Side-Chain Flexibility in Protein-Ligand Binding:

The Minimal Rotation Hypothesis

Maria I. Zavodszky[1] and Leslie A. Kuhn[1,2]

[1]Department of Biochemistry and Molecular Biology and [2]Quantitative Biology and Modeling Initiative, Michigan State University, East Lansing, Michigan

Correspondence to: Leslie A. Kuhn, Protein Structural Analysis and Design Lab, Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48824-1319.

E-mail: KuhnL@msu.edu

Telephone: (517) 353-8745

Fax: (517) 353-9334

URL: http://www.bch.msu.edu/labs/kuhn and http://qbmi.msu.edu.

Keywords: conformational sampling, flexibility, induced fit, docking, side-chain rotamers, SLIDE, stereochemistry, χ angles.

Abstract

The goal of this work is to learn from nature about the magnitudes of side-chain motions that occur when proteins bind small organic molecules and to test a computational model of sampling these motions for its ability to improve the prediction of protein-ligand complexes. Following analysis of protein side-chain motions upon ligand binding in 63 complexes, we tested the ability of the docking tool SLIDE to model these motions appropriately without restricting them to rotameric transitions or deciding which side chains should be modeled as flexible. The model encoded and tested is that side-chain conformational changes involving more atoms or larger angles are likely to be more costly and less prevalent than small motions due to energy barriers between rotamers and the potential of large motions to cause new steric clashes. Accordingly, SLIDE adjusts the protein and ligand side groups as little as necessary to achieve steric complementarity. We tested the hypothesis that small motions are sufficient to achieve good dockings for 63 ligands into the apo structures of 20 different proteins and compared SLIDE side-chain rotations to those experimentally observed between apo and ligand-bound structures. None of these proteins undergo major main-chain conformational change upon ligand binding, ensuring that side-chain flexibility modeling is not required to compensate for main-chain motions. Although more frugal in the number of side-chain rotations performed, this minimum rotation model substantially mimics the experimentally observed motions. In fact, most side chains do not shift to a new rotamer, and small motions are both necessary and sufficient to sample the correct binding orientation and most interactions between protein and ligand for the 20 proteins we have analyzed.

Introduction

It is widely accepted that flexibility is indispensable for protein function. The questions are: how much flexibility is needed, in general, for protein-ligand interactions, and how does this flexibility partition between the protein and its ligand? Molecular flexibility can contribute to a favorable change in the Gibbs binding free energy in two ways: by optimizing the non-covalent interactions between the protein and ligand (contributing to a favorable change in enthalpy) or by increasing the entropy (or minimizing the decrease in entropy) of the two molecules upon binding, by releasing interfacial water and increasing flexibility in parts of the protein or ligand. For instance, in the Ras-Raf complex, the switch 2 region of Ras actually becomes more flexible upon Raf binding (Gohlke et al. 2004). Strategies for ligand design can be based on enthalpic or entropic considerations (Velazquez-Campoy et al. 2000); however, improving the enthalpy of binding is more likely to improve the specificity of ligands and delay the onset of drug resistance due to protein mutations. Structure-based evaluation of entropic changes upon complex formation remains computationally challenging, beyond simple counts of the number of rotatable bonds buried in the interface. Recent progress has been made via molecular framework and molecular dynamics approaches (Gohlke and Case 2004; Jacobs and Dallakyan 2004; Swanson et al. 2004). However, molecular dynamics has not yet proven practical for docking, given the computational complexity of predicting the orientation of the two molecules as well as the details of their conformations. To maintain a reasonable computational time, the compromise often made is that the small-molecule ligand is modeled as fully flexible, via a series of often random, discrete dihedral rotations of its relatively few rotatable bonds. On the other hand, the protein partner has a vast number of rotatable bonds and is typically approximated as rigid or very

selectively flexible (e.g., by specifying in advance one or a few side chains that can rotate, and the states they can adopt).

The first docking tools, the most widely known of them being DOCK (Kuntz et al. 1982), were designed based on the key-and-lock mechanism of protein-ligand recognition, considering both the ligand and the protein as rigid bodies. The next generation methods model the protein as rigid while allowing ligand flexibility (Burkhard et al. 1998; Ewing et al. 2001; Goodsell et al. 1996; Kramer et al. 1999; Taylor and Burnett 2000). DOCK evolved to become more realistic, too, by handling ligands totally flexibly (Ewing et al. 2001; Lorber and Shoichet 1998). The rationale behind this treatment is that small molecule ligands have fewer degrees of freedom than the protein, so it is computationally less expensive to handle them flexibly. On the other hand, studies of conformational changes accompanying protein-protein (Betts and Sternberg 1999) and protein-ligand (Najmanovich et al. 2000) associations show that even in the case of proteins with conserved main-chain conformations across crystallographic complexes with various ligands, there are side-chain conformational changes in at least 60% of the proteins upon ligand binding. However, side chain conformations in these studies were considered to be different only if the side chains changed to different rotamers, generally involving single-bond rotations of 60º or more. Nevertheless, they point to the necessity of also modeling protein side-chain flexibility in docking.

Docking and screening tools reach various levels of sophistication trying to achieve this goal. Soft docking (Jiang and Kim 1991) handles protein flexibility implicitly by allowing a certain degree of interpenetration between the protein and the docked ligand, making the reasonable assumption that the exactly correct conformers of the protein and ligand are not sampled. The docking tool GOLD (Jones et al. 1997) allows rotation of terminal hydrogen atoms on the

proteins to optimize fit and hydrogen bonding. The next level of sophistication is reached by using rotamer libraries (Dunbrack, Jr. and Karplus 1993; Lovell et al. 2000; Tuffery et al. 1991) to sample the low energy conformations available to each side chain while optimizing the shape complementarity between the protein and the docked ligand (Kallblad and Dean 2003; Leach 1994; Leach and Lemon 1998). However, our experience is that even very complete rotamer libraries, such as backbone-dependent ones (Dunbrack, Jr. and Karplus 1993), are insufficient to sample the conformations of protein side chains finely enough to yield a collision-free complex. To improve the efficiency of selecting side-chain rotamers, one group implemented a post-docking side-chain optimization procedure using the dead-end elimination (DEE) algorithm, followed by local minimizations and energy evaluations of all generated DEE solutions (Schaffer and Verkhivker 1998). This is a computationally affordable method for fine docking but not for screening. Assuming that crystal structures show the protein side chains in favorable conformations, an alternative approach to sample the available side-chain conformational space is the use of side-chain conformers from multiple x-ray structures (Claussen et al. 2001; Knegtel et al. 1997). A similar approach was taken to account for protein side-chain motions by combining multiple target structures within a single grid-based look-up table of interaction energies for docking with AutoDock (Osterberg et al. 2002).

Two recent studies assess the effect of protein flexibility on docking accuracy and efficiency. One group has tested the performance of four docking algorithms, DOCK, FlexX, GOLD and CDOCKER, as a function of the conformation of the protein binding site used as the target (Erickson et al. 2004). While each of the tested docking tools performs quite well in redocking experiments, when the ligand is docked into the bound conformation of the target protein, their performance drops significantly when the docking is done into the unbiased apo structure. The

decrease in docking accuracy was correlated with the degree of flexibility of the active site, a result that points to the importance of modeling protein flexibility for protein-ligand interactions. Another group has compared the efficiency of "soft docking" using a rigid protein structure with softened van der Waals potential in the scoring function, allowing some intermolecular penetration to account for small atomic motions, with flexible docking using crystal structure alternative conformations as well as modeled ones (Ferrari et al. 2004). Flexible docking resulted in better bound ligand conformations and superior enrichment in database screens when compared to soft docking against a single receptor conformation. An excellent overview of the state of the art in flexible docking appears in Halperin et al. 2002.

SLIDE models flexibility by allowing protein side-chain rotations and ligand flexibility, assuming that the protein changes its apo conformation as little as necessary to result in an overlap-free docked orientation of the ligand (Schnecke and Kuhn 1999; Schnecke and Kuhn 2000). In practice, this has worked well for diverse protein complexes: subtilisin, cyclodextrin glycosyltransferase, uracil-DNA glycosylase, rhizopuspepsin, HIV-1 protease, human estrogen and progesterone receptor, Asn tRNA synthetase, bacterial aspartyl protease, thrombin, cyclophilin A, and glutathione S-transferase (Schnecke et al. 1998; Schnecke and Kuhn 1999; Schnecke and Kuhn 2000; Zavodszky et al. 2002; Zavodszky et al. 2004). The goal of this paper is to assess the extent to which minimal motion accounts for protein side-chain flexibility in proteins for which multiple ligand-bound structures are available (thrombin and glutathione S-transferase, GST), plus 18 other proteins. These proteins were selected because they are diverse and do not undergo major main-chain conformational change upon ligand binding, ensuring that side-chain flexibility modeling is not required to compensate for main-chain motions. The series of thrombin and GST structures present a particular challenge for side-chain flexibility modeling,

because different side-chain positions are observed in the crystal complexes with different ligands. In each case, the known ligands were docked into the apo protein structure with SLIDE. Minimal side-chain rotations performed by SLIDE, necessary to accommodate the ligand, were compared to dihedral angle differences between the x-ray structures of the apo and ligand-bound protein. Comparing side-chain conformations of the binding site residues before and after ligand binding allows us to assess whether the minimal rotation approximation is generally sufficient for sampling the side-chain conformations observed in the biological complex. Beyond testing SLIDE's approach, the results of this analysis provide guidance on how to improve side-chain conformational sampling for docking and ligand design methods in general.

Methods

SLIDE

Version 2.3 of the docking and screening software SLIDE (Screening for Ligands by Induced-fit Docking, Efficiently) (Schnecke and Kuhn 1999; Schnecke and Kuhn 2000) was used to dock known ligands obtained from the x-ray structures of 63 complexes (Tables 1-3). SLIDE models protein-ligand interactions based on steric complementarity combined with matching hydrophobic and hydrogen-bonding sites between the protein and ligand. The binding site is represented by a template with hydrophobic and hydrogen bonding points (Zavodszky et al. 2002). Multi-step indexing quickly tests all possible matches of hydrophobic and hydrogen bonding interaction centers on each ligand candidate with the protein template. Upon finding an appropriate match, the ligand is transformed into the binding site. In most cases, this will result in some van der Waals collisions between atoms of the ligand and the apo protein structure. Because both the protein side chains and typically some functional groups in the ligand can be

rotated around single bonds, most of these collisions can be resolved. All protein-ligand atomic overlaps, together with all the possible side-chain rotations that could solve them, are tabulated after the initial transformation of the whole ligand into the binding site. The elimination of the overlaps starts by selecting the side-chain rotation that can resolve the largest number of overlaps with the lowest cost. The cost of rotation is modeled as proportional to the product of the angle and the number of the atoms being rotated, like a moment of inertia, and in practice reflects the reality that rotating more atoms through a larger angle is more likely to result in new steric clashes. After each step, the table containing the remaining overlaps and rotations to resolve them is updated, and the next most favorable rotation is selected. The process runs for up to 10 iterations or until the remaining overlaps fall below an accepted cutoff (usually set to 2 Å). If a solution is not found, this ligand orientation is rejected (Schnecke and Kuhn 2000).

Scoring of the docked protein-ligand complex by SLIDE is based on the number of intermolecular hydrogen bonds and the hydrophobic complementarity between the ligand and its protein environment. A more detailed scoring function is being implemented in version 3.0 of SLIDE which improves the detection of the best ligand binding mode (manuscript in preparation).

Testing the minimal rotation hypothesis

To examine whether or not side-chain flexibility modeling is necessary for successful docking into the apo protein structures, known ligands obtained from crystallographic complexes deposited into the Protein Data Bank (PDB (Berman et al. 2000); Tables 1-3) were docked into the active site of the corresponding ligand-free structure both by rigid and flexible docking. The two approaches were evaluated by comparing the number and the root mean square atomic positional deviation (RMSD) of successful dockings relative to the crystal structure orientation.

A successful docking was defined as one with a ligand RMSD of 2.5 Å or less from the crystal structure orientation.

To evaluate the realism of induced fit modeling by SLIDE, the side-chain rotations produced by SLIDE upon docking known ligands into the apo structures of their target proteins were compared to the dihedral-angle differences of binding-site residues between corresponding ligand-free and ligand-bound x-ray structures of the proteins. A set of 32 human thrombin (Table 1) and 13 human glutathione S-transferase (Table 2) crystallographic complexes with known ligands were used in addition to the active-site ligand-free structures of thrombin (PDB code 1vr1 (Dekker et al. 1999)) and GST (PDB code 16gs (Oakley et al. 1998)). To ensure the validity of the conclusions across a wide range of proteins, a dataset of 18 additional ligand-free protein structures with corresponding ligand-bound complexes was also analyzed (Table 3).

To test whether the observed side-chain conformational changes are ligand-induced rather than just small thermal fluctuations or positional errors in the x-ray structures, side chain dihedral angle differences between ligand-free and bound protein pairs were also calculated for surface residues. For this analysis, only those 6 pairs (out of the 18) were used for which both the ligand-free and bound structures are biological monomers, to avoid including some side-chains whose motions are constrained by biological interfaces. The 3cox/1coy pair was excluded because of the large number of incomplete surface residues. The program MSMS (Sanner et al. 1996) was used to calculate the solvent-accessible surface area (SAS) for each atom. Surface residues were identified as having at least one side-chain atom with SAS > 5.0 $Å^2$. Incomplete residues or those with multiple occupancies were not included in the calculations.

Since this study focused on modeling side-chain flexibility in systems with no significant backbone changes following ligand binding, only ligand-bound and ligand-free protein pairs with

backbone superpositional RMSD values of ≤ 0.5 Å and pairwise backbone atom positional deviations of ≤ 1 Å were used. Binding site residues were defined as those having at least one atom within 4.0 Å of any ligand atom. To reduce the possibility of significant crystallographic errors in side chain positions, only crystal structures with resolution of 2.5 Å or better were included in the analysis.

By default, all template points are assigned as key points by the template generator of SLIDE. Defining a subset of template points as key points can be used, alternatively, to focus docking or screening on parts of the binding site of particular interest, e.g., regions known to be critical for function. In SLIDE runs, 3 template points, including at least one key point must be matched during docking. This option typically reduces the number of dockings and selectively eliminates ligand candidates not filling the critically important binding pockets of the target protein. The template describing the binding site of ligand-free thrombin consisted of 139 points, with 24 of these points assigned as key points. Key points were selected as template points at a distance of 6.5 Å or less from the CG side-chain carbon atom of Asp189 in the specificity pocket of thrombin (Figure 1A). The template representing the binding site of GST consisted of 120 template points with no key points assigned. The templates used to represent the binding sites of the diverse set of 18 proteins varied in size from 37 to 153 points. No key points were assigned for these cases.

Results

Thrombin

From the set of 32 known thrombin ligands, only 13 could be successfully docked (RMSD ≤ 2.5 Å) with rigid docking, while all 32 were correctly docked when the protein side chains and

ligand were considered flexible (Table 1). Most of the side chain rotations performed by SLIDE upon docking these 32 known ligands to thrombin were small (Figure 1B), 83% of them being 15º or less, and 94% of them 45º or less (Figure 2A). The dihedral angle differences between protein side chains from the ligand-free and ligand-bound crystal structures of these ligands had a very similar distribution (Figure 2B), with 67% of all dihedral angle differences being 15º or less, and 85% of the differences being 45º or less. Comparing the distributions indicates that SLIDE was making rotations of appropriate magnitudes for active-site side chains upon ligand binding approximately 80% of the time. About 15% of side-chain rotations in crystallographic complexes were rotamer changes, of which SLIDE reproduced 5%, missing the other 10%.

GST

Due to the relatively large and open binding site of GST, 10 of the 13 known GST ligands (Table 2) could be docked into the binding site of the ligand-free crystal structure (PDB code 16gs) with rigid docking. The RMSD values of 5 of these dockings were slightly higher with rigid than with flexible docking. All 13 ligands could be docked with side-chain flexibility allowed, with no docking failures observed. The side chain rotations performed by SLIDE for the 13 successful dockings are shown in Figure 3A. Only 4% of the angles rotated by SLIDE were larger than 45º, with 91% of them being smaller than 15º. This result is similar to the observed crystal structure dihedral angle differences between the ligand-free protein and ligand-bound complexes (Figure 3B), where 96% of the angle differences were 45º or smaller and 85% were 15º or smaller.

Eighteen Pairs of Ligand-free and Ligand-bound Proteins

Only six of the ligand orientations could be correctly predicted with rigid docking for the other 18 complexes, and all but one of these dockings had a higher RMSD value than the best docking with side-chain flexibility (Table 3). As in case of thrombin and GST, most side chains (94%)

from the binding sites of the apo structures were rotated by SLIDE with 45º or less, with 75% or the rotations being smaller than 15º (Figure 4A). The distribution of the SLIDE-performed side-chain rotations was found to be very similar to the distribution of dihedral-angle differences observed between the apo and ligand-bound crystal structures (Figure 4B), out of which 95% were 45º or smaller and 83% were 15º or smaller. Again, significantly more very small (less than 7.5º) rotations were observed in nature (via analyzing the ligand-free to bound dihedral changes), possibly due to thermal fluctuations in side chain orientations rather than the influence of the ligand.

To test whether these small differences are indeed characteristic of thermal fluctuations of unconstrained side chains, dihedral angle changes of surface residues outside the binding sites were computed for 6 ligand-free and ligand-bound pairs occurring as biological monomers (marked with ^{M} in Table 3). This choice was made to avoid including constrained side chains buried in natural protein-protein interfaces. Figure 5 shows the difference between binding-site and other surface residue dihedral angle changes for the 6 ligand-bound and free protein structural pairs (counts normalized to 100 for both cases to allow comparison). Surprisingly, binding-site residues were found to undergo small angle rotations about 20% more frequently than other surface residues, while there was no significant difference in the number of larger dihedral angle rotations.

Side-chain conformational analysis

The side chain conformations generated using the minimal rotation hypothesis in SLIDE for the best RMSD docking of each of the 63 known ligands were compared to these side chains' conformations in the apo structures as well as the crystallographic complexes using χ_1-χ_2 plots generated by PROCHECK (Laskowski et al. 1993; Morris et al. 1992). For 94 cases out of 101,

the side chain conformations generated by SLIDE, as judged by the χ_1-χ_2 values, were favorable. Figure 6 shows representative plots for four residue types. Even those conformations labeled as unfavorable by PROCHECK were close to favorable regions. Good qualitative agreement between the SLIDE conformations (column 2, Figure 6) and x-ray structure conformations from ligand-bound structures (column 3, Figure 6) were found for most residues (representative data shown in Figure 6). When necessary, SLIDE also performed large rotations, equivalent to switching from one rotameric state to another, as seen on the χ1-χ2 plot of Ile (Figure 6, column 2). Similar "χ-hopping" to a new rotamer was also observed for Ile and Lys between the apo and ligand-bound crystal structures (Figure 6, columns 1 and 3).

Another way to measure the favorability of the protein side-chain conformations is by using the G-factors computed with PROCHECK. Low (more negative) G-factors correspond to low probability conformations. The G-factors of SLIDE conformations versus the G-factors of the corresponding side chains from the apo structures were plotted and colored by the G-factors of the holo form for all side-chain transitions in the 63 complexes (Figure 7). Based on these G-factors, 31.1% of the new side-chain conformations generated by SLIDE were more favorable than the apo conformation (data points above the diagonal), 27.4% were similarly favorable (on the diagonal), while 41.5% were less favorable (below the diagonal). For the side-chain conformational changes observed between apo and complex crystal structures, the results were comparable: 41.9% became more favorable, 14.7% were similar, and 43.4% became worse.

Discussion

This work on defining the typical side-chain motions of proteins upon ligand binding was motivated by the failure of our early attempts to use a comprehensive backbone-dependent

rotamer library to model ligand-induced changes in SLIDE. In many cases, there were no rotamers that resolved the van der Waals collisions between a side chain from the binding site of the apo protein structure and the ligand, even when the ligand was in its bioactive conformation and docked correctly. This encouraged us to learn what types of side-chain motions actually occur upon ligand binding, rather than assuming that the motions involve transitions between rotamers.

This study was restricted to proteins with no major main-chain movements in order assess how side-chains moved upon ligand binding and influenced docking. Including proteins with main-chain flexibility might well result in the movement of side-chains to compensate for main-chain motions that were not modeled. Despite not including proteins with significant main-chain flexibility, results on the 18 ligand-bound thrombin-ligand complexes showed that side chains do undergo large movements occasionally, even though small motions are preferred. In fact, large motions occur just as often in the binding sites as they do in perhaps less tightly packed surface-exposed regions elsewhere in the structure (Figure 5). This indicates that side chain movements are not unusually constrained in proteins that undergo limited backbone movements. Furthermore, another group has found no correlation between the degree of backbone movement and side-chain flexibility for 980 pairs of apo- and complex structures (Najmanovich et al. 2000).

The importance of modeling protein flexibility in docking is illustrated by the fact that most known thrombin ligands and many of the other proteins' ligands failed to dock with rigid docking, while a correct binding orientation could be predicted by allowing minimal rotations (Tables 1 and 3). Although no strong correlation between the ligand size and the success or failure of docking can be detected, there was some tendency for the larger ligands to fail rigid docking in the case of thrombin. We noted for GST that ligands binding only to the xenobiotic

or X site tended to fail in rigid docking. This might be due to the less specific character of the hydrophobic interactions predominant in this site. Also, the current lack of modeling directional aromatic interactions in SLIDE tended to generate dockings closer to the wall of the pocket than the ligand is observed to bind in the complex, resulting in irresolvable van der Waals overlaps in the case of rigid docking. Including modeling favored directions as well as distances of aromatic interactions in SLIDE is expected to further improve the quality of docking.

Across the 20 proteins, there is a good agreement between the pattern of side-chain rotations that emerges from comparing ligand-free and ligand-bound protein structures with side-chain motions made by minimal rotations within SLIDE (Figures 2-4). On average, across the 63 complexes, 95% of side-chain rotations were smaller than 45º, not large enough to allow changes between rotamers, but necessary to correctly dock about 54% of the ligands that could not be docked into a rigid binding site. While docking could be achieved with a rigid protein structure for the remaining 46% of the ligands, the resulting ligand orientations were typically less accurate, compared to the orientations when protein flexibility was included. Studies of ligand-induced changes in side-chain conformations in protein binding sites usually consider only those side-chain rotations larger than 45º, or even 60 or 75º (Betts and Sternberg 1999; Najmanovich et al. 2000), which correspond to changes in rotameric states of the side chains. Heringa and Argos on the other hand, observed that ligand binding induces non-rotamericity in the preferred side-chain conformations (Heringa and Argos 1999). The model of protein side-chain flexibility implemented in SLIDE provides results that are more in agreement with the latter observations, as reflected by the χ_1-χ_2 distributions and G-factors (Figures 6 and 7). While SLIDE tends to shift the side-chains off the perfect χ_1 and χ_2 values to achieve good dockings, only seven out of 101 SLIDE-generated conformations were flagged as unfavorable by PROCHECK. One of

these unfavorable conformations (Phe 198), for example, was very close to that in the ligand-free structure, which was also labeled unfavorable. This residue comes from the 1.95 Å resolution structure of a mutant human carbonic anhydrase II (PDB code 1ydc (Nair et al. 1995)) in which the native Leu198 was replaced by Phe 198. In light of this bulky mutation, it is not surprising that Phe 198 takes on a strained conformation both in the apo structure and in the SLIDE-generated complex.

An interesting observation made when analyzing the Leu χ_1-χ_2 distribution plots (Figure 6) was that Leu 99 from the active site of thrombin can be found in one of two conformations in the apo structure and the 32 complexes with inhibitors. Some inhibitor-bound thrombin structures showed only small deviations from the apo conformation of this residue, while others showed a change in rotamer, requiring χ_1 and χ_2 rotations of 35-45º and 135-145º, respectively (encircled squares in columns 1 and 3 of the Leu χ_1-χ_2 plots, Figure 6). However, these two rotamers are virtually isosteric, with a maximal deviation of 0.75 Å in atomic positions. The two rotamers may actually reflect ambiguity during the side-chain fitting stage in crystallographic structure determination. While the minimal rotation hypothesis allowed successful modeling of small deviations for Leu 99, it did not reproduce the motion into the second, isosteric rotameric state.

Comparing the left and right panels of the plots showing the distributions of side-chain rotations (Figures 2-4) it is apparent that SLIDE is more parsimonious than nature, producing a smaller number of rotations in the protein upon ligand binding. One reason for this is that the larger the number of side chains rotated by SLIDE, the larger the possibility of creating new intramolecular overlaps in the protein. In addition, many of the small differences in side-chain conformations from free to bound structures are likely thermal fluctuations similar to the ones observed at the surfaces of proteins, rather than ligand-driven conformational changes.

The difference between binding site and surface rotations (Figure 5, counts normalized to 100 in both cases to allow comparison) showed that binding-site residues tended to undergo small angle rotations about 20% more frequently that other surface residues, while there is no significant difference in the frequency of larger dihedral angle rotations. The small changes in the binding site are, on average, somewhat energetically unfavorable but necessary to accommodate the ligand. However, surface residues, which may be subject to fewer steric constraints, may choose to make a few larger moves that are favorable in energy, rather than many small moves. Overall, predominant small changes in side chain dihedral angles are characteristic for binding sites, likely reflecting steric constraints upon ligand binding as well as optimization of interactions with the ligand.

In summary, the assumption that both protein side chains and ligands move as little as necessary in order to achieve a collision-free complex proved to be both necessary and sufficient to dock all known ligands from 63 complexes into the apo binding sites of their target proteins to within 2.5 Å RMSD, and 78% of ligands within 1 Å RMSD. Comparing the ligand-free and ligand-bound crystal structures underscore that side chain conformational changes upon ligand binding are typically not changes between rotamers, but instead mostly involve modest (<15°) changes in side-chain angles relative to the original (typically rotameric) position. Employing rotameric sampling of side-chain conformations to model induced fit in docking therefore would usually cause changes that are too large to maintain key interactions, whereas the minimal rotation approach in most cases assures good shape complementarity between the ligand and the protein binding site and also mimics the motions found in nature.

Acknowledgments

We thank Volker Schnecke for his insight in developing the minimal-rotation approach and encoding it in SLIDE, and NIH for their support of this research through Mathematics in Biology grant GM067249 and Partnerships for Novel Therapeutics grant AI053877.

References

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res.* **28**: 235-242.

Betts, M.J. and Sternberg, M.J. 1999. An analysis of conformational changes on protein-protein association: Implications for predictive docking. *Protein Eng.* **12**: 271-283.

Burkhard, P., Taylor, P., and Walkinshaw, M.D. 1998. An example of a protein ligand found by database mining: Description of the docking method and its verification by a 2.3 Angstrom X-ray structure of a thrombin-ligand complex. *J. Mol. Biol.* **277**: 449-466.

Claussen, H., Buning, C., Rarey, M., and Lengauer, T. 2001. FlexE: Efficient molecular docking considering protein structure variations. *J. Mol. Biol.* **308**: 377-395.

Dekker, R.J., Eichinger, A., Stoop, A.A., Bode, W., Pannekoek, H., and Horrevoets, A.J. 1999. The variable region-1 from tissue-type plasminogen activator confers specificity for plasminogen activator inhibitor-1 to thrombin by facilitating catalysis: release of a kinetic block by a heterologous protein surface loop. *J. Mol. Biol.* **293**: 613-627.

Dunbrack, R.L., Jr. and Karplus, M. 1993. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J. Mol. Biol.* **230**: 543-574.

Erickson, J.A., Jalaie, M., Robertson, D.H., Lewis, R.A., and Vieth, M. 2004. Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. *J. Med. Chem.* **47**: 45-55.

Ewing, T.J., Makino, S., Skillman, A.G., and Kuntz, I.D. 2001. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aided Mol. Des.* **15**: 411-428.

Ferrari, A.M., Wei, B.Q., Costantino, L., and Shoichet, B.K. 2004. Soft docking and multiple receptor conformations in virtual screening. *J. Med. Chem.* **47**: 5076-5084.

Gohlke, H. and Case, D.A. 2004. Converging free energy estimates: MM-PB(GB)SA studies on the protein-protein complex Ras-Raf. *J Comput. Chem.* **25**: 238-250.

Gohlke, H., Kuhn, L.A., and Case, D.A. 2004. Change in protein flexibility upon complex formation: analysis of Ras-Raf using molecular dynamics and a molecular framework approach. *Proteins* **56**: 322-337.

Goodsell, D.S., Morris, G.M., and Olson, A.J. 1996. Automated docking of flexible ligands: Applications of AutoDock. *J. Mol. Recognit.* **9**: 1-5.

Halperin, I., Ma, B., Wolfson, H., and Nussinov, R. 2002. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **47**: 409-443.

Heringa, J. and Argos, P. 1999. Strain in protein structures as viewed through nonrotameric side chains: II. Effects upon ligand binding. *Proteins* **37**: 44-55.

Jacobs, D.J. and Dallakyan, S. 2004. Elucidating Protein Thermodynamics from the Three Dimensional Structure of the Native State Using Network Rigidity. *Biophys. J.* (Epub ahead of print).

Jiang, F. and Kim, S.H. 1991. "Soft docking": Matching of molecular surface cubes. *J. Mol. Biol.* **219**: 79-102.

Jones, G., Willett, P., Glen, R.C., Leach, A.R., and Taylor, R. 1997. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **267**: 727-748.

Kallblad, P. and Dean, P.M. 2003. Efficient conformational sampling of local side-chain flexibility. *J. Mol. Biol.* **326**: 1651-1665.

Knegtel, R.M., Kuntz, I.D., and Oshiro, C.M. 1997. Molecular docking to ensembles of protein structures. *J. Mol. Biol.* **266**: 424-440.

Kramer, B., Rarey, M., and Lengauer, T. 1999. Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. *Proteins* **37**: 228-241.

Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R., and Ferrin, T.E. 1982. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **161**: 269-288.

Laskowski, R.A., MacArthur, M.W., Moss, D.S., and Thornton, J.M. 1993. Procheck - A program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography* **26**: 283-291.

Leach, A.R. 1994. Ligand docking to proteins with discrete side-chain flexibility. *J. Mol. Biol.* **235**: 345-356.

Leach, A.R. and Lemon, A.P. 1998. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Proteins* 33: 227-239.

Lorber, D.M. and Shoichet, B.K. 1998. Flexible ligand docking using conformational ensembles. *Protein Sci.* **7**: 938-950.

Lovell, S.C., Word, J.M., Richardson, J.S., and Richardson, D.C. 2000. The penultimate rotamer library. *Proteins* **40**: 389-408.

Morris, A.L., MacArthur, M.W., Hutchinson, E.G., and Thornton, J.M. 1992. Stereochemical quality of protein structure coordinates. *Proteins* **12**: 345-364.

Nair, S.K., Krebs, J.F., Christianson, D.W., and Fierke, C.A. 1995. Structural basis of inhibitor affinity to variants of human carbonic anhydrase II. *Biochemistry* **34**: 3981-3989.

Najmanovich, R., Kuttner, J., Sobolev, V., and Edelman, M. 2000. Side-chain flexibility in proteins upon ligand binding. *Proteins* **39**: 261-268.

Oakley, A.J., Lo, B.M., Ricci, G., Federici, G., and Parker, M.W. 1998. Evidence for an induced-fit mechanism operating in pi class glutathione transferases. *Biochemistry* **37**: 9912-9917.

Osterberg, F., Morris, G.M., Sanner, M.F., Olson, A.J., and Goodsell, D.S. 2002. Automated docking to multiple target structures: Incorporation of protein mobility and structural water heterogeneity in AutoDock. *Proteins* **46**: 34-40.

Sanner, M.F., Olson, A.J., and Spehner, J.C. 1996. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers* **38**: 305-320.

Schaffer, L. and Verkhivker, G.M. 1998. Predicting structural effects in HIV-1 protease mutant complexes with flexible ligand docking and protein side-chain optimization. *Proteins* **33**: 295-310.

Schnecke, V. and Kuhn, L.A. 1999. Database screening for HIV protease ligands: The influence of binding-site conformation and representation on ligand selectivity. Proc. Int. Conf. *Intell. Syst. Mol. Biol.* 242-251.

Schnecke, V. and Kuhn, L.A. 2000. Virtual screening with solvation and ligand-induced complementarity. *Perspectives in Drug Discovery and Design* **20**: 171-190.

Schnecke, V., Swanson, C.A., Getzoff, E.D., Tainer, J.A., and Kuhn, L.A. 1998. Screening a peptidyl database for potential ligands to proteins with side-chain flexibility. *Proteins* **33**: 74-87.

St Charles, R., Matthews, J.H., Zhang, E., and Tulinsky, A. 1999. Bound structures of novel P3-P1' beta-strand mimetic inhibitors of thrombin. *J. Med. Chem.* **42**: 1376-1383.

Swanson, J.M., Henchman, R.H., and McCammon, J.A. 2004. Revisiting free energy calculations: a theoretical connection to MM/PBSA and direct calculation of the association free energy. *Biophys. J.* **86**: 67-74.

Taylor, J.S. and Burnett, R.M. 2000. DARWIN: A program for docking flexible molecules. *Proteins* **41**: 173-191.

Tuffery, P., Etchebest, C., Hazout, S., and Lavery, R. 1991. A new approach to the rapid determination of protein side chain conformations. *J. Biomol. Struct. Dyn.* **8**: 1267-1289.

Velazquez-Campoy, A., Todd, M.J., and Freire, E. 2000. HIV-1 protease inhibitors: enthalpic versus entropic optimization of the binding affinity. *Biochemistry* **39**: 2201-2207.

Zavodszky, M.I., Lei, M., Thorpe, M.F., Day, A.R., and Kuhn, L.A. 2004. Modeling correlated main-chain motions in proteins for flexible molecular recognition. *Proteins* **57**: 243-261.

Zavodszky, M.I., Sanschagrin, P.C., Korde, R.S., and Kuhn, L.A. 2002. Distilling the essential features of a protein surface for improving protein-ligand docking, scoring, and virtual screening. *J. Comput. Aided Mol. Des.* **16**: 883-902.

Table 1. Thrombin crystallographic complexes used in testing the minimal rotation hypothesis. These ligands were docked to the apo form of the thrombin active site (PDB entry 1vr1) with both flexible and rigid docking. The "–" sign indicates that the ligand could not be docked with RMSD ≤ 2.5 Å given the same set of parameters used for all dockings. Note that all ligands could be docked with flexible docking, while 19 out of 32 could not be docked with a rigid protein structure.

#	PDB code	Ligand name	# Non-H atoms	Resolution (Å)	Best ligand RMSD (Å)	
					Flexible docking	Rigid docking
1	1a3b	BOROLOG1	34	1.8	0.30	0.76
2	1a46	BETA-STRAND MIMETIC INHIBITOR	38	2.1	0.35	0.88
3	1a4w	ANS-ARG-2EP-KTH	42	1.8	0.52	–
4	1a5g	BIC-ARG-EOA	42	2.1	0.97	0.97
5	1a61	MOL-ARG-LOM	36	2.2	0.96	–
6	1ad8	MDL103752	40	2.0	0.78	–
7	1ae8	EOC-D-PHE-PRO-AZALYS-ONP	30	2.0	0.65	0.31
8	1afe	CBZ-PRO-AZALYS-ONP	24	2.0	1.05	–
9	1aht	P-AMIDINO-PHENYL-PYRUVATE	15	1.6	0.81	1.11
10	1aix	$PHCH_2OCO$-D-DPA-PRO-BOROVAL	42	2.1	1.40	–
11	1awf	GR133487	52	2.2	2.11	–
12	1ay6	HMF-PRO-ARG-HHO	42	1.8	0.71	–
13	1b5g	BCC-ARG-THZ	40	2.1	0.40	–
14	1ba8	PMS-RON-GLY-ARG	32	1.8	0.51	–
15	1bb0	PMS-RON-GLY-3GA	35	2.1	0.65	0.65
16	1bcu	PROFLAVIN	16	2.0	2.16	2.16
17	1bhx	SDZ 229-357	31	2.3	0.53	0.78
18	1fpc	DAPA	35	2.3	0.90	–
19	1lhc	AC-(D)PHE-PRO-BOROARG-OH	33	2.0	0.67	0.75
20	1lhd	AC-(D)PHE-PRO-BOROLYS-OH	31	2.3	0.74	–
21	1lhe	AC-D-PHE-PRO-BORO-N-BUTYL-AMIDINO-GLY-OH	33	2.2	0.71	0.74
22	1lhg	AC-D-PHE-PRO-BOROHOMOORNITHINE-OH	30	2.2	1.30	–
23	1nrs	LEU-ASP-PRO-ARG	34	2.4	0.75	–
24	1ppb	PPACK	30	1.9	0.71	–
25	1tbz	DPN-PRO-ARG-BOT	38	2.3	1.10	–
26	1tmb	CYCLOTHEONAMIDE A	53	2.3	1.00	–
27	1tmt	PHE-PRO-ARG	30	2.2	0.54	0.77
28	1tom	METHYL-PHE-PRO-AMINO-CYCLOHEXYLGLYCINE	28	1.8	0.74	–
29	1uma	N,N-DIMETHYLCARBAMOYL-ALPHA-AZALYSINE	15	2.0	0.94	0.94
30	3hat	RNG-GLY-VAL-ARG	32	2.5	0.89	–
31	7kme	SEL2711	39	2.1	0.38	–
32	8kme	SEL2770	54	2.1	0.79	1.09

Table 2. GST crystallographic complexes used in testing the minimal rotation hypothesis. The second column denotes whether the ligand binds in the glutathione/peptidyl ("P") binding site, the hydrophobic/xenobiotic ("X") binding site, or extends into both ("B") sites.

#	Binding	PDB code	Ligand name	# Non-H atoms	Resolution (Å)	Best ligand RMSD (Å)	
						Flexible docking	Rigid docking
1	B	10gs	BENZYLCYSTEINE PHENYLGLYCINE	33	2.2	0.36	0.36
2	B	12gs	S-NONYL-CYSTEINE	29	2.1	0.36	0.62
3	X	13gs	SULFASALAZINE	28	1.9	1.78	1.78
4	B	18gs	1-(S-GLUTATHIONYL)-2,4-DINITROBENZENE	32	1.9	0.61	0.61
5	B	1aqv	P-BROMOBENZYLGLUTATHIONE	28	1.9	0.37	0.37
6	P	1aqw	GLUTATHIONE	20	1.8	0.46	0.54
7	B	1aqx	S-(2,3,6-TRINITROPHENYL)CYSTEINE	35	2.0	0.78	0.93
8	B	1pgt	S-HEXYLGLUTATHIONE	26	1.8	0.53	0.73
9	X	20gs	CIBACRON BLUE	22	2.5	0.52	-
10	X	2gss	ETHACRYNIC ACID	19	1.9	2.48	-
11	B	2pgt	(9R,10R)-9-(S-GLUTATHIONYL)-10-HYDROXY9,10DIHYDROPHENANTHRENE	35	1.9	0.26	1.10
12	B	3gss	ETHACRYNIC ACID-GLUTATHIONE CONJUGATE	39	1.9	0.52	0.52
13	B	3pgt	GLUTATHIONE CONJUGATE OF (+)-ANTI-BPDE	43	2.1	0.59	-

Table 3. Ligand-free structures and their corresponding ligand-bound complexes used in testing the minimal rotation hypothesis. The template sizes are given as the number of template points to indicate the differences in the sizes of the binding sites of the proteins.

#	PDB code		Protein/ligand complex	Resolution (Å)		Best ligand RMSD (Å)		Template size
	Free	Bound		Free	Bound	Flexible docking	Rigid docking	
1[M]	1ahc	1ahb	ALPHA-MOMORCHARIN/ FORMYCIN 5'-MONOPHOSPHATE	2.0	2.2	0.94	–	88
2[M]	1ajz	1aj2	DIHYDROPTEROATE SYNTHASE/ DIHYDROPTERINE-DIPHOSPHATE	2.0	2.0	0.75	–	79
3	3cox	1coy	CHOLESTEROL OXYDASE/3-BETA-HYDROXY-5-ANDROSTEN-17-ONE	1.8	1.8	1.61	–	74
4	1gmq	1gmr	RNASE SA/GUANOSINE-2'-MONOPHOSPHATE	1.8	1.8	1.28	1.63	87
5	3grs	1gra	GLUTATHIONE REDUCTASE/ GLUTATHIONE DISULFIDE	1.5	2.0	0.69	1.88	139
6	1kem	1kel	CATALYTIC ANTIBODY 28B4 FAB FRAGMENT/AAH*	2.2	1.9	0.46	–	74
7[M]	2hvm	1llo	HEVAMINE(ENDOCHITINASE)/ N-ACETYL-D-ALLOSAMINE	1.8	1.9	0.67	–	150
8	1nsb	1nsc	NEURAMINIDASE/N-ACETYL NEURAMINIC ACID(SIALIC ACID)	2.2	1.7	0.40	0.67	74
9	1swa	1swd	STREPTAVIDIN/BIOTIN	2.0	1.9	0.62	0.67	37
10[M]	2ptn	1tps	TRYPSIN/INHIBITOR A90720A	1.5	1.9	0.93	–	143
11	1xib	1xid	D-XYLOSE ISOMERASE/ L-ASCORBIC ACID	1.6	1.7	2.28	–	45
12[M]	1ydc	1ydb	CARBONIC ANHYDRASE II/ ACETAZOLAMIDE	2.0	1.9	1.42	–	50
13	2chs	2cht	CHORISMATE MUTASE/ENDO-OXABICYCLIC INHIBITOR	1.9	2.2	1.02	–	39
14	2apr	3apr	ACID PROTEINASE/REDUCED PEPTIDE INHIBITOR	1.8	1.8	0.54	–	153
15	1tli	3tmn	THERMOLYSIN/VAL-TRP	2.0	1.7	0.99	0.99	75
16	2ctv	5cna	CONCANAVALIN A/ALPHA-METHYL-D-MANNOPYRANOSIDE	2.0	2.0	1.99	–	57
17	2sga	5sga	PROTEINASE A/TETRAPEPTIDE ACE-PRO-ALA-PRO-TYR	1.5	1.8	0.59	1.90	126
18[M]	6taa	7taa	FAM. 13 ALPHA AMYLASE/ MODIFIED ACARBOSE HEXASACCHARIDE	2.1	2.0	0.82	–	133

* AAH = 1-[N-4'-nitrobenzyl-N-4'-carboxybutylaminomethylphosphonic acid.
[M] Proteins existing as biological monomers both in their ligand-free and ligand-bound states.

Figure Legends

Figure 1. (A) The active site of thrombin filled with template points colored according to type: blue for donor, red for acceptor, white for donor and/or acceptor, green for hydrophobic. The template points in the bottom of the S1 specificity pocket (circled in figure) were marked as key points, meaning that each docked ligand was required to match at least one of these points. (B) An example of a set of typical side chain rotations by SLIDE (shown in green) in the active site of thrombin upon docking a known beta-strand mimetic inhibitor (red spheres) from the PDB complex 1a61 (St Charles et al. 1999). The original positions of the side chains in the ligand-free crystal structure (1vr1) are shown in white. Residues in white without a corresponding green conformation were not moved by SLIDE.

Figure 2. Side-chain rotations performed by SLIDE (A) upon docking 32 known ligands into the ligand-free active site of thrombin (PDB code 1vr1), compared to (B) the dihedral angle differences observed between ligand-free and ligand-bound crystal structures.

Figure 3. Side-chain rotations performed by SLIDE (A) upon docking 13 known ligands into the ligand-free active site of GST (PDB code 16gs), compared to (B) the dihedral angle differences observed between ligand-free crystal structure and the corresponding side chains from the ligand-bound structures.

Figure 4. Side-chain rotations performed by SLIDE (A) upon docking 18 known ligands into the corresponding ligand-free target structures, compared to (B) the dihedral angle differences

between corresponding side chains from the binding sites of ligand-free and ligand-bound structures.

Figure 5. Difference between binding-site and non-binding-site surface side-chain changes from ligand-free to ligand-bound X-ray conformations. The number of occurrences was normalized to 100 for both the binding-site (484 total dihedral changes computed) and the surface rotations (1537 total dihedral changes computed) before subtracting the non-binding-site distribution from the binding-site distribution to generate the difference histogram shown here.

Figure 6. Favorability of protein side-chain conformations for four representative residue types (His, Ile, Leu, Lys) as reflected by χ_1 and χ_2 dihedral angle values for the 63 proteins analyzed. The plots were generated with PROCHECK (Laskowski et al. 1993; Morris et al. 1992). Only those residues rotated by SLIDE are shown. The number of side chains, shown in brackets, varies between the three columns because this number depends on how many instances of that side chain occurred in the binding sites of the thrombin, GST, and 18 other apo structures (first column), as well as how often SLIDE or nature moved that side chain in the 63 ligand-bound structures (second and third columns). Coloring of data squares indicates the favorability of conformations from yellow (most favorable) to red (least favorable) relative to the favorable conformations obtained from an analysis of 163 structures at resolution 2.0 Å or better, shown as the green background (data provided with the program PROCHECK). Those in unfavorable conformations (score < -3.00 given by PROCHECK) are labeled by residue number.

Figure 7. The G-factors of SLIDE conformations versus the G-factors of the corresponding side-chains from the apo structures colored by the G-factors of the holo form, where yellow represents most favorable and dark red the least favorable side-chain conformations in the ligand-bound crystal structures. The dashed lines at the G-factor value of -3.0 represents the cutoff below which the conformation is labeled as unfavorable by PROCKECK.

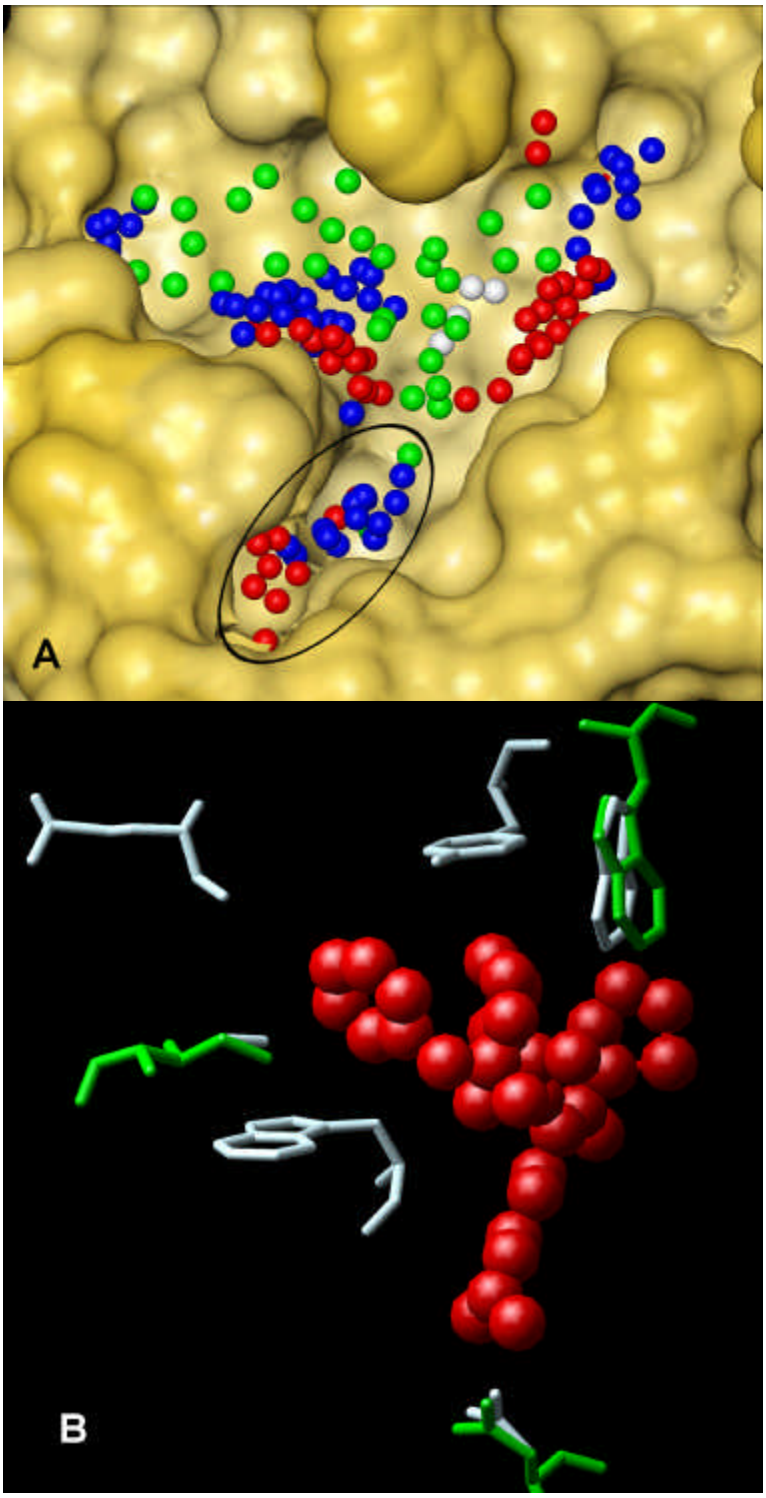


Figure 1. Zavodszky et al. PROTSCI/2004/011536

Fig. 2

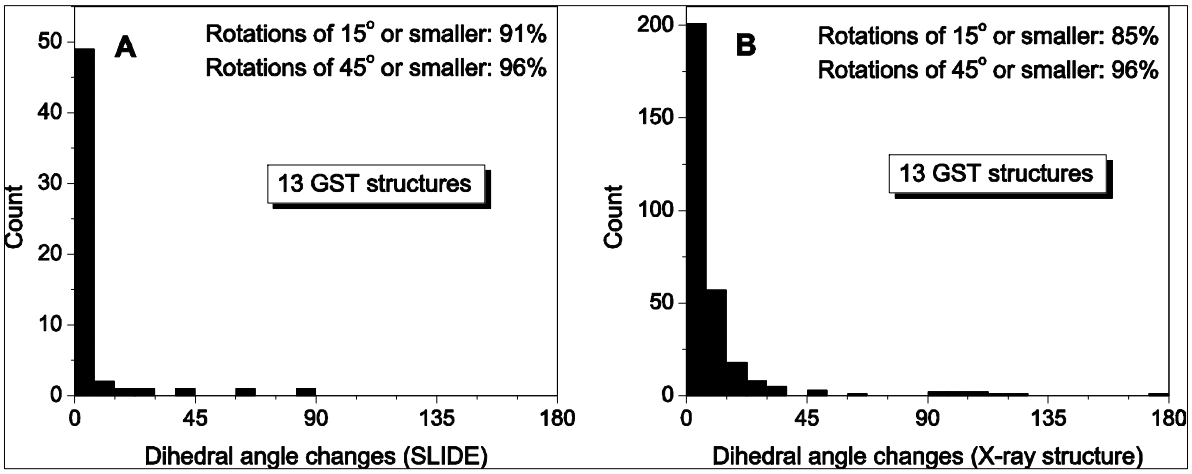


Fig. 3

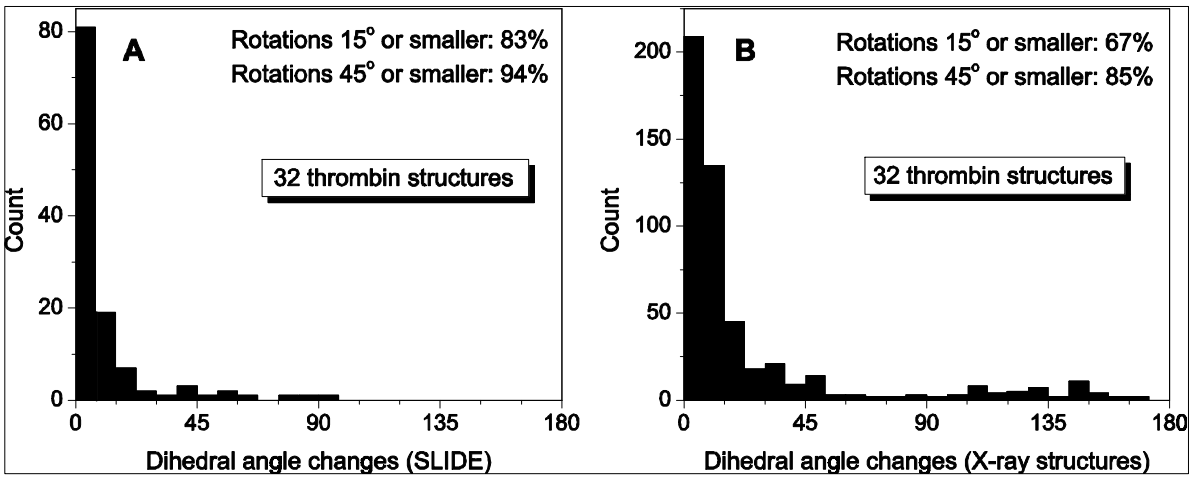
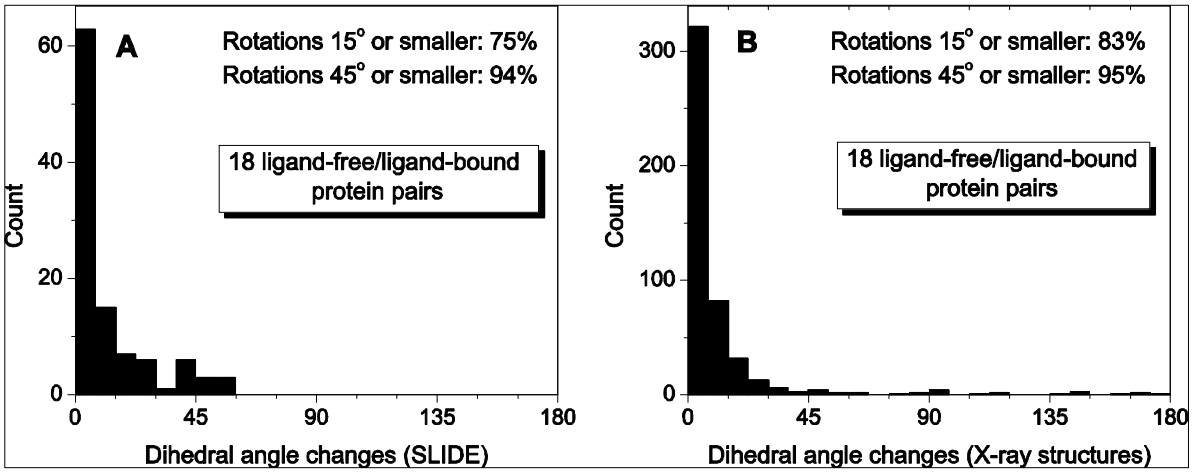


Fig. 4



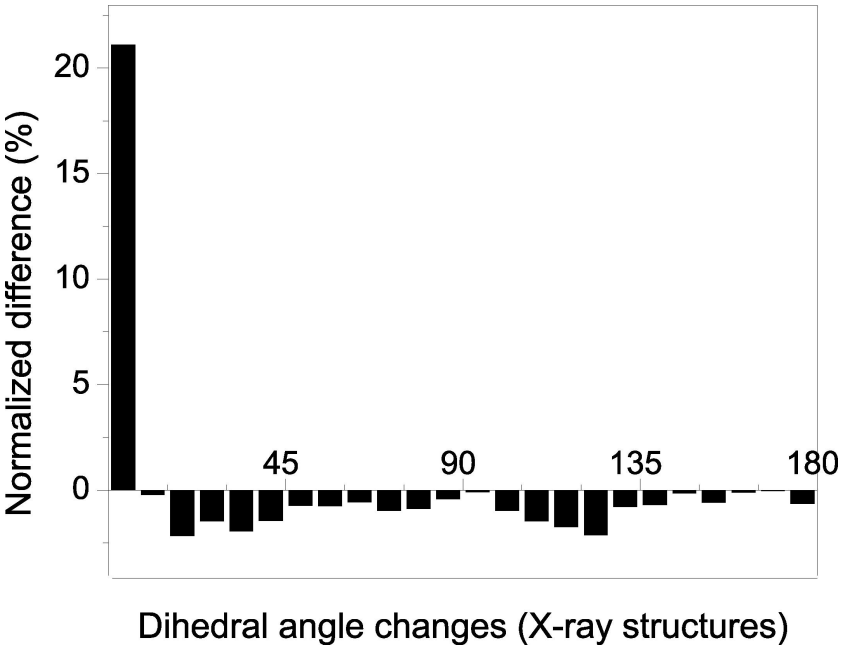


Figure 5. Zavodszky et al. PROTSCI/2004/011536

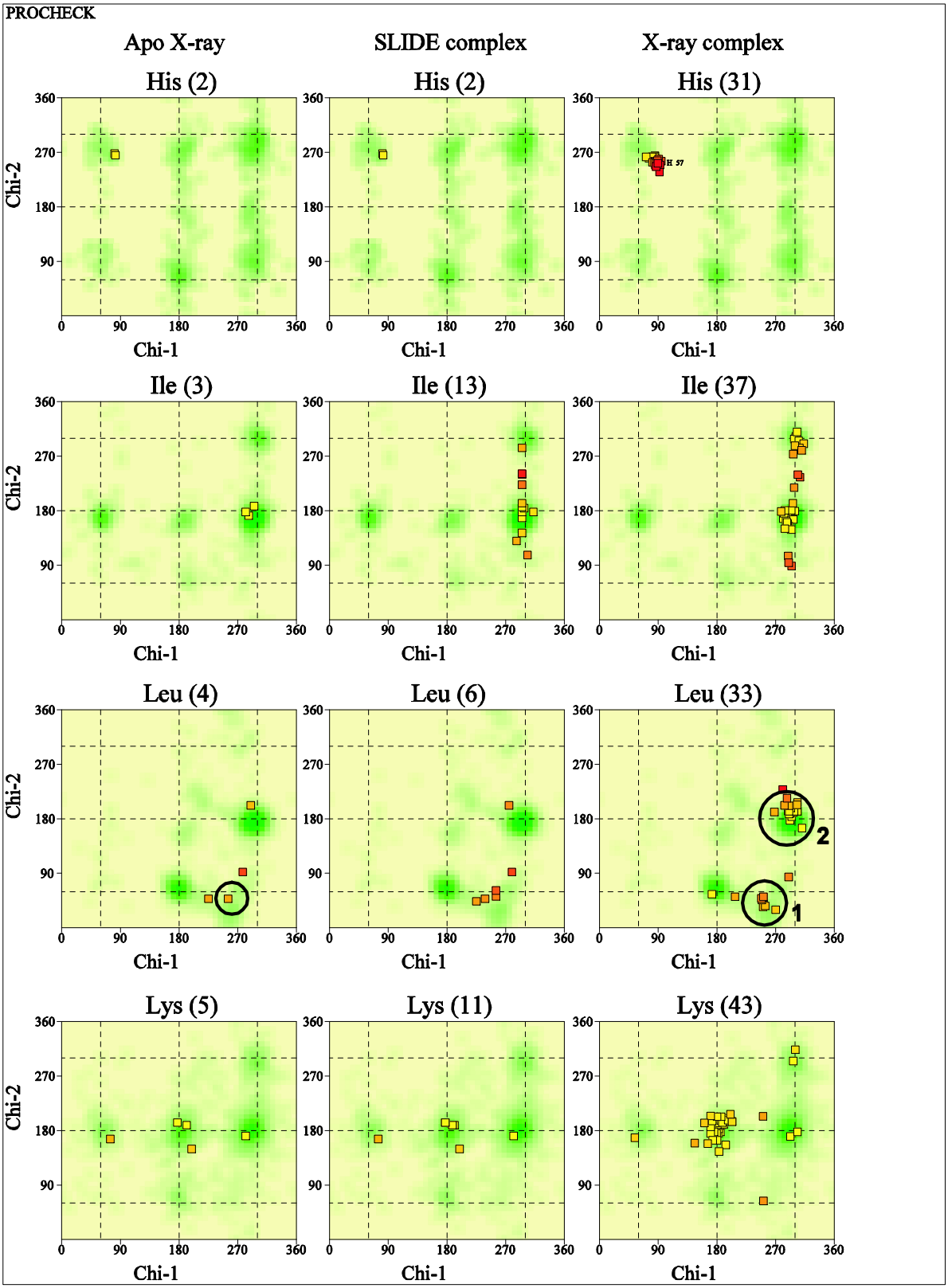


Figure 6. Zavodszky et al. PROTSCI/2004/011536

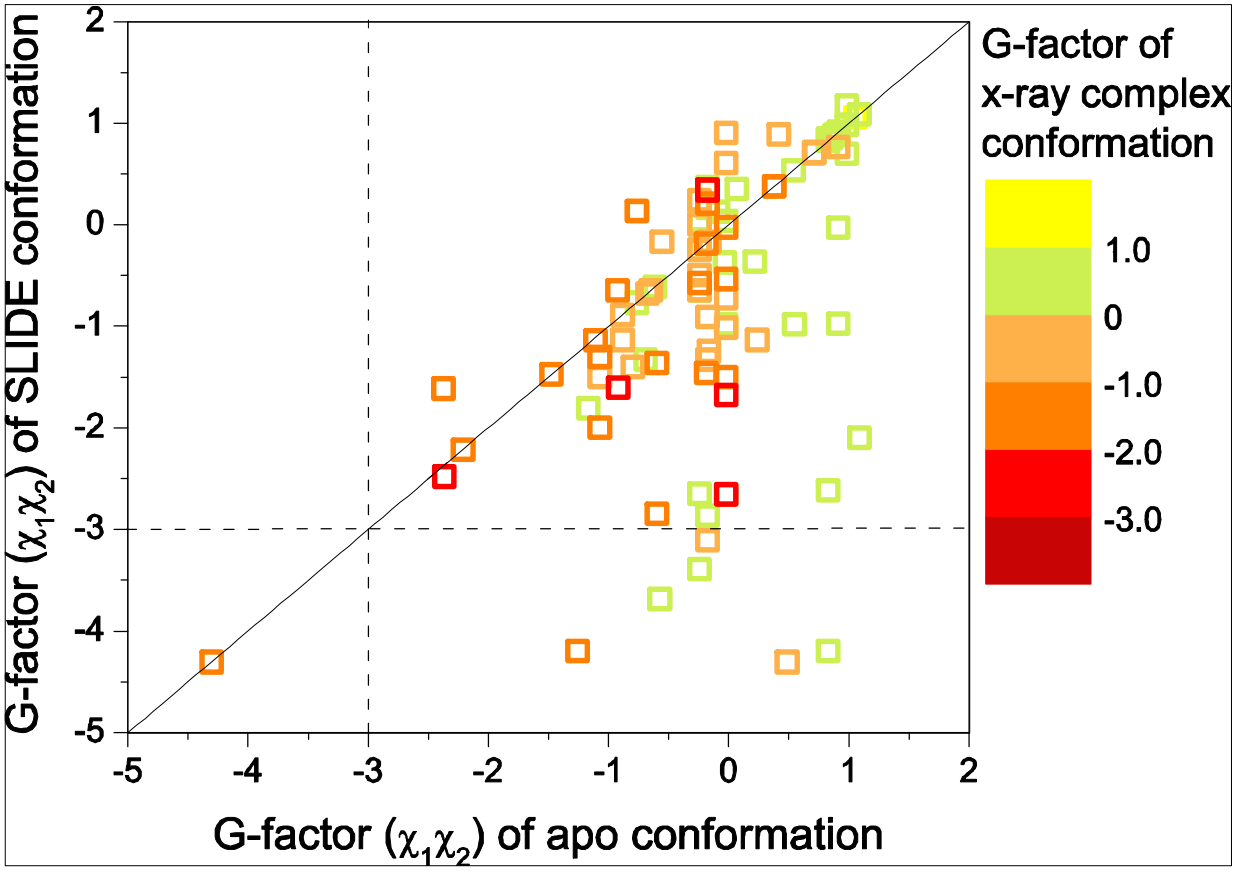


Figure 7. Zavodszky et al. PROTSCI/2004/011536