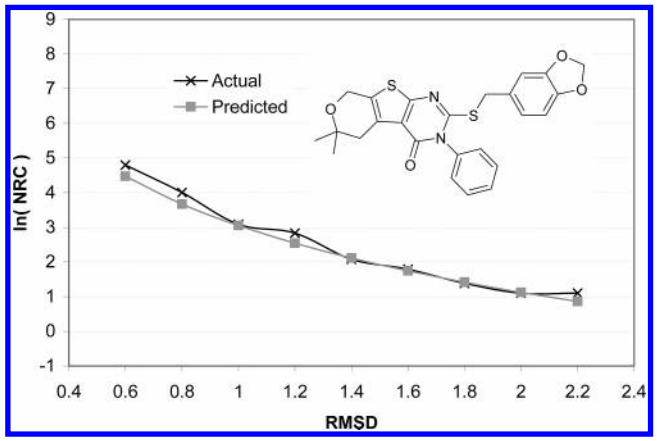

Figure 4. Comparison of ln(NRC), actual and predicted, using eq 7, as a function of RMSD for PubChem CID 1212923, depicted above, which contains 33 non-hydrogen atoms, 4 rotors, and 7 "nonaromatic" ring atoms. Effective rotors: 5.4.

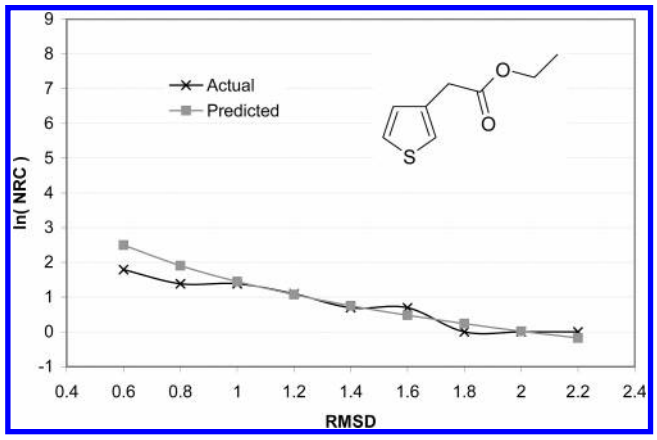

Figure 5. Comparison of ln(NRC), actual and predicted, using eq 7, as a function of RMSD for PubChem CID 520865, depicted above, which contains 11 non-hydrogen atoms, 4 rotors, and 0 "nonaromatic" ring atoms. Effective rotors: 4.0.

obtain a 90% prediction interval. For eq 7, RSE = 0.56 and $t_{0.1;6763} = 1.645$, therefore, the 90% prediction interval is

$$\ln(\mathrm{NRC}) = \ln(\mathrm{NRC})_{\mathrm{calc}} \pm 0.92 \quad (9)$$

3.2. Analysis of NRC and Coverage for MMDB Data Set. The diversity of the MMDB data set in terms of molecular size and flexibility is shown in Figure 6.

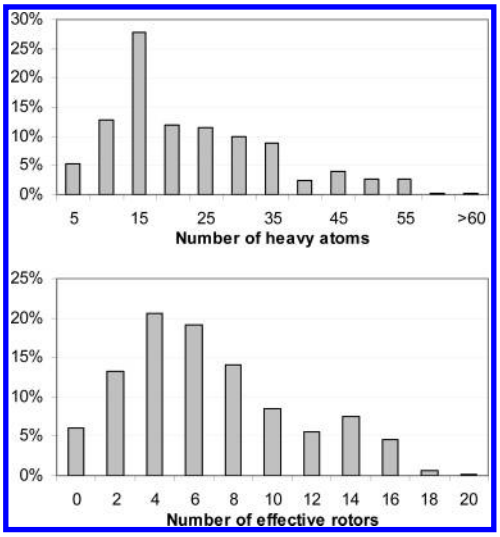

Figure 6. Distributions of the number of heavy atoms and effective rotors for the 14 504 MMDB ligands.

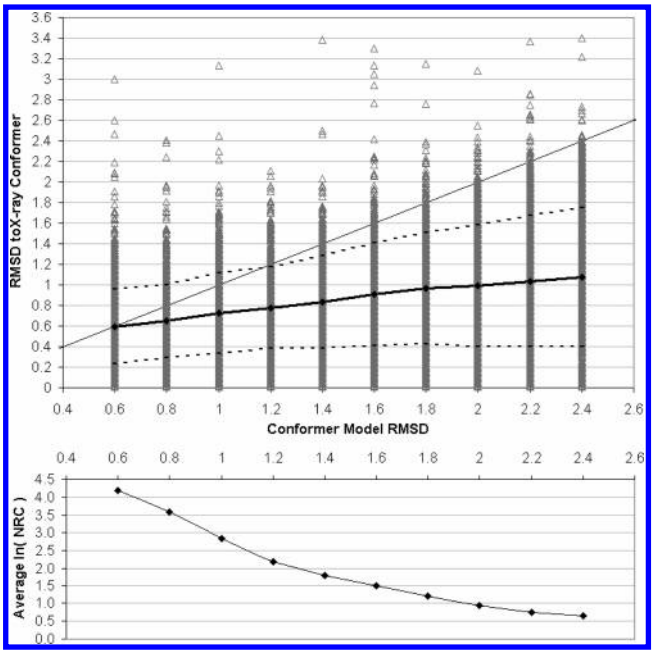

Figure 7. Upper graph: Plot of the RMSD resolution of all 14 504 conformer models in the MMDB data set versus the smallest RMSD from the experimentally derived 3-D "biologically active" conformer to a single conformer in the conformer model. Each triangle denotes one of the conformer models. Triangles above the solid straight line represent cases where the "bioactive" conformation is not within the RMSD resolution of the conformer model. The thick solid line denotes the average RMSD distance to the "bioactive" conformation. The dotted lines denote the standard deviation from the average RMSD distance. **Bottom graph:** Average size of conformational ensemble (ln(NRC)) generated at different RMSD resolutions.

Figure 7 shows the distribution of $RMSD_{X\text{-}ray}$ values for the 14 504 MMDB ligands versus RMSD resolution used.

The graphs suggest a statistical dependence between the average RMSD of X-ray structure reproduction and the average ensemble size across the range of ensemble RMSD resolution of coverage. The correlation between average ln(NRC) and average $RMSD_{X\text{-}ray}$ is −0.98. Therefore, conformational coverage as represented by the ensemble size is deemed to be a determining factor in providing reproduction of the X-ray structure. Figure 7 also shows that, despite the average $RMSD_{X\text{-}ray}$ being lower than the RMSD value, in many cases $RMSD_{X\text{-}ray}$ is greater than RMSD. In these cases, the conformational space coverage is insufficient in the region of the X-ray conformation.

Figure 8 shows the success rates of X-ray structure reproduction for the MMDB ligands falling into the indicated effective rotor ranges. It also shows prediction intervals of the average ln(NRC) using eq 9 and the actual average ln(NRC) values. The numbers of ligands for which the averages were calculated are shown in Table 2.

For compounds with 0.2−12 effective rotors, the graphs in Figure 8 show that the predictions are of reasonable accuracy, given that the average actual ln(NRC) data lie within the prediction interval. However, for compounds with 12.2−20 effective rotors, the predicted NRC values significantly exceed the actual NRC values at low RMSD resolutions, as indicated by the average actual ln(NRC) being below the prediction interval. Furthermore, this effect appears to correspond to a significant decrease in the reproducibility of X-ray structures within generated conformer ensembles. These major differences can be explained by limitations placed on and within the Omega application, relative to the maximum number of conformers in ensembles. In addition, this may suggest that one may need to consider different parameter settings when attempting to generate high quality ensembles at low RMSD values for flexible compounds.

The Omega parameters, MAXCONFS = 10 000 and NUMCONFS = KEEPCONFS = 25, used in this study allow for a maximum of 250 000 conformers (ln(NRC)=12.43). However, this maximal total conformer value is an upper limit and can only be reached: if all initial 25 conformers generated by random geometry are unique; if there is no overlap with respect to minimum RMSD spacing of conformations between the conformations produced during torsion driving of the initial 25 conformers; and if the torsion driving step reaches the maximum of 10 000 produced conformations.

For highly flexible molecules, including those with more than 12.0 effective rotors, it is possible that the 10 000 limit on the maximum number of conformers during torsion driving is reached. For example, among compounds in the range of 12−20 effective rotors, there are many dinucleotides, such as flavin-adenine dinucleotide (FAD). These molecules have several hundred conformers at an RMSD = 2.0 Å, and, with decreasing RMSD value, their NRC grows extremely fast, almost certainly hitting the 10 000 conformation numeric limitation at the lower RMSD values examined in this study. Hitting the maximum conformer limit during torsion driving implies a truncation of the explored conformational space, likely creating gaps in coverage, which decrease the reproducibility of X-ray structures within generated conformer ensembles and decrease the NRC.

Conformational exploration is inherently combinatoric with respect to the number of degrees of freedom. It is conceivable that a limit of 100 000, 1 000 000, or even a much larger value of maximum conformations will still be inadequate as an upper limit of the number of conformations produced during torsion driving. Consider the earlier example of FAD that has 13 rotatable bonds. If only three torsion values are considered for each rotatable bond, this yields 3^{13} = 1 594 323 potential conformations to be explored. A simplistic algorithmic approach that iterates over the allowed torsion angles will rapidly hit a "reasonable" conformation

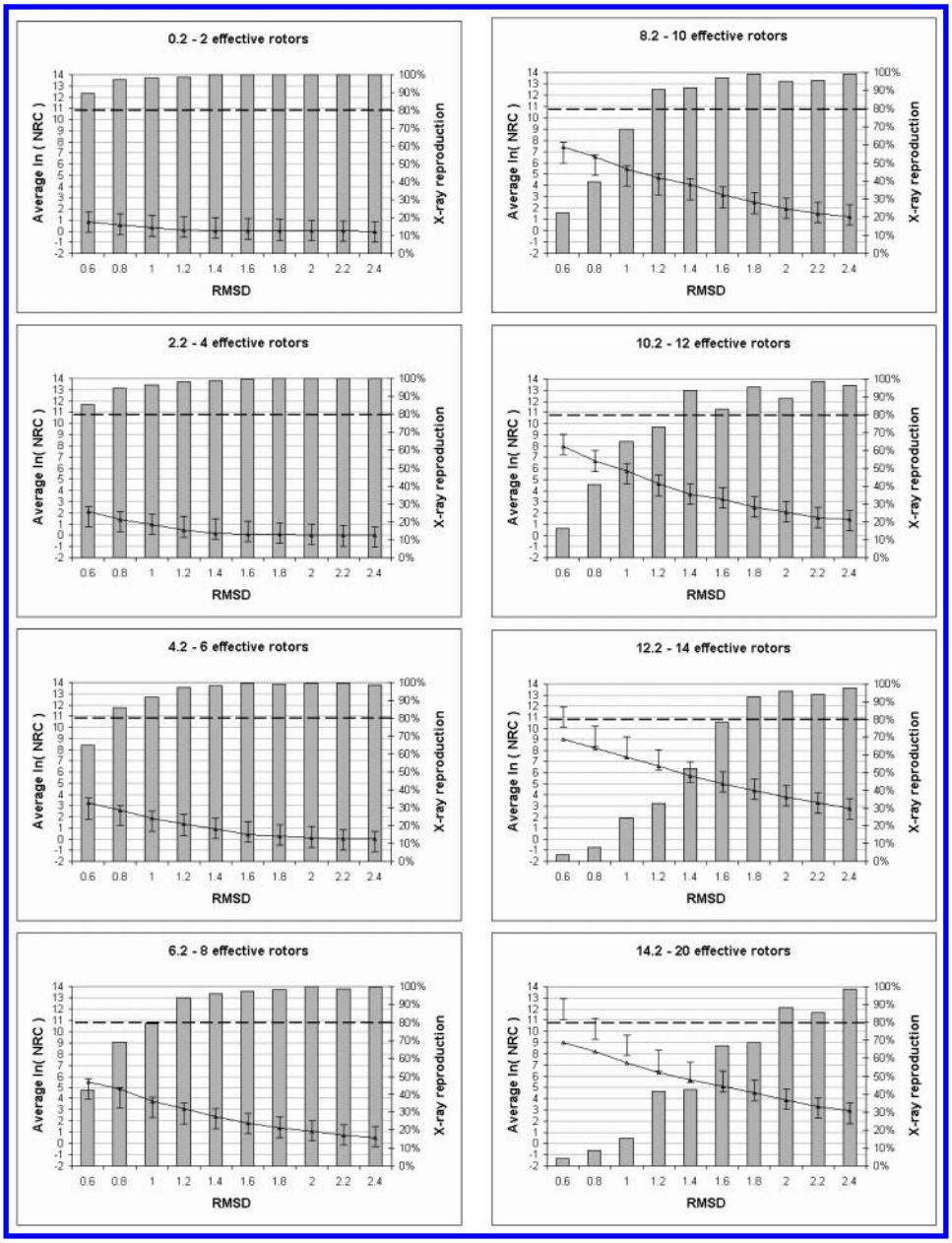

Figure 8. Curves: actual average ln(NRC) values. Intervals: prediction intervals of the average ln(NRC). Bars: success rates of X-ray structure reproduction. The dotted horizontal line marks the 80% threshold of X-ray structure reproduction. Each individual panel, which covers the effective rotor range shown at the top, displays the results as a function of the RMSD of coverage of conformational space.

Table 2. Count of Structures in the MMDB Data Set Binned by Effective Rotor Range and RMSD Value[a]

	range of effective number of rotors								
RMSD (Å)	0	0.2−2	2.2−4	4.2−6	6.2−8	8.2−10	10.2−12	12.2−14	14.2−20
0.6	85	194	291	261	183	125	80	114	76
0.8	71	185	286	280	197	129	88	103	81
1.0	100	175	318	285	222	121	69	119	77
1.2	87	209	303	255	186	127	75	98	70
1.4	99	189	297	251	201	120	78	84	85
1.6	97	195	297	303	216	132	78	115	94
1.8	96	190	278	294	205	123	92	122	77
2.0	87	166	325	267	207	120	81	96	76
2.2	81	212	285	296	213	132	81	120	62
2.4	73	209	313	271	198	113	84	127	76

[a] The counts indicate the set sizes for the individual ordinate values in all of the panels in Figure 8.

limit for molecules like FAD. More advanced approaches and techniques are clearly needed to rigorously explore the expansive conformational space of highly flexible molecules.

The results in the graphs of Figure 8 clearly demonstrate the strong effect of compound flexibility on the reproducibility of bound ligand X-ray structures, with respect to

Table 3. Prediction Intervals for Average NRC as a Function of RMSD and Effective Rotor Range for the MMDB Data Set[a]

	range of effective number of rotors (number of compounds)							
RMSD (Å)	0.2−2 (1924)	2.2−4 (2993)	4.2−6 (2763)	6.2−8 (2028)	8.2−10 (1242)	10.2−12 (806)	12.2−14 (1098)	14.2−20 (774)
0.6	1−6	2−13	*6−38*	*53−336*	*413−2603*	*1245−7837*	***25197−158654***	***74532−469292***
0.8	1−5	1−9	3−21	*22−136*	*129−810*	*301−1894*	***4573−28791***	***10515−66206***
1.0	1−4	1−7	2−13	*11−67*	*52−327*	*100−629*	***1217−7662***	***2302−14493***
1.2	1−4	1−5	1−9	6−38	25−156	*41−256*	*413−2598*	***665−4189***
1.4	1−3	1−4	1−6	4−23	13−83	19−120	*165−1041*	*233−1467*
1.6	1−3	1−4	1−5	2−15	8−49	10−62	*75−471*	*94−591*
1.8	1−3	1−3	1−4	2−11	5−30	5−35	37−234	*42−265*
2.0	1−3	1−3	1−3	1−8	3−20	3−21	20−125	21−129
2.2	1−2	1−2	1−2	1−6	2−13	2−13	11−71	11−68
2.4	1−2	1−2	1−2	1−4	1−9	1−8	7−43	6−37

[a] Values in italics represent combinations of $nr_{effective}$ and RMSD values where the success rate of X-ray reproduction is less than 80%. Values in italics and boldface indicate where the prediction overestimates the actual NRC.

RMSD. If one sets a statistical threshold of an 80% success rate (four out of five) of reproduction for a given RMSD value (indicated in the graphs by a dotted line), then X-ray conformation reproduction will be successful at virtually any RMSD value for compounds in the flexibility range of 0−4 effective rotors using the methodology employed in this study. A transition occurs at 4.2−6.0 effective rotors where the success rate dips below 80% for the lowest RMSD value studied, 0.6 Å. As the compound flexibility further increases, the success rate of X-ray conformation reproducibility becomes progressively worse, even at more "moderate" RMSD values of 1.0 Å or greater. The apparent systematic deterioration in the ability to routinely reproduce the X-ray conformation as a function of increasing flexibility and decreasing RMSD may suggest an aggregate lower bound RMSD resolution possible for a molecule of given flexibility due to the methodology employed within Omega. Considering that this trend is consistent across the full range of flexibility and our choice of parameters does not restrict the conformation exploration of moderately flexible structures by Omega, this is likely the result of a combination of insufficient sampling of the torsion angles used in the Omega ring and torsion knowledge base and the use of an inappropriate or inadequate force field, suboptimal energy thresholds, or other inherent additive or accumulative systematic errors specific to the implementation of Omega.

Table 3 shows the prediction intervals for the average NRC as a function of effective rotor range for the entire MMDB data set. The italicized values show the progressive breakdown in the ability to routinely reproduce the X-ray structure as a function of RMSD and of flexibility, i.e., where conformational coverage "holes" at a particular RMSD resolution may be an issue. The values in italics and boldface indicate where prediction overestimates actual NRC, presumably due to Omega conformation production limits being reached, discussed previously.

As shown in Table 3, the conformational space of less flexible molecules is small and may be represented by a relatively small number of conformers for RMSD values all the way down to 0.6 Å. Obviously, the lower bound for any estimated average NRC is 1, which is indeed found for all RMSD values for the least flexible compounds ($nr_{effective}$= 0.2−2). This implies that, in many cases, a single conformer may be sufficient to reproduce the biologically active structure of the least flexible molecules within the highest resolution of coverage studied.

In contrast, the conformational space of more flexible molecules becomes large quickly, especially at smaller RMSD values. The NRC prediction intervals given in Table 3 provide only conservative estimates of the actual interval size, considering eq 7 is not based on the results of an exhaustive conformational search. However, even these estimates show that, to represent the conformational space of flexible molecules with high resolution of coverage, a large number of conformers may need to be generated. For those italicized domains in Table 3 where the success rate of X-ray structure reproduction was low, NRC values reported are obviously too low for regular coverage. In other words, larger numbers of conformations should be expected to achieve regular coverage at those resolutions.

Interestingly, Table 3 can help estimate the accuracy of bioactive structure reproduction expected if one uses conformer ensembles of a particular size. For example, using ensembles of up to 500 conformers, one might expect the accuracy of reproduction for compounds to be as follows: 0.6 Å with 0.2−8 effective rotors; 0.8 Å with 8.2−10 effective rotors; ≥1.0 Å with 10.2−12 effective rotors; ≥1.2 Å with 12.2−14 effective rotors; and ≥1.4 Å with more than 14 effective rotors.

These accuracy estimates are in line with results by Kirchmair et al.[25] Analyzing X-ray structure reproduction of 778 high-resolution PDB ligands by two different conformer generators, Omega and Catalyst,[26] these authors demonstrated that ensembles of a size of 50−500 conformers provide an average $RMSD_{X\text{-}ray}$ of <0.7 Å for compounds with less than 3 rotors; 0.6−0.8 Å for compounds with 3−5 rotors; 0.9−1.2 Å for compounds with 6−8 rotors, 1.1−1.6 Å for compounds with 9−11 rotors; 1.3−1.7 Å for compounds with 12−14 rotors; and 1.6−2.3 Å for compounds with more than 14 rotors.

In a number of studies, authors have used ensembles limited to 50−500 conformers[25,27−29] or used RMS deviation from X-ray structures averaged over all investigated compounds, regardless of flexibility, as a criterion of the ensemble quality.[27−28,30] We would contend that this is a questionable practice. Achieving low $RMSD_{X\text{-}ray}$, for example, 0.6 Å, using small-sized ensembles is trivial for inflexible structures and practically unreachable, other than by sheer luck, for compounds with more than eight effective rotors. By the same token, our findings may seem to contradict earlier conclusions[31,32] that a small number of conformations are always sufficient to represent conforma-

tional spaces of druglike molecules. As this work shows, this is only true if the minimum RMSD spacing between conformations is increased as a function of flexibility, implying a decrease in the degree of accuracy. We emphasize that the sufficiency of the conformational ensemble size very much depends on the accuracy required of the ensemble.

Our results suggest that, for the assessment of the quality/effectiveness of conformational models generated using different methodologies and applications, it may be useful to compare these models with conformer ensembles providing regular coverage of the entire conformational space of molecules, i.e., those produced by an "ideal" systematic search where there are no "holes" in conformational coverage. A methodology that would allow a significant decrease of the number of conformers without any loss of the bioactive structure reproduction accuracy should be considered to be effective in the above sense. For example, it can certainly be argued that use of an energetic threshold may significantly decrease the number of conformers in a regularly spaced ensemble without any loss in the X-ray structure reproduction accuracy by eliminating biologically irrelevant parts of the conformational space.

4. CONCLUSIONS

This study has demonstrated and, to some degree, quantified the interdependence between the resolution of coverage, the size of conformational ensembles, and molecular flexibility. We have shown that the coverage of a molecule's conformational space plays an important role in the reproduction of biologically active conformations. As such, to produce conformational models for arbitrary chemical structures with a particular accuracy of bioactive structure reproduction, a particular RMSD resolution of coverage should be used, and, consequently, the conformational ensemble sizes will vary significantly depending on molecular size and flexibility. For molecules with little flexibility, a fairly small number of conformers is sufficient to cover their entire bioactive conformational space at RMSD resolutions down to 0.6 Å. For compounds with more than eight effective rotors, however, ensembles of up to 500 conformers may be insufficient to provide an adequate conformational description at high or even medium RMSD resolution of coverage.

ACKNOWLEDGMENT

We would like to thank Marc Nicklaus and Dmitrii Filimonov for interesting discussions and OpenEye Scientific Software, Inc., for helpful insights. This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

Supporting Information Available: TXT files with PubChem[16] "CID" identifiers for 3414 and 3352 structures of the training and test sets and TXT file with PubChem[16] "SID" identifiers for 14 504 structures of the MMDB set. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Murral, N. W.; Davies, E. K. Conformational freedom in 3-D databases. 1. Techniques. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 312−316.
(2) Hurst, T. Flexible 3D searching: The directed tweak technique. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 190−196.
(3) Klebe, G.; Mietzner, T. A. Fast and Efficient Method to Generate Biologically Relevant Conformations. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 583−606.
(4) Renner, S.; Schwab, C. H.; Gasteiger, J.; Schneider, G. Impact of Conformational Flexibility on Three-Dimensional Similarity Searching Using Correlation Vectors. *J. Chem. Inf. Model.* **2006**, *46*, 2324−2332.
(5) Greene, J.; Kahn, S.; Savoj, H.; Sprague, P.; Teig, S. L. Chemical Function Queries for 3D Database Search. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1297−1308.
(6) Kolossvary, I.; Guida, W. C. Low Mode Search. An Efficient, Automated Computational Method for Conformational Analysis: Application to Cyclic and Acyclic Alkanes and Cyclic Peptides. *J. Am. Chem. Soc.* **1996**, *118*(21), 5011−5019.
(7) Schwab, C. H. Conformational Analysis and Searching. In *Handbook of Chemoinformatics;* Gasteiger, J., Ed.; Wiley-VCH: Weinheim, Germany, 2003; Vol. 1, pp 262−301.
(8) Nicklaus, M. C.; Wang, S.; Driscoll, J. S.; Milne, G. W. Conformational Changes of Small Molecules Binding to Proteins. *Bioorg. Med. Chem.* **1995**, *3*, 411−428.
(9) Marshall, G. R.; Motoc, I. Approaches To the Conformation of the Drug Bound To the Receptor. In *Molecular Graphics and Drug Design;* Burgen, A. S. V., Roberts, G. C. K., Tute, M. S., Eds.; Elsevier: Amsterdam, 1986; pp 115−156.
(10) Perola, E.; Charifson, P. S. Conformational Analysis of Drug-Like Molecules Bound to Proteins: An Extensive Study of Ligand Reorganization upon Binding. *J. Med. Chem.* **2004**, *47*, 2499−2510.
(11) Bostrom, J.; Norrby, P. O.; Liljefors, T. Conformational Energy Penalties of Protein-Bound Ligands. *J. Comput.-Aided. Mol. Des.* **1998**, *12*, 383−396.
(12) Diller, D. J.; Merz, K. M. Can We Separate Active from Inactive Conformations? *J. Comput.-Aided. Mol. Des.* **2002**, *16*, 105−112.
(13) *Omega, version 1.8.1*; OpenEye Scientific Software, Inc.: Santa Fe, NM, U.S.A., 2004.
(14) Beusen, D. D.; Shands, E. F. B.; Karasek, S. F.; Marshall, G. R.; Dammkoehler, R. A. Systematic Search in Conformational Analysis. *J. Mol. Struct.* **1996**, *370*, 157−171.
(15) http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml (accessed Mar 1, 2007).
(16) http://pubchem.ncbi.nlm.nih.gov (accessed Mar 1, 2007).
(17) Spellmeyer, D. C.; Wong, A. K.; Bower, M. J.; Blaney, J. M. Conformational Analysis Using Distance Geometry Methods. *J. Mol. Graphics Modell.* **1997**, *15*, 18−36.
(18) Halgren, T. A. Merck Molecular Force Field: I. Basis, Form, Scope, Parameterization and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.
(19) Halgren, T. A. Merck Molecular Force Field: VI. MMFF94s Option for Energy Minimization Studies. *J. Comput. Chem.* **1999**, *20*, 720−729.
(20) Mayo, S. L.; Olafson, B. D.; Goddard, W. A. Dreiding: a generic force field for molecular simulations. *J. Phys. Chem.* **1990**, *94*(26), 8897−8909.
(21) *OEChem, version 1.3.4*; OpenEye Scientific Software, Inc.: Santa Fe, NM, U.S.A., 2005.
(22) Ivakhnenko, A. G.; Ivakhnenko, G. A. The Review of Problems Solvable by Algorithms of the Group Method of Data Handling (GMDH). *Pattern Recognit. Image Anal.* **1995**, *5*(4), 527−535.
(23) Chen, J.; Anderson, J. B.; DeWeese-Scott, C.; Fedorova, N. D.; Geer, L. Y.; He, S.; Hurwitz, D. I.; Jackson, J. D.; Jacobs, A. R.; Lanczycki, C. J.; Liebert, C. A.; Liu, C.; Madej, T.; Marchler-Bauer, A.; Marchler, G. H.; Mazumder, R.; Nikolskaya, A. N.; Rao, B. S.; Panchenko, A. R.; Shoemaker, B. A.; Simonyan, V.; Song, J. S.; Thiessen, P. A.; Vasudevan, S.; Wang, Y.; Yamashita, R. A.; Yin, J. J.; Bryant, S. H. MMDB: Entrez's 3D-structure database. *Nucleic Acids Res.* **2003**, *31(1)*, 474−477.
(24) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(25) Kirchmair, J.; Wolber, G.; Laggner, C.; Langer, T. Comparative Performance Assessment of the Conformational Model Generators Omega and Catalyst: A Large-Scale Survey on the Retrieval of Protein-Bound Ligand Conformations. *J. Chem. Inf. Model.* **2006**, *46*, 1848−1861.
(26) *Catalyst, version 4.11*; Accelrys: San Diego, CA, U.S.A., 2006.
(27) Bostrom, J. Reproducing the Conformations of Protein-Bound Ligands: A Critical Evaluation of Several Popular Conformational Searching Tools. *J. Comput.-Aided. Mol. Des.* **2001**, *15*, 1137−1152.
(28) Kirchmair, J.; Laggner, C.; Wolber, G.; Langer, T. Comparative Analysis of Protein-Bound Ligand Conformations with Respect to

Catalyst's Conformational Space Subsampling Algorithms. *J. Chem. Inf. Model.* **2005**, *45*, 422−430.

(29) Kristam, R.; Gillet, V. J.; Lewis, R. A.; Thorner, D. Comparison of Conformational Analysis Techniques To Generate Pharmacophore Hypotheses Using Catalyst. *J. Chem. Inf. Model.* **2005**, *45*, 461−476.

(30) Bostrom, J.; Greenwood, J. R.; Gottfries, J. Assessing the Performance of Omega with Respect to Retriving Bioactive Conformations. *J. Mol. Graphics Modell.* **2003**, *21*, 449−462.

(31) Smellie, A.; Kahn, S. D.; Teig, S. L. Analysis of Conformational Coverage. 1. Validation and Estimation of Coverage. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 285−294.

(32) Smellie, A.; Kahn, S. D.; Teig, S. L. Analysis of Conformational Coverage. 2. Applications of Conformational Models. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 295−304.

CI7000956