# Strength in flexibility: Modeling side-chain conformational change in docking and screening

Leslie A. Kuhn

Departments of Biochemistry & Molecular Biology, Computer Science & Engineering, and Physics & Astronomy, and the Quantitative Biology and Modeling Initiative

502C Biochemistry Building, Michigan State University, East Lansing, Michigan, 48824, USA

Tel. (517) 353-8745. Fax: (517) 353-9334

E-mail: KuhnL@msu.edu.

URLs: http://www.bch.msu.edu/labs/kuhn *and* http://QBMI.msu.edu

Modeling protein flexibility in structure-based drug design and virtual screening remains a strong challenge due to the number of degrees of freedom involved and the co-optimization of protein and ligand shape and chemistry. However, there is a growing trend towards incorporating some protein side-chain flexibility modeling in docking, which enables better ligand positioning and scoring, which in turn can enhance the success of virtual screening. Here, we present several methods used for side-chain flexibility modeling in docking and recent insights gained from analyzing conformational transitions between ligand-free and bound crystal structures.

## BACKGROUND

### Improving Docking and Screening through Side-chain Flexibility Modeling

Several studies have shown that better sampling of motions during docking, including sampling ligand orientations more finely and modeling induced fit between the protein and ligand, improves the ability of protein-ligand complementarity scoring functions to detect the most accurate docking (1-4). For instance, when using the docking and screening tool SLIDE (Screening for Ligands with Induced-fit Docking, Efficiently) to dock 42 known thrombin ligands and 15 glutathione S-transferase (GST) ligands into the apo protein structures (reflecting the ligand-free binding site conformations), only 9 of the 42 thrombin ligands and 9 of the 15 GST ligands could be docked without modeling protein conformational change, even when the ligands were provided in their protein-bound conformations (3). Modeling modest conformational change – by choosing the single bond(s) in the protein or ligand that could resolve the steric overlaps with the smallest rotation - allowed 86% of the thrombin ligands and 93% of the GST ligands to be docked accurately (to within 1.3 Å ligand RMSD, on average). The same approach for modeling side-chain flexibility allowed SLIDE to identify 9 out of 10 known thymidine kinase ligands within the 40 top-scoring compounds (and 6 in the top 25 compounds), when using the unbiased, ligand-free conformation of thymidine kinase as the screening template and a database of 80,000 conformers of ligand candidates (representing low-energy conformations of the known ligands mixed with a set of 1,000 drug-like molecules) (5). In a virtual screening project to discover inhibitors for asparaginyl-tRNA synthetase from *Brugia malayi* (a human parasite causing elephantiasis), this approach resulted in a 15% hit rate; 7 out of 45 compounds identified by SLIDE were experimentally confirmed to be micromolar inhibitors (6).

,

The need for side-chain flexibility modeling is not so obvious from the numerous redocking studies that have been published, which tend to emphasize the ability of existing docking tools to predict ligand binding modes to within 2 Å RMSD across a range of protein structures and ligand chemical classes. However, even in this easy case of redocking, in which the protein and ligand structures are provided in their bound conformations, the best of these methods (5,7) currently fail to dock 45% of the ligands. The problem becomes much more complex in a predictive mode, in which virtual screening is used to identify new classes of ligands, given a protein binding site that is not pre-conformed to fit any of them. Thus, docking and screening studies using apo protein structures are likely to represent a much more realistic test. Fortunately, ligand conformations, at least for the cases with relatively few rotatable bonds (which tend to be favorable, in any case, due to the decreased entropic cost of binding), can be reasonably well sampled by existing Monte Carlo, genetic algorithm, and exhaustive torsional search methods in tools such as GOLD (8), AutoDock (9), FLEXX (10), DOCK (4,11), and Omega (12-13).

**Enhancing Target Specificity through Flexibility Modeling**

For many protein targets of interest, the druggable binding site (e.g., the ATP site in protein kinases), is highly conserved in homologous proteins. Inhibitors that bind to highly conserved sites present a risk of serious side effects or toxicity, which is typically evaluated through costly *in vitro* screening of compounds against a broad panel of homologous proteins to assess cross-reactivity, followed by extensive pharmacological testing. Many otherwise promising, high-affinity compounds discovered by screening and improved by structure-based design are lost in the process.

Structural plasticity among human protein kinases (14) and differences in flexibility and dynamics of bacterial thymidylate kinases (15) have been proposed as the basis for designing more specific inhibitors. For the folate biosynthetic enzyme and antibiotic drug target, 6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase (HPPK), differences in active-site loop and side-chain conformations between three bacterial enzymes have been identified by crystallography (16), framework dynamics (17-19), and molecular dynamics analysis (L. Yao, M. Tonero, L. A. Kuhn, and R. I. Cukier, unpublished results). We are now representing these conformational differences as a series of design templates to screen for species-selective inhibitors. A similar approach has elucidated the specificity of long side-chain pyrrolopyrimidines for asparaginyl-tRNA synthetase from *Brugia malayi* relative to the human enzyme. Their ligand-binding residues are absolutely conserved. However, a single side-chain difference (Thr to Ala) near the base of an active-site loop, and facing away from the site, apparently allows the loop to open more in the Brugia protein, allowing the ligand to bind preferentially to *Brugia* (6). In general, we propose that low-energy protein conformations that differ from the closed, catalytic conformation are likely to present greater differences between species than the closed conformation. The existence of these unique, low-energy conformations can reflect sequence variation that occurs outside, but near, the binding site, and they are likely to be subject to decreased evolutionary selection relative to the catalytically productive, closed conformation. Thus, beyond improving docking and screening, the ability to accurately model side-chain (and main-chain) conformations in and around ligand binding sites is expected to open a range of new possibilities for gaining specificity between closely homologous enzymes.

## APPROACHES

### The State of the Art in Modeling Protein Side-chain Flexibility

We are fortunate that side-chain flexibility modeling can largely be decoupled from modeling main-chain flexibility (which involves many additional degrees of freedom). A study of almost 1,000 pairs of ligand-free and bound protein structures found no correlation between the degree of main-chain and side-chain movement (20). As a result, some groups have used rotamer libraries for protein side chains (21) to allow efficient sampling of their conformations in the ligand interface (22-24). This is computationally feasible for docking but not for screening. Additional drawbacks of rotamer sampling will be discussed below.

Some methods for modeling side-chain flexibility effectively couple side-chain and main-chain motion by using as docking targets an ensemble of experimentally-observed structures of the protein, often reflecting crystal structures solved in different space groups or with different ligands bound (9,25-26,33). An advantage of this approach is that all the known, low-energy protein conformations can be considered. One of these methods considers combinations of conformations from different crystal structures that are mutually compatible (26), but it is not clear whether this has an advantage over considering the different target conformations individually. Disadvantages of these approaches are that only existing protein conformations are sampled, and they do not reflect all the possible conformations, particularly when the protein binds to a substantially different class of ligands.

Soft docking is a more conservative approach which accepts that not all protein and ligand accommodations upon binding can be accurately predicted, and thus allows some degree of overlap between protein and ligand atoms during docking. Several methods either allow small van der Waals overlaps or dock the molecules using a smoothed representation of the protein surface (27-29). This strategy can be combined with any of the others (e.g., side-chain sampling, or docking into ensembles of structures), and often is. A complementary approach used at the end of docking is energy minimization to ensure that any van der Waals overlaps between protein and ligand atoms can be resolved. Energy minimization adjusts the atoms' positions to energetically improve interactions in the interface but does not attempt to overcome the energy barriers that would be involved in significant rearrangements; therefore, the motions are typically quite small. One successful application of energy minimization is GOLD, in which only polar, terminal hydrogen atoms on protein side chains are considered flexible, and the penultimate bond rotational angle is chosen to optimize hydrogen bond interactions (9). However, as with other methods that perform detailed energy calculations in the course of optimization, this approach (which also includes genetic algorithm sampling of ligand conformations and orientations) proves too computationally intensive for large-scale high-throughput screening. At the far end of this spectrum, in terms of fineness of sampling and scoring, are Monte Carlo and molecular dynamics (MD) techniques that consider all atoms as free to move within a force field including van der Waals, electrostatics, bond torsion and bending, and solvation energy terms (30-35). These methods are generally appropriate for docking single protein-ligand complexes once there is a reasonably accurate initial placement of the ligand. However, MD simulations typically cannot surmount large conformational or orientational energy barriers within a reasonable timeframe.

SLIDE (3,27) represents an intermediate approach, in which all interfacial protein side chains and all single bonds in ligands are free to rotate during docking, but these motions are designed to remove van der Waals overlaps rather than thoroughly search the conformational space. As such, SLIDE's motions tend to be small, similar to those in energy minimization. However, because the bond angles to resolve collisions are calculated geometrically (Figure 1), rather than with respect to an energy function, the process is very fast.  Protein side-chain and ligand flexibility have been modeled while screening and docking 150,000-800,000 compounds or 3D conformers using SLIDE (6,27), with about 100,000 candidates screened per day on a two-processor workstation. How SLIDE selects the bond(s) to rotate during flexibility modeling is described in Figure 2. The motions performed within SLIDE are typically disseminated throughout the interface, as shown for a protease-peptidyl ligand complex (with somewhat larger-than-typical motions; Figure 3).  To complement its balanced protein and ligand flexibility modeling, SLIDE is typically combined with Omega to fully sample low-energy ligand conformations (12-13).  SLIDE can also be combined with ROCK (Ring-Optimized Conformational Kinetics;18-19) to sample protein main-chain conformations that preserve the native non-covalent bond network, reflecting moderate to large-scale motions that tend to be low in energy.  Therefore, full protein and ligand flexibility can be modeled by providing a database of low-energy ligand conformations as input to SLIDE for screening against a panel of protein conformations sampled by ROCK or by molecular dynamics, crystallography, or NMR (6,19).

## Learning from Nature: Observing Side-chain Motions Upon Ligand Binding

An important question is: what kinds of motions do protein side chains actually undergo when binding ligands?  Are they minimal motions, rotameric transitions, or something more complex?  Studies of conformational changes associated with protein-protein (36) and protein-ligand (20) binding in crystal structures show that even for proteins with conserved main-chain conformations upon ligand binding, there are side-chain conformational changes in at least 60% of the cases. However, side chain conformations in these studies were considered to differ only if they reflected a rotameric transition, generally involving single-bond rotations of 60º or more. Other results, however, indicate that ligand binding induces strain or non-rotamericity in the preferred side-chain conformations (37). To address this question without biasing towards a rotameric or non-rotameric interpretation, we analyzed protein side-chain rotations in 63 complexes - 32 of thrombin with different ligands (some highly flexible), 13 of glutathione S-transferase (GST) with both hydrophobic (xenobiotic binding site) and polar (glutathione binding site) ligands, and 18 other, diverse complexes (38). The goal was to observe protein side-chain motions in response to a variety of ligands, across a variety of proteins. 90% of interfacial protein side-chain rotations in the 63 structures were less than 45% upon ligand binding, and the remaining larger rotations were distributed broadly between 45º and 180º. Thus, most side-chain motions upon ligand binding are not rotameric transitions, instead reflecting small adjustments relative to existing rotamers that generate some degree of strain, consistent with (37). The same general trend holds for protein-protein interfaces (M. I. Zavodszky and L. A. Kuhn, unpublished results), although rotations of greater than 45º are found to occur more frequently (28% of the time). Ligand binding-site motions are similar to those elsewhere on protein surfaces, except that small-scale (<15º) side-chain rotations are 20% more common in binding sites. These results indicate that side-chain motions upon ligand binding typically involve small-scale induced fit, which is found to be modeled appropriately by SLIDE (38). The 10% of rotations that are larger cannot be classified simply as rotameric transitions, since they are distributed approximately evenly in dihedral angles between 45 and 180º.

We now understand that side-chain rotations upon ligand binding are not well represented by rotameric transitions or by energy minima within a rotameric state. Beyond not typically changing to a new rotamer, interfacial side chains observe favored dihedral angles only for $X_1$ ($C_\alpha$-$C_\beta$), and, to a lesser extent, $X_2$ ($C_\beta$-$C_\gamma$) bond angles. Figure 4 shows the closeness with which even a very detailed main-chain-dependent rotamer library (39) can sample observed ligand-bound conformations for 25 cases in which side chains were observed to undergo large rotations (40). In 17 out of the 25 cases, there exist no reasonable rotamer matches, even when considering just the side chains' $X_1$ and $X_2$ angles. This is consistent with interfacial side chains adopting strained conformations due to packing in a tight interface and a new chemical environment.

### The Future: Knowledge-based Modeling of Side-chain Motions

Given that we cannot afford to thoroughly sample low-energy side-chain conformations during docking and screening, how can we intelligently identify which side chains move significantly, and how to move them? Some guidance is provided by diagnosing why particular side chains experience large rotations. In one-third of the 25 cases mentioned above, large-scale rotations were needed to avoid steric collisions with the ligand (40). In an additional 50% of the cases, the side chains apparently moved to satisfy hydrogen-bonding groups that could not be satisfied in the ligand-free orientation. This is a complementary picture to that derived from analyzing 30 protein-ligand complexes, indicating that a majority (75%) of interfacial, intra-protein hydrogen bonds are preserved upon ligand binding, and these side chains tend to move very little (40). Together, these results create a relatively tractable scenario in which minimal (or greater) motion can be used to resolve steric overlaps between atoms during docking, then a conformational search can be performed to satisfy the hydrogen-bonding potential of buried polar side chains. As it turns out, most of the large-rotation cases actually involve small displacements (side-chain RMSDs of 1 Å or less), due to compensatory rotations in successive $X_i$ angles (40). Thus, starting with the ligand-free conformation of each buried, unsatisfied polar side chain (of which there are typically just one or two per interface) and performing a local conformational search to optimize hydrogen bonding is expected to generate more accurate side-chain positions and allow better scoring of interactions during docking and screening.

## ACKNOWLEDGMENTS

This work was supported by NIH grants GM 67249 and AI 53877 to L.A.K. and colleagues and reflects the thesis and postdoctoral work of several talented and enthusiastic researchers, particularly Volker Schnecke, Maria Zavodszky, and Sameer Arora. We appreciate the generosity of OpenEye Software (Santa Fe, NM) in providing Omega for our use.

## REFERENCES

1. M. Kontoyianni, G. S. Sokol, and L. M. McClellan, J. Comput. Chem., 2005, 26, 11.

2. E. Perola. W. P. Walters, and P. S. Charifson, Proteins, 2004, 56, 235.

3. M. I. Zavodszky, P. C. Sanschagrin, R. S. Korde, and L. A. Kuhn, J. Comput.-Aided Mol. Des., 2002, 16, 883.

4. D. M. Lorber, M. K. Udo, and B. K. Shoichet, Protein Sci., 2002, 11, 1393.

5. M. I. Zavodszky and L. A. Kuhn, Proteins, 2006, in review.

6. S. C. K. Sukuru, T. Crepin, Y. Milev, L. C. Marsh, J. B. Hill, R. J. Anderson, J. C. Morris, A. Rohatgi, G. O'Mahony, M. Grøtli, F. Danel, M. G. P. Page, M. Härtlein, S. Cusack, M. A. Kron, and L. A. Kuhn, J. Comp.-Aided Molec. Des., 2006, in press.

7. E. Kellenberger, J. Rodrigo, P. Muller, and D. Rognan, Proteins, 2004, 57, 225.

8. G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor, J. Mol. Biol., 1997, 267**,** 727.

9. F. Osterberg, G. M. Morris, M. F. Sanner, A. J. Olson, and D. S. Goodsell, Proteins, 2002, 46, 34.

10. B. Kramer, M. Rarey, and T. Lengauer, Proteins, 1999, 37, 228.

11. T. J. Ewing, S. Makino, A. G. Skillman, and I. D. Kuntz, I.D., 2001, J. Comput.-Aided Mol. Des. 15, 411.

12. OpenEye Software, Santa Fe, New Mexico; http://www.eyesopen.com/products/applications/omega.html

13. J. Bostrom, P. O. Norrby and T. Liljefors, J. Comput.-Aided Mol. Des., 1998, 12: 383.

14. M. Huse and J. Kuriyan, Cell, 2002, 109, 275.

15. J. S. Finer-Moore, A. C. Anderson, R. H. O'Neil, M. P. Costi, S. Ferrari, J. Krucinski, and R. M. Stroud, Acta Crystallogr. D, 2005, 61, 1320.

16. B. Xiao, G. Shi, J. Gao, J. Blaszczyk, Q. Liu, X. Ji and H. Yan, J. Biol. Chem., 2001, 276, 40274.

17. D. J. Jacobs, A. J. Rader, L. A. Kuhn, and M. F. Thorpe, Proteins, 2001, 44, 150.

18. M. Lei, M. I. Zavodszky, L. A. Kuhn, and M. F. Thorpe, J. Comp. Chem., 2004, 25, 1133.

19. M. I. Zavodszky, M. Lei, M. F. Thorpe, A. R. Day, and L. A. Kuhn, Proteins, 2004, 57, 243.

20. R. Najmanovich, J. Kuttner, V. Sobolev, and M. Edelman, Proteins, 2000, 39, 261.

21. R. L. Dunbrack, Jr. and M. Karplus, J. Mol. Biol., 1993, 230, 543.

22. L. Schaffer and G. M. Verkhivker, Proteins, 1998, 33, 295.

23. A. R. Leach and A. P. Lemon, Proteins, 1998, 33, 227.

24. P. Kallblad and P. M. Dean, J. Mol. Biol., 2003, 326, 1651.

25. R. M. Knegtel, I. D. Kuntz, and C. M. Oshiro, J. Mol. Biol., 1997, 266, 424.

26. H. Claussen, C. Buning, M. Rarey, and T. Lengauer, J. Mol. Biol. 2001, 308, 377.

27. V. Schnecke, C. A. Swanson, E. D. Getzoff, J. A. Tainer, and L. A. Kuhn, Proteins, 1998, 33, 74.

28. F. Jiang and S. H. Kim, J. Mol. Biol., 1991, 219, 79.

29. A. M. Ferrari, B. Q. Wei, L. Costantino, and B. K. Shoichet, J. Med. Chem., 2004, 47, 5076.

30. M. L. Lamb and W. L. Jorgensen, Curr. Opin. Chem. Biol., 1997, 1, 449.

31. B. A. Luty, Z. R. Wasserman, P. F. W. Stouten, C. N. Hodge, M. Zacharias, and J. A. McCammon, J. Comp. Chem., 1995, 16, 454.

32. H. A. Carlson, K. M. Masukawa, W. L. Jorgensen, R. D. Lins, J. M. Briggs, and J. A. McCammon, J. Med. Chem., 2000, 43, 2100.

33. D. Bouzida, P. A. Rejto, S. Arthurs, A. B. Colson, S. T. Freer, D. K. Gehlhaar, V. Larson, B. A. Luty, P. W. Rose, and V. M. Verkhivker, Int. J. Quantum. Chem., 1999, 72, 73.

34. M. Totrov and R. Abagyan, Proteins, 1997, Suppl. 1, 215.

35. J.-H. Lin, A. L. Perryman, J. R. Schames, and J. A. McCammon, J. Am. Chem. Soc., 2002, 124, 5632.

36. M. J. Betts and M. J. Sternberg, M.J, Protein Eng., 1999, 12, 271.

37. J. Heringa and P. Argos, Proteins, 1999, 37, 44.

38. M. I. Zavodszky and L. A. Kuhn, Protein Sci., 2005, 14, 1104.

39. R. L. Dunbrack, Jr., Curr. Opin. Struct. Biol., 2002, 12, 431.

40. S. Arora, “Optimizing Side-chain Interactions in Protein-Ligand Interfaces”, M.S. Thesis, Michigan State University, East Lansing, MI, 2005.

## FIGURE LEGENDS

**Figure 1.** SLIDE performs directed rotations, calculated geometrically, to resolve protein-ligand van der Waals overlaps. Directed rotations are performed around a rotatable bond preceding the mobile atom. (Choice of the bond to be rotated, and therefore which atom is fixed versus mobile, is explained in Figure 2.) The rotatable bond is aligned along the Y-axis, with the hinge atom of the bond at Y=0. The molecular system is rotated such that the fixed atom falls on the X axis, in the positive quadrant of the X-Y plane. The rotation angle, **a,** is calculated such that the center of the mobile atom is displaced from the center of the fixed atom by the magnitude of their van der Waals overlap. This resolves the collision by performing a minimal rotation.

**Figure 2.** Intermolecular van der Waals collisions are resolved in SLIDE by directed rotations of single bonds in either the ligand or protein side chains. There are typically several possible rotations to resolve an intermolecular collision. An approach based on mean-field theory is used to decide which rotations are the most efficient for resolving one or more van der Waals overlaps in the current conformation of the complex.
For all pair-wise intermolecular collisions, the bonds that can be rotated to resolve a particular collision are identified. They are stored in a matrix together with the corresponding rotation angle and the number of non-hydrogen atoms that will be displaced by the rotation. The product of the angle and the number of atoms (similar to a moment of inertia) provides a basis for the force that represents the cost of a rotation. A probability is associated with each rotation in the system. All rotations that can be used to resolve a particular collision are initialized with equal probabilities. During the

optimization, the probabilities are updated to converge to an approximately optimal set of values, which assigns the highest probabilities to those rotations that solve the most collisions with the least overall cost.

**Figure 3.** Conformational changes of the protein and the top peptidyl ligand (Tyr-Ser-Met-Ser-Phe) modeled by SLIDE after screening a set of 140,000 peptides (all five-residue peptide structures extracted from a low-homology subset of the Protein Data Bank) against the ligand-free structure of the aspartic protease rhizopuspepsin (PDB 2apr). The docking of the ligand anchor fragment was based on a match of three polar ligand atoms onto the template points represented by spheres (hydrogen-bond acceptor, red; donor, blue). The native, ligand-free conformations of binding-site side chains are shown by blue tubes, and the initial conformation of the peptidyl ligand side chains are shown by white tubes; SLIDE's final conformation of the complex, involving induced fit of several aromatic residues in the protein and ligand, is shown in green.

**Figure 4.** The number of rotamers approximating the dihedral angles of the ligand-bound conformations of 25 interfacial side chains undergoing large rotations (>60 degrees) upon ligand binding. The May 2002 Dunbrack backbone-dependent rotamer library (http://dunbrack.fccc.edu/bbdep) was used to identify all rotamers matching the observed X values to within one standard deviation of the average value for that rotamer bin; all rotamer bins within +/- 10 degrees of the interfacial residue's main-chain Φ and Ψ values were searched. Side chain labels on the Y-axis include the Protein Data Bank code, chain ID (if present), residue type, and residue number. Rotamer searches were

done in incremental fashion, first comparing only $X_1$, then $X_1$ and $X_2$, and so on, up to $X_4$, depending on the number of rotatable bonds in the side chain.  Only rotamers with probability values at least 5% as large as the probability of the most common rotamer within the same bin were considered. This helps exclude rotamers that are very rare, many of which may represent poorly resolved side-chain conformations in the Protein Data Bank.

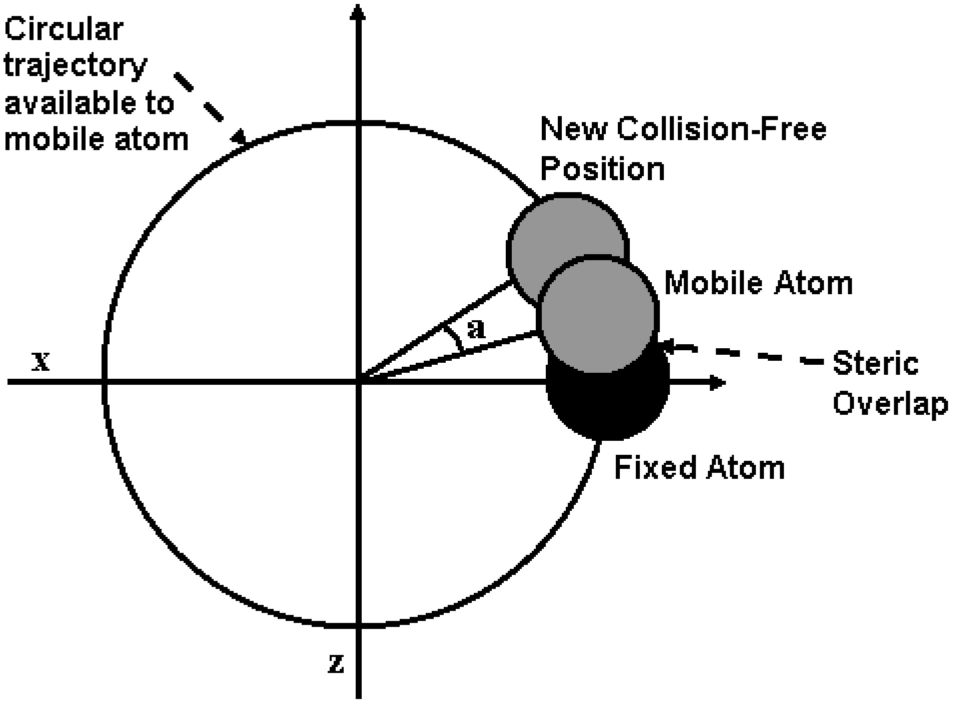
Circular
trajectory
available to
mobile atom
New Collision-Free
Position
Mobile Atom
a
x
Steric
Overlap
Fixed Atom
z

**Figure 1**

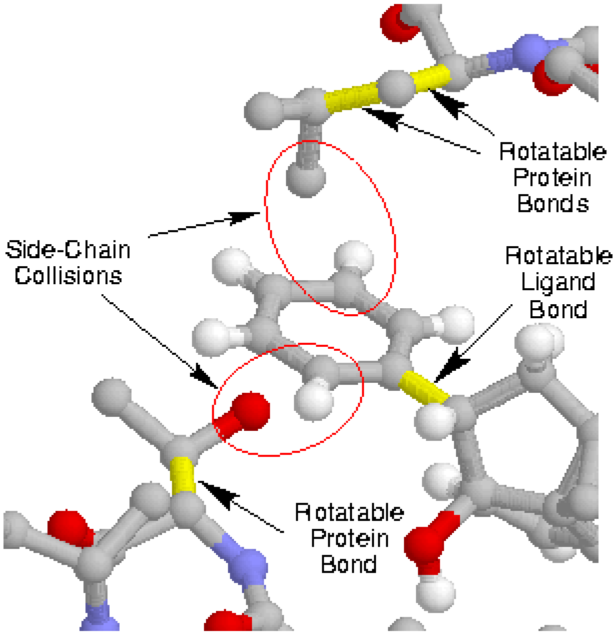
Rotatable
Protein
Bonds
Side-Chain
Collisions
Rotatable
Ligand
Bond
Rotatable
Protein
Bond

**Figure 2**

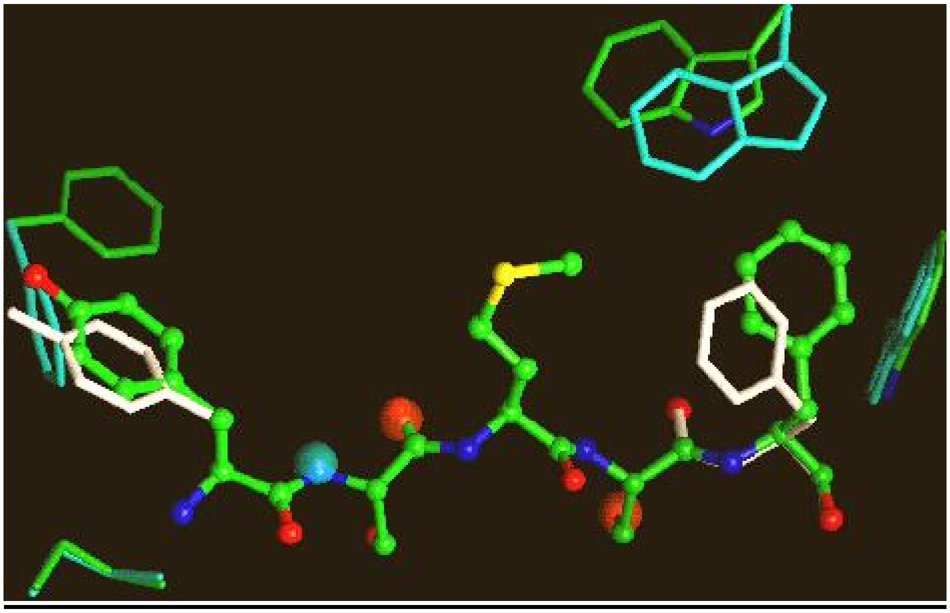

**Figure 3**

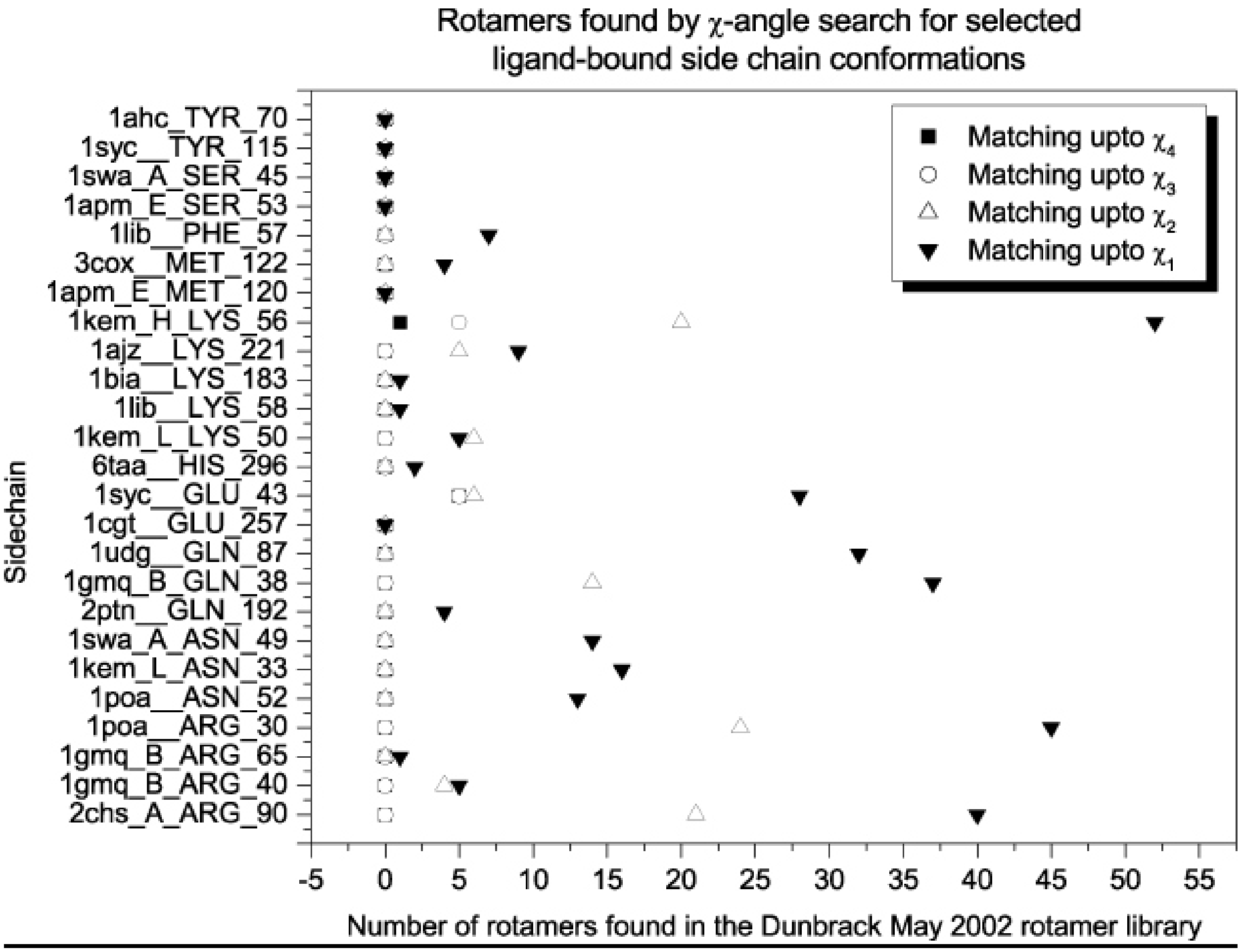
Rotamers found by χ-angle search for selected ligand-bound side chain conformations
Sidechain
1ahc_TYR_70
1syc__TYR_115
1swa_A_SER_45
1apm_E_SER_53
1lib__PHE_57
3cox__MET_122
1apm_E_MET_120
1kem_H_LYS_56
1ajz__LYS_221
1bia__LYS_183
1lib__LYS_58
1kem_L_LYS_50
6taa__HIS_296
1syc__GLU_43
1cgt__GLU_257
1udg__GLN_87
1gmq_B_GLN_38
2ptn__GLN_192
1swa_A_ASN_49
1kem_L_ASN_33
1poa__ASN_52
1poa__ARG_30
1gmq_B_ARG_65
1gmq_B_ARG_40
2chs_A_ARG_90
■ Matching upto χ4
○ Matching upto χ3
△ Matching upto χ2
▼ Matching upto χ1
-5
0
5
10
15
20
25
30
35
40
45
50
55
Number of rotamers found in the Dunbrack May 2002 rotamer library

**Figure 4**