The Impact of Side-Chain Packing on Protein Docking Refinement

Mohammad Moghadasit,‡,‡ Hanieh Mirzaei,‡,‡ Artem Mamonov,§ Pirooz Vakili,‖‡ Sandor Vajda,§ Ioannis Ch. Paschalidis,⊥,⊥,⊥ and Dima Kozakov⊥,§

†Division of Systems Engineering, ‡Center for Information and Systems Engineering, §Department of Biomedical Engineering, ‖Department of Mechanical Engineering, and ‡Department of Electrical and Computer Engineering, Boston University, Boston, Massachusetts 02215, United States

Supporting Information

ABSTRACT: We study the impact of optimizing the side-chain positions in the interface region between two proteins during the process of binding. Mathematically, the problem is similar to side-chain prediction, which has been extensively explored in the process of protein structure prediction. The protein–protein docking application, however, has a number of characteristics that necessitate different algorithmic and implementation choices. In this work, we implement a distributed approximate algorithm that can be implemented on multiprocessor architectures and enables a trade-off between accuracy and running speed. We report computational results on benchmarks of enzyme–inhibitor and other types of complexes, establishing that the side-chain flexibility our algorithm introduces substantially improves the performance of docking protocols. Furthermore, we establish that the inclusion of unbound side-chain conformers in the side-chain positioning problem is critical in these performance improvements. The code is available to the community under open source license.

The prediction of the tertiary structure of proteins is an important problem in computational structural biology with applications in protein structure design, protein association, and homology modeling. In general, side chains are more flexible than the backbone, and positioning them is a critical component of protein structure prediction.1–3

It is therefore not surprising that side-chain prediction has received significant attention during the last few decades. Most of the existing literature views the problem as an optimization/search problem over possible side-chain conformations. Several works first attempt to reduce the search space by applying the idea of dead-end elimination (DEE), which eliminates all side-chain conformations that cannot possibly be in the optimal solution.4,5 Lee and Subbiah4 proposed an approach based on a simulated annealing search. Lee6 also suggested a similar approach using a mean-field optimization search. Roitberg and Elber7 proposed a method that combined the latter two approaches. Bower et al.8 introduced heuristics to search over the space of specific energy functions implemented in the SCWRL package. The latest version of the package, SCWRL4.0,9 implements a tree decomposition algorithm,10 which is an exact method using dynamic programming. Side-chain prediction has also been formulated as a mathematical programming problem. Specifically, it has been formulated as an integer linear programming (ILP) problem,11,12 and several strategies have been proposed to solve it.12,13 A semidefinite programming relaxation of the ILP problem was developed by Chazelle et al.14 and a second-order cone programming relaxation was proposed by Kingsford et al.11 The primary application of the work surveyed above is in side-chain prediction in the context of protein folding. In fact, some works consider the joint folding and side-chain prediction problem (see Loose et al.15). Side-chain prediction algorithms attempt to resolve the uncertainty in the position of side chains (especially the ones on the protein surface) that computational or experimental determinations of the tertiary structure of proteins leave unresolved.

Side-chain prediction is, however, extremely important in the context of protein–protein association. As the two partner proteins approach each other, side chains in the interface region between the proteins tend to reorient so as to avoid steric clashes and facilitate the process of binding. Capturing this effect algorithmically has the potential to enhance docking protocols, which is the main motivation behind the work in this paper.

This problem of side-chain prediction in the course of protein docking has a number of characteristics—distinct from its application to folding—that enable the development of specialized and more efficient algorithms. First, side chains need to be repacked many times in the process of iterative docking algorithms, and hence, speed is a primary consideration. Second, the accuracy does not have to be extremely high. In fact, it was shown by Wolfson and co-workers16 that docking results can be substantially improved even by a very approximate adjustment of side chains that removes steric clashes. Third, the unbound protein structure provides a good approximation for the bound conformation of many side chains; it has been shown

Received: June 30, 2014

DOI: 10.1021/ci500380a
J. Chem. Inf. Model. XXXX, XXX, XXX--XXX
that over 60% of surface side chains retain the unbound conformation upon association with the partner protein. Thus, as will be discussed, considering the unbound conformer as one of the potential states substantially improves the results. Fourth, prediction is performed in the presence of a second protein that, in many cases, significantly reduces the potential joint conformations. In this light, the approach we have developed can be seen as accounting for these special conditions.

More formally, we will consider the so-called problem of side-chain packing (SCP) defined as follows: given the unbound structures of the receptor and the ligand, and assuming that the backbones remain rigid, predict the interface side-chain conformations that minimize the overall energy of the complex. SCP has been shown to be NP-hard and inapproximable (i.e., there is no polynomial-time algorithm to obtain solutions that are arbitrarily close to optimal).

Some forms of SCP have already been incorporated into docking procedures. In our docking protocol, first a large set of unbound receptor—ligand conformations are sampled using a rigid-docking technique called PIPER. Low-energy conformations are retained for further refinement. Refinement techniques iteratively move the ligand while keeping the receptor fixed in order to minimize an approximate energy function. This iterative search aims to find the rotation and orientation of the ligand that locally minimize the ligand—receptor interaction energy. SCP then becomes a component of energy evaluation for each ligand move.

SCP is a combinatorial problem, assuming that the side-chain positions are selected from a discrete set of conformations called rotamers. This is generally a good approximation within the framework of the required accuracy, particularly because in docking we generally work with relatively smooth scoring functions that do not heavily penalize minor steric overlaps. In this work, we formulate SCP as a maximum weighted independent set (MWIS) problem on an appropriately defined graph. We have developed a fully distributed algorithm that can output near-optimal solutions. We have established that our algorithm obtains an optimal solution to SCP for a special class of problem instances motivated by the structure of SCP arising in docking. In contrast to the aforementioned related work in the context of folding, our method is based on an approximate algorithm and forgoes optimality since state-of-the-art interaction energy models are also approximate. However, our method is fully distributed and requires only message passing between neighboring nodes of the graphical model of the SCP problem, which will be illustrated in Figure 1. Distributed algorithms are algorithms designed to run over multiple processors with no tight centralized control. This is appealing in our docking framework since, as mentioned earlier, one has to solve many instances of SCP in the course of docking two proteins. In some large instances of SCP involving numerous residues in a protein complex, the distributed implementation of our algorithm allows us to position the side chains with near-optimal accuracy yet with an average running time significantly smaller than for the state-of-the-art centralized algorithms.

The approach we have developed further enables the user to parametrize the method so as to trade-off the quality of the solution against the running time. In the docking application, and especially in the early stages of docking, we are looking not necessarily for the most near-native set of rotamers but for a good feasible solution that resolves the steric clashes of the interface. In such cases, the accuracy of the algorithm can be set such that desirable timing constraints are met.

Following an earlier observation, we test the impact of including the unbound conformations of side chains in the set of possible conformers. Our study of large benchmarks of enzyme— inhibitor and other types of complexes (as defined in Chen et al.) establishes that this inclusion substantially improves the side-chain prediction accuracy and the effectiveness of docking protocols. Essentially, we find that the unbound protein structure contains substantial information about the side chains of the bound state.

Our discussion thus far suggests that SCP in the process of docking exhibits significant special structure that provides us with a number of algorithmic and implementation choices (e.g., exact vs approximate, distributed vs centralized, inclusion of unbound conformers, etc.). In this light, our approach is not directly comparable to the existing and well-established side-chain prediction methods we surveyed. Still, we do report results comparing the side-chain prediction accuracy of our approach and that of SCWRL4, which is considered the state of the art. Several considerations need to be taken into account when interpreting such results. First, SCWRL4 is available in binary form and does not include the unbound rotamers. Moreover, it is an exact and centralized algorithm designed with folding applications in mind, and it does not benefit from a multi-processor environment. Our findings can potentially guide the development of alternative approaches for docking applications, including the adaptation of tools like SCWRL4.

The remainder of the paper is organized as follows. Section 1 outlines our method for solving the SCP problem. Computational results on benchmark sets and an extensive discussion are included in Section 2. Section 3 contains some concluding remarks.

1. METHODS

1.1. SCP Formulation. In the context of our docking application, we are interested only in positioning the side chains located in the interface between the receptor and the ligand. Side chains buried within the proteins are typically well-packed, and non-interface surface side chains have no significant effect on docking. We fix the position and orientation of the ligand with respect to the receptor and define the interface residues \( I \) as the set of all receptor and ligand residues whose \( C_\alpha \) atoms are within a small distance (10 Å) from a \( C_\beta \) atom located on the partner molecule. Let \( U_i \) denote the set of rotamers for each residue \( i \in I \) and \( |I| \) denote the cardinality (the number of elements) of \( I \).

The goal of SCP is to choose one rotamer per residue to minimize the free energy of the complex. Let \( i \) and \( j \) denote the rotamers selected for residues \( i \) and \( j \in I \). Then, the overall energy takes the form:

\[
E = E_0 + \sum_{i \in I} E(i) + \sum_{i,j \in I, i < j} E(i, j)
\]

where \( E_0 \) is the self-energy of the two backbones, \( E(i) \) is the energy of the interaction between rotamer \( i \), of residue \( i \) and the two backbones, including the self-energy of rotamer \( i \), and \( E(i, j) \) is the pairwise interaction energy between the selected rotamers \( i \) and \( j \), for \( i \neq j \).

We next formulate SCP as an MWIS problem on an appropriately defined graph \( G = (V, E) \) whose nodes are assigned weights. The MWIS problem amounts to selecting a set of nodes of \( G \) that form an independent set (i.e., no two nodes selected are connected by an edge) of maximal total weight. We construct \( G \) as follows. The node set of the graph, \( V \), consists of

DOI: 10.1021/ci500380a
two types of nodes: single-rotamer nodes and pair-rotamer nodes. To each rotamer \( i \) of each interface residue \( i \) we assign a single-rotamer node, and to each pair of rotamers \( (i_r, j_r) \) from two different residues \( i \) and \( j \) we assign a pair-rotamer node. We associate an energy value with each node: \( E(i) \) with single-rotamer nodes and \( E(i_r, j_r) \) with pair-rotamer nodes. We also assign to each node a non-negative weight such that higher weights correspond to nodes with lower energies; this can be done by reversing the signs of the energy values and shifting them equally to make them non-negative. Turning to the edge set of \( G \), each edge represents a conflict between a set of rotamers. The term “conflict” means that the nodes incident to the edge correspond to two different rotamers of the same residue, e.g., nodes \( (i_r, j_r) \) and \( (i_r', j_r') \). Since in SCP we need to select exactly one rotamer per residue, both nodes connected by an edge cannot be selected. From the construction, it follows that SCP is equivalent to the MWIS problem for graph \( G \). A graphical representation of such modeling is shown in Figure 1 for a system of two residues \( i \) and \( j \) that have three and two rotamers, respectively.

![Figure 1. Construction of the graphical model of a system with two residues (i) and (j) with sets of rotamers: \( \mathcal{U}_i = \{i_1, i_2, i_3\} \) and \( \mathcal{U}_j = \{j_1, j_2\} \). The optimal set of rotamers of the residues \( i \) and \( j \) can be obtained by finding the MWIS of this weighted graph. If the triple \( \{i_1, j_2, j_2\} \) is the MWIS, then the solution to the SCP problem will be rotamers \( i_k \) and \( j_i \) for residues \( i \) and \( j \) respectively.](image)

1.2. Rotamer Selection. We use the backbone-dependent rotamer library\(^5\) to derive the initial set of rotamers. In addition to the rotamers listed in the rotamer library, we generate more rotamers by considering the standard deviation value \( \sigma_1 \) (also available in the rotamer library) of the dihedral angle \( \chi_1 \) for each rotamer. Specifically, we split each rotamer of the library into three rotamers with the following first dihedral angles: \( \chi_1 - \sigma_1, \chi_1 \), and \( \chi_1 + \sigma_1 \). We keep the rest of the dihedral angles \( \{\chi_2, \chi_3, \chi_4\} \), if any, as they are, and assign to each new rotamer a probability equal to one-third of the original rotamer probability. As discussed in the Introduction, we also add one more conformer from the unbound structure of the protein to the set of rotamers. The set of rotamers gained from the expansion of the original library spans the conformational space of the side chains better and gives the algorithm a broader search space to seek the optimal side-chain configuration.

Before solving the MWIS formulation, we run a preprocessing subroutine called *rotamer refinement* that refines the set of rotamers for each residue and excludes any infeasible rotamers from the set. This subroutine consists of two phases. (i) First, we find the atomic coordinates of each rotamer and define its distance to the backbone as the nearest distance between its heavy atoms and the backbone heavy atoms. We remove from consideration rotamers whose distance to the backbone is smaller than a predefined threshold. These rotamers form steric clashes with the backbone and cannot belong to the optimal solution. (ii) Next, we implement another preprocessing step to reduce the number of rotamers for each interface residue. We use a DEE algorithm,\(^5\) which is based on a refinement of the elimination criterion known as the *Goldstein criterion*. The idea is as follows: a rotamer \( i \) of residue \( i \) can be eliminated from the set if there exists some other rotamer \( i_k \) of the same residue such that

\[
E(i) - E(i_k) + \sum_{j \neq k} \min_{i \in \mathcal{U}_j} \{E(i, j) - E(i_k, j)\} > 0
\]

for some other residue \( j \) with a set of rotamers \( \mathcal{U}_j \). This indicates a situation in which the “best” conformation that includes \( i_k \) is in \( \mathcal{U}_j \) has a higher total energy than the “worst” conformation that includes \( i_k \) is in \( \mathcal{U}_i \). In other words, for any feasible solution of SCP that includes rotamer \( i_k \) is in \( \mathcal{U}_j \), replacing \( i_k \) by \( i \) gives us a new feasible solution with lower total energy. In this case, we can eliminate \( i_k \) from \( \mathcal{U}_j \). DEE stops when it finds no more rotamers to remove. These preprocessing phases can reduce the size of \( G \) drastically, thereby speeding up the process of finding an MWIS without sacrificing optimality.

1.3. Our Distributed Algorithm To solve the MWIS. The MWIS problem is NP-hard. We have developed a two-phase algorithm\(^24,27,28\) to find effective solutions: the first phase solves a relaxation of MWIS, and the second phase leverages the relaxed solution to construct an effective MWIS feasible solution. This feasible solution indicates which rotamer to pick for each interface residue.

In the first phase, we employ a stronger relaxation than the standard linear programming relaxation of MWIS. In particular, we introduce constraints on the cliques of the graph that are redundant but make the linear programming relaxation of the integer programming problem tighter. We have developed a gradient projection (GP) method (see ref 24) for solving the (linear programming) dual problem of this relaxation. Our algorithm involves message passing only among adjacent nodes of the graph and uses local information. This message-passing approach allows us to solve the problem in a distributed fashion using multiple processors. As we discussed earlier, benefiting from a multiprocessor architecture can be useful because of the many and large problem instances one has to tackle in the course of docking two proteins.

Since we solve a relaxation of MWIS in the first phase of the algorithm, we have to round up the solution to a feasible solution for the original problem. To that end, the second phase of the algorithm consists of a greedy estimation procedure that constructs a feasible MWIS solution based on the solution of the relaxation. Our estimation phase is also distributed and works on the basis of passing messages between the nodes of the problem graph.

Our two-phase algorithm produces a near-optimal solution to the problem and has several parameters (e.g., accuracy of the relaxation phase) that can be tuned to trade-off the accuracy of the final solution against the running time. This is useful in the
context of our docking application; for instance, in early phases of the docking protocol a less expensive and less accurate version can be used, and the accuracy can be tightened in the final stages of docking.

1.4. Partitioning the Interface Residues. The number of nodes in the graph $G$ increases quadratically with the number of interface residues. This can lead to a very large $G$ that is computationally expensive to handle. To reduce the size of the graphs we have to partition, we partition the set of interface residues into non-overlapping clusters on the basis of their interaction energy values. We first compute the interaction energy between each pair of residues in the set. If the interaction energy between a pair of residues is greater than a small threshold $\varepsilon$, we say that those two residues interact. If, however, two residues are too far away, there would be no interaction energy between them. We consider a subset of the residues as a cluster if interactions involving these residues are exclusively confined within the cluster. In this sense, the clusters are non-overlapping, and the union of clusters forms the whole interface set.

After partitioning the interface set into several clusters, we solve the SCP problem using the MWIS formulation on each cluster separately and in parallel. It should be noted that since the clusters do not energetically interact, breaking the main SCP problem into smaller subproblems does not change the overall solution, yet it speeds up the procedure notably.

On the basis of our statistical analysis over a docking benchmark set composed of tens of receptor–ligand complexes with thousands of conformations each, we conclude that a significant portion of the clusters contain only two residues. Even though our algorithm is an approximate method in general, it does find the exact solution for clusters of size 2 because of the special structure of the MWIS graph.24 For larger clusters, it can find an effective feasible solution which is near-optimal.24

1.5. Off-Grid Minimization with an Optional SCP Step. To study the role of SCP in protein docking, we have incorporated our side-chain packing approach into off-grid refinement, where it is typically used.

We have implemented a standard Monte Carlo minimization (MCM)-based off-grid refinement protocol, which is used in many refinement approaches.19,29,30 Off-grid refinement seeks the lowest-energy configuration in the vicinity of the initial conformation. The protocol we use performs iterations, each consisting of four main steps. (I) In step 1, the ligand position and orientation with respect to the receptor are slightly back into contact. (II) In step 2, we slide the proteins back into contact. (III) In step 3, SCP is applied. This step is optional; to assess SCP’s role in refinement, we will show results for runs without SCP that leave side chains in their unbound positions, runs with SCP in which the whole interface is rearranged using the algorithm we presented and the standard rotamer library, and runs with SCP using the standard rotamer library to which we add the unbound side-chain conformers. (IV) The final step in each iteration of the refinement protocol locally minimizes the energy of the resulting complex using a rigid-body minimization algorithm22 and allowing the side chains to slightly move to off-rotamer positions in order to alleviate potential steric clashes. After these four steps are performed, we have a new candidate complex. We decide either to accept or reject this candidate based on the Metropolis criterion, namely, if the total energy of the candidate complex is lower than the energy of the complex in the beginning of the iteration, we accept the candidate; otherwise, we accept the candidate with a probability that is inversely exponentially proportional to the energy difference. If the candidate is accepted, then it becomes the complex used to initialize the next iteration; otherwise, we discard the candidate complex and start the next iteration with exactly the same complex we had at the beginning of the current iteration.

1.6. Refinement Set Generation. To study the effect of off-grid minimization with SCP, we generated sets of near-native structures using a soft rigid-body approach.26 For our study set, we used enzyme–inhibitor and other types of complexes from the Protein–Protein Docking Benchmark.26 The following steps were performed for each of the enzyme–inhibitor complexes: (1) We systematically sampled mutual receptor–ligand orientations using a fast Fourier transform (FFT)-based approach (PIPER20) and obtained 70 000 lowest-energy structures. (2) The 1000 lowest-scoring structures were clustered31 using a greedy algorithm, and the clusters were ranked on the basis of their size (a larger cluster corresponds to a higher rank). (3) The highest-ranking cluster whose center had a root-mean-square deviation (RMSD) of all atom positions under 10 Å from the native was selected for refinement. The top 1000 lowest-energy structures out of the 70 000 generated in step 1 that were also within 12 Å RMSD from the selected cluster center were selected as the refinement set. A similar protocol was used for other types of complexes, with the exception that clusters were chosen from three FFT sampling runs, each with different weights in the energy function. Details are described in Kozakov et al.32 We used all of the complexes in the Protein–Protein Docking Benchmark26 for which PIPER20 was able to produce at least 50 solutions within 5 Å RMSD to the native. Our study set consisted of 35 cases of enzyme–inhibitor complexes and 34 cases of other types of protein complexes.

1.7. Energy Function. For the energy function terms referenced thus far, we use a state-of-the-art, high-accuracy docking energy potential that combines force-field and knowledge-based energy terms.19,29,33 In particular, interaction energies are computed as a weighted sum (the $w$'s are the corresponding weights):

$$E = w_{VDW}E_{VDW} + w_{SOL}E_{SOL} + w_{DARS}E_{DARS} + w_{COUL}E_{COUL} + w_{HB}E_{HB} + w_{RP}E_{RP}$$

where $E_{VDW}$ is the Lennard-Jones potential, $E_{SOL}$ is an implicit-solvation term, $E_{COUL}$ is the Coulomb potential, $E_{HB}$ is a knowledge-based hydrogen-bonding term, $E_{RP}$ is a statistical energy term associated with a specific selection of rotamers from the backbone-dependent rotamer library, and $E_{DARS}$ is a structure-based intermolecular potential derived from the nonredundant database of native protein–protein complexes using a novel decoys as reference state (DARS)35 reference set. The DARS reference set is formed by generating a large decoy set of docked conformations based only on a shape-complementarity scoring function; we compute the potential by observing the frequency of interactions in these decoys.

In order to calculate $E_{RP}$ we need to know the probability $p_{\alpha}$ of each rotamer $\alpha$, which can be approximated by the fraction of time that amino acid residue $i$ is found in rotamer $\alpha$ in a large data set. These probabilities are given in the rotamer library. The statistical energy value of such a rotamer is given by $-\log(p_{\alpha})/p_{\alpha}$ and thus, the more frequent a rotamer, the lower the energy assigned to it. The weights in the energy function are chosen according to the selections in Gray et al.19
2. RESULTS AND DISCUSSION

2.1. Accuracy of Side-Chain Positioning.

We use SCP in predicting the bound-state side-chain conformations of an unbound receptor−ligand complex. To assess the accuracy of our algorithm, we test it against a benchmark set consisting of 48 unbound enzyme−inhibitor (EI) complexes and 67 other-type (OT) protein complexes and compare our predictions to the native-state conformers observed using experimental techniques. We also compare the accuracy of our algorithm with that of the SCWRL4.0 package,23 which, as we commented earlier, is the state-of-the-art side-chain prediction tool. We refer the reader to our earlier discussion of the differences between SCWRL4.0 and our approach and on how the results should be interpreted.

Figure 2. Comparison of SCWRL4.0 and MWIS to native complexes. We compare the performance of side-chain positioning of three modes: (i) *scwrl* shows the prediction accuracy of SCWRL4.0; (ii) *MWIS−UB* denotes the performance of our MWIS algorithm without considering the unbound conformers; and (iii) *MWIS + UB* indicates the performance of MWIS including the unbound conformers. Moreover, we report the numbers of interface residues whose predicted conformations are considered accurate on the basis of the $\chi^1$ and $\chi^1+2$ criteria. Also, #res indicates the number of interface residues for each system.
We use standard criteria to evaluate our side-chain prediction approach.\textsuperscript{9,17} The first criterion, called $\chi_1$, is based on the difference between the first dihedral angle ($\chi_1$) of the sets of residues in the predicted structure and the native structure. The second criterion, called $\chi_{1,2}$, is based on the differences between the first two dihedral angles ($\chi_1$ and $\chi_2$) of the residues in the predicted structure and the corresponding dihedral angles in the native structure. For the $\chi_1$ criterion, an accurate prediction of a residue occurs when the $\chi_1$ angle of a predicted residue is within 40° of its native-state value. For the $\chi_{1,2}$ criterion, the prediction of a residue is considered accurate when both the $\chi_1$ and $\chi_2$ angles of the predicted structure are within 40° of their native-state values. Although the 40° deviation may appear to be large, it is the size of the error considered in standard criteria used for evaluating side-chain prediction algorithms. In addition, as

<table>
<thead>
<tr>
<th>Enzyme-Inhibitor Benchmark</th>
<th>Others Benchmark</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PDB</strong></td>
<td><strong>Input</strong></td>
</tr>
<tr>
<td>2b42</td>
<td>106</td>
</tr>
<tr>
<td>1udi</td>
<td>163</td>
</tr>
<tr>
<td>1n8o</td>
<td>220</td>
</tr>
<tr>
<td>1f34</td>
<td>104</td>
</tr>
<tr>
<td>2abz</td>
<td>106</td>
</tr>
<tr>
<td>1avx</td>
<td>225</td>
</tr>
<tr>
<td>1cgi</td>
<td>129</td>
</tr>
<tr>
<td>1bvn</td>
<td>210</td>
</tr>
<tr>
<td>1mah</td>
<td>271</td>
</tr>
<tr>
<td>2jot</td>
<td>44</td>
</tr>
<tr>
<td>2sic</td>
<td>338</td>
</tr>
<tr>
<td>2sni</td>
<td>327</td>
</tr>
<tr>
<td>1dfj</td>
<td>214</td>
</tr>
<tr>
<td>1ezu</td>
<td>119</td>
</tr>
<tr>
<td>1fle</td>
<td>128</td>
</tr>
<tr>
<td>1jtg</td>
<td>344</td>
</tr>
<tr>
<td>1oc0</td>
<td>39</td>
</tr>
<tr>
<td>1ro</td>
<td>382</td>
</tr>
<tr>
<td>1ay7</td>
<td>73</td>
</tr>
<tr>
<td>1oyv</td>
<td>228</td>
</tr>
<tr>
<td>1g1l</td>
<td>169</td>
</tr>
<tr>
<td>1ijk</td>
<td>209</td>
</tr>
<tr>
<td>2mta</td>
<td>231</td>
</tr>
<tr>
<td>2uuy</td>
<td>197</td>
</tr>
<tr>
<td>1eaw</td>
<td>222</td>
</tr>
<tr>
<td>1tmq</td>
<td>347</td>
</tr>
<tr>
<td>7cei</td>
<td>241</td>
</tr>
<tr>
<td>1e6e</td>
<td>160</td>
</tr>
<tr>
<td>1acq</td>
<td>89</td>
</tr>
<tr>
<td>1ewy</td>
<td>62</td>
</tr>
<tr>
<td>1yvb</td>
<td>376</td>
</tr>
<tr>
<td>2pcc</td>
<td>62</td>
</tr>
<tr>
<td>2oul</td>
<td>333</td>
</tr>
<tr>
<td>1ppe</td>
<td>518</td>
</tr>
<tr>
<td>4cpa</td>
<td>333</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>7319</td>
</tr>
</tbody>
</table>
already mentioned, side-chain prediction generally requires relatively limited accuracy in applications to docking.

To show the effect of including the unbound conformers of the side chains in side-chain prediction, we consider two different cases: (i) solving the SCP problem without including the unbound conformers (−UB) and (ii) solving the SCP problem with unbound conformers (UB). We compare the overall packing results in the absence and in the presence of the side chains’ unbound conformations to show how the inclusion of the unbound conformers in the rotamer set can affect the side-chain prediction results. We also provide the SCWRL4.0 predictions to determine the accuracy of our algorithm in comparison with that method. As mentioned in the Introduction, in SCWRL4.0 the unbound side chains are not considered as possible rotameric states of the residues.

For each complex, we run each algorithm over exactly the same interface set of residues obtained from the unbound structure of the complex. We report the numbers of interface residues whose predicted conformations are considered accurate on the basis of the $\chi_1$ and $\chi_{1,2}$ criteria.

The detailed results are presented in Figure 2. We provide the side-chain prediction accuracies of the aforementioned methods for the two different types of protein structures (EI and OT) separately. In the last row of the table, we compare the performances of these methods over the full benchmark by calculating the percentage of all interface residues that are predicted within the accuracy range. A couple of observations are in order. First, our method produces slightly less accurate results compared with SCWRL4.0 when the unbound side-chain conformations are not included in the rotamer set. Essentially, this is the small price to pay for an approximate algorithm (vs the exact approach of SCWRL4.0), which, however, has a number of characteristics that are useful in docking applications (distributed, scalable, tunable speed–accuracy trade-off). The second, and more important, observation is that the inclusion of the unbound rotamers improves the accuracy of the predictions. This shows that the unbound structure of proteins carries substantial information about their native docked structure, and hence, considering them in the side-chain prediction method is of great importance in docking applications.
2.2. SCP as a Protein Docking Component. As discussed earlier, our main motivation for this work is to apply SCP in protein docking refinement protocols. Next, we analyze the effectiveness of our SCP algorithm when it is used as a component of our protein docking refinement procedure. We report on the impact of SCP in the overall performance of the off-grid optimization refinement procedure and, more specifically, in increasing the number of near-native predictions.

For this purpose, we refine the PIPER outputs using three different modes of the off-grid optimization refinement protocol outlined in section 1.5: (i) REF−SCP, in which the conformations are refined without employing the SCP algorithm, and (ii) REF+SCP−UNB and (iii) REF+SCP+UNB, in which the conformations are refined by the off-grid optimization procedure that uses SCP as a component of the energy evaluation without and with, respectively, consideration of the unbound side-chain conformers in SCP.

For each mode, we calculate the RMSD of each predicted conformation in the set from the native structure. A prediction is considered “accurate” when this RMSD is below 5 Å from the native structure. For 35 EI and 34 OT complexes, the tables in Figure 3 (left for EI, right for OT) as well as Figures 4 (for EI) and 5 (for OT) report the numbers of accurate predictions in the refinement sets (see section 1.5). The first column of each table in Figure 3 lists the PDB codes of the complexes. The second column reports the numbers of accurate refined conformations (within 5 Å RMSD from the native) out of the top 1000 PIPER outputs in the refinement set. These conformations are the input to the refinement stage. The three following columns specify the numbers of accurate refined conformations generated by the three different modes of off-grid optimization described above, denoted as R-SP, R+SP-UB, and R+SP+UB, respectively. The last two rows of each table report the total numbers of accurate predictions over all complexes and the percentage improvements over PIPER. The latter metric is computed by averaging over all of the complexes the per-complex percentage improvement, and it reflects a view of performance that is not biased by the number of accurate complexes for each refinement set. The results show that adding side-chain packing and including the unbound conformers can improve the overall refinement performance by increasing the number of accurate predictions.

In Figures 4 and 5 for the EI and OT protein benchmarks, respectively, we plot three curves that indicate the increase/decrease in the number of PIPER accurate conformations using the three settings of the refinement procedure described above: REF−SCP (green), REF+SCP−UNB (blue), and REF+SCP+UNB (red). As an example, consider the EI complex 1yvb, which has 376 accurate conformations generated by PIPER (as shown in the left table in Figure 3) and is the first protein shown in Figure 4. The green, blue, and red data points show values of 117, 93, and 117 respectively, for 1yvb, reflecting the respective gains of the three modes over the PIPER result. As shown in Figures 4 and 5, in most cases the red curve is superior to the other two curves, indicating that the REF+SCP+UNB method works better than the other two methods. It follows that the use of SCP including the unbound conformers increases the number of near-native predictions and improves the refinement performance.

2.3. Effect of Parallelization on Running Time. To validate the effect of parallelization on improving the running time of our SCP algorithm, we study how the average running time over the benchmark set of 48 EI and 67 OT unbound proteins (listed in the table in Figure 2) changes as we increase the number of processors. To get a better sense of the improvement, we categorize the benchmark set on the basis of the size of the interface into two subsets labeled as “Large” and “Small”. The size of the interface refers to the number of interface residues of each protein complex and is reported in the third column of the table in Figure 2. The size of the MWIS optimization problem associated with our SCP algorithm increases quadratically with the size of the interface. Therefore, the parallel approach is of great importance when it comes to large problem instances. In our setting, the protein complexes with interface size greater than 20 are considered in the Large category, and the ones whose interface size is in the range of 20 or less are considered in the Small category. We also evaluate the running time over the entire benchmark (labeled as “All” in Figure 6).
Our results were obtained on a desktop workstation with an Intel Core i7-950 processor (8 M cache, 3.06 GHz, 4.80 GT/s Intel QPI) and 4 GB of RAM. We report the speedup values in Figure 6 for the cases of two-, four-, six-, and eight-processor runs of the algorithm with respect to the single-processor running time. The speedup value of the n-processor setting is computed by dividing the average running time of the algorithm when using n processors by the average running time of the single-processor run. Figure 6 shows that using the multiprocessor architecture is generally beneficial in speeding up the packing process, especially for large systems.

3. CONCLUSION

We have considered the problem of side-chain packing in the process of protein–protein docking. Specifically, this is the problem of appropriately positioning side chains in the interface region between the two proteins. The problem exhibits significant special structure that makes it notably different from the side-chain prediction problem extensively explored in the context of protein folding. These differences motivated our development of a new approximate but fully distributed approach.

We have tested this approach against benchmark sets of enzyme–inhibitor and other types of complexes. We have found that the incorporation of side-chain packing in each iteration of protein docking refinement protocols facilitates the docking process and leads to improved performance. We have also established that the inclusion of the unbound conformers as an option in the side-chain packing optimization improves the side-chain positioning accuracy and docking performance. The latter can potentially motivate the adaptation of alternative side-chain prediction approaches. Our side-chain packing software is available under an open source license at https://github.com/StructuralBioinformaticsLab/sidechain_example.

■ ASSOCIATED CONTENT

3 Supporting Information

Detailed description of our methods, including our ILP-based formulation of SCP and our two-phase distributed algorithm to solve the problem (Figures S1 and S2), running time analysis for different numbers of processors (Figure S3), more details on clustering of the interface residues (Figure S4), our choice of energy terms and their corresponding weights (Figure S5), and discussion of the optimality and convergence rate of our two different formulations of the SCP problem. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: yannisp@bu.edu.
*E-mail: midas@bu.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This research was partially supported by the NIH/NIGMS (Grants GM093147 and GM061867), the NSF (Grants CNS-1239021, DBI-1147082, and IIS-1237022), the ARO (Grants W911NF-11-0-0227 and W911NF-12-1-0390), and the ONR (Grant N00014-10-1-0952).

■ REFERENCES


