Torsion Angle Preference and Energetics of Small-Molecule Ligands Bound to Proteins

Ming-Hong Hao,* Omar Haq,† and Ingo Muegge

Department of Medicinal Chemistry, Boehringer Ingelheim Pharmaceuticals, Inc., Ridgefield, Connecticut 06877

Received June 1, 2007

Small organic molecules can assume conformations in the protein-bound state that are significantly different from those in solution. We have analyzed the conformations of 21 common torsion motifs of small molecules extracted from crystal structures of protein−ligand complexes and compared them with their torsion potentials calculated by an ab initio DFT method. We find a good correlation between the potential energy of the torsion motifs and their conformational distribution in the protein-bound state: The most probable conformations of the torsion motifs agree well with the calculated global energy minima, and the lowest torsion-energy state becomes increasingly dominant as the torsion barrier height increases. The torsion motifs can be divided into 3 groups based on torsion barrier heights: high (>4 kcal/mol), medium (2−4 kcal/mol), and low (<2 kcal/mol). The calculated torsion energy profiles are predictive for the most preferred bound conformation for the high and medium barrier groups, the latter group common in druglike molecules. In the high-barrier group of druglike ligands, >95% of conformational torsions occur in the energy region <4 kcal/mol. The conformations of the torsion motifs in the protein-bound state can be modeled by a Boltzmann distribution with a temperature factor much *higher* than room temperature. This high-temperature factor, derived by fitting the theoretical model to the experimentally observed conformation occurrence of torsions, can be interpreted as the perturbation that proteins inflict on the conformation of the bound ligand. Using this model, it is calculated that the average strain energy of a torsion motif in ligands bound to proteins is ∼0.6 kcal/mol, a result which can be related to the lower binding efficiency of larger ligands with more rotatable bonds. The above results indicate that torsion potentials play an important role in dictating ligand conformations in both the free and the bound states.

INTRODUCTION

Accurate prediction of the protein-bound conformations of small molecules is an essential tool in structure-based drug design. A popular approach is to use sets of discrete rotamers of small motifs to build up complete ligands in search of docking poses or diverse conformations.[1−4] A rotamer library is built based on the preferred conformations of torsion motifs,[5,6] i.e. molecular fragments with a rotatable bond. Knowledge of the preferred conformations of basic torsion motifs of ligands can help in finding the bioactive conformations of drug molecules. The information on the preferred conformations of torsion motifs can be obtained either from experimental structures (such as the Cambridge Crystallographic Database (CSD)[7] and the Protein Data Bank (PDB))[8] or from theoretical calculations. When one applies the knowledge of conformations in torsion motifs to drug design, it is important to understand two basic questions: (1) How well does the rotamer state of torsion motifs—derived either from empirical structures or by theoretical calculation—represent the protein-bound state of ligand molecules? (2) How predictable is the preferred conformation of the torsion motifs in the bound state of a drug molecule?

Conformational preferences of the torsion motifs of small organic molecules bound to proteins have been studied in the past. Several groups have studied the relationship between the conformations of small molecules in solution and in the protein-bound state.[9−14] The prevalent view is that the protein-bound conformation of a small molecule usually does not correspond to its lowest-energy conformation in solution as calculated by theoretical methods,[13,14] with the caveat that current searching methods are not necessarily always able to identify the global minimum of medium to large flexible ligands. It has been reported that the calculated strain energy of a ligand structure in the bound state does not correlate with the binding potency of the ligand.[13] Hence, predicting the bioactive conformation of drug molecules based on simple conformational-energy calculations is not expected to be successful. It is nevertheless important to investigate what a more complicated role the conformational energy plays in the bound state of a ligand. We believe that information about torsion motifs, including the potential energy profile and experimentally observed conformational preference, can shed light on the nature of complete ligands.

There is a large number of publications devoted to high-level quantum chemistry analyses of the conformation of chemical motifs. The problems having been investigated range from thermodynamics[15−18] through material electronic properties[19,20] to spectroscopy.[21−23] However, to date, there has been no systematic analysis of the correlation between ab initio calculated torsion potential energies and the conformation of small-molecule ligands binding to proteins that are broadly applicable to medicinal chemistry and drug discovery.

* Corresponding author e-mail: mhao@rdg.boehringer-ingelheim.com.
† Current address: BioMaPS Institute for Quantitative Biology, Rutgers University, Piscataway, NJ 08854.

10.1021/ci700189s CCC: $37.00
Published on Web 09/19/2007

Table 1. List of Torsion Motifs

no.	motif class	no. of occurrences[a]	comments
1	sulfonamide	141	include alkylated sulfonamide
2	phenyl-sulfone	158	include phenyl-sulfonamides
3	thiazole-sulfone	39	
4	ester	424	
5	amide	678	exclude secondary amides
6	urea	76	
7	benzamide	302	exclude ligands with ortho substitutions on the phenyl
8	phenyl acetamide	73	exclude ligands with ortho substitution on the phenyl
9	thiazole acetamide	31	
10	anisole	571	exclude diortho substitutions on the ring
11	*N*-Me-aniline	464	include 6-member aromatics; exclude diortho substitutions on the ring
12	styrene	75	exclude structures with ortho substitutions on the phenyl group
13	biphenyl	53	exclude ortho substituted phenyl
14	ortho-substituted biphenyl	34	X is a non-hydrogen atom
15	phenylimidazole	21	include phenyls with ortho-substituted OH or F
16	diphenyl amine	169	include mono ortho-substituted phenyl in the class
17	diphenyl ether	111	include mono ortho-substituted phenyl in the class
18	benzophenone	40	include mono ortho-substituted phenyl in the class
19	phenyl ethyl	170	X is substituted non-H atoms, i.e., C, N, O.
20	3-ethylindole	131	include α-Me substitution on the ethylene
21	dimethoxyethane	578	

[a] The PDB structure codes of the ligands containing the torsion motifs are provided in the Supporting Information

In the drug design process, a torsion-angle scan by either a force field or an ab initio method is a common approach for the assessment of the likelihood that a ligand binds in a given conformation to a target protein. For practical reasons, such calculations typically do not take into account the interactions between protein and ligand. To understand the usefulness and limitation of the calculated structures for ligand design, it is imperative to formally assess the correlation between the calculated torsion potential and the protein-bound conformations of such molecules.

Here we survey the conformational distribution of 21 torsion motifs of small-molecule ligands in the protein-bound state by analyzing the X-ray structures of protein−ligand complexes in the PDB. The torsion potential of these motifs is calculated using a quantum chemistry Density Function Theory (DFT) method. The calculated torsion potential energy profiles are compared with the conformational distribution of the torsion motifs found in the protein−ligand complexes. Focusing on the torsion motifs rather than on complete ligands allows us to determine the complete potential energy surface for the torsion motifs using ab initio methods at higher accuracy than force field methods. A statistically significant comparison is made based on the large number of experimental observations of the torsion motifs because many different ligands can have the same motifs. From this study, we have determined the relationship between the torsion angle preferences and energetics of small-molecule ligands bound to protein. We believe that such information can shed light on the interactions between proteins and ligands and guide the prediction of the bioactive conformation of drug molecules.

MATERIALS AND METHODS

Selection of Common Torsion Motifs of Small-Molecule Ligands. Torsion motifs relevant to drug design have been identified by others. For example, the MIMUMBA program[5,6] compiles a large number of torsion motifs based on the Cambridge crystallographic small-molecule structure database. In the commercial software Omega,[6,12] 123 torsion rules are compiled for frequently encountered torsion motifs. From an examination of the torsion motifs compiled in the past, we observed that the majority of the motifs are analogs or derivatives of a smaller number of basic chemical fragments. When a variation or substitution of the basic motif does not change the conformational preference, the torsion potential energy of the basic motif can be used to describe the conformations of the derivatives. It is preferable to consider as many close analogs together in one torsion motif as possible so that more examples can be found from the available experimental structures of protein−ligand complexes.

Based on our own experience and using the Omega torsion library as a reference, we have selected 35 torsion motifs in our structure survey that occur in druglike molecules. However, from the PDB we have found only 16 of these torsion motifs with >50 *independent* conformational occurrences, i.e., conformations not related by crystallographic units in a given X-ray structure. An additional 5 torsion motifs yield 20−50 independent occurrences with well-recognized conformational preferences. We limit our analysis to these 21 motifs, because motifs without good sampling statistics cannot be subjected to a robust statistical analysis. While the selected 21 motifs may not cover the full range of relevant torsion motifs, we believe they are sufficient for the purpose of analyzing the relationship of torsion angle preference and energetics. Table 1 lists the selected motifs, indicates the number of occurrences of each motif observed in the crystal structures, and provides a chemical specification of the motifs. The chemical structures of these motifs are shown in Figure 1 in the Results section. Note that for all the 21 torsion motifs described here, all chemically feasible (heavy-atom) substitutions on the terminal atoms of the motifs are allowed except for those specified in Table 1. The following section details the process of identifying the selected torsion motifs from the PBD.

Compilation of the Torsion Angle Statistics of the Motifs. The Relibase program[24] was used to identify the small-molecule ligands in the PDB that contain the selected torsion motifs. The August 2006 version of the PDB was used in this study. The hit structures from the Relibase search were further filtered; only PDB structures that have a resolution <2.5 Å and that are not nucleic acids were extracted. A total of 55 013 raw ligand structures were extracted from the PDB. A Relibase search did not always extract all the ligands that contain a selected motif. When less than 50 ligands were found for a given motif by a Relibase search, a second substructure search on the same motif was carried out on the Ligand Depot database. The Ligand Depot is a database derived from the PDB and is available from the same source.[8] New structures found in the Ligand Depot database were added to the Relibase search results. Multiple copies of a ligand in a given PDB structure are considered as independent conformation occurrences at this step.

To calculate the torsion value for the large number of structures, the downloaded structures were first converted into molecular graphs using Marvin libraries.[25] Each molecular graph consists of nodes that represent atoms and vertices that represent bonds in the chemical structure of a torsion motif. Atoms were further annotated as chain or ring atoms, aromatic or aliphatic types. Bonds were differentiated between rotable or nonrotable. The rules for a match between a part of the molecular graphs and a query motif were written as script files incorporating the Marvin libraries. With these scripts, matches between the part of the ligand and the query motif could be automatically identified. Once a match was established, the torsion angle of the motif in a ligand was calculated by a script using the 3D coordinates of the ligand molecule and exported into a table. For further inspection, the output also provided information on the PDB structure from which the ligand had been extracted, the name and chain identifier of the ligand in the PDB structure, and the serial numbers of the 4 atoms that define the torsion angle. The script files incorporating the Marvin libraries described most of the rules for extracting the ligands and calculating the torsion motifs correctly. However, there were some incorrect structures that were included into a motif. To remove such incorrect structures, the structures identified for each motif by scripts were subjected to visual inspections, and the mismatched structures were removed.

The final step in compiling the torsion angle statistics was to remove nonindependent conformations among the multiple copies of a protein−ligand structure. In the majority of cases, multiple copies of a structure corresponded to an identical structure. For all such cases, multiple copies were counted as one conformation occurrence. However, there were a significant number of structures in which different copies of the same ligand exhibited large differences in torsion angle values that reflected true conformational variations of the torsion motifs. Consequently, we counted copies of a ligand that exhibited >10° torsion-angle difference as independent conformational occurrences. This approach retained on average ∼20% of the copies of a ligand in multicopy X-ray structures.

The Supporting Information of this paper provides the PDB structures, ligand identities, and the torsion angle values for each of the motifs studied here. Please note that the same motif can occur multiple times in a ligand; these occurrences are counted as independent conformations of the motif.

Calculation of the Torsion Energy Profile. All potential energies were calculated using the quantum chemistry software Jaguar version 6.5.[26] To test the proper computational procedure, the energy profiles of four motifs, sulfonamide, urea, dimethoxyethane, and *N*-methylaniline, have been calculated using the DFT/B3LYP and LMP2 methods. The basis sets of 3-21G*, 6-31+G*, 6-311++G**, and 6-311G-3DF-3PD* have been tested. We have found that the energy profiles calculated by the DFT/B3LYP and LMP2 methods are essentially similar for all the four motifs when the basis set of 6-311++G** is used. Because the LMP2 method is much slower than the DFT method, for consistency and practicality reasons, we have used the DFT/B3LYP method for all energy calculations reported in this work. We expect that, for the majority of torsion motifs practically studied by the torsion-scan procedure, the DFT method will give results comparable to those obtained with the LMP2 method. As to the effects of basis sets on the calculated torsion potential energy profiles, we found that calculations by DFT/B3LYP with 6-31+G*, 6-311++G**, and 6-311G-3DF-3DP basis sets give similar results for the selected motifs. Only the 3-21G* basis set gives sometimes quantitatively different results. Based on these tests, we believed that using the medium-sized 6-311++G** basis set produces results compatible with those obtained with larger basis sets. The torsion energy profiles for all the motifs reported here have been calculated with the 6-311++G** basis set.

To calculate the torsion energy profile, each of the motifs is first energy minimized with full geometry optimization. A torsion scan is then carried out on the optimized structure by varying the variable torsion angle between −180 and 180°, in 15° intervals. For each generated structure at a given torsion angle, a constrained energy minimization is carried out with a single constraint on the torsion angle to restrain the molecule to the specific torsion angle value, while the rest of the molecule is geometrically optimized. The energy of the molecule in the optimized structure at the convergence is taken as the energy of the specified conformational torsion. When the energy of the optimized structure of a motif strongly depends on two or more coupled torsion angles, e.g., motifs of the form R-X-R′ in which both bonds R-X and X-R′ are rotatable, multiple values of the secondary variables are chosen to generate the starting structures. For such cases, multiple torsion scans of the interested torsion are carried out with the different starting structures. The conformational series with the lowest energy is chosen as the torsion potential energy profile for the specific motif. In the torsion energy profiles of a motif, the energy of all conformations is expressed as the difference relative to the lowest energy; the lowest energy is taken customarily as zero.

For torsion motifs in which a hydrogen atom is directly attached to a nitrogen atom that defines the bond of rotation, the calculated potential energy profiles often show cusps at the energy maxima. Such cusps always associate with an abrupt configuration change of the sp^3 character of the nitrogen atom from one stereoisomer to another during the geometry minimization. The cusps do not affect the energy minima; however, they may slightly influence the estimate of the energy barrier height.

RESULTS AND DISCUSSION

Comparison of Experimental Conformations and Calculated Torsion Potentials. The torsion angle distributions of the 21 torsion motifs from protein−ligand structures and the calculated potential energy profiles of the motifs by the DFT method are shown in Figure 1. Each panel of Figure 1 corresponds to one motif: the chemical structure of the motif is shown on the left-hand side, and the four atoms that define the bond of rotation are specified. The torsion distribution of the motif is represented in a histogram showing the abundance (count) of torsion angle occurrences at different torsion values binned in 15° intervals. Because all the torsion motifs studied here show C_2 symmetry with respect to the bond of rotation, the conformational distributions of a motif at torsion angles $\pm\tau$ are statistically indistinguishable when averaged over a large number of structures. To gain better statistics for the torsion distribution from a limited number of experimental structures, the occurrences of conformations at torsion angles $\pm\tau$ of each motif are binned together in the histogram. The range of the torsion angle is therefore between 0 and 180°.

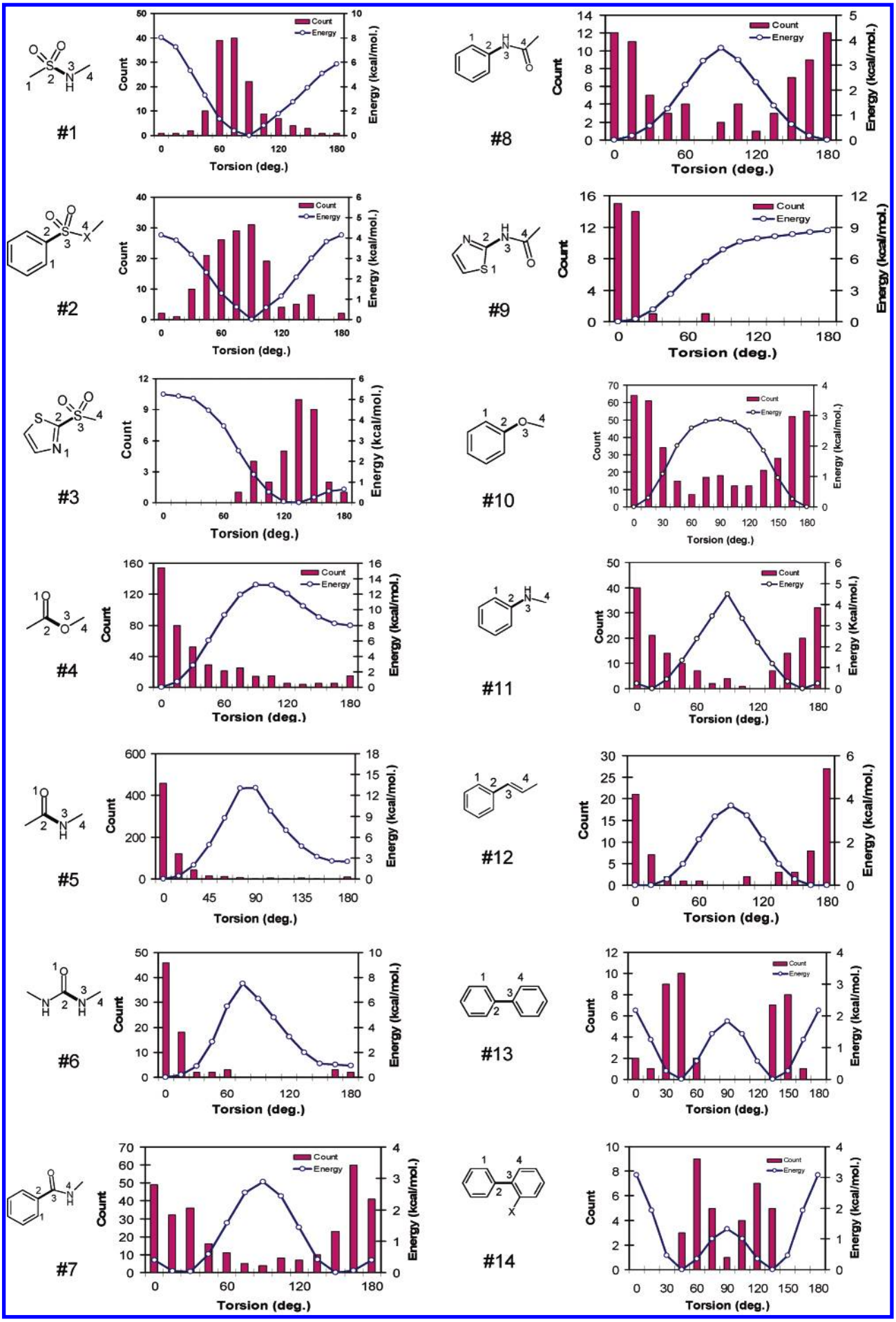
To compare the X-ray derived conformational distribution and the theoretical torsion energy profile, the histogram and energy are plotted together as a function of the torsion angle in the same graph (Figure 1). The potential energy is shown on the right axis, while the number of conformational occurrences is shown on the left axis. The 21 graphs of Figure 1 illustrate that for the majority of torsion motifs the potential energy minimum and the most populated conformation of each motif coincide. The conformational histogram is anticorrelated with the variations of the torsion energy profile. Bins holding conformations with lower torsion energies are more populated than those with higher torsion energies. It should be commented that, although the DFT method calculated the gas-phase energy of the motifs, this energy should be relevant for ligands bound in the low dielectric environment of proteins binding site.

The properties of the 21 torsion motifs can be better described through dividing the motifs into 3 groups based on the torsion potential barrier heights: high (>4 kcal/mol), medium (2−4 kcal/mol), and low (<2 kcal/mol).

The high torsion-barrier group includes the ester (#4), amide (#5), urea (#6), thiazole-amide (#9), and sulfonamide (#1) motifs. The conformation distribution of this group of motifs is characterized by a large population in a predominant state and few occurrences in high energy torsion-angle regions. Based on the data for the high-barrier group shown in Figure 1, 4 kcal/mol is an approximate cutoff value above which the chance of the occurrence of a conformational torsion is very small, i.e., <5%. The ratio of the number of conformations with an energy >4 kcal/mol to the total number of conformations for the motifs is the following: sulfonamide (#1): 6/141 = 0.042; amide (#5): 44/678 = 0.064; urea (#6): 3/76 = 0.039; thiazole amide (#9): 1/31 = 0.032; ester (#4): 138/424 = 0.32. The last motif, ester, appears to be an exception for which in 30% of the experimental structures the motif has an energy >4 kcal/mol. Examining the structures of the high-energy ester motifs (their PDB codes can be found in the Supporting Information, Supplement-1) indicated that most of such conformations occur in long-chain aliphatic esters and large lipidlike ligands. Such ligands are intrinsically flexible and have extensive packing and interaction with the protein. It is possible that the packing and interaction with the protein suppresses the torsion energy barrier of these ester motifs. The structures of such ester molecules are fundamentally different from conventional druglike molecules. The about 5% high-energy (>4 kcal/mol) conformations among the 4 other motifs occur in more druglike ligands. Examination of those high-energy conformations indicated that such a motif may (1) have a high thermal factor, (2) connect to charged groups and therefore may be subjected to strong interaction from the protein, and (3) be poorly resolved end groups (such as PDB structures 1AJ8, 1HV7, 1HLH, and 1DMY). Considering the majority of the conformations (>95%), it is reasonable to state that 4 kcal/mol is a torsion barrier above which there are only rare occurrences of conformational torsions. The bioactive conformation of high torsion-barrier motifs therefore can be predicted with high confidence.

The second group includes torsion motifs with the torsion barrier heights between 2 and 4 kcal/mol. Examples of this group include aryl amides (#7, 8), aromatic amines and ethers (#10, 11), biaryls (#13, 15), styrene (#12), and potentially many moderately substituted alkane motifs not covered in this study. For this group of motifs, we observe from Figure 1 that there are well recognizable, preferred conformations that are predictable using DFT calculated torsion potentials. For aromatic amine and ether motifs, as shown in panels 10 and 11 of Figure 1, the experimental structures show a preference for a planar conformation between the amine/ether and the aromatic ring in the protein-bound state. Such a preference extends to aromatic amine and ether structures with a single ortho substitution on the aromatic group. However, diortho substitutions in the aromatic ring force the amine or ether group out of plane with respect to the ring. The experimental structures of aromatic amides and aromatic acetamides show a preference for planar conformations also. In these two motifs, a single substitution at the ortho position of the aromatic ring or alkylation of the amide significantly changes the potential energy profile and the conformation preference. Based on the results shown in Figure 1, it appears that for the medium barrier-height motifs calculated potential energy profiles can predict the most probable conformation assumed by the ligand in the bound state most of the time. However, there is a high abundance of conformations at higher energies for a number of motifs in this group. Therefore, the proper conformation predictions in practice may be achieved by a Boltzmann energy weighted probabilistic approach as elaborated in a later section.

The low energy-barrier class of motifs includes ortho-substited biphenyl (#14) and ortho-substituted phenyl amides and moderately substituted alkanes (#20, 21). For these motifs, the torsion potential is <2 kcal/mol, and there are multiple energy shoulders and shallow valleys in the torsion space. Although calculated energy profiles for such motifs show energy minima, the low energy-barrier motifs do not exhibit a clear conformation preference in the protein-bound state, and the energy profiles do not correlate with the experimental conformational distributions. As an example, when an ortho substitution (#14) is introduced to the biphenyl (#13) motif, the barrier height at a torsion angle of 90° is reduced, and there is a significant increase in the population of conformational torsions near the torsion value of 90°.


#1
#2
#3
#4
#5
#6
#7
#8
#9
#10
#11
#12
#13
#14
Count
Energy
Energy (kcal/mol.)
Energy (Kcal/mol.)
Torsion (deg.)

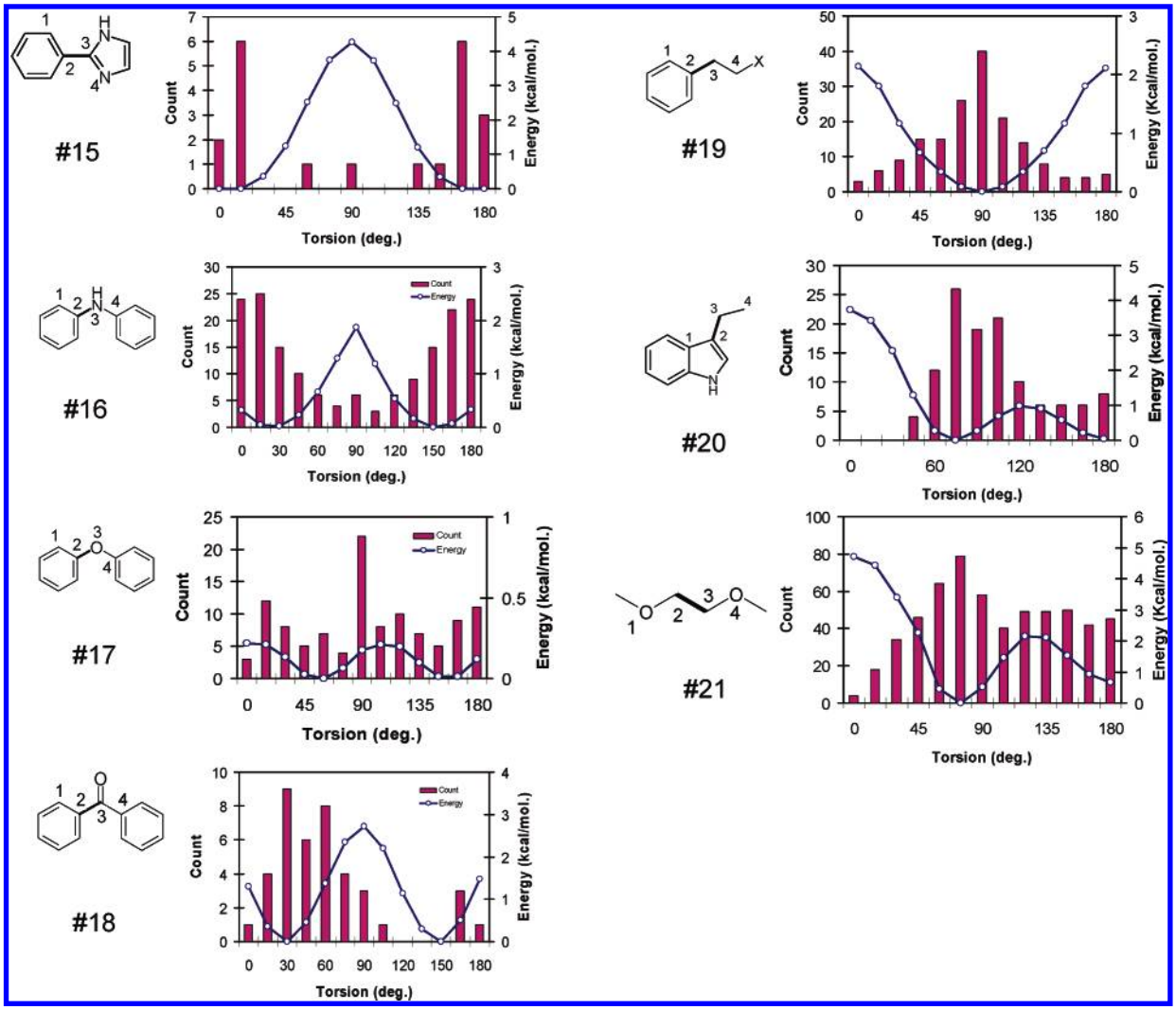


Figure 1. The conformational histograms and potential energy profiles of 21 chemical motifs are depicted as a function of the torsion angle. Each panel represents one motif; its chemical structure is depicted on the left-hand side. The conformational histogram derived from PDB X-ray structures is represented as a bar diagram, and the value of the bar is given on the left axis, indicating the number of occurrences of the torsional motif within the given torsion range of the indicated value ± 7.5°. The torsion potential energy calculated by the DFT B3LYP/6-311++G** method as a function of the torsion angle is shown as the curve; the energy value is given on the right axis. The lowest-energy conformation is set to zero.

Two cases from different groups are highlighted here: the conformational preference of sulfur-containing heterocyclic amides, exemplified by the thiazole acetamide motif (#9) and the dimethoxyethane motif (#21).

The experimentally found conformation distribution of the thiazole acetamide indicates a preference for the sulfur atom in the heterocycle to be in a cis conformation with respect to the carbonyl oxygen atom in the amide group. The DFT calculated potential energy profile is in agreement with the experimental structure of the thiazole acetamide motif. Such a preference extends to thiophene amides for which a DFT calculation predicts that the carbonyl oxygen has a preference to be in the cis conformation with respect to the sulfur atom. We have found only 7 independent occurrences of the thiozole acetamide motif in the PDB (codes: 1nmq, 1i00, 1hvy, 2kce, 1utz, 1mq5, and 2tsr). All the thiozole acetamide motifs show a cis conformation between the sulfur and oxygen atoms.

The calculated energy profile of the dimethoxyethane motif (#21) shows three energy minima at gauche (torsion angle ±60°) and trans (torsion angle 180°) conformations. The calculated potential energy of the gauche conformation is 0.67 kcal/mol lower than the trans conformation. From the 578 conformational occurrences extracted from the PDB for this motif, it is found that the gauche conformation indeed occurs about twice as often as the trans conformation. Based on this example, we expect that even for low barrier-height (<2 kcal/mol) motifs, one can detect a preference for lower-energy states when there is a sufficiently large number of experimental structures.

The discussion up to this point has been based on conformational rotations of a single bond. This is a valid approximation for motifs consisting of two rigid chemical groups, such as aromatic rings or conjugated systems, linked by a single bond of rotation. However, in many ligands, the conformation preferences depend on two or more rotable bonds that are coupled. An accurate description of the conformation preference for such motifs requires the calculation of a high-dimensional potential surface. The three linked diphenyl motifs, i.e., diphenyl amine (#16), diphenyl ether

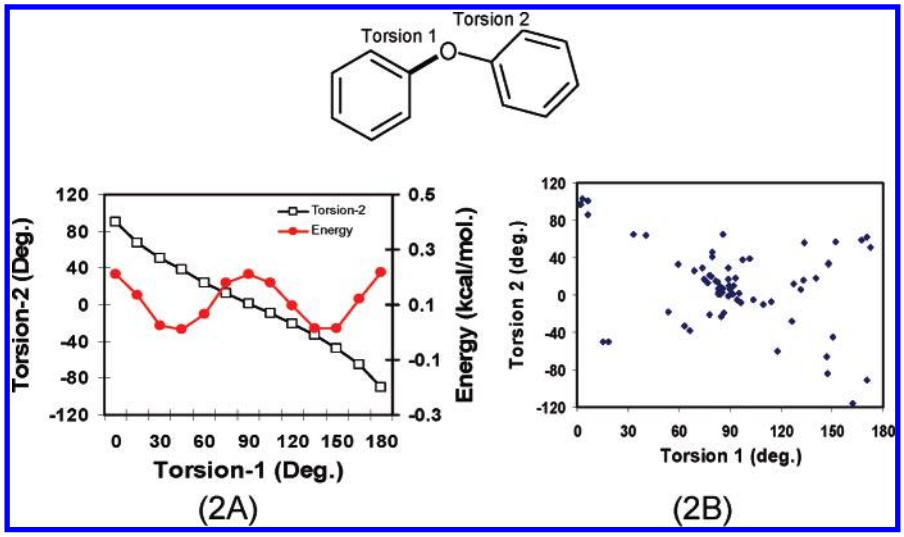


Figure 2. (A) The black curve referring to the left axis shows the correlated variation of the second torsion angle (torsion 2) with respect to the change of the first torsion angle in the minimized structure of diphenyl ether. The red curve referring to the right axis depicts the energy of the molecule as a function of torsion 1. (B) Distribution of the conformational occurrences of the diphenyl ether motifs in PDB structures. Each dot represents one structure with two torsion values specified on the *x* and *y* axes. The density of the dots reflects the probability of the conformations to be assumed as defined by the two torsions.

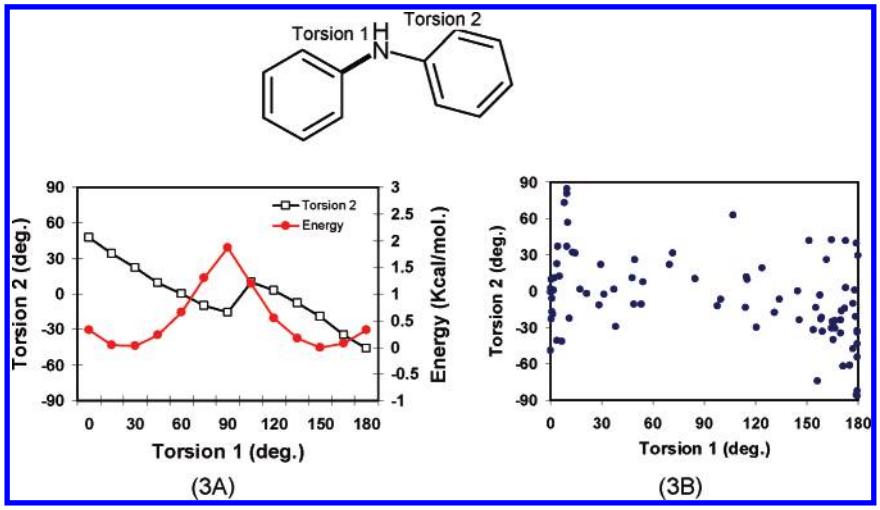


Figure 3. (A) The black curve referring to the left axis shows the correlated variation of the second torsion angle (torsion 2) with respect to the change of the first torsion angle in the minimized structure of diphenyl amine. The red curve referring to the right axis shows the energy of the molecule as a function of torsion angle 1. (B) Distribution of the conformation of the diphenyl amine motif in PDB structures.

(#17), and benzophenone (#18), exemplify correlated conformation preferences between two connected bonds. As one varies one of the torsion angles in these linked diphenyl motifs, the second torsion angle changes in a concerted way to optimize the entire molecule keeping it on a low-energy trajectory on the potential energy surface. Figures 2−4 show the correlated variations between two torsion angles and associated energy variations for diphenyl ethers (2A), diphenyl amines (3A), and benzophenones (4A), respectively. All three motifs show a similar correlation; as one torsion angle increases, the second torsion angle decreases. The low-energy conformations of these linked diphenyl systems lie on two symmetry-related trajectories in which one torsion angle assumes an in-plane conformation, while the second torsion assumes an out-of-plane conformation with either positive or negative torsion angle values.

Figures 2B−4B show the correlated torsion angle variations of diphenyl ethers (2B), diphenyl amines (3B), and benzophenones (4B) observed in the crystal structures of protein−ligand complexes. Each point in Figures 2B−4B corresponds to one occurrence of the motif in the crystal structures. A comparison of panels A with B in Figures 2−4 reveals that the distributions of conformational torsions resemble the calculated low-energy trajectories. The one-dimensional conformation histogram of the diphenyl ether motif (Figure 1, #17) shows no conformational preference because the potential energy profile is almost flat (±0.2 kcal/mol) over the entire torsion angle range. However, in the two-dimensional distribution (Figure 2A,B), one can see that in fact most conformations lie in two diagonal lines (X shaped) that correspond to two low-energy trajectories. For the diphenyl amine (3B) and benzophenone (4B) motifs, the conformation distribution concentrates in the low-energy areas in the two-dimensional space. The experimental data do not provide adequate sampling over the entire conformational space.

The linked diphenyl systems suggest that an analysis of the high-dimensional conformational surface is more revealing than that of a single torsion angle space. However, the available experimental structures are sparse when projected

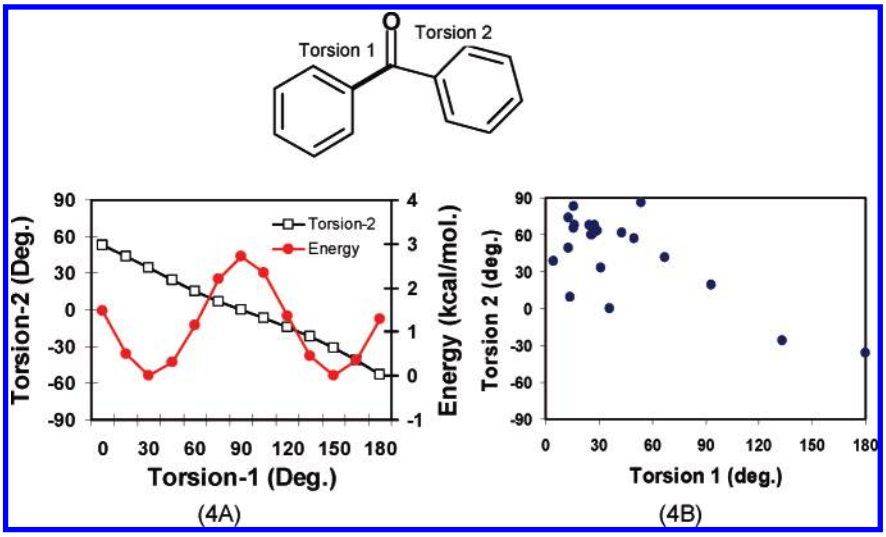


Figure 4. (A) The black curve referring to the left axis shows the correlated variation of the second torsion angle (torsion 2) with the change of the first torsion angle in the minimized structure of benzophenone. The red curve referring to the right axis shows the energy of the molecule as a function of torsion angle 1. (B) Distribution of the conformation of benzophenone motif in PDB structures.

onto high dimensions making a quantitative correlation analysis of the experimental conformation distributions and the theoretical calculation unreliable. Therefore, we have carried out the correlation analysis in single torsion angle space which is relevant for the majority of the torsion motifs studied here.

A Model for the Conformational Distribution of Torsion Motifs in the Protein-Bound State. Based on a correlation analysis of the experimental data and the calculated torsion potentials, we develop a statistical model describing the conformation distribution of the torsion motifs of ligands in the protein-bound state. There are 12 motifs studied here for which there are more than 100 independent conformational occurrences. The correlation analyses of these motifs are considered statistically sound. But there are 5 motifs for which there are less than 50 conformational occurrences. We have formally analyzed all these motifs applying the same procedure although only the results for motifs with sufficient sampling can be considered statistically significant.

To understand the physical basis of the correlation between the conformational distribution of the motifs and the torsion energy, we have utilized a Boltzmann model. The Boltzmann distribution function is the product of two components: a prefactor and an exponential function of energy over temperature. Based on this model, a linear function between the logarithm of the conformational count (number of occurrences) at torsion τ, denoted as $N(\tau)$), and the torsion energy $E(\tau)$ can be obtained

$$-\mathrm{Log}\, N(\tau) = bE(\tau) + b0 \qquad (1)$$

where b and $b0$ are two constants which can be interpreted as the inverse temperature and logarithm of the prefactor, respectively. To determine the two parameters of this model, a least-squares fitting has been carried out between the conformational histogram and the torsion energy profile for each of the motifs. This fitting approach is complementary to a potential of mean force approach[5,6,28−30] where the distribution of conformational probability is used to derive a Helmholtz free energy that can be interpreted as the energy required for a torsion angle to adopt a given value. The present approach reconvolutes the experimental conformational distribution into the molecular energetic components that govern the distribution probability through the Boltzmann model.

Table 2. Statistical Parameters of the Boltzmann Models for each Torsion Motif[a]

no.	motif class	b	$b0$	r^2	s	$F_{1,11}$	E_s
1	sulfonamide	0.48	3.23	0.82	0.592	49.51	0.96
2	phenyl-sulfone	0.69	3.52	0.66	0.741	21.47	0.61
3	thiazole-sulfone	0.30	1.37	0.54	0.606	12.78	1.03
4	ester	0.19	4.39	0.55	0.763	13.60	3.79
5	amide	0.20	3.34	0.28	1.451	4.24	2.69
6	urea	0.26	1.58	0.29	1.017	4.58	1.37
7	benzamide	0.74	3.53	0.74	0.461	31.93	0.25
8	phenyl-acetamide	0.52	2.20	0.66	0.501	21.47	0.56
9	thiazole amide	0.21	1.58	0.50	0.713	10.81	2.71
10	anisole	0.58	4.08	0.85	0.279	61.70	0.67
11	*N*-Me-aniline	0.65	3.04	0.65	0.726	20.33	0.41
12	styrene	0.61	1.87	0.52	0.813	12.07	0.41
13	biphenyl	0.95	1.73	0.58	0.642	14.93	0.19
14	ortho-substituted biphenyl	0.51	1.34	0.37	0.706	6.41	0.34
15	phenyl imidazole	0.26	0.81	0.36	0.558	6.31	0.92
16	diphenyl amine	0.92	2.82	0.54	0.492	13.14	0.10
17	diphenyl ether	0.11	0.97	0.01	0.496	1.01	0.66
18	benzophenone	0.41	1.79	0.24	0.743	1.87	0.55
19	phenyl ethyl	0.92	3.15	0.83	0.336	52.07	0.17
20	3-ethyl indole	0.37	3.71	0.74	0.335	30.8	0.89
21	dimethoxyethane	0.32	3.59	0.66	0.425	23.97	0.89

[a] b and $b0$ are the constants of a least-square fitting of the logarithm of the torsion population to the torsion energy. $b = 1/KT$ in the Boltzmann model. r^2 is the squared correlation coefficient, s is the standard deviation, and $F_{1,11}$ is the value of the F test at 13 torsion points (intervals) for the least-square fitting. E_s, in units of kcal/mol, is the calculated strain energy of the torsion motifs bound to proteins.

Table 2 shows the parameters and statistics of the Boltzmann model for each of the motifs. The quality of fit of the model for each motif is measured by the correlation coefficient r^2, the standard deviation s, and the F value. These are two-parameter models (b and $b0$) fitting over 13 torsion points (intervals). A 95% of probability of significance corresponds to an F value of 4.84. Based on this measure, the models for 4 of the motifs: amide (#5), urea (#6), diphenyl ether (#17), and benzophenone (#18) are not statistically significant. The amide and urea motifs are very

stiff and have high torsion barriers. They are rather considered rigid groups (as aromatic rings) than rotable motifs. Diphenyl ether and benzophenone belong to correlated motifs that, as discussed earlier, should be analyzed in a 2D torsion space rather than in single-torsion space. The models for most other models have an F value >10 and therefore are statistically significant models. Especially, for motifs with large sample size, the models tend to show large r^2 values (3 of them > 0.8) and small standard deviations; the models for these motifs are therefore considered robust.

The Boltzmann model provides a physical explanation of the conformational distribution of the torsion motifs. In reference to eq 1, the most interesting parameter is b, which is related to the absolute temperature: $b = 1/KT$, with K being the Boltzmann constant. From Table 1 we observe that for most motifs, $b < 1$. At room temperature (300 K) the conformational torsions in the free state are expected to be distributed according to a Boltzmann model with the constant $b = 1.69$. For 9 motifs that have good sampling size and robust statistics ($F > 20$, $r^2 > 0.65$), the b parameter varies between 0.32 and 0.9, corresponding to temperatures between 1584 and 563 K. The ligands appear to be in a heat bath under this environment. This result expresses a simple physical phenomenon. When a ligand binds to a protein, in addition to the normal thermal fluctuations, the conformation of the ligand is subjected to a variety of perturbations originating from protein−ligand interactions. In some complexes, the protein−ligand interactions are stronger than in others. The interactions can be either attractive or repulsive that push or pull the conformation of the ligand to higher conformation-energy regions. Averaging over all these independent conformational states assumed by the ligand in different protein−ligand complexes, the conformation of the ligand will appear to fluctuate in a much higher temperature environment. It is this phenomenon that is captured in the parameter b.

The above analysis can be used to understand the physical meaning of the statistical preferences, sometimes interpreted as potentials of mean force, derived from collections of experimental structures.[5,6,28−30] First, the present analysis demonstrates that the conformational distribution of ligands in the protein-bound state fits a Boltzmann model. Second, the model suggests that the potential of mean force actually consists of two components: one coming from the intrinsic torsion potential; the other arising from statistically random interactions between the ligand and a variety of receptors. As a side comment, it shall be noted that because the potential of mean force has absorbed these two contributions, one should not add the conformational energy of ligands to a scoring function derived by potentials of mean force to avoid calculating the conformational energy twice.

The model developed here can be further used to estimate the strain energy of ligands bound to proteins. For free ligands at room temperature, the internal torsion energy of a motif can be calculated as

$$<E> = \Sigma_\tau[E(\tau)\exp(-E(\tau))/KT]/\Sigma_\tau[\exp(-E(\tau))/KT] \quad (2)$$

where T is the room temperature. When the ligand binds to the protein, the model derived above indicates that the conformational distribution of the torsion motif occurs at a higher temperature T^*. The internal energy of the torsion motif in the protein-bound state, $<E^*>$, can be calculated using eq 2 with T^* replacing T. The strain energy of the motif in the bound state is then calculated as

$$E_s = <E^*> - <E> \quad (3)$$

The property E_s calculated by eq 3 is a statistical-mechanical internal energy. Hence it is related to the strain energy of the torsion motif. E_s does not include the effects of conformational entropy on a bound ligand. Column 8 of Table 2 shows the calculated strain energy of the torsion motifs according to eq 3. Including the 9 motifs that have robust statistical models (#1, 2, 7, 8, 10, 11, 19−21), the average strain energy of these torsion motifs due to binding to the protein is ∼0.6 kcal/mol. The strain energy of a complete ligand bound to a protein can be estimated by summing up the strain energy of individual torsion motifs. For a ligand with 5 torsion motifs, the strain energy of the ligand in the bound state would be ∼3 kcal/mol. Ligands with more rotable bonds are expected to have higher strain energy in the bound state. Perola and Charifson[13] have reported that the median strain energy for bound ligands with 4−6 rotable bonds is ∼3 kcal/mol and that ligands with more rotable bonds show on average higher strain energy in the bound state. Although the strain energies are calculated differently by Perola and Charifson (energy minimization) and by us (statistical mechanics), the qualitative picture of Perola and Charifson's conclusion is similar to ours.

The approximation that the total strain energy of a ligand is calculated by summing up the strain energy of individual torsion motifs provides a partial explanation of the low binding efficiency of many large ligands.[31,32] Kuntz et al.[31] have observed that as the size of ligands increases the binding affinity of most ligands reaches a plateau despite the fact that larger ligands form more extensive interactions with the proteins. Part of the cause for the low binding efficiency of large ligands can be attributed to the low probability of a matching interaction between the binding-site residues of a protein and the functional groups of a ligand. A large ligand needs to adjust its conformation to form favorable interactions and/or to avoid unfavorable interactions with the protein in the bound state. In the course of conformational adjustment, the strain energy of larger ligand builds up that can partially cancel the binding interaction between the protein and the ligand. Furthermore, if the ligand strain energy becomes so high, the ligand will not be able to assume certain bioactive conformations that might have better binding energy. This scenario is consistent with the calculated strain energy of the torsion motifs obtained above.

CONCLUSION

We have studied the relationship between the torsion potentials and the conformation preference of torsion motifs of druglike molecules bound to proteins. Correlation analyses have been carried out between potential energy profiles calculated by a DFT method and the conformation distribution of 21 torsion motifs extracted from crystal structures of protein−ligand complexes. For the majority of the torsion motifs studied here, there are a good number of experimental structures available so that the comparison between calculated energy and experimental conformation is statistically robust.

The most probable conformation of torsion motifs in the protein-bound state is generally consistent with the energy minima of the potential energy. The higher the torsion energy barrier the more predictable is the experimentally observed protein-bound conformation of a torsion motif. The torsion motifs have been classified into 3 groups: high-barrier, medium-barrier, and low-barrier motifs. Conformational torsions with energies higher than 4 kcal/mol as calculated by the present DFT method are rarely assumed in bound ligands, with the estimated probability being <5% from the experimental structures analyzed in this study. For the high torsion-barrier motifs the bioactive conformation is easier to predict. The majority of the torsion motifs of drug molecules are expected to fall in the medium barrier region (2−4 kcal/mol), where the potential energy can still be a good predictor for the most likely bound conformation.

The conformations of torsion motifs in the protein-bound state are subject to strong perturbations. The distribution of the torsion angles can be described by a Boltzman model in which the dispersion of the conformations occurs at much higher temperature than what would be expected for free ligands at room temperature. The higher temperature is a quantitative parameter to characterize the perturbation effects of proteins on the conformation of bound ligands. This perturbation is responsible for the strain energy of the ligand in its protein-bound state. It is found that ligands bound to a protein typically sustain approximately 0.6 kcal/mol of strain energy per rotatable bond.

The importance of the torsion potential in determining the conformation of ligands bound to proteins is characterized in this study by two quantities: a cutoff torsion energy value, 4 kcal/mol above which it is unlikely for a ligand to assume such torsion, and the average strain energy per torsion motif of 0.6 kcal/mol. These two quantities are related to but different from each other. In a given ligand, one or two torsion motifs can have higher strain energy, while most other torsion motifs have minor strains. In practical drug design, we expect that ligand conformations are accessible when all the torsion motifs of the ligand have energies below the cutoff value 4 kcal/mol. However, due in part to the specifics of the protein binding site, most designed ligands may still sustain a strain energy of 0.6 kcal/mol per torsion motif on average.

The torsion energy profiles calculated in this study take into account all bonded and nonbonded intramolecular interactions within a torsion motif. For relatively small or reasonably rigid ligands, the torsion energy profile represents a major portion of the potential energy surface of the molecule. The close correlation between the experimental conformations of torsion motifs of protein-bound ligands and the calculated torsion profiles presented in Figure 1 indicates that such a calculation has predictive value in determining the most likely bioactive conformation at the torsion motif level. However, the present approach has limitations. It is not clear that this approach can be straightforwardly extended into generating the complete potential energy surface of flexible ligands; the computational cost increases exponentially with the number of torsion motifs in a ligand. Another limitation of the present study is the analysis of the high-energy (>4 kcal/mol) conformations of torsion motifs. Such an analysis will be useful for setting up a rigorous cutoff criterion for acceptable ligand conformations in drug design. However, such work will require detailed information about the electron density of the ligands in the X-ray structures and a study of the computational accuracy with newer generation DFT and/or higher level ab initio methods, which can be pursued in the future.

ACKNOWLEDGMENT

We would like to thank Drs. Bennett T. Farmer and John Regan for advice and support of this study. Helpful suggestions by anonymous referees are also gratefully acknowledged.

Supporting Information Available: PDB codes of the protein−ligand complex structures with the torsion motifs. Data in two files: torsion motifs 1−11 (Supplement-1) and motifs 12−21 (Supplement-2). For each motif, the PDB code, the identity of the ligand from which the motif has been extracted, and the measured torsion angle are provided, respectively, in three columns. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Klebe, G.; Mietzner, T. A Fast and Efficient Method to Generate Biologically Relevant Conformations. *J. Comput.-Aided Mol. Des*. **1994**, *8*, 583−606.

(2) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A Fast Flexible Docking Method Using an Incremental Construction Algorithm. *J. Mol. Biol*. **1996**, *261*, 470−489.

(3) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved Protein-Ligand Docking Using GOLD. *Proteins* **2003**, *52*, 609−623.

(4) Boström, J. Reproducing the Conformations of Protein-bound Ligands: A Critical Evaluation of Several Popular Conformational Searching Tools. *J. Comput.-Aided Mol. Des*. **2001**, *15*, 1137−1152.

(5) Klebe, G.; Mietzner, T.; Weber, F. J. Methodological Developments and Strategies for a Fast Flexible Superposition of Drug-size Molecules. *J. Comput.-Aided Mol. Des*. **1999**, *13*, 35−49.

(6) Sadowski, J.; Boström, J. MIMUMBA Revisted: Torsion Angle Rules for Conformer Generation Derived from X-ray Structures. *J. Chem. Inf. Model*. **2006**, *46*, 2305−2309.

(7) 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 Version 3 and 4 of the Cambridge Structural Database System. *J. Chem. Inf. Comput. Sci*. **1991**, *31*, 187−204.

(8) (a) 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. (b) 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*, 2153−2155.

(9) 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.

(10) Boström, J.; Norrby, P. O.; Liljefors, T. Conformational Energy Penalties of Protein-Bound Ligands. *J. Comput.-Aided Mol. Des*. **1998**, *12*, 383−396.

(11) Vieth, M.; Hirst, J. D.; Brooks, C. L., III Do Active Site Conformations of Small Ligands Correspond to Low Free-energy Solution Structures? *J. Comput.-Aided Mol. Des*. **1998**, *12*, 563−572.

(12) (a) Boström, J.; Greenwood, J. R.; Gottfries, J. Assessing the Performance of OMEGA With Respect to Retrieving Bioactive Conformations. *J. Mol. Graphics Modell*. **2003**, *21*, 449−462. (b) *OMEGA, version 2.0*; OpenEyes Scientific Software: Santa Fe, NM, 2006.

(13) 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.

(14) Tirado-Rives, J.; Jorgensen, W. L. Contribution of Conformer Focusing to the Uncertainty in Predicting Free Energetics for Protein-Ligand Binding. *J. Med. Chem.* **2006**, *49*, 5880−5884.

(15) Allinger, N. L.; Grev, R. S.; Yates, B. F.; Schaefer, H. F., III The syn Rotational Barrier in Butane. *J. Am. Chem. Soc*. **1990**, *112*, 114−118.

(16) Murphy, R. B.; Beachy, M. D.; Friesner, R. A.; Ringnalda, M. N. Pseudo-Spectral Localized Moller-Plesset Methods: Theory and Calculation of Conformational Energies. *J. Chem. Phys*. **1995**, *103*, 1481−1490.

(17) Murcko, M. A.; Castejon, H.; Wiberg, K. B. Carbon-Carbon Rotational Barriers in Butane, 1-Butene, and 1,3-Butadiene. *J. Phys. Chem*. **1996**, *100*, 16162−16168.

(18) Wiberg, K. B.; Bohn, R. K. Conformations of Ethyl Esters versus Thiolesters. *Theor. Chem. Acc.* **1999**, *10*, 272−278.

(19) Cacelli, I.; Prampolini, G. Torsional Barriers and Correlations between Dihedrals in p-Polyphenyls. *J. Phys. Chem. A* **2003**, *107*, 8665−8670.

(20) Gatti, C.; Frigerio, G. Steric and Electronic Effects in Methyl-Substituted 2,2′-Bipyrroles and Poly(2,2′-bipyrrole)s: Part II Theoretical Investigation on Monomers. *Chem. Mater*. **2000**, *12*, 1490−1499.

(21) Herrebout, W. A.; van der Veken, B. J.; Wang, A.; Durig, J. R. Enthalpy Difference Between Conformers of n-Butane and the Potential Function Governing Conformational Interchange. *J. Phys. Chem.* **1995**, *99*, 578−585.

(22) Cheeseman, J. R.; Frisch, M. J.; Devlin, F. J.; Stephens, P. J. Hartree-Fock and Density Functional Theory *ab Initio* Calculation of Optical Rotation Using GIAOs: Basis Set Dependence. *J. Phys. Chem. A* **2000**, *104*, 1039−1046.

(23) Wiberg, K. B.; Vaccaro, P. H.; Cheeseman, J. R. Conformational Effects on Optical Rotation. 3-Substituted 1 Butenes. *J. Am. Chem. Soc.* **2003**, *125*, 1888−1896.

(24) *Relibase+, version 2.1.1*; The Cambridge Crystallographic Data Center: Cambridge, U.K., 2006.

(25) *Marvin, version 4.0.6*; Chemaxon. www.chemaxon.com (accessed July, 2006).

(26) *Jaguar, version 6.5*; Schrödinger LLC: Portland, OR, 2006.

(27) Hansch, C.; Leo, A. *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology*; ACS Professional Reference Book, American Chemical Society: Washington, DC, 1995; pp 535−541.

(28) Bohm, H.-J. The Development of a Simple Empirical Scoring Function to Estimate the Binding Constant for a Protein Ligand Complex of Known 3-Dimensional Structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243−256.

(29) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical Scoring Functions: 1. The Development of a Fast Empirical Scoring Function to Estimate the Binding Affinity of Ligands in Receptor Complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425−445.

(30) Muegge, I. PMF Scoring Revisited. *J. Med. Chem*. **2006**, *49*, 5895−5902.

(31) Kuntz, I. D.; Chen, Z.; Sharp, K. A.; Kollman, P. A. The Maximum Affinity of Ligands, *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9997−10002.

(32) Hajduk, P. J. Fragment-based Drug Design: How Big is too Big? *J. Med. Chem*. **2006**, *49*, 6972−6976.

CI700189S