Published in final edited form as:

Proteins. 2012 February ; 80(2): 591–601. doi:10.1002/prot.23222.

Minimal ensembles of side chain conformers for modeling protein-protein interactions

Dmitri Beglov^{1,2,*}, David R. Hall^{1,2}, Ryan Brenke^{1,2}, Maxim V. Shapovalov³, Roland L. Dunbrack Jr.³, Dima Kozakov^{1,2,*}, and Sandor Vajda^{1,2,*}

¹BioMolecular Engineering Research Center, Boston University, Boston, Massachusetts

²Department of Biomedical Engineering, Boston University, Boston, Massachusetts

³Institute for Cancer Research, Program in Molecular and Translational Medicine, Fox Chase Cancer Center, Philadelphia, Pennsylvania

Abstract

The goal of this paper is to reduce the complexity of the side chain search within docking problems. We apply six methods of generating side chain conformers to unbound protein structures, and determine their ability of obtaining the bound conformation in small ensembles of conformers. Methods are evaluated in terms of the positions of side chain end groups. Results for 68 protein complexes yield two important observations. First, the end group positions change less than 1 Å upon association for over 60% of interface side chains. Thus, the unbound protein structure carries substantial information about the side chains in the bound state, and the inclusion of the unbound conformation into the ensemble of conformers is very beneficial. Second, considering each surface side chain separately in its protein environment, small ensembles of low energy states include the bound conformation for a large fraction of side chains. In particular, the ensemble consisting of the unbound conformation and the two highest probability predicted conformers includes the bound conformer with an accuracy of 1 Å for 78% of interface side chains. Since more than 60% of the interface side chains have only one conformer and many others only a few, these ensembles of low energy states substantially reduce the complexity of side chain search in docking problems. This approach was already used for finding pockets in protein-protein interfaces that can bind small molecules to potentially disrupt protein-protein interactions. Side chain search with the reduced search space will also be incorporated into protein docking algorithms.

Keywords

rotamer libraries; side chain flexibility; protein binding; structure prediction; preexisting ensemble of conformers

*Corresponding authors: Dmitri Beglov, Department of Biomedical Engineering, Boston University, 44 Cummington street, Boston, MA 02215, USA, dbeglov@bu.edu, phone: 617-358-4069, fax: 617-353-6766. Dima Kozakov, Department of Biomedical Engineering, Boston University, 44 Cummington street, Boston, MA 02215, USA, midas@bu.edu, phone: 617-353-4842, fax: 617-353-6766. Sandor Vajda, Department of Biomedical Engineering, Boston University, 44 Cummington street, Boston, MA 02215, USA, vajda@bu.edu, phone: 617-353-4757, fax: 617-353-6766.
The first two authors contributed equally to this article.

INTRODUCTION

Protein-protein interactions are fundamental to a large variety of biological processes, including protein localization, competitive inhibition, allosteric regulation, gene regulation, and signal transduction. The stable and specific binding of two proteins requires shape, electrostatic, and chemical complementarity of the interacting surfaces, primarily determined by the conformation of the side chains in the interface. In non-obligatory complexes these side chains serve double duty, as they interact with a protein partner when it is present, and with the solvent otherwise.

Proteins are dynamical objects, and it is expected that many of the side chains significantly change conformation when adapting to a different environment. Proteins are found to increase the nonpolar area of their interfaces upon the association.¹ Such conformational changes have traditionally been described in terms of the induced-fit model.² In many systems there is also evidence^{3–5} for the conformational selection model, involving selection from an ensemble of preexisting states.^{6–8} According to this model, a folded protein does not have a single unique structure but is regarded as an ensemble of similar structures having similar energy contents that are in rapid fluctuation with each other. Thus, the free energy landscape of the protein consists of many local minima of similar energies, and the heights of the energy barriers separating the minima define the timescale of conformational exchange.⁹ If the barriers are low relative to the Boltzmann energy ($k_B T$), thermal fluctuations can lead to a significant population of the visited conformational states. Following binding, the ensemble of conformations undergoes a population shift, redistributing the states, and both conformational selection and induced fit appear to play roles.^{8, 10}

It is easy to show that many surface side chains of a protein move among a number of conformers, thereby forming an ensemble of substates. Even static X-ray crystallography reveals alternative conformations for a number of side chains. In addition, different conformations are frequently observed when the same protein is crystallized in other space groups under different conditions,¹¹ and alternative conformations for important side chains are sometimes observed in the asymmetric unit of the same crystal. For example, the Arg 288 side chain has two substantially different conformations in the ligand-free structure (PDB code 1prg) of the PPAR- γ ligand binding domain.¹² Recently it was emphasized that many side chains may have conformers that are not seen in the final X-ray structure. Alber and co-workers¹³ systematically sampled the electron density around each dihedral angle in a test set of 402 structures determined at 1.5 Å resolution or higher, and showed that 15% of the residues with unbranched side chains had unmodelled secondary electron-density peaks below the normal noise threshold. These peaks are significantly enriched at low energy rotameric positions, supporting the interpretation that they reflect true minor populations. NMR methods developed over the past decade also allow the structural characterization of weakly populated states in the conformational ensemble.¹⁴

As shown by the various rotamer libraries,^{15, 16} some of the longer side chains have many conformers. However, there is substantial evidence that a given protein environment reduces the number of conformational states. Moreover, the evidence also suggests that relatively

small ensembles of the low energy states in the unbound protein also include the conformation seen in a protein-protein complex. For many side chains the “ensemble” consists of a single conformer, i.e., the conformation does not change between unbound and bound states. For example, Stroud and co-workers have shown that the conformational changes in proteins interacting with small ligands can be relegated to a few specific residues and that it is possible to find a minimal manifold of structures that can be used as input to a more exactly defined docking target.¹⁷ Edelman and co-workers analyzed a large database of complexes formed between proteins and small ligands, and showed that in 85% of cases only three or fewer side chains in the ligand binding pocket change conformation upon binding.¹⁸ Kimura et al.¹⁹ suggested that in protein-protein binding a number of “key” side chains act as ready-made recognition motifs by acquiring bound-like conformations before any physical interaction with the receptor. This hypothesis was found to be consistent with molecular dynamics (MD) simulations of three proteins in explicit solvent as well as by the structural conservation of key residues in some protein families. These results were confirmed by Rajamani et al.²⁰ who performed nanosecond scale MD simulations of 11 proteins, showing that the side chains most important for recognition spend substantial fractions of the simulation time (30–90%) in rotameric states that are close to their conformations in the complex. Thus, the set of high probability conformers that are compatible with the unbound protein also includes the bound state. Supporting this observation, Li et al.²¹ recently noted that the residues that form the energetic hot spots in protein-protein interfaces are frequently located in preexisting pockets. In 16 out of the 18 analyzed complexes, the root mean-squared deviation (RMSD) between the bound and unbound states of the atoms lining these pockets was found to be as small as 0.9 Å, implying that the unbound and bound forms of the pockets are very similar. Smith et al.²² studied the extent to which the conformational fluctuations of proteins in solution reflect the conformational changes that they undergo when they form binary protein-protein complexes. According to extensive MD simulations, residues that get buried in the core of the interface upon binding are less mobile in the free protein, and assume conformations similar to those in the bound state. All of these results indicate that the ensemble of potential conformers in the unbound protein carry substantial information about the conformers of the bound state.

The main goal of this paper is to reduce the complexity of the side chain search within docking problems. As shown by Baker and co-workers,²³ accounting for side chain flexibility can substantially improve the accuracy of docking results, but repacking all interface side chains in an iterative docking algorithm is computationally very expensive. These computational costs can be substantially reduced if the search is restricted to certain side chains and to a limited number of conformers even for such moveable side chains. To achieve the goal of reducing the number of side chain conformers to be considered in docking we show that relatively small ensembles of side chain conformers, generated from unbound protein structures, include the bound conformations with specified accuracy. We explored six different methods of conformer generation, four of which rely on the backbone dependent rotamer library of Dunbrack and Karplus,¹⁵ and the remaining two on an end group position library that was developed for this paper (see Methods). In all cases, the initial conformers are adjusted to account for the local protein environment, either by

removing the conformers that lead to steric clashes, or by performing energy minimization. The ensembles generated by the six different methods are compared to the bound and unbound states, primarily in terms of side chain end group positions. The rationale for using this measure, rather than torsional angles or RMSD (Root Mean Square Deviations) is that molecular recognition is determined by the shape of the interacting proteins and by the positions of hydrophobic, polar, and charged patches on the interface. Nevertheless, additional comparison of methods in terms of torsional angles is provided as Supporting Information.

For the comparison of the six methods we consider a large set of protein complexes for which the structures of unbound component proteins are also available. For each side chain in the unbound structure we check for the presence of the bound conformation in small ensembles of low energy conformers. Based on this data we can calculate the number of conformers required to achieve some required accuracy. The analysis expands on our earlier studies^{19, 20} that used nanosecond scale molecular dynamics (MD) simulations of the individual proteins in explicit solvent to characterize potential side chain conformational states, but due to the high computational costs of the MD simulations were restricted to 11 complexes. In contrast, here we study proteins from 68 complexes, and the results provide the statistics required to make substantially more general statements on the predictive performance of minimal ensembles of side chain conformers. As will be shown, for more than 60% of the interface side chains the ensemble includes only the unbound conformer, whereas for others only a few low energy states. Thus, our calculations can substantially reduce the complexity of side chain search in docking problems. This approach was already used for finding pockets in protein-protein interfaces that can bind small molecules to potentially disrupt protein-protein interactions.²⁴ Side chain search with the reduced search space will also be incorporated into protein docking algorithms, but due to the larger number of side chains this will be somewhat more difficult.

METHODS

Side chain end group definitions

Assuming that the backbones do not significantly change, the interactions between two proteins are primarily determined by the positions of certain side chain functional groups, generally located close to the end of the side chains. Here we used the positions of the end groups, shown in column 2 of Table I, to describe the side chain conformations and to measure conformational changes. Given an interface side chain, these groups have a high probability of being buried in the protein-protein interface and being at least partially exposed to solvent in the unbound proteins. Some of the end groups are defined as a middle point between two side chain atoms to account for ambiguity of placement or symmetry of these atoms.

Test set for determining the minimal ensemble of conformers

We selected a set of 68 binary protein complexes from the protein-protein docking benchmark version 3.0.²⁵ Protein pairs that were parts of a larger than binary complex were excluded. The set includes 32 enzyme-inhibitor complexes, 24 antibody-antigen complexes

and 12 “other” type complexes (Supplementary Table I). In each complex we identified the interface side chains as those being less than 5 Å from any atom of the partner protein. While structures of both unbound proteins are available for all enzyme-inhibitor and “other” type complexes, the antigen-free antibody structure is not known for about half of the antibody-antigen complexes, and hence only the interface side chains of the antigens are considered in our study. The total number of selected interface side chains was 2,294. The composition of the set by residue type is shown in the third column of Table I. Starting from the unbound structures we generated the ensembles of side chain conformers by 6 different methods and tested these conformers for the closeness to the known bound conformation. The closeness was measured as a distance between the corresponding end groups after placement of the generated and bound conformers to a local backbone based coordinate system, in which C_α is at the origin, the C_α -N vector defines the x -axis, the y -axis is along the component of the C_α -C vector perpendicular to the x axis. The z -axis is obtained by the vector product of x and y . A side chain conformer satisfies the 1 Å accuracy criterion if the distance between its end group and the end group in the bound protein structure is less than 1 Å. We present the average accuracy of the six methods for generating conformer ensembles in the Results section for 1 Å and 2 Å accuracy thresholds.

Methods of conformer generation

We use six different methods to generate ensembles of conformers for the interface side chains. In all methods the side chains are considered one-by-one, while keeping the rest of the protein in its unbound conformation. In the first method we select rotamers in the order of their probability from the May 2002 version of the backbone dependent rotamer library.¹⁵ Rotamers that lead to a steric clash, defined as having a van der Waals energy in excess of 50 kcal/mol, are removed. We will use the abbreviation RL (rotamer library) to refer to this method. Energy calculations are performed using version 19 of the Charmm potential with polar hydrogens only. In the second method we use the conformation from the unbound structure as the first conformer. The remaining conformers are obtained from the backbone dependent rotamer Library, excluding both the sterically clashing rotamers and the ones that are closer than 1 Å to the unbound conformation in terms of the end group distance. The method is named RLU (rotamer library with unbound conformer).

In the third method we perform energy minimizations, starting from the rotamers of the backbone dependent rotamer library, to reach local minima. The method is therefore named RLM (rotamer library with minimization). The interface side chains are considered one-by-one in the minimizations, with all others being fixed in their unbound state. The energy function for the minimization includes the bonded and van der Waals terms of the Charmm potential,²⁶ and the electrostatic/solvation term based on the Analytical Continuum Electrostatics model.²⁷ A minimization run either succeeds by ending in a local minimum, or fails if the initial steric clashes cannot be resolved. To each minimized conformation from a successful run is assigned the probability of the corresponding initial rotamer. After finishing all minimizations, the resulting conformations are clustered based on their end group positions. We use the clustering radius of 1 Å and a simple greedy algorithm. For each end group position we first count the number of neighbors within the 1 Å radius. The position with the highest number of neighbors is selected as a first cluster center and the

structures within 1 Å (i.e., the members of the first cluster) are removed. Next, we again select the position with the highest number of neighbors within the 1 Å radius to form the second cluster. This is repeated until the entire set is exhausted. For each cluster we define a non-normalized probability by adding the probabilities of all cluster members. Finally, we rank the clusters according to their cumulative probabilities, and consider the centers of the few highest ranking clusters as predicted conformers.

In the fourth method called RLUM (rotamer library with unbound conformer and minimization) the conformation from the unbound structure is selected as the first conformer. The rest of the method is the same as the RLM method, but cluster centers within 1 Å of the unbound conformer are excluded. For the fifth method we developed a conformer library based on side chain end group positions (see below). Apart from selecting conformers from the end group library, this method, called EGLM (end group library with minimization), is the same as the RLM method. EGLM uses local minimization with subsequent clustering to identify separate local minima. For each side chain type, Column 6 of Table I shows the maximum number of such minima, obtained by applying the method to the interface side chains in our protein test set. The highest numbers are obtained for Arg (38) and Lys (36) side chains. Method 6 is developed from method 5 by adding the unbound conformer as the first conformer before the minimization. Thus, the method is named EGLUM, i.e., end group library with unbound conformer and minimization.

Development of a conformer library based on side chain end group positions

End group libraries of rotamers, specified in Cartesian coordinates, have been used previously for conformational studies.²⁸ Our library is based on the non-redundant high resolution subset of the Protein Data Bank (PDB), selected in the Dunbrack lab. This set was selected from all secondary structure types. Side chains with electron density in the bottom 10 percentile were removed. All instances of side chain conformations were considered without differentiating surface and core positions. Table I (column 4) shows the number of such end groups for 18 amino acid types. Glycine and alanine side chains are excluded from consideration since each has only one possible conformer. To obtain a statistics of the end group positions we placed each amino acid residue into a local coordinate system based on the N - C α - C atoms. The aligned conformations are clustered based on the positions of their end groups. We use a clustering algorithm similar to the one described for the RLM method. Again, for each end group position we first count the number of neighbors within the 1 Å radius. The first cluster center has the highest number of neighbors. The second cluster center has the highest number of members among the structures which do not belong to the first cluster. This procedure is repeated until all end groups are clustered. For each cluster we define an occupancy by counting the number of end group positions within the 1 Å radius around the cluster center. In this step we account for all positions, even for the ones that are members of a higher ranked cluster, and thus the sum of cluster occupancies generally exceeds the total number of end group positions. The resulting clusters define the potential conformational states for each side chain (Table I, column 5).

RESULTS AND DISCUSSION

Bound Conformations in the Ensemble of Unbound States

The two central hypotheses of this paper are that (1) for a large percentage of interface side chains a small ensemble of low energy conformers, generated for the unbound protein, also includes the bound conformation, and that (2) the required low energy conformers can be obtained for each side chain separately, while keeping all others in their unbound states. We study the validity of these hypotheses for a test set of 68 protein complexes that include enzyme-inhibitor and antigen-antibody pairs, as well as “other”, primarily signal transduction, complexes. We first investigated the ability of 6 side chain conformation generation methods to reproduce the bound conformation in terms of the end group positions. Since in the modeling of protein complexes it is advantageous to consider the minimum number of side chain conformers required to achieve a specified accuracy, we focused on generating small ensembles. Table II summarizes the average capability of the methods to generate a conformer, which is closer than 1 Å or 2 Å to the bound state, in ensembles of 1 to 10 conformers, averaged over all 2,294 interface side chains of the protein test set.

The clash filtered backbone dependent rotamer library (RL method) generates a close-to-bound top rotamer in 55% and 67% of cases, respectively, for 1 Å and 2 Å accuracy. If the unbound conformation is used as the top rotamer (RLU method), the average success rate increases to 63% and 76%. Thus, 63% of interface side chains do not change their conformation more than 1 Å, and 76% more than 2 Å, upon binding. The use of the unbound conformation is also beneficial when several conformers are considered, and the RLU method consistently outperforms the RL method. When only the first rotamer from the backbone dependent rotamer library is used with local minimization (RLM method), the success rates for 1 Å and 2 Å accuracy are 52% and 65%, respectively. However, RLM performs better than RL when 2 to 5 rotamers are considered with accuracy 1 Å, and significantly better than RL for more than 2 rotamers with accuracy 2 Å. The inclusion of the unbound conformation as a top rotamer above the locally minimized conformations (RLUM method) gives advantage over the previous three methods for small ensembles of 3 to 6 predicted rotamers.

With the 1 Å accuracy, the performance of the EGLM method, which uses the end group library with local minimization, is slightly lower than that of the RL method. However, EGLM outperforms RL for the 2 Å accuracy when generating ensembles of several conformers. The EGLM method also has the ability to generate conformations different from the unbound one. Adding the unbound conformation as the first conformer to the ones generated by the EGLM method yields the EGLUM (End Group Library with Unbound structure and Minimization) method, which provides slightly better performance than the other five methods. The advantage of the EGLUM method over RLU is small (1–2%), but is consistently present for all ensemble sizes and for both 1 Å and 2 Å accuracy thresholds.

The success rates somewhat differ between the three different types of complexes. The overall superiority of the EGLUM method is retained for enzyme-inhibitor and antibody-antigen protein complexes (Supplementary Tables SIV and SV). Here we summarize the

results of the EGLUM method with 1 Å accuracy. For enzyme-inhibitor pairs the unbound conformer alone predicts the bound state for 65% of the side chains, and this is increased to 79% when we add the two highest probability predicted states. For enzymes the starting accuracy using the unbound state is 71%, which improves to 83% when adding two predicted states. For inhibitors the starting value is only 58%, and adding two more states yields only 75% accuracy. These differences are easy to explain, as inhibitors are generally more convex in shape than enzymes, and the interface side chains have more freedom to move in the unbound state, making their bound position more difficult to predict. For antibodies and antigens adding two predicted conformers to the unbound conformation yields a 17% improvement, resulting in the overall accuracy of 76%. The antibody side chains are predicted with 83% accuracy versus 72% for the antigens. For the “other” type protein complexes (Supplementary Table SVI) the best performing methods at the ensemble size of 3 and accuracy 1 Å are the backbone dependent, unbound enriched methods RLU and RLUM. They show 78% success rate with 16% improvement when adding the two highest ranked states. The EGLUM method trails them with a slightly lower success rate of 77%.

Distribution of End Groups in Proteins

Table I shows the definition of the side chain end groups we monitor in our study, and the number of side chains extracted from the high resolution non-redundant subset of the Protein Data Bank (PDB). The end groups of 8 side chains in Table I are defined as a middle point between two atoms of the corresponding residue. These two atoms are either symmetrically equivalent, or are somewhat ambiguous in terms of placement into crystallographic electron density. For each residue type we also show the number of clusters obtained by clustering the observed end group positions in a local coordinate system aligned to backbone atoms. The centers of these clusters constitute our initial end group library that was used to provide the starting points for the minimization runs in the EGLM and EGLUM methods. Similar to a rotamer library, the cluster centers and the corresponding occupancy numbers define the discrete distribution of the *a priori* end group positions, i.e., without accounting for the protein environment. Although restricting considerations to end group positions rather than considering all torsional angles removes some of the redundancies, the numbers of states are generally not much smaller than the ones in the standard rotamer library¹⁵ used in the RL, RLU, RLM and RLUM methods. In fact, we have many conformers for the long side chains Lys and Arg, for which the end group positions distribute over a relatively large volume. As an illustration, Figure 1A shows the distribution of observed positions of the NZ atoms (shown as cyan spheres) of Lys residues, selected from the PDB and aligned to a common N - C α - C frame. Red spheres show the centers of the 1 Å clusters obtained from this distribution, representing the members of the end group library for Lys. Figure 1A also shows (in green) the conformation of the Lys 549 side chain in the structure of the unbound von Willebrand Factor A1 domain (PDB code 1AUQ), as well as the two highest ranked (i.e., highest probability) cluster centers (in magenta and yellow, respectively). Note that the unbound conformation of the side chain is outside of the densely populated regions of the original distribution. Figure 1B shows the same three Lys 549 conformers, as well as the conformation in the bound state (shown in cyan). Notice that the highest ranked predicted conformer (magenta) is much closer to the bound conformation than to the unbound one.

Prediction of Bound Conformers for Different Side Chain Types

Figure 2 and 3 show the success rates of the six side chain generation methods for each of the 18 amino acid types and for small ensembles with up to 3 conformers. For the 2 Å accuracy (Figure 3) and for an ensemble size of 3 conformers, the end group based EGLUM method (shown in green) performs as well or better than the other methods for all amino acids except for Gln, Glu, Lys and Met. These four side chains are better generated by the RLUM method based on backbone dependent rotamer library (shown in blue). For Pro and Val with 2 Å accuracy, the success rate of all 6 methods is 100%. As can be also seen in Table II, the local minimization is beneficial for approaching the bound conformation in the 2 Å range.

For 1 Å accuracy (Figure 2) and an ensemble size of 3 conformers, the EGLUM method (shown in green) is as good as the others for Arg, Asn, Asp, Cys, Gln, Glu, Ser, Thr, and Tyr. It is better than the others for Leu, Pro and Trp. For His, Met and Phe the best performer is the RL method (shown in magenta). The unbound conformer enriched RLU method (shown in cyan) performs best for Lys and Val side chains. The RLUM method (shown in blue) performs better for Ile.

In this work we use the end group distance metrics to calculate the accuracy of the methods. To provide comparison with other methods, however, we also calculated the accuracy in terms of the more commonly used χ_{1+2} metrics. According to this measure, a conformer is close to the bound conformation when the difference in χ_1 (and, if present, in χ_2) is less than 40 degrees. The results for this χ_{1+2} metrics are presented in Supplementary Figure S1.

Removal of Clashes for Long Side Chains

Long side chains are known to undergo significant conformational changes upon the protein-protein association.¹ Here we emphasize that considering several rotamers is generally very useful for docking even when the accuracy of the long side chain prediction is relatively limited. This was first demonstrated using the fast side chain refinement program FireDock, which adjusts the conformations only for some of the interface side chains.²⁹ The analysis of FireDock results has shown that although side chain accuracy was not improved, by removing some clashes the method generally improved scoring and provided better overall docking results. Although the long side chains have limited success rate of the prediction with the 1 Å accuracy, adding the highest ranked conformers helps to remove clashes. To show that this is the case, we calculated the rate of finding a non-clashing conformer for Arg and Lys side chains (Figure 4). The unbound protein structures and their bound states were globally superimposed to detect steric clashes, defined as resulting in a van der Waals energy in excess of 50 kcal/mol. Using only unbound conformers, Arg and Lys have no clashes with the partner protein in 70% and 73% of cases. After adding the two highest probability conformers generated by the EGLUM method, the percentage of clash-free side chains increases to 83% for Arg and to 82% for Lys. These numbers increase to 92% when considering 21 conformers for Arg and to 93% when considering 17 conformers for Lys.

Independence of “Moving” Side Chains

The simplicity of our analysis is based on the assumption that the highly populated states of each side chain can be determined while keeping all other side chains in their unbound conformational states. Thus, we assume that the side chains that have multiple alternative conformers, and hence can change conformation upon binding, mostly interact with side chains that are fixed in their unbound conformational states. We think that results of the previous two sections justify this approximation. Indeed, the small ensembles based on this assumed non-cooperativity provide good predictions of the bound conformation for a very large percentage of the side chains, particularly if we allow for 2 Å deviations from the true positions of the end groups.

Biophysical Implications

The results of this paper confirm that 63% of the surface side chains do not change conformation upon binding (possibly apart from a minor re-adjustment of less than 1 Å), and that the rest of the side chains select conformations close to one of their low energy (high probability) states in the unbound protein. The limited complexity of side chain rearrangement upon binding is in good agreement with both thermodynamic and kinetic expectations. From a thermodynamic perspective the important implication is that the side chain entropy loss upon binding is substantially lower than the values calculated using generic conformer libraries for determining the degrees of freedom in the unbound proteins.

The reduced complexity is even more important for explaining the relatively high observed values of protein-protein association rates. The maximum rate constant for this process, $10^9 M^{-1} s^{-1}$, is given by the Smoluchowski equation that describes the collision rate for two uniformly reactive spherical molecules in solution.³⁰ A successful reaction between two proteins must meet the additional constraint that small reactive patches on a particular face of each protein are properly aligned.³¹ The probability of satisfying this condition in a random collision is very small, suggesting that the reaction rates should be several orders of magnitude lower than the diffusion-limited collision rates. However, it was shown that long-range electrostatic effects can heavily bias the approach of the molecules to favor reactive conditions, increasing the rates as high or higher than $10^9 M^{-1} s^{-1}$.^{31, 32} Rates that are two to three orders of magnitude higher than expected can also be observed for complexes without strong electrostatic interactions, which was explained in terms of desolvation effects.³³ In both cases the fast association is at least partly due to the fact that key side chains of the unbound proteins have conformations similar to those observed in the bound complex, and thus form ready-made recognition motifs that facilitate the formation of stable encounter complexes. The need for reduced side chain rearrangement upon binding also seems to be important. Schlosshauer and Baker³⁴ developed a simple diffusional model that predicts protein-protein association rates in the correct range, but only when neglecting any possible free energy barrier caused by side-chain freezing during complex formation, and any possible slowing down of diffusion within the binding funnel caused by the potential ruggedness of the landscape.³⁴ Our results demonstrating that the majority of side chains do not have to move, while the rest have only a few potential conformational states and are largely independent of each other, make meeting these conditions much more likely, assuring diffusion-controlled association.

Implications for protein-protein docking

Our results are in good agreement with a number of observations concerning protein-protein docking. First, the fact that a large percentage of side chains do not change conformations upon binding explains why rigid body docking methods still compete with approaches that account for side chain flexibility.³⁵ It appears that the ability to perform rigid but global sampling, e.g., by using a Fast Fourier Transform (FFT) correlation method may be more important than capturing the correct side chain conformations, particularly if a smooth (i.e., reduced sensitivity) scoring function is used for the docking. Even in Rosettadock,²³ where the interface side chains are re-packed in the docking process, the packing algorithm varies the conformation of only one residue at a time while keeping the other side-chains fixed; a full, combinatorial rotamer optimization is performed only once every eight cycles. Our results, obtained at the assumption of relative independence of the “moving” side chains, fully support this approach. Moreover, our results also imply that, given two unbound protein structures, the burden of side chain search can be further reduced. In fact, it is possible to identify, prior to docking, the side chains that have multiple potential high probability states, and restrict the conformational search to such “moveable” side chains and a set of discrete conformers for each of them. Pre-calculating an ensemble of protein structures is a well established and effective way of dealing with receptor flexibility in protein-small molecule docking.^{36–40} The approach is eminently feasible in that case because the protein-ligand interface generally includes only a few “moveable” side chains.¹⁷ As mentioned in the introduction, this approach was already used for finding pockets in protein-protein interfaces that can bind small molecules to potentially disrupt protein-protein interactions.²⁴ Since the interface and the number of “moveable” side chains are generally much larger in protein-protein complexes, docking all pre-generated alternative structures is not computationally effective. However, it is possible to design effective side chain search algorithms that restrict consideration to the “moveable” side chains and their potential conformational states.

CONCLUSIONS

There is an inherent contradiction concerning the behavior of side chains in protein-protein association. On one hand, the large number of conformers for many side chains suggests that there is a huge number of potential conformational states, which would imply both substantial loss of conformational entropy upon binding and would present a very difficult combinatorial problem for structure prediction. On the other hand, there is evidence that the number of feasible side chain arrangements is much smaller and that the conformational changes between unbound and bound conformations are restricted to a few side chains, which implies smaller entropy loss and an easier structure prediction problem. In addition, the smaller conformational changes also mean that the unbound protein structure provides substantial information on the bound state, which can be exploited in protein-protein docking. Although a number of studies support this second alternative,^{17, 29, 39, 40} we believe that the present paper provides the most systematic analysis of the problem so far.

We have developed six different methods of generating side chain conformers to unbound protein structures, and determined their ability of obtaining the bound conformation in small

ensembles of conformers. Each method was evaluated in terms of the positions of side chain end groups. Two methods emerged as particularly promising. The method called EGLUM (End Group Library with Unbound conformer and Minimization) is slightly more accurate than all the others, but involves energy minimizations from a number of initial states provided by our end group library. The method RLU (Rotamer Library with Unbound conformer) is almost as accurate, and since it requires only the removal of steric clashes involving a few energy evaluations rather than minimizations, it is substantially faster. Applied to the unbound structures of the component proteins prior to docking, both methods determine the high probability conformers for surface side chains. Results for 68 protein complexes show that more than 60% of the interface side chains have only one conformer and many others only a few, and thus predicting the ensembles of low energy states substantially reduces the complexity of side chain search in docking problems.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work has been supported by grants GM061867, GM093147, and GM064700 from the National Institute of General Medical Sciences.

References

1. Ruvinsky AM, Kirys T, Tuzikov AV, Vakser IA. Side-chain conformational changes upon Protein-Protein Association. *J Mol Biol.* 2011; 408:356–365. [PubMed: 21354429]
2. Koshland DE. Application of a theory of enzyme specificity to protein synthesis. *Proc Natl Acad Sci USA.* 1958; 44:98–104. [PubMed: 16590179]
3. Leder L, Berger C, Bornhauser S, Wendt H, Ackermann F, Jelesarov I, Bosshard HR. Spectroscopic, calorimetric, and kinetic demonstration of conformational adaptation in peptide-antibody recognition. *Biochemistry.* 1995; 34:16509–16518. [PubMed: 8845380]
4. Cao Y, Musah RA, Wilcox SK, Goodin DB, McRee DE. Protein conformer selection by ligand binding observed with crystallography. *Protein Sci.* 1998; 7:72–78. [PubMed: 9514261]
5. Berger C, Weber-Bornhauser S, Eggenberger J, Hanes J, Pluckthun A, Bosshard HR. Antigen recognition by conformational selection. *FEBS Lett.* 1999; 450:149–153. [PubMed: 10350075]
6. Ma B, Kumar S, Tsai CJ, Nussinov R. Folding funnels and binding mechanisms. *Protein Eng.* 1999; 12:713–720. [PubMed: 10506280]
7. Tsai CJ, Kumar S, Ma B, Nussinov R. Folding funnels, binding funnels, and protein function. *Protein Sci.* 1999; 8:1181–1190. [PubMed: 10386868]
8. Boehr DD, Nussinov R, Wright PE. The role of dynamic conformational ensembles in biomolecular recognition. *Nat Chem Biol.* 2009; 5:789–796. [PubMed: 19841628]
9. Frauenfelder H, Sligar SG, Wolynes PG. The energy landscapes and motions of proteins. *Science.* 1991; 254:1598–1603. [PubMed: 1749933]
10. Csermely P, Palotai R, Nussinov R. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends Biochem Sci.* 2010; 35:539–546. [PubMed: 20541943]
11. Lee AY, Gulnik SV, Erickson JW. Conformational switching in an aspartic proteinase. *Nat Struct Biol.* 1998; 5:866–871. [PubMed: 9783744]
12. Nolte RT, Wisely GB, Westin S, Cobb JE, Lambert MH, Kurokawa R, Rosenfeld MG, Willson TM, Glass CK, Milburn MV. Ligand binding and co-activator assembly of the peroxisome proliferator-activated receptor-gamma. *Nature.* 1998; 395:137–143. [PubMed: 9744270]

13. Fraser JS, Clarkson MW, Degnan SC, Erion R, Kern D, Alber T. Hidden alternative structures of proline isomerase essential for catalysis. *Nature*. 2009; 462:669–673. [PubMed: 19956261]
14. Mittermaier A, Kay LE. New tools provide new insights in NMR studies of protein dynamics. *Science*. 2006; 312:224–228. [PubMed: 16614210]
15. Dunbrack RL, Karplus M. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J Mol Biol*. 1993; 230:543–574. [PubMed: 8464064]
16. Dunbrack RL. Rotamer libraries in the 21st century. *Curr Opin Struct Biol*. 2002; 12:431–440. [PubMed: 12163064]
17. Anderson AC, O'Neil RH, Surti TS, Stroud RM. Approaches to solving the rigid receptor problem by identifying a minimal set of flexible residues during ligand docking. *Chem Biol*. 2001; 8:445–457. [PubMed: 11358692]
18. Najmanovich R, Kuttner J, Sobolev V, Edelman M. Side-chain flexibility in proteins upon ligand binding. *Proteins*. 2000; 39:261–268. [PubMed: 10737948]
19. Kimura SR, Brower RC, Vajda S, Camacho CJ. Dynamical view of the positions of key side chains in protein-protein recognition. *Biophys J*. 2001; 80:635–642. [PubMed: 11159432]
20. Rajamani D, Thiel S, Vajda S, Camacho CJ. Anchor residues in protein-protein interactions. *Proc Natl Acad Sci USA*. 2004; 101:11287–11292. [PubMed: 15269345]
21. Li X, Keskin O, Ma B, Nussinov R, Liang J. Protein-protein interactions: hot spots and structurally conserved residues often locate in complemented pockets that pre-organized in the unbound states: implications for docking. *J Mol Biol*. 2004; 344:781–795. [PubMed: 15533445]
22. Smith GR, Sternberg MJ, Bates PA. The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking. *J Mol Biol*. 2005; 347:1077–1101. [PubMed: 15784265]
23. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, Baker D. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J Mol Biol*. 2003; 331:281–299. [PubMed: 12875852]
24. Kozakov D, Hall DR, Chuang GY, Cencic R, Brenke R, Grove LE, Beglov D, Pelletier J, Whitty A, Vajda S. Structural conservation of druggable hot spots in protein-protein interfaces. *Proc Natl Acad Sci USA*. 2011; 108:13528–13533. [PubMed: 21808046]
25. Hwang H, Pierce B, Mintseris J, Janin J, Weng Z. Protein-protein docking benchmark version 3.0. *Proteins*. 2008; 73:705–709. [PubMed: 18491384]
26. Brooks BR, Brooks CL, Mackerell AD, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kucsera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M. CHARMM: the biomolecular simulation program. *J Comput Chem*. 2009; 30:1545–1614. [PubMed: 19444816]
27. Schaefer M, Karplus M. A comprehensive analytical treatment of continuum electrostatics. *J Phys Chem*. 1996; 100:1578–1599.
28. Rainey JK, Goh MC. Statistically based reduced representation of amino acid side chains. *J Chem Inf Comput Sci*. 2004; 44:817–830. [PubMed: 15154746]
29. Andrusier N, Nussinov R, Wolfson HJ. FireDock: fast interaction refinement in molecular docking. *Proteins*. 2007; 69:139–159. [PubMed: 17598144]
30. Smoluchowski MV. Versuch einer mathematischen theorie der koagulationskinetik kolloider loeschungen. *Z Phys Chem*. 1917; 92:129168.
31. Berg PH, von Hippel OG. Diffusion-controlled macromolecular interactions. *Annual Review of Biophysics and Biophysical Chemistry*. 1985; 14:131–158.
32. Schreiber G, Haran G, Zhou HX. Fundamental aspects of protein-protein association kinetics. *Chem Rev*. 2009; 109:839–860. [PubMed: 19196002]
33. Camacho CJ, Kimura SR, DeLisi C, Vajda S. Kinetics of desolvation-mediated protein-protein binding. *Biophys J*. 2000; 78:1094–1105. [PubMed: 10692300]
34. Schlosshauer M, Baker D. Realistic protein-protein association rates from a simple diffusional model neglecting long-range interactions, free energy barriers, and landscape ruggedness. *Protein Sci*. 2004; 13:1660–1669. [PubMed: 15133165]

35. Vajda S, Kozakov D. Convergence and combination of methods in protein-protein docking. *Curr Opin Struct Biol.* 2009; 19:164–170. [PubMed: 19327983]
36. Knegtel RM, Kuntz ID, Oshiro CM. Molecular docking to ensembles of protein structures. *J Mol Biol.* 1997; 266:424–440. [PubMed: 9047373]
37. Lorber DM, Shoichet BK. Flexible ligand docking using conformational ensembles. *Protein Sci.* 1998; 7:938–950. [PubMed: 9568900]
38. Carlson HA, McCammon JA. Accommodating protein flexibility in computational drug design. *Mol Pharmacol.* 2000; 57:213–218. [PubMed: 10648630]
39. Totrov M, Abagyan R. Flexible ligand docking to multiple receptor conformations: a practical alternative. *Curr Opin Struct Biol.* 2008; 18:178–184. [PubMed: 18302984]
40. Huang SY, Zou X. Ensemble docking of multiple protein structures: considering protein structural variations in molecular docking. *Proteins.* 2007; 66:399–421. [PubMed: 17096427]

Figure 1.

End group positions for the Lys side chain. (a) Distribution of NZ atom positions of lysine are shown in semi-transparent cyan spheres. Red spheres show the centers of the 1 Å clusters obtained from the set of initial positions. The cluster centers are the members of the end group library for Lys. Figure 1A also shows (in green) the conformation of the Lys 549 side chain in the structure of the unbound von Willebrand factor A1 domain (PDB code 1AUQ), as well as the centers of the first and second highest probability clusters shown in magenta and yellow, respectively. Note that the unbound conformation of Lys 549 is on the periphery of the densely populated region of the original distribution. (b) The same three Lys 549 conformers shown in 1A, plus the conformation in the bound state (PDB code 1M10, chain X), shown in cyan. The highest ranked predicted conformer (magenta) is substantially closer to the bound conformation than to the unbound one.

Figure 2.

Success rates of finding the bound conformation with 1 Å accuracy in small ensembles of side chain conformers. Results are shown for the six side chain conformer generation methods and for each side chain type separately, averaged over the 68 protein complexes in the test set. Vertex-down triangles, vertex-up triangles, and circles show, respectively, performance for ensembles with one, two, and three conformers. For each side chain, results for the six methods are shown from left to right as follows: RL - magenta, RLU - cyan, RLM - brown, RLUM - blue, EGLM - red, and EGLUM - green.

Figure 3.
Success rates of finding the bound conformation with 2 Å accuracy in small ensembles of side chain conformers. All notations are the same as in Fig. 2.

Figure 4. Success rates of finding a non-clashing conformer for Arg and Lys side chains (shown in red and green, respectively) as functions of the number of conformers generated by the EGLUM method.

Table 1
Side Chain End Group Definitions and Statistics of Test Set and End Group Library

Side chain	End group	Number of side chains			Number of	
		In test set	In PDB set	Initial clusters	Max. local minima	
ARG	NH1-NH2	150	37649	237	38	
ASN	OD1-ND2	155	32078	14	5	
ASP	OD1-OD2	173	43775	14	5	
CYS	SG	69	10138	9	4	
GLN	OE1-NE2	125	27002	48	10	
GLU	OE1-OE2	161	47601	55	10	
HIS	CE1-NE2	91	17952	27	5	
ILE	CD1	85	45044	28	9	
LEU	CD1-CD2	144	71705	18	6	
LYS	NZ	143	39098	203	36	
MET	CE	47	13252	93	20	
PHE	CZ	88	32213	37	6	
PRO	CG	126	34781	3	2	
SER	OG	184	43994	8	4	
THR	OG1	166	42370	7	4	
TRP	CH2	83	11863	111	10	
TYR	CZ	196	28499	43	7	
VAL	CG1-CG2	108	57177	3	3	

Table II
Average Rate (%) of Finding the Bound Conformation in Small Ensembles of Side Chain Conformers

Number of conformers	Method					
	RL ^a	RLU ^b	RLM ^c	RLUM ^d	EGLM ^e	EGLUM ^f
1 Å accuracy						
1	55	63	52	63	43	63
2	66	72	67	72	63	73
3	72	76	73	77	71	78
4	74	78	75	79	73	79
5	75	79	76	79	74	80
6	77	80	76	80	75	81
7	78	81	77	80	76	81
8	78	81	77	81	76	82
9	79	81	77	81	77	82
10	79	82	77	81	77	82
2 Å accuracy						
1	67	76	65	76	61	76
2	79	86	83	86	80	87
3	84	89	90	91	88	91
4	87	91	91	93	91	93
5	89	92	93	94	93	94
6	90	93	94	95	94	95
7	91	94	95	95	95	96
8	92	95	95	96	95	96
9	92	95	95	96	96	96
10	93	95	96	96	96	97

^aRotamer Library
^bRotamer Library with the Unbound conformer.
^cRotamer Library with Minimization.
^dRotamer Library with Unbound conformer and Minimization.

$e_{\text{End Group Library with Minimization.}}$

$f_{\text{End Group Library with Unbound conformer and Minimization.}}$