JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Prediction of Long Loops with Embedded Secondary Structure Using the Protein Local Optimization Program

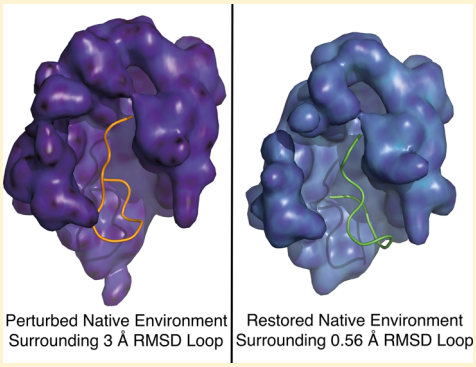
Edward B. Miller,[†] Colleen S. Murrett,[†] Kai Zhu,[‡] Suwen Zhao,[§] Dahlia A. Goldfeld,[†] Joseph H. Bylund,[†] and Richard A. Friesner*[,†]

[†]Department of Chemistry, Columbia University, New York, New York

[‡]Schrödinger, Inc., New York, New York

[§]Department of Pharmaceutical Chemistry, University of California, San Francisco, California

ABSTRACT: Robust homology modeling to atomic-level accuracy requires in the general case successful prediction of protein loops containing small segments of secondary structure. Further, as loop prediction advances to success with larger loops, the exclusion of loops containing secondary structure becomes awkward. Here, we extend the applicability of the Protein Local Optimization Program (PLOP) to loops up to 17 residues in length that contain either helical or hairpin segments. In general, PLOP hierarchically samples conformational space and ranks candidate loops with a high-quality molecular mechanics force field. For loops identified to possess α-helical segments, we employ an alternative dihedral library composed of (ϕ, ψ) angles commonly found in helices. The alternative library is searched over a user-specified range of residues that defines the helical bounds. The source of these helical bounds can be from popular secondary structure prediction software or from analysis of past loop predictions where a propensity to form a helix is observed. Due to the maturity of our energy model, the lowest energy loop across all experiments can be selected with an accuracy of sub-Ångström RMSD in 80% of the cases, 1.0 to 1.5 Å RMSD in 14% of the cases, and poorer than 1.5 Å RMSD in 6% of the cases. The effectiveness of our current methods in predicting hairpin-containing loops is explored with hairpins up to 13 residues in length and again reaching an accuracy of sub-Ångström RMSD in 83% of the cases, 1.0 to 1.5 Å RMSD in 10% of the cases, and poorer than 1.5 Å RMSD in 7% of the cases. Finally, we explore the effect of an imprecise surrounding environment, in which side chains, but not the backbone, are initially in perturbed geometries. In these cases, loops perturbed to 3 Å RMSD from the native environment were restored to their native conformation with sub-Ångström RMSD.



INTRODUCTION

Continual advances in loop prediction have yielded accurate modeling from 12-residue loops[1] up to loops as long as 20 residues.[2] These methods have managed to achieve near-atomic accuracy performing loop prediction in the presence of the crystal structure environment—a necessary but not sufficient condition for realistic homology modeling.

Historically, loop prediction was first approached analytically by Go and Scheraga[3] in 1970. Demonstrated was the ability to predict, by solving a set of equations, the conformation of peptide fragments containing up to six rotatable torsions. This analytical method was updated 21 years later by Palmer and Scheraga.[4] Here, the authors relax constraints on the original formulation by permitting each residue in the loop to adopt independent bond lengths or bond angles. However, the analytical method still remained limited to six torsion angles—three residues assuming the backbone ω torsion remained fixed. To accommodate larger loops, Palmer and Scheraga extend the method by permitting additional torsions, beyond the six that can be analytically determined, so long as they are independently set prior to the calculations. Thus, their method requires that the algorithm be repeated numerous times over a conformational search of these additional independent torsions. Hence, for larger loops combinatorics must be considered.

Moult and James in 1986 proposed one of the first combinatorial searches through a discrete set of torsions.[5] Here, the authors described the use of a systematic search through torsion angles obtained from a Ramachandran plot. For loops as small as five residues, their method yields about 10^{10} conformations, already an intractable number. To cope with the combinatorial explosion, the authors employ the use of rules and filters to restrict and prune the number of conformations to a manageable subset before performing more expensive scoring. Loops are scored using a simple pairwise electrostatic energy function and a surface area based hydrophobic term.

Later methods vary in both the sampling rules and scoring function. Bruccoleri and Karplus in 1987 released CONGEN, from which our algorithm draws some similarity.[1,6] There, the authors use the CHARMM energy function[7] to score loops. In 1992, Bassolino-Kilmas and Bruccoleri advance CONGEN to

Received: December 10, 2012
Published: February 7, 2013

   dx.doi.org/10.1021/ct301083q | J. Chem. Theory Comput. 2013, 9, 1846−1864

permit directed loop buildup which takes into account information from partially built structures.[8] In 2003, DePristo et al.[9] and de Bakker et al.[10] used the AMBER force field[11] and generalized Born solvation model[12] for scoring loops. Loop buildup is performed using, among other modifications, a fine-grained torsion library that is residue-specific. Like CONGEN, our work draws similarities to this last method.[1] We note that this historical review is not exhaustive but is intended to highlight the origins of loop prediction as it relates to this work.

In general, the use of combinational exploration of torsion space for loop buildup has within it two subproblems, sampling problems where coping with the combinatorics of loop buildup requires the development of clever pruning strategies and energy problems where the minimization, scoring, and ranking of the resultant loops must be computationally affordable yet accurate enough to identify the best conformation among those produced.

Throughout the literature, the functional definition of a loop has been a local segment of the protein that is free of secondary structure other than, perhaps, three-residue 3_{10} helices but lies between large, likely well-conserved, secondary structure elements.[2b,13] Indeed, initial homology models are often constructed on the assumption that secondary structure elements are conserved between the template and the target.[14] However, this loop definition has not always been strictly followed. Notable cases of loops containing secondary structure are the ECL2 loops of the human β2-adrenergic receptor[15] and turkey β1-adrenergic receptor,[16] both G-protein coupled receptors (GPCRs). These loops are actually loop−helix−loops (LHLs) containing an eight-residue α-helix. Spinach rubisco is another example. The active site is composed of a highly conserved α/β barrel. Lying between each α/β pair are loops, of which loop 5 contains a five-residue α-helix and two residues that form part of βF, a β-strand external to the active site, and loop 8 which contains a four-residue α-helix.[17]

Recent attempts have been made to model the GPCR LHLs and have been met with significant success reaching an accuracy as high as a 1.59 Å RMSD.[18] As the method we provide here exists along a continuum of protein structure prediction methods, one that shares significant applicability to secondary structure-free loops, we retain the loose definition of the word "loops," and here refer to loops as a region of the protein that may contain secondary structure but is flanked by even larger secondary structure elements. Presented in greater detail below is a precise definition, which was strictly enforced, to select a set of test cases.

Throughout the literature, predictions performed on loops containing secondary structure are scant. Zhu et al. presented a refinement protocol that addresses loop−helix−loops and loop−hairpin−loops, referred to more generally as protein segments in the paper, using a knowledge-based potential.[19] What is explored is the refinement of these segments, rather than the prediction of the segments *de novo*. Consequently, the success of their refinement is dependent on the difficulty of the initial structure. For hairpins and loop−helix−loops, close to 70% of their refinements yield predictions with an RMSD of 2.0 Å or better. In these cases, the secondary structure elements are kept fixed with their native torsions and moved as a rigid body. However, as our method discussed in this paper is independent of the conformation of the input loop (although it is dependent on the conformation of the surrounding environment), results cannot be directly compared.

Alternatively, Rohl et al. described *de novo* loop construction using the Rosetta algorithm.[20] Included in their test set are predictions of 10 loops, referred to as structurally variable regions, of 13 to 34 residues in length. These predictions were done in the crystal structure environment and do include loops containing secondary structure. Although some of the members of their test set include, for example, loop−helix−loops, only 10 cases were done in the context of the native protein—too few to permit comparisons between our method without relying on anecdotal information. Instead, the authors concentrate on the more ambitious task of loop prediction in an unrefined homology model. Finally, we note in a previous study our attempt to address the challenges of helix packing.[21] In Li et al., we explored placement of a helix in a loop−helix−loop but treated the helix as a rigid body. Although the method relies on prior knowledge of the presence of a helix, for large helices, this is not unreasonable, as is stated above, because significant segments of secondary structure tend to be conserved across homologous structures. Indeed, the smallest helix considered in this previous study was eight-residues.

To the best of our knowledge, no studies have been performed that systematically address the challenges of *de novo* prediction of loops containing secondary structure, particularly for cases when *a priori* knowledge about the presence of small secondary structure is noisy at best. As loop prediction matures to accurate prediction of larger and larger loops, it becomes awkward to exclude cases of secondary structure-embedded loops. In this work, we propose a method to predict long loops containing possibly multiple helices or a hairpin. Our initial test set is composed of loops containing between 8 and 17 residues. The secondary structure length explored ranges from 3 to 13 residues, although in principle, prediction of loops containing larger secondary structure segments remains tractable.

For loop−helix−loops, we constructed a separate dihedral library taken from a nonredundant set of high-resolution Protein Data Bank[22] structures containing α-helices. The user is required to specify which residues this helical dihedral library is to be applied to, termed the helical bounds. Results with exact helical bounds taken from the crystal structure were used as an initial validation. More relevant to actual structure prediction and refinement, we then concentrated on accurate loop prediction using helical bounds supplied by either sequence-based secondary structure prediction algorithms or previous loop predictions performed without the use of our helical dihedral library. That is, in many cases, nascent helices were predicted without supplying any expectation of a helix. This suggested a propensity for this loop to include a helix and allow us to repredict the loop using our helical dihedral library. Throughout all sampling methods explored, what remains crucial is that purely from our energy model, we are able to pick out the loop with the lowest or near lowest RMSD relative to the native structure. Finally, for loops containing either helices or hairpins, we explored loop reprediction in a perturbed local environment, similar to an environment encountered in full homology models, although without deviations of the backbone from the native structure, and established success in restoring the native loop conformation. The results are generally satisfactory with loop−helix−loop predictions from imprecise helical bounds routinely reaching sub-Ångström RMSD and hairpin predictions reaching similar atomic accuracy.

MATERIALS AND METHODS

Selection of Test Cases. All PDB structures that were available as of August 30, 2010 were searched. Global criteria were used to select structures that satisfy the following properties:

1. A sequence identity between any two proteins must be ≤50%.
2. Only crystal structures were selected.
3. The resolution of the crystal structure must be <2.0 Å.
4. Structures reporting only C_α coordinates were excluded.
5. A minimum R_{work} of 0.25 was enforced.
6. The pH of the crystal structure was restricted to lie between 6.0 and 8.0.

The exclusion of proteins due to sequence identity was performed using the PISCES Web server[23] (http://dunbrack.fccc.edu/PISCES.php). Loops were selected using a local criterion that satisfies the following:

1. The average temperature factor of atoms within the loop must be ≤35.
2. The real-space *R*-factor[24] of any residues in a selected target loop must not be greater than 0.200.
3. All residues within the loop or interacting with any residues within the loop must be free of alternate conformations.
4. To reduce effects due to loop−ligand interactions, the minimum distance between any loop atom and any atom as part of a neutral ligand must be >4 Å. For charged ligands, this cutoff is increased to 6.5 Å.

The real-space *R*-factor was found by reference to the Uppsala Electron Density Server[25] (http://eds.bmc.uu.se/eds/). The above criteria are similar to what was used to create test sets in our past publications.[2,26]

Identification of Secondary Structure-Containing Loops. In our most recent publications, loops were defined as being a segment of the protein absent of secondary structure.[2b,26] To identify loops containing secondary structure, an alternative definition was proposed. For loops containing secondary structure, the loop must be bounded by a span of secondary structure larger than the greatest contiguous span of secondary structure within the loop. For example, if a loop contained, at most, a six-residue α-helix, then flanking the loop must be residues that are a part of a secondary structure element of at least seven residues in length. Furthermore, the first and last residue of a loop must also not display secondary structure. Assignment of secondary structure on a per residue basis was done using the DSSP program.[27]

A loop was defined as a loop−helix−loop only if there were no other types of secondary structure present other than turns and helices (including 3_{10} and α-helices); i.e., any loop containing both β-bridges and helical residues was discarded from this study. A total of 35 loop−helix−loop regions were identified which were either 16 or 17 residues in length in all. This loop length was chosen to select cases that were considered sufficiently difficult to demonstrate the efficacy of our approach. In our previous publication, loops free of secondary structure were successfully predicted up to 17 residues in length.[2b]

For loops containing β-hairpins, it became necessary to distinguish between a β-hairpin and a segment that is part of a larger β-sheet. To make such a distinction, the following criteria were used:

1. The loop must contain the secondary structure pattern strand−turn−strand.
2. However, the turn residues need not be immediately adjacent to a strand residue.
3. The loop must be free of helices.
4. The strand residues comprising part of the pattern in criterion 1 must be forming backbone hydrogen bonds only to other residues within the loop.
5. The hydrogen-bonding pattern must be antiparallel.

For hairpins, requiring loops to be either 16 or 17 residues in length yielded too few test cases. Thus, a loop was accepted so long as it was not greater than 17 residues. A total of 41 cases satisfying the above hairpin criteria were identified.

Single-Loop Prediction. Single loop prediction is performed through individual runs of the Protein Local Optimization Program (PLOP). Briefly, PLOP operates through four stages: buildup, closure, clustering, and scoring. Full details can be found in Jacobson et al.[1] However, the salient features will be presented here, and the modifications of the PLOP protocol utilized in this work will be described.

Loop buildup is begun with a backbone dihedral angle library constructed from rotamers frequently observed in crystal structures. Initially, the library contained a set of dihedrals on a single amino acid basis.[1] As larger loops were explored, efficient exploration of conformational space dictated the use of a dipeptide dihedral library.[2] In this approach, a library is constructed from each of the 400 (20 × 20) possible dipeptide pairs and used in a sequence specific manner during buildup. For example, a loop containing an arginine−alanine dipeptide would explore sampling from a different rotamer library than an arginine−valine dipeptide. This implicitly treats the individual amino acid torsions as coupled.

In helices, the backbone torsions are highly coupled to form the necessary hydrogen-bonding network. It was therefore natural to extend the use of a dipeptide dihedral library to exploit coupled backbone torsions across the four residues, or greater, of an α-helix. As such, for residues considered to be helical, a separate *n*-residue α-helical library was used for loop buildup, where *n* is four or larger. The aspects of this α-helical library are discussed in greater detail below. In β-hairpins, nonlocal torsional coupling is present, and so to enforce torsional coupling during loop buildup would heavily constrain both the coupled, hydrogen-bonding residues as well as the intervening turn residues. Although such an approach may still be fruitful, we found that for β-hairpins, our previous dipeptide torsional library was effective, and so we did not explore further the use of an alternative β-hairpin library.

Loop buildup is performed simultaneously from both ends of the loop up to the C_α atom on the closure residue. In our prior publications, the closure residue was simply picked as the midpoint of the loop.[1,2b,26] For the loop−helix−loops described in this work, the closure residue, shared by both halves of a loop, cannot be permitted to bisect a helix. As is described further below, the helical library is based on the construction of entire helices, and not helical fragments. If the closure residue of the loop was a part of a helix, the helix would be split between both halves of the loop. Thus for this work, we were forced to alter the designation of the closure residue. The closure residue is initially set with the equation

$$C_{a,\text{closure}} = N_{\text{term,LHL}} + (\text{Length}_{\text{LHL}} - 1 \pm \text{Length}_{\text{Helix}})/2$$

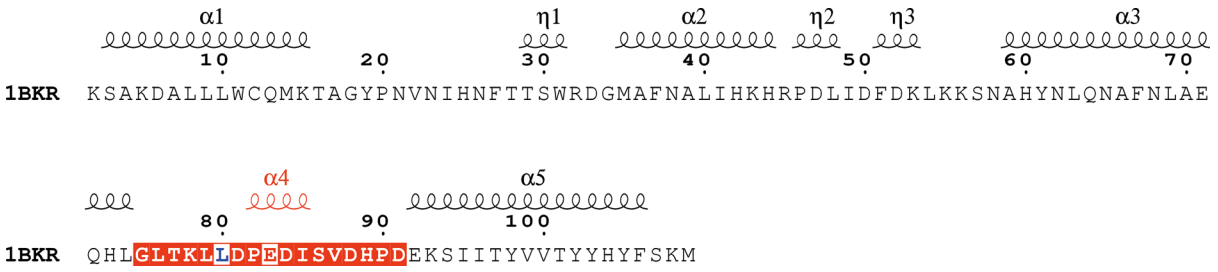


Figure 1. Loop−helix−loop predicted in PDB 1BKR. The target loop−helix−loop residues are highlighted red from residues 75−91. The helix of interest, labeled α4, spans residues 82−85. Loop prediction without the helical library would assign the closure residue to be residue 83, highlighted in white. The LHL method places the closure residue at position 80. This figure was generated using ESPript.[45]

where + is used when the C-terminus loop is the longer loop and − is used when the N-terminus loop is longer or if both flanking loops are of equal length. $N_{\text{term,LHL}}$ refers to the residue number of the N-terminus of the loop−helix−loop. Should the closure lie adjacent to the helix, the closure residue is shifted one residue farther away from the helix. This is to afford extra flexibility to the residues that precede loop closure.

Clarifying by example, consider the LHL predicted in PDB 1BKR (Figure 1). Predicted was the 17-residue loop−helix−loop from G75−D91 containing a four-residue α helix from P82 to I85. When predicting this loop without the helical library, the closure residue is at the midpoint of the LHL, residue 83, highlighted in white in Figure 1. This residue intersects the helix and so cannot serve as the closure residue when employing the helical dihedral library from segments 82−85. Application of the above equation places the closure residue adjacent to the helix at residue D81, but for further flexibility, the closure residue is assigned to be residue L80 on the N-terminus loop, two residues away from the start of the helix. As in our previous work, the Cartesian positions of the two closure C_α atoms are averaged, and the remaining atoms of the loop backbone are generated using standard geometry algorithms to close the loop.

During loop buildup, nascent loops undergo preliminary screening through the use of a parameter termed the overlap factor (ofac). The ofac is defined as the ratio of the distance between two atom centers to the sum of their atomic radii. A lower ofac cutoff allows for a higher overlap between the van der Waals radii. If during loop buildup, a backbone atom is placed with a smaller ofac than permitted by the threshold, then that candidate loop is discarded.

Three additional screens are used to reject unreasonable loops early in their construction:

1. For the current residue(s) being predicted, there must exist at least one acceptable side-chain conformation, based on sampling a 30° side-chain rotamer library.
2. The loop must not travel further than 6.32 Å away from every C_α atom in the protein. This is an empirically determined value and is meant to reject loops that fail to form contacts with the rest of the protein.
3. The distance between the latest residue predicted and the closure residue must be less than a threshold beyond which closure is not considered possible. For example, a statistical analysis of a set of >500 proteins found that the maximum C_α−C_α distance that can be spanned by four residues is 13.97 Å.

Full details of these screening methods are given in Jacobson et al.[1]

An additional screening method is also employed to enforce broad sampling of conformational space. During loop buildup via single dihedrals, all pairs of states must obey the relationship $\Delta\phi^2 + \Delta\psi^2 > R_{\text{eff}}^2$, where R_{eff} is the "effective resolution" of (ϕ, ψ) space. The effective resolution is adaptively set during loop buildup. The total number of loop candidates is constrained to lie between a minimum of 512 loops up to a maximum of 10^6 loops. This constrains the number of loop candidates to a tractable size. We achieve this by initially setting the effective resolution to a coarse value of 300° and then gradually improve the resolution to finer values down to a minimum of 5° (the resolution limit of the dihedral library). For loop buildup using the dipeptide dihedral library, the effective resolution relationship becomes:

$$\Delta\phi_1^2 + \Delta\psi_1^2 + \Delta\omega^2 + \Delta\phi_2^2 + \Delta\psi_2^2 > R_{\text{eff}}^2$$

Loop buildup using the helical dihedral library did not utilize any effective resolution relationship. Principally, this was because the size of the helical dihedral library is significantly smaller than the single peptide or dipeptide dihedral library. Due to a "lever effect," a small change in the dihedrals at one end of a helix can significantly alter the coordinates of the opposite end of the helix. This effect becomes more dramatic for larger helices. To exclude what few candidate loops are produced during buildup because of a resolution cutoff would be to ignore this lever effect. Greater detail about the construction and composition of the helix dihedral library is presented below.

To prevent expensive optimization of similar loop candidates, the k-means clustering algorithm[28] is employed, and only one representative loop per cluster is passed onto side chain sampling and optimization. The number of clusters is set to be 4 times the number of residues in a loop, excluding residues initially flagged as helical during input to loop prediction, up to a preset maximum of 50 clusters. The number of clusters determines the number of representative loops passed onto side chain sampling/loop optimization and is empirically set to balance the conformational space that must be accurately scored against computational expense. Since the entire helix is constructed as a whole from the helical library, it would seem awkward to count the helical residues the same as the nonhelical ones, and so helical residues are excluded when determining the number of clusters to optimize. For the loops described in this paper, this often had little consequence. For a 17-residue loop with a four-residue helix, the maximum number of clusters, set at 50, is reached. The most common helical size was four residues (see Figure 2, below). For a 16-residue loop with a four-residue helix, the number of clusters is 48. Only for the few cases, such as PDB 2JA2, where a 16-residue loop

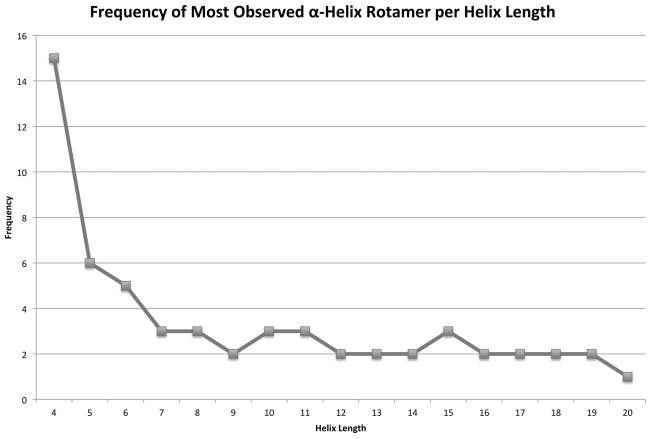


Figure 2. Plot of the greatest frequency observed of an α-helix rotamer per helix length. After a six residue α-helix, rotamers were only observed no more frequently than three times.

contains an eight-residue helix were the number of clusters, set to 32, significantly different from the maximum value of 50. These cases are the exception, and as is described later, the results from these cases, despite the reduced number of clusters, were excellent.

Side chain sampling is performed using a 10°-resolution rotamer library constructed by Xiang and Honig.[29] The algorithm for side-chain optimization works by initially placing side-chains in a random rotamer state onto the backbone. Self-consistent optimization is then performed where all side chains but one are held fixed while the free side chain is minimized. With the exception of loop prediction in a perturbed native environment, the default of one round of side-chain randomization per entire loop minimization was found to be sufficient. When considering perturbed native environments, where the surrounding side chains are included in refinement, additional rounds of side-chain randomization/self-consistent optimization are performed separately to compare to predictions done without this extra sampling. The lowest energy side-chain rotamers are selected across any additional rounds of side-chain randomization. After self-consistent side chain rotamers are selected, the complete loop, with both side chains and backbone atoms, is then energy minimized. Full details about side-chain optimization are described in our past publications.[1,30]

Scoring is done using an augmented form of the Optimized Potential for Liquid Simulations (OPLS) all-atom force field.[31] For solvation, an implicit model was used based on the surface generalized Born model as described initially in Ghosh et al.[32] A variable dielectric approach is used to treat polarization from protein side chains.[33] Additional corrections were added to the energy model to better account for $\pi-\pi$ interactions, self-contact interactions, and hydrophobic interactions. The force field, solvation model, and all correction terms are discussed in greater detail in Li et al.[2a] The protonation state of all titratable residues was set using the independent cluster decomposition algorithm of Li et al.[34]

Since we evaluate our loop prediction method against published crystal structures, crystal-packing effects were taken into consideration. The crystallographic asymmetric unit, as well as all atoms from other surrounding unit cells that are within 30 Å, are included in the simulation. The coordinates of all copies of the asymmetric unit are updated for steric clash checking and energy calculation throughout the course of the loop prediction.

Construction of the Helical Dihedral Library. As a natural extension to the dipeptide dihedral library, we constructed a helical dihedral library to exploit the coupled torsions present in an α-helix. An initial set of PDB structures was obtained from the precompiled culled PDB lists from the PISCES Web server.[23] The parameters used to cull the structures were a percentage identity cutoff of 30%, a resolution cutoff of 2.0 Å or better, and an *R*-factor cutoff of 0.25. The PDB list was obtained on October 16, 2007. The list contained 3900 PDB structures. Using an internal PLOP implementation of the DSSP algorithm,[27] α-helices were identified with lengths ranging from four to 24 residues. The (ϕ, ψ) angles for the helical residues were extracted. We ignored values for the ω dihedral and instead used 180° during loop buildup. Deviations from the *trans* conformation are permitted during loop minimization. The dihedral angles were rounded and binned to a 10° resolution. The frequency of each binned helical rotamer was counted per helix length. In structures containing homomultimeric proteins, the helix was only counted once. We did not include helical fragments from larger helices as part of the set of dihedrals for smaller helices. That is, the torsions in a six-residue α-helix are kept separate from the torsions in a four-residue α-helix. This adherence to the use of only complete helices was rigidly followed throughout loop prediction. Specifically, loop buildup from both ends of the loop was done such that the helix was not divided between both loop halves. When predicting a subsection of a loop, as is done during hierarchical loop prediction, in any instance where a subsection of the helix was predicted, the dipeptide dihedral library from Zhao et al.[2b] was used instead.

Initially, we sought to include all rotamers observed with a frequency above a set cutoff. However, this approach was problematic. Despite the large number of PDB structures, for large helices, many rotamer sets do not appear more than once. For example, in a nine-residue helix containing 18 dihedral angles (ϕ, ψ), a single 10° difference in any (ϕ, ψ) angle would place that rotamer in a new bin. For helices of this length, a helical rotamer was not observed with a frequency greater than twice (Figure 2). Beyond a six-residue α-helix, rotamers were observed no more frequently than three times. We therefore felt that there was no suitable frequency cutoff to use. Ultimately, we arbitrarily decided to set the library to contain $2 \times \text{Length}_{\text{Helix}}$ rotamers and populated the library with the most frequent rotamers that conformed closest to ideal helical dihedral angles of $(\phi, \psi) = (-60°, -40°)$. Any nonideality in a helix was left to be predicted during loop minimization and the multiple stages of loop refinement described in the following section.

Hierarchical Loop Prediction. Hierarchical loop prediction was first described by Jacobson et al.[1] in 2004 and then expanded by Zhu et al.[26] in 2006. In short, multiple runs of PLOP are performed where increasing constraints are applied to subsequent rounds of loop predictions. The lowest energy loops from each PLOP run are passed onto subsequent, constrained rounds of refinement. The lowest energy loop across all PLOP runs and all constraints is considered the final structure.

Hierarchical loop prediction is begun with an initial set of candidate loops that are predicted by running PLOP at discrete values of the overlap factor (ofac). In this work, we permitted the ofac to vary from 0.3 to 0.7 in increments of 0.05. The best

15 loops, in terms of energy, are passed onto a Ref stage. A Ref stage constrains the C_α atoms of any new prediction to lie within a set radius of the C_α coordinates of the previous stage. In this case, the Ref1 stage used a 4 Å radius. The best 20 loops from this stage are passed onto a Fix-*n* stage. In a Fix-*n* stage, we repredict a subset of the original target loop but use the output from a previous stage as the scaffold, holding a total of *n* terminal residues fixed. For example, in a Fix3 stage, we hold three terminal residues fixed and repredict the interior loop residues that remain. There are a total of four possible ways to fix three terminal residues:

1. Fix three N-terminal residues
2. Fix three C-terminal residues
3. Fix two N-terminal residues and one C-terminal residue
4. Fix one N-terminal residue and two C-terminal residues

All four possibilities are explored when selecting the lowest energy loop from the Fix3 stage. In general, there are $n + 1$ possible combinations for a given Fix-*n* stage. We ran a total of eight Fix stages from Fix1 to Fix8. The Fix1 stage passed the top 10 loops onto Fix2. Each subsequent Fix stage passed one less loop onto a subsequent stage so that the Fix8 stage passed only the top three predictions. Finally, a second Ref stage is run, Ref2, where a 6 Å C_α constraint is used. In total, taking into account all permutations in the Fix stages as well as the initial stage and Ref stages, there is a minimum of 334 PLOP runs per hierarchical loop prediction. The minimum number of PLOP runs can be exceeded by adaptively varying the ofac during hierarchical loop prediction, described in greater detail below.

To accommodate our helical dihedral library, we modified the hierarchical loop prediction method in two ways:

1. The generation of our helical library was based on complete helices. To be precise, the helical library for four-residue helices is taken only from the coordinates of helices that are exactly four residues. We do not include in our four residue helical library segments of, for example, an eight-residue helix spanning four residues in length. As such, we do not construct our loops using a separate set of "partial" secondary structural elements. As a result of this, Fix stages that would constrain part of a helix instead revert to using our general dihedral library for the individual PLOP run.

2. The use of a helical library also resulted in a large number of individual PLOP runs that failed to produce any candidate helices. This can happen under normal circumstances, say, during a late Fix stage where the majority of the loop is kept constrained and only a small subset of the loop is resampled. Loop construction in these late Fix stages requires the residue buildup to occur without violating our ofac criterion despite being in an environment made all the more crowded by the unconstrained segments of the loop. This problem becomes compounded when working with a helical library. Since loop buildup with a helical library appends the helix onto a nascent loop in a single step, a slight displacement of the preceding residue leads to a large displacement of the terminal end of the helix—a sort of lever effect. If this crude displacement of the terminal residue of a helix places the loop in a steric clash with the surrounding environment, the loop candidate could be rejected due to the ofac criterion. In these cases, the outcome of a loop prediction becomes all the more sensitive to the ofac parameter. To further decouple the effect the ofac has on a successful loop prediction, any individual PLOP run beyond the initial stage that fails to succeed past loop buildup is automatically rerun with a lower ofac down to the lowest ofac sampled during this initial stage. In a PLOP run, the rate-limiting factor is during side chain optimization/minimization, rather than during loop buildup. Restarting a PLOP job after a failed buildup stage is on an order of magnitude of one minute. Since this procedural augmentation can apply to loop−helix−loops as much as it can to other loops, this improved sampling adjustment was applied to all cases studied in this work, regardless of the dihedral library used.

Calculation of RMSD. The success of loop prediction was gauged by using the backbone RMSD calculated against the native, crystal structure conformation of the loop. RMSD was calculated by superimposing the protein backbone, excluding the loop, and using the N, C_α, and C coordinates of the loop to compute the deviations. Unless otherwise stated, we report the RMSD for the lowest energy predicted loop.

Calculation of the Relative Energy. Similar to RMSD, at the conclusion of complete hierarchical loop prediction, we report the relative energy of our predicted structure against the energy of the minimized native. This relative energy is defined as $\Delta E = E_{\text{prediction}} - E_{\text{native}}$. A final structure that has a poor RMSD but a calculated energy that is erroneously superior to the native would thus have a negative ΔE and would indicate a failure of our energy model. Minimization of the target for comparison against predictions is necessary to permit a fair comparison between structures but is particularly important when comparing to crystal structures as the PDB structures obtained have, in all the structures examined in this paper, no explicit hydrogen atoms. The minimization of the native was performed similarly to minimization/optimization of candidate loop structures as described above in the Single-Loop Prediction subsection of the methods. For the native, the target loop is first minimized followed by side chain sampling using the protocol described above in the Single-Loop Prediction section. For predictions done in a perturbed native environment, ΔE reports are still against the energy of the minimized native. For these cases, all additional surrounding residues that are included in the prediction are also minimized in the native to permit an accurate comparison. In instances when we used additional rounds of side chain sampling, the native loop, during minimization, was also permitted an identical number of additional side chain sampling.

Sequence Based Secondary Structure Prediction. Loop prediction using the helical dihedral library requires the user to provide a range of loop residues, known as the helical bounds, over which to apply this library. To serve as an initial test of our method without the complication of uncertainty in the existence and size of a helix, we predicted loop−helix−loops from previously published crystal structures. In these experiments, the helical bounds were known *a priori*. After we had observed success using exact helical bounds, we tested the robustness of this method in a more realistic setting where the helical bounds were supplied by popular sequence-based secondary structure prediction software. Specifically, we ran local copies of the secondary structure prediction packages SSPro4[35] and PSIPRED.[36] The output of either of these programs is a secondary structure assignment across each of the residues contained in the protein chain of interest. We examined the secondary structure assignments only for the residues that spanned our particular loops. Often times, these assignments labeled more than one set of intraloop residues as helical. In particular, the loops discussed in this paper are sometimes bounded by larger helices, and these secondary

structure assignment algorithms had occasionally assigned the terminal residues of the loop to be a part of that larger flanking helix. In other cases, three, two, or even a single intraloop residue was assigned as helical. As the loop–helix–loop prediction method described in this paper is intended for α-helices (helices of four residues or larger), assigning less than four residues as helical is not useful for our purposes. Thus, for simplicity, the largest intraloop helical segment predicted by SSPro4 or PSIPRED, spanning at least four residues, was used as the inputted helical bounds. When both PSIPRED and SSPro4 offered usable helical bounds, we performed loop prediction with both bounds separately and compared the results.

Loop Prediction in an Inexact Environment. Unless otherwise noted, all loop predictions in this work were done by deleting the loop residues but leaving all surrounding side chains intact, thereby preserving the crystal structure environment. In an actual homology modeling experiment, the surrounding side chains are unlikely to be placed *a priori* in their correct native conformation. To test the effectiveness of our method in refining loops in an inexact environment, we followed the approach of Sellers et al.[37] to perturb the surrounding side chains to a reasonable but non-native conformation. To do this, we ran multiple rounds of PLOP to predict the loop of interest in the crystal structure and selected a loop with a backbone RMSD of no better than 3 Å. A list of surrounding residues is obtained by noting all residues that are within 7.5 Å of any candidate predicted loop, not just the one loop with a 3 Å RMSD. The union of the side chains from the surrounding residue list as well as the loop side chains is minimized with the 3 Å backbone RMSD loop held in place. At this point, the surrounding side chains are "biased" toward the 3 Å RMSD loop. This structure then provides the surrounding environment for subsequent tests of our loop prediction methods.

Dipeptide Rotamer Frequency Score. For a number of challenging cases, we experimented with the use of a new addition to our energy model that penalizes loop conformations that are constructed with seldom-observed dipeptide dihedrals. The dipeptide rotamer frequency-based scoring term employed a greatly expanded dipeptide rotamer library (garnered from ~7500 high-quality PDB structures) that incorporated the frequency of each rotamer in this subset of the PDB. This information was used to penalize loop dipeptides whose combination of (ϕ, ψ) angles fall in an extremely unpopulated region of the five-dimensional dipeptide analogue to the well-known Ramachandran plot. The set of five angles for each dipeptide in the predicted loop, using a "sliding window" scheme, is compared against the new library to find the nearest dipeptide rotamer. Two criteria determine whether a penalty will be applied to the dipeptide:

1. if the Euclidean distance between the loop dipeptide and the nearest rotamer in the library is greater than a certain, empirically determined cutoff
2. if the total population of rotamers within a set radius of the loop dipeptide is below a certain threshold

The form of this penalty term, its implementation, and its successes in improving loop prediction in crystal structure and homology model environments will be discussed in detail in an upcoming publication. This term was used in two situations:

1. for all of the predictions in inexact environments—this is a substantially more challenging sampling and scoring problem, and the information contained in the dipeptide score can be expected to improve results systematically
2. for a small subset of the predictions in the native environment where difficulties in the standard prediction approach were encountered

To date, we have not found any cases where this term worsens results. However, more extensive tests are underway and will be presented in a subsequent publication

RESULTS AND DISCUSSION

Description of Test Cases. Application of the discriminating criteria used to select suitable test cases yielded a set of 35 loop–helix–loops of 16 or 17 residues in length. These loops exhibited a distribution of helix size as shown in Figure 3.

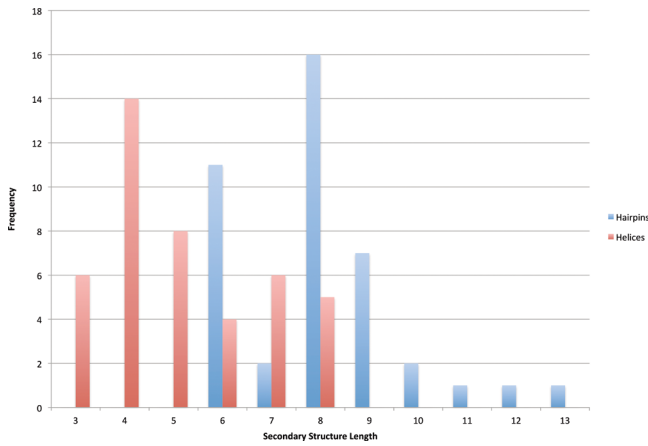


Figure 3. Distribution of secondary-structural elements within the test set of loops. Helices of length 3 were from 3^{10} helices found in loops already containing an α-helix. Hairpin length includes the terminal hydrogen bonded residues as well as all residues in between.

The distribution indicates a diversity of helix sizes within a 16- or 17-residue loop. Although the helical library described in this work is only for α-helices, loops were included that contained 3^{10} helices, either separate from an α-helix already present in the loop or as the sole secondary structure of the loop. It is these former cases where a loop contains a 3^{10} helix with an α-helix that led to the nonzero frequency for helices of length three (Figure 3).

PDB 1W27 contains a noteworthy example of a multihelical loop. The 17-residue loop contains a four-residue 3^{10} helix and five-residue α-helix separated by a single residue, D302 (Figure 4). Evidently, residue D302 permits flexibility in the backbone to transition from one helical type to another. We explored the use of our α-helical library in three approaches: (1) loop prediction given the α-helix as the helical bounds, (2) loop prediction given the 3^{10}-helix as the helical bounds, and (3) loop prediction where the 3^{10} and α-helix bounds are combined to yield a 10-residue "α-helix." The results of these approaches are described in greater detail below.

PDB 2VPN was another case of a multihelical loop. The 16-residue loop of interest is composed of a four-residue α-helix and a seven-residue α-helix separated by a single residue, E102 (Figure 5). Residue E102 is kinked, according to DSSP, failing to form the periodic hydrogen bond expected of an α-helix. As in the 1W27 case, we tried three approaches to predicting this loop.

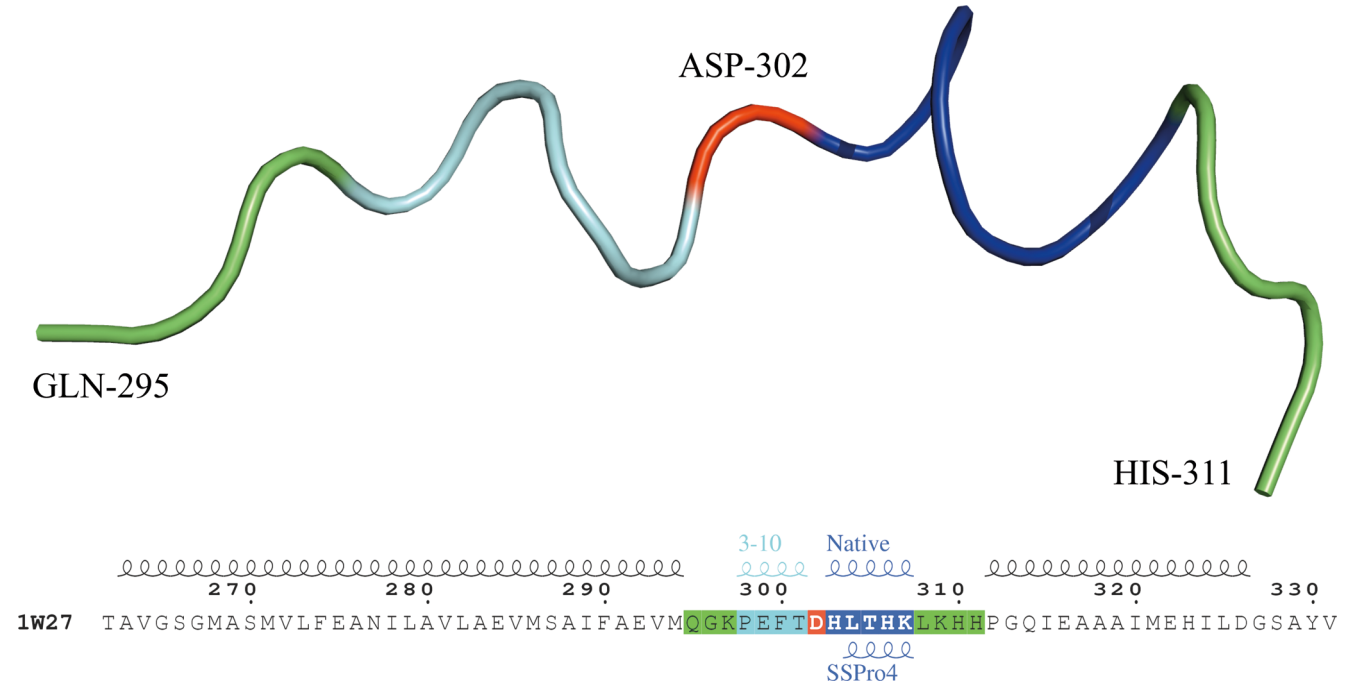


Figure 4. Multihelical loop in PDB 1W27. The loop bounds are Q295 to H311. Residues preceding and following the helices are colored green. The five-residue α-helix is colored blue, while the four-residue 3^{10}-helix is colored cyan. Residue D302, the kinked residue dividing the two helices, is colored red. We attempted separately to use the helical bounds of either the α-helix or 3^{10}-helix or treated all 10 residues as one "α-helix." SSPro4, a sequence-based secondary structure prediction program, assigned the four residues from L304–K307 as helical. The sequence annotation was generated using ESPript.[45] This loop confirmation and all other similar illustrations were produced using Pymol.[46]

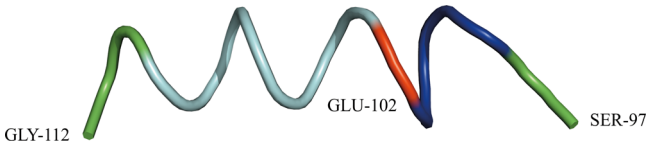


Figure 5. Multihelical loop in PDB 2VPN. The loop bounds are S97 to G112. Residues preceding and following the helices are colored green. The seven-residue α-helix is colored cyan, while the four-residue α-helix is colored blue. Residue E102, the kinked residue dividing the two helices, is colored red. We attempted separately to use the helical bounds of either the seven-residue helix or the four-residue helix or treated all 12 residues as one "α-helix."

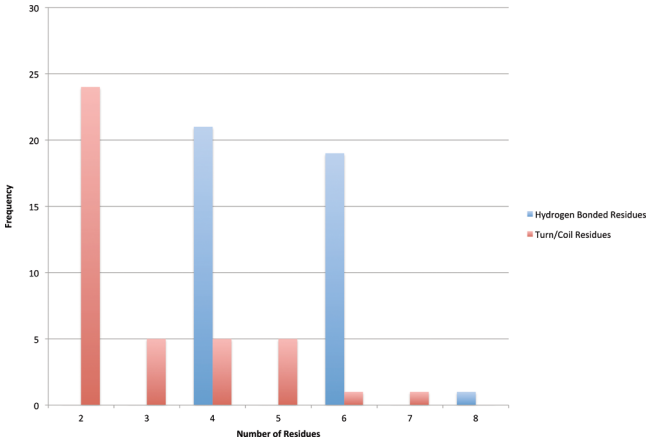


Figure 6. Distribution of hairpin characteristics. Hairpins contained from four to eight hydrogen bonded residues and with the internal turn/coil residues spanning a length from two to seven residues.

For β-hairpins, a set of 41 cases was collected, satisfying the criteria described in the Materials and Methods section. The size of the hairpin region ranged from 6 to 13 residues within loops up to 17 residues in length. Hairpin size is defined to be the number of residues from the start of the first β-strand to the end of the second β-strand, including all non-β residues in between. Hairpins occurred most frequently as either six or eight residues in length (Figure 3). However, since the formation of the coordinated hydrogen bonds is what is most challenging in loop–hairpin–loop prediction, we feel it is useful to describe the distribution of hydrogen bonds across our set of β-hairpins. Hairpins contained from four to eight hydrogen bonded residues with the number of coil/turn residues contained within the hairpin ranging from two to seven residues (Figure 6). Thus, this test set of β-hairpin containing loops required the successful prediction of at least one specific hydrogen bond spanning at most seven residues.

Predictions Performed in the Crystal Structure Environment. A total of 35 loop–helix–loop (LHL) cases and 41 beta-hairpin cases were predicted in the crystal structure environment. In the crystal structure environment, the loop of interest is deleted and rebuilt while the surrounding residues remain fixed. In this work, we compare the predictions done using a helical dihedral library versus predictions performed using the standard PLOP dihedral library.[2b]

Loop–Helix–Loops Predicted Using the Dipeptide Dihedral Library versus the Helical Dihedral Library with Exact Helical Bounds. As a first test of the helical dihedral library, we performed loop prediction on the set of 35 LHL cases either with the previous dipeptide dihedral library[2b] or with the helical library described in this work. Experiments such as these were primarily meant to ensure that in the absence of uncertainty in the size and location of the helix, our helical library method could succeed. A prediction performed where

Table 1. Comparison of Loop−Helix−Loop Predictions with the Dipeptide Dihedral Library versus the Helical Dihedral Library[a]

		dipeptide dihedral library				helical dihedral library with exact helical bounds			
		RMSD (Å)		ΔE (kcal/mol)		RMSD (Å)		ΔE (kcal/mol)	
helix length	number of cases	median	mean	median	mean	median	mean	median	mean
4	12	0.53	1.29	3.89	12.39	0.55	0.99	−0.02	−3.18
5	7	0.91	1.09	−7.94	−6.19	0.51	0.80	−3.74	−3.41
6	4	1.00	0.95	2.06	11.11	0.62	0.77	2.47	2.75
7	5	0.55	1.94	0.51	5.19	0.79	0.91	2.52	1.59
8	5	0.81	0.98	4.33	6.66	0.36	0.41	−4.22	−2.18

[a]The two noteworthy multihelical loops found in PDB 1W27 and 2VPN are excluded in this table. The ΔE value compares the energy of the lowest energy loop against the crystal structure loop coordinates, minimized using our energy function. The RMSD reported is of the lowest energy loop prediction and corresponds with the ΔE.

the helix is postulated from secondary structure prediction software is our primary methodological algorithm to be used in realistic prediction situations and is discussed later. Table 1 provides a summary of the results as a function of helix length. Compared to the dipeptide dihedral library, the helical dihedral library consistently displays improved accuracy, with mean and median RMSD always below 1 Å. No strong correlation is noted between the size of the internal helix and the results from either dihedral library. This suggests, consistent with past results,[2,26] that the difficulty in loop prediction lies with the size of the loop, rather than the secondary structure contained in the loop, at least for helices up to eight residues in length.

For LHLs containing a four-residue helix, both dihedral libraries appear to perform similarly. As might be expected, the helical library shows the greatest advantage for predictions containing an eight-residue helix with superior median and mean RMSD values by around 0.5 Å. It is likely that the coordinated hydrogen bonds that need to be formed are easily generated when explicit helical dihedrals spanning the precise residues are deliberately introduced during sampling. This seems particularly relevant for the LHL in PDB 2YR5. This is a 16-residue loop containing a seven-residue α-helix (Figure 7).

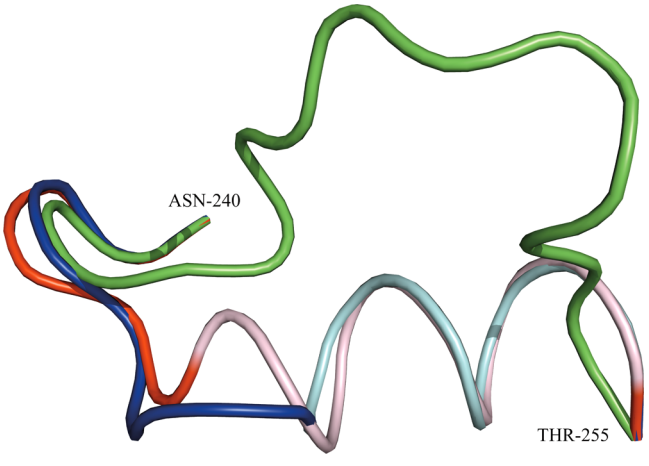


Figure 7. Loop−helix−loop predicted in PDB 2YR5. The native loop coordinates are colored blue with the seven-residue α-helix colored teal. The prediction using the helical dihedral library is shown in red with the resulting nine-residue α-helix colored in pink. The loop prediction performed using the dipeptide dihedral library is shown in green. Despite supplying the exact seven-residue helical bounds during loop prediction with the helical library, what resulted was a slightly larger helix, evidently "seeded" by the smaller seven-residue α-helix.

The dipeptide dihedral library produces a 7.26 Å RMSD loop with a ΔE of −0.9 kcal/mol relative to the minimized crystal structure, while the helical dihedral library leads to a 1.11 Å RMSD loop with a ΔE of −18.34 kcal/mol. The dipeptide dihedral library clearly fails to form the native helix, forming instead a loop that protrudes out in solution. The prediction with the helical library is dramatically superior but forms a larger nine-residue α-helix. Evidently, the shorter seven-residue α-helix "seeds" the larger helix. Considering the large negative ΔE relative to the native, these additional two helical residues may be the result of an energy error incorrectly favoring formation of additional helical residues. While slightly detrimental to the accuracy of this particular loop prediction, as is discussed in greater detail below, the use of a shorter helix to "seed" a larger one is later exploited to find the lowest energy loop.

Two PDB structures, 1W27 and 2VPN, each contain a multihelical loop−helix−loop that still satisfied the criteria stated above for selecting loops (Figures 4 and 5). These cases provided an opportunity to explore the effect of the helical dihedral library in complex situations. We attempted to predict the loop by supplying as helical bounds either of the two helices or treated the helices as combined, disregarding the nonhelical residues dividing the helices. Table 2 describes the result of these loop predictions. In both cases, the helical library produced the lowest energy conformation with sub-Ångström RMSD.

Loop−Helix−Loop Prediction Based on Helical Bounds Derived from SSPro4 and PSIPRED. In the previous section, exact helical bounds were used which were taken from the output of DSSP when applied to the crystal structure. Such accurate information will not be known *a priori*. Indeed, significant variability in the definition of secondary structure assignment has been known to affect the precise bounds of secondary structure, especially as the number of secondary structure assignment definitions is now legion.[38] To simulate the effectiveness of using the helical dihedral library in more realistic computational experiments, and to further gauge the sensitivity of our method to accurate knowledge of the helical bounds, we applied the popular sequence-based secondary structure prediction packages SSPro4[35] and PSIPRED[36] to our set of 35 loop−helix−loops and attempted loop prediction using these predicted helical bounds. The results from these secondary structure prediction packages, excluding the multihelical loops of PDB 1W27 and 2VPN, are presented in Table 3.

Comparing the two packages, it would appear that SSPro4 could more reliably find exact or overlapping helical bounds

Table 2. Prediction of Multihelical Loops Using Various Loop Bounds[a]

PDB	1W27					2VPN			
helical bounds supplied	none	4-res 3^{10}-helix	5-res α-helix	combined 10-res "helix"	SSPro truncated α-helix	none	4-res α-helix	7-res α-helix	combined 12-res "helix"
RMSD (Å)	2.69	1.50	0.77	1.98	0.34	0.42	0.41	0.37	0.38
ΔE (kcal/mol)	38.96	22.27	−3.43	24.32	−12.19	2.01	11.08	−9.69	0.23

[a]When no helical bounds were supplied, loop prediction was performed using the dipeptide dihedral library. The 1W27 prediction using the 4-res 3^{10}-helix for helical bounds still employed the α-helix dihedral library described in this work. The combined helical bounds of 1W27 and 2VPN consider both helices to be one large α-helix during loop buildup. The truncated SSPro helix is equivalent to the 5-res α-helix but truncated one residue at the helical N-terminus. ΔE refers to the change in energy of the predicted loop relative to the native conformation.

Table 3. Results of Sequence-Based Secondary Structure Prediction Packages PSIPRED and SSPro4 on Our Set of LHLs, Excluding Cases 1W27 and 2VPN, the Multihelical Loops[a]

helical bounds predicted	PSIPRED	SSPro4
exact	2	14
truncated	6	2
overlapping	6	9
nonoverlapping	1	1
no helix	18	7
total	33	33

[a]Exact helical bounds are those that are in perfect agreement with the bounds assigned by DSSP on the crystal structure. Truncated helical bounds are those that lie within the DSSP assigned bounds. Helical bounds are considered overlapping if the secondary structure predicted helix has at least a single residue overlapping the exact bounds. No helix is considered predicted if the entire loop−helix−loop lacks any helical assignments greater than three residues.

Table 4. LHL Prediction Using the Helical Bounds Available from PSIPRED and SSPro4[a]

		RMSD (Å)		ΔE (kcal/mol)	
method	number of successful cases	median	mean	median	mean
PSIPRED	13	0.44	0.49	−1.37	−1.54
SSPro4	25	0.60	0.91	1.05	0.65

[a]Multihelical cases 1W27 and 2VPN are included in these statistics. Cases where the helical bounds provided by sequence-based secondary-structure prediction are not usable in our method are excluded. Further, cases where no loops were able to be predicted with the supplied helical bounds are also excluded.

compared to PSIPRED; however, the two methods are complementary. For example, SSPro4 fails to find any helix in the LHL in PDB 3LY0, while PSIPRED found a truncated helix whose bounds are contained within the DSSP results. We must caution the reader that we do not attempt here to perform a rigorous evaluation of secondary structure prediction algorithms. For that, we refer the reader to Koh et al.[39] and Pirovano and Heringa.[40] Rather, we simply selected two popular and easily available packages for our study. Alternative secondary structure prediction algorithms may be just as valid, as is using more than two packages to find the helical bounds. However, the fact that in a large set of cases, the exact, DSSP helical bounds were identified provides some legitimacy in interpreting the results from the previous section—accurate knowledge of a helix within an LHL is not unreasonable.

For the two multihelical loops in PDB 1W27 and 2VPN, the two secondary structure prediction methods contrast. For the LHL in PDB 1W27 (Figure 4), PSIPRED correctly identifies the five-residue α-helix but fails to predict the four-residue 3^{10}-helix. SSPro4 also fails to identify the 3^{10}-helix, but the α-helix is incorrectly predicted to be four residues, truncated at the N-terminus. In 2VPN (Figure 5), PSIPRED predicts a combined helix that spans both α-helices and extends one residue further toward the C-terminus. Contrastingly, SSPro4 considers the entire LHL to be one large helix—a result that is inadequate for our helical dihedral library approach. In both of these cases, PSIPRED offers a reasonable set of helical bounds for use in our method.

Table 4 summarizes the results of LHL prediction using the helical bounds, when available, from PSIPRED and SSPro4. In general, the helical bounds provided by the sequence-based secondary structure prediction methods SSPro4 and PSIPRED are effective in loop−helix−loop prediction. Although the statistics might suggest that the fewer cases afforded by PSIPRED result in higher quality predictions, we refrain from making such a conclusion, as it may be necessary to also take into account the size of the exact helix studied. This does illustrate, however, that sequence based secondary structure assignments are useful to our method when performing three-dimensional loop prediction.

It should be mentioned that five cases were found where the helical bounds offered by either PSIPRED or SSPro4 resulted in failed loop predictions where not a single predicted loop was constructed. In four of these five cases (PSIPRED bounds: PDBs 1N45, 1OAO, 2YR5; SSPro4 bounds: PDB 3GWI), the sequence-based secondary structure assignment places the helix as part of the N or C terminus. It would appear that in these cases, the sequence-based assignment is extending the larger helix that forms the boundary of the loop−helix−loop into what DSSP, and the criteria used in this paper, consider to be part of the loop. Although in practice, assigning the terminal residues of a loop to be helical is not fatal—PSIPRED and SSPro4 both place a helix on the C-terminus of the LHL in PDB 1HN0, and yet a sub-0.5 Å RMSD loop is predicted—loop prediction without any nonhelical residues to precede the helix is extremely difficult.

In these situations, the lever effect, described previously in the Single-Loop Prediction section of the Materials and Methods, becomes very pronounced. As PLOP constructs the loop in a tree-based method, where the tree is split into additional branches as more loop residues are predicted, placing the helix at a loop terminus means there are no preceding branches to rely upon. Whatever few positions the leading residue of the helix is placed at are set entirely by the sparse number of helical rotamers present in our library. In practice, this means that all the rotamers in our helical rotamer library for a given helix size are easily rejected. Although in principle one could reduce the ofac parameter to permit greater steric overlap between a loop residue and the surrounding environment, in practice, the ofac was rarely seen as the limiting factor.

The one case that permitted loop prediction after adjusting the ofac was the PSIPRED bounds for 1N45; however, we had to set the ofac to an abnormally low value of 0.20, meaning enormous steric clashes were permitted. Even still, the output of this loop prediction only produced a 5.69 Å RMSD loop with a ΔE of 9.30 kcal/mol.

In all cases, nascent loop segments were screened out when the helix placed a residue too far from the body of the protein to what has been empirically observed across published crystal structures containing protein loops. Or instead, loops were screened when the distance between the loop segment containing the helix and the opposing end of the loop was considered too great to be spanned by whatever intermediate residues remain. In other words, the helix places one-half of the loop too far away for loop closure to be possible. These loop screening methods are described briefly in the Materials and Methods section, and in greater detail in Jacobson et al.[1] Setting the ofac to an arbitrary low value has no effect on these screens—the helical rotamer library simply does not contain a suitable rotamer to permit loop prediction with the supplied helical bounds. Although there is certainly an argument to be made for increasing the size of the helical library, as evidenced from our other successes, the size of the library does not appear to be an impediment to loop–helix–loop prediction. Rather, the practitioner of our method might gain insight by noting that if no suitable rotamer is present in the library, it may be prudent to consider alternative helical bounds. Indeed, none of these terminus-bounded helices are the crystal structure helical bounds—we avoided such cases by our definition of loop–helix–loops. Determining the helical bounds from the output of our previous dipeptide-dihedral library method, as discussed in greater detail below, may be a fruitful alternative. The multihelical loop of 2VPN (Figure 5) is one slight exception. In this case, PSIPRED combines the four-residue α-helix and the adjacent seven-residue α-helix into one large helix and even extends the helical bounds further by one additional residue to produce a 13-residue helix. SSPro4 simply considers the entire loop–helix–loop to be one large helix, an outcome useless for our helical dihedral library. In this case, the helical bounds provided by PSIPRED produce independent N- and C-terminus loop segments, but closure is not achieved. This result occurs regardless of how low we set the ofac. Again, extending the size of the helical library may offer a solution to this case, but more likely, the helical bounds provided deviate too greatly from the native structure to permit reasonable loop prediction.

Truncated Helical Bounds from Sequence-Based Secondary Structure Prediction or Derived from Inspection of Coordinates Predicted with the Standard PLOP Dihedral Library. In a few cases, sequence-based secondary structure prediction methods produced a helix that was truncated relative to the native helical bounds, yet these cases performed as well, if not better, than the native bounds. For example, PDB 1W27, one of the multihelical loops, is composed of a four-residue 3^{10}-helix and an adjacent five-residue α-helix (Figure 4). SSPro4 fails to identify the 3^{10}-helix but predicts the α-helix to be truncated by one residue at the helical N-terminus, relative to the exact helical bounds (Figure 4). PLOP was able to predict this LHL with an RMSD of 0.77 Å and a ΔE of −3.43 kcal/mol when using the native, five-residue α-helix. However, the SSPro4 bounds led to a predicted LHL with a superior RMSD of 0.34 Å and a ΔE of −12.19 kcal/mol. Table 2 summarizes these results. These loop predictions are illustrated in Figure 8.

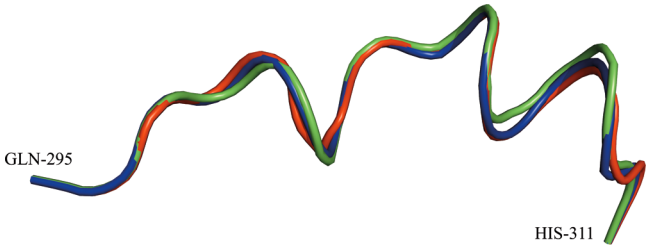


Figure 8. Loop–helix–loop prediction for the multihelical loop in PDB 1W27. The native loop is shown in red. Loop prediction using the exact five-residue α-helix is shown in green. Loop prediction using the truncated, four-residue α-helix provided by SSPro4 is shown in blue. Loop prediction using the truncated four-residue α-helical bounds appears to permit improved sampling of the α helix. Notice that the greatest discrepancy between the two loop predictions occurs along the α-helix near the C-terminus.

Consistent with our past discussion, the smaller helix may permit less of a lever effect and thereby enable finer sampling of the α-helix. It should be noted, however, that the absence of any helical bounds, that is, using the previous dipeptide dihedral library from our previous work, results in a 2.69 Å RMSD prediction (Table 2). Thus, the small helix is shown to also seed our hierarchical sampling method to more heavily explore conformational space near α-helices.

The LHL in PDB 2YR5 is another case where truncated helical bounds led to a superior prediction. However, this is one of the cases where the helical bounds provided by both PSIPRED and SSPro4 were attached to the LHL C-terminus, and no loops emerged from our attempts at predicting this LHL with such helical bounds. Rather, we attempted LHL prediction using as helical bounds all possible four-residue α-helices that lie within the 10 residue α-helix suggested by PSIPRED and SSPro4—a set of seven possible helical bounds. Both PSIPRED and SSPro4 suggested identical helical bounds. The results from these predictions are shown in Table 5.

The predictions indicate that nearly every possible four-residue α-helix attempt produces results that are nearly identical to the LHL prediction performed using the native,

Table 5. Prediction Results from the LHL in PDB 2YR5[a]

helical bounds	RMSD (Å)	ΔE (kcal/mol)
none	7.26	−0.9
B:248–B:254 (native bounds)	1.11	−18.34
bounds derived from PSIPRED/SSPro4 truncation		
B:246–B:249	1.11	−6.2
B:247–B:250	*1.11*	*−18.72*
B:248–B:251	1.11	−18.49
B:249–B:252	1.11	−18.43
B:250–B:253	1.10	−18.18
B:251–B:254	1.10	−18.28
B:252–B:255	4.27	28.57

[a]LHL prediction without helical bounds refers to the use of the dipeptide dihedral library exclusively. The native bounds are those provided by DSSP analysis on the crystal structure. The PSIPRED/SSPro helical bounds are from B:246 and B:255 and bracket the seven truncation attempts shown. The lowest energy prediction across all helical bounds is shown in italics.

seven-residue α-helix. While knowledge of the precise, native helical bounds may not be available, we demonstrate that we can still exploit information provided by sequence-based secondary structure prediction, even if that information does not perfectly match the DSSP secondary structure identification obtained from the crystal structure of the native conformation.

In total, we attempted all possible four-residue α-helix bounds for all LHL cases where the lowest energy loop was found only by using the native helical bounds. This was performed in order to discount the concern that precise *a priori* information about a helix must be known. In many cases, information about a helix was provided by sequence-based secondary-structure prediction. However, as we show in Table 1, providing no helical bounds and using the dipeptide dihedral library can still lead to low RMSD predictions and the formation of a helix. From these cases where a helix four-residues or larger was produced *ab initio*, we also applied our truncation sampling method across the predicted helix and took the lowest energy loop. When the dipeptide-dihedral library simply produced a four-residue helix, we reattempted loop prediction using the helical dihedral library with this previously found four-residue helix as bounds. The lowest energy loops predicted from these experiments are shown in Table 6. In general, the truncation method produces helices that, on their own, are quite accurate with sub-Ångström RMSD routinely reported.

Table 6. Result of LHL Prediction Using Truncated Helical Bounds[a]

PDB	RMSD (Å)	ΔE (kcal/mol)
1HN0	0.31	−2.77
1Q1R	0.30	−8.07
1WOV	0.95	−6.08
2EX0	1.74	0.91
2FHF	0.62	−8.02
2II2	0.35	−3.4
2J9O	1.55	2.65
2QMC	0.49	−3.05
2VPN	0.22	−11.94
2YR5	1.11	−18.72
3GWI	0.53	3.77
mean	0.80	−4.28
median	0.58	−3.23

[a]All possible four-residue helical bounds that lie within bounds provided by sequence-based secondary structure prediction or by analyzing the results from the dipeptide−dihedral based predictions were used. What is shown is the lowest energy prediction across all helical bounds attempted.

Creation of a Systematic Method for Predicting Loop−Helix−Loop Regions. We have described above a number of different approaches to predicting LHL regions, each of which exhibits significant success for a subset of test cases. We briefly enumerate these methods below:

1. normal loop prediction, without any use of the helical rotamer library
2. use of the rotamer library with helical bounds specified by the results of either SSPro or PSIPRED secondary structure prediction (this leads to two separate calculations)
3. reprediction of the loop subsequent to normal loop prediction, using as helical bounds helical regions forming spontaneously in the normal loop prediction simulation
4. truncated helix loop prediction where all possible four-residue helices that can fit within previously obtained helical bounds are explored

Our final algorithm is a composite method in which all of the above calculations are performed for each loop and the lowest energy prediction is selected as the predicted result. The computational cost of this composite method is roughly four times that of one normal loop prediction. In return, one achieves a remarkably high level of reliability as is shown in Table 7. The vast majority of predictions are sub-Ångström, an exceptionally low level of error for loops of this length and complexity. Only one loop has an RMSD greater than 2 Å, the loop in PDB 2O70. We discuss this case further below, but in essence neither normal loop prediction nor any of the secondary structure prediction methods predict a helix in the relevant region. When the native helix is seeded into the calculation, a superior prediction is returned. Thus, this is a sampling problem, which we can hope to solve by improving the sampling algorithm. However, with the current approach, such sampling errors are very infrequent.

Arguably, the results from predictions with the native helical bounds rely on information that may not be precisely known in a homology modeling experiment. As such, we report in Table 7 the RMSD of the lowest energy loop prediction across all sampling methods. For comparison, results of LHL prediction using helical bounds taken only from the native PDB are shown in the right half of Table 7.

Overall, by exploring helical bounds provided by sequence-based secondary-structure prediction methods, as well as using the truncation method, we were able to predict LHLs with slightly superior accuracy than if we were to rely on the DSSP identified helical bounds. However, there were four cases where we were unable to produce a prediction that was superior to the approach using the DSSP-based bounds. Three of the four predictions are 0.11 Å worse in RMSD than the DSSP results and can be left as acceptable.

The only egregiously inferior prediction was for the LHL in PDB 2O70. Here, the use of the DSSP-based helical bounds led to a 1.71 Å RMSD prediction compared to a 3.24 Å RMSD prediction performed solely using the dipeptide dihedral library—that is, without any supplied helical bounds (Table 7). Evidently, this LHL is a challenge for sequence-based secondary-structure prediction as well since neither PSIPRED nor SSPro4 predict there being any helix at all within the LHL. Cendron et al. argue that the sequence of PDB 2O70, an OHCU decarboxylase from *Danio rerio* (zebrafish), lacks homology with other known amino acid sequences.[41] This may have been the case in early 2007 but evidently is now no longer so. In June 2007, the crystal structure of *Arabidopsis thaliana* OHCU decarboxylase was published (PDB: 2Q37), and in 2010, the *Klebsiella pneumoniae* structure (PDB: 3O7I) was deposited in the PDB.[42] However, in these two more recent structures, the five residues comprising the α-helix are not conserved, and the more homologous eukaryotic 2Q37 structure fails to form a helix at this position. It seems reasonable then that PSIPRED and SSPro4 would fail to identify this helix.

With respect to the size of our helical dihedral library, the LHL in PDB 1O7E posed the only challenge. In Table 7, we report the prediction results when using an augmented helical

Table 7. Results of All LHL Predictions Independent of Helical Bounds Derived from Analysis of the Crystal Structure As Well As the Results Using Bounds Derived Exclusively from the Crystal Structure[a]

	helical bounds identified without DSSP			exclusively DSSP identified helical bounds	
PDB	method	RMSD (Å)	ΔE (kcal/mol)	RMSD (Å)	ΔE (kcal/mol)
1BKR	SSPro4	0.55	1.05	0.55	1.05
1E3D	SSPro4	0.55	−1.1	0.55	−1.10
1HN0	PSIPRED	0.35	−5.51	0.3	0.22
1L5W	SSPro4	0.4	−3.74	0.4	−3.74
1LLF	PSIPRED	0.44	−5.53	0.45	−5.5
1N45	dipeptide	0.36	−4.22	2.05	12.55
1N7O	SSPro4	0.4	−1.18	0.4	−1.18
1O7E*	SSPro4	0.37	3.34	0.37	3.34
1OAO	SSPro4	0.49	−80.01	0.49	−80.01
1OX0	dipeptide	1.35	−13.23	0.58	−8.7
1Q1R	truncate	0.3	−8.07	0.3	−8.07
1QMY	SSPro4	1.27	−2.67	1.27	−2.67
1SU8	dipeptide	0.43	2.24	1.45	18.65
1W27	SSPro4	0.34	−12.19	0.77	−3.43
1WOV	truncate	0.95	−6.03	0.67	−2.88
1ZX0	dipeptide	1.04	11.2	1.81	20.12
2DEB	dipeptide	1.35	−10.14	1.55	8.93
2EX0	*PSIPRED*	*0.44*	*−0.86*	*0.33*	*−4.4*
2FHF	*dipeptide*	*0.54*	*−8.7*	*0.54*	*−10.16*
2II2	*truncate*	*0.35*	*−3.4*	*0.36*	*−4.22*
2J9O	truncate	1.55	2.65	1.55	2.65
2JA2	dipeptide	0.81	0.13	0.72	1.15
2JDI	SSPro4	0.51	15.02	0.51	15.02
2O70	*dipeptide*	*3.24*	*−12.42*	*1.71*	*−15.09*
2P0W	dipeptide	0.47	−7.94	0.51	−6.96
2QMC	truncate	0.49	−3.05	0.49	−3.05
2RJ2	PSIPRED	0.31	−1.37	0.57	4.72
2V36	dipeptide	0.18	−1.22	0.25	−0.37
2VPN	truncate	0.22	−11.94	0.37	−9.69
2WEU	dipeptide	0.91	−10.24	1.73	12.19
2YR5	truncate	1.11	−18.72	1.11	−18.34
3CWW	SSPro4	0.28	−12.04	0.27	−10.71
3GWI	truncate	0.53	3.77	0.38	7.28
3HL0	PSIPRED	0.93	−2.04	0.39	2.52
3LY0	PSIPRED	0.54	1.28	0.79	9.14
mean		0.70	−5.91	0.76	−2.31
median		0.50	−3.57	0.55	−1.74

[a]By sampling with alternate helical bounds derived from sequence-based secondary-structure prediction and/or the truncation method, the LHL prediction statistics are slightly superior to predictions using helical bounds derived from the output of DSSP. The four cases that are inferior to LHL prediction with exact DSSP helical bounds are shown in italics. Only one case, 2O70, has an egregiously poor RMSD. The LHL in PDB 1O7E was an exception in that the low Å RMSD reported herein was only produced by introducing the native helical dihedrals into our helical dihedral library.

dihedral library containing the native dihedrals for the helix. In the absence of this addition to our library, the LHL prediction led to a sampling error with an RMSD of 2.09 Å and a 16.99 kcal/mol ΔE compared to a 0.37 Å RMSD and 3.34 kcal/mol ΔE with the augmented library. As discussed in the Materials and Methods section, our helical dihedral library is populated with rotamers that conform close to ideality. This approach fails here and seems likely due to the large discrepancy from ideal (ϕ, ψ) angles for the two terminal residues of the helix. While we expect angles near $(\phi, \psi) = (-60°, -40°)$, the torsions for two of the N-terminus residues of the helix, A223 and G224, are $(\phi_{A223}, \psi_{A223}) = (-68°, -20°)$ and $(\phi_{G224}, \psi_{G224}) = (-104°, 1°)$. In particular, the terminal glycine residue poses the largest problem. From this limited case, there may indeed be utility in further expanding our helical dihedral library, but even in its current implementation, the difficulty in this LHL case appears anecdotal.

The ability of the energy model to robustly pick out the correct loop as being lowest in energy provides new confirmation of the quality of our latest generation model, supporting the results obtained in Li et al., for long loop regions without secondary structure elements embedded.[2a] It is true that phase space available to the loop is significantly restricted when the native environment is (as here) retained; nevertheless, previous results from our group and others show that it is quite easy to generate grossly incorrect predictions (with substantial energy errors) with an inferior scoring function. The results discussed below, in which surrounding side chains are allowed to move, and sub-Ångström results are uniformly obtained, provide further evidence of scoring function accuracy and robustness.

Hairpins Predicted Using the Standard PLOP Dihedral Library. In addition to loop−helix−loops, we also attempted prediction of what could be termed loop−hairpin−loops as another challenge of loop prediction containing local secondary structure.

The results from loop−hairpin−loop prediction, arranged by hairpin length, are shown in Table 8, and the complete results for all 41 hairpin predictions are provided in Table 9.

Table 8. Results of Loop−Hairpin−Loop Predictions Using the Dipeptide Dihedral Library[a]

		dipeptide dihedral library			
		RMSD (Å)		ΔE (kcal/mol)	
hairpin length	number of cases	median	mean	median	mean
6	11	0.41	1.07	−5.61	−5.05
7	2	1.13	1.13	−21.38	−21.38
8[b]	16	0.64	0.90	−6.47	−6.77
9	7	0.51	0.89	−5.00	−5.74
10	2	0.42	0.42	−7.32	−7.32
11	1	0.53	0.53	−10.55	−10.55
12	1	0.30	0.30	−3.06	−3.06
13	1	0.44	0.44	−0.04	−0.04

[a]The ΔE value compares the energy of the lowest energy loop against the crystal structure loop coordinates, minimized using our energy function. The RMSD reported is of the lowest energy loop prediction. [b]Of the eight-residue hairpins, one of the cases, the loop−hairpin−loop as part of PDB 2ZBX, initially reported the best structure as that with a 17.29 Å RMSD. The results for this prediction were rescored, using the dipeptide rotamer frequency score, leading to a 1.02 Å prediction being considered the lowest in energy and was used in the statistics reported in this table. This rescoring is discussed in detail in the text.

Similar to the results for loop−helix−loop predictions, we observe no correlation between the size of the hairpin and the RMSD of the predicted loop−hairpin−loop. We note however that one of the eight-residue hairpin cases produced a large discrepancy between the median and the mean (Table 8). This case is part of PDB 2ZBX and led to an RMSD of 17.29 Å with

Table 9. Results of All Loop−Hairpin−Loop Predictions[a]

PDB	loop length	hairpin length	RMSD (Å)	ΔE (kcal/mol)
1C7N	13	8	0.69	−3.34
1F0L	11	6	0.41	−1.71
1GWI	15	8	0.64	−9.05
1GYH	14	9	0.38	−3.18
1LLF	11	6	1.69	−5.61
1NVM	15	9	0.51	−9.55
1O5K	11	6	0.17	−10.79
1TC5	15	8	1.08	1.69
1U60	14	8	0.47	−12.09
1U8V	13	9	0.33	−1.05
2BS2	15	9	0.6	0.03
2C0D	11	7	0.29	−10.5
2CIU	15	10	0.29	−7.11
2IJ2	16	9	0.61	−7.83
2O36	12	9	3.61	−5
2OKX	16	8	2.88	−6.87
2PB2	13	9	0.21	−13.6
2R2N	8	6	0.24	−6.21
2RFG	11	6	0.63	−5.61
2SLI (A: 177−190)	14	6	0.26	−0.93
2SLI (A: 236−249)	14	8	0.47	−8.88
2WIY	16	8	0.63	−2.36
2WM5	15	8	1.14	−18.43
2YR5	13	6	0.63	−10.95
2YWN	17	13	0.44	−0.04
2ZBX	15	8	1.02	4.70
2ZWA	16	11	0.53	−10.55
2ZYO	8	6	0.36	−1.32
3A9S	12	6	0.18	−4.65
3BF7	11	6	0.98	−9.1
3BJE	12	8	0.34	−3.33
3CSS	17	12	0.30	−3.06
3CU2	11	8	0.38	−3.71
3EGW	12	8	0.49	−10.12
3EI9	15	8	2.12	−2.04
3EJA	15	7	1.97	−32.25
3F8T	14	10	0.54	−7.52
3FAU	13	6	6.21	1.33
3GW9	15	8	0.51	−6.06
3HVW	16	8	0.47	−11.81
3LID	10	8	1.02	−16.62

[a]For PDB 2SLI, two hairpins satisfying the criteria described in the Materials and Methods were found. Those predictions occurred for the chain A residues 177−190 and 236−249.

a surprising ΔE of −177.74 kcal/mol. It should be noted that the second best case has an acceptable RMSD of 1.02 Å and a ΔE of −10.91 kcal/mol. Of course, we cannot choose this 1.02 Å loop as the best case *a priori*, as determination of the best loop is made purely on energetic grounds. The apparent lowest-energy loop and the native are shown in Figure 9.

However, it was observed that the dihedrals in the predicted loop occupy regions of dipeptide-dihedral space (ϕ_1, ψ_1, ω, ϕ_2, ψ_2) that are poorly populated across a set of high quality PDB structures. It became possible in this case, and in other cases not discussed in this work, to identify the more "native-like" loop by introducing a dipeptide-dihedral rotamer frequency-based scoring (RFS) term that penalizes structures with non-native dipeptide confirmations. The details of the RFS will be discussed in a future publication. We applied this penalty term to this loop−hairpin−loop case.

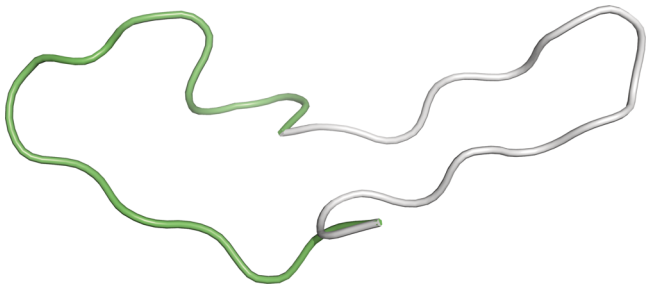


Figure 9. Loop−hairpin−loop prediction for PDB 2ZBX. The native loop is shown in gray while the predicted loop is shown in green.

Application of the penalty term ranks the 1.02 Å RMSD prediction lower in energy than the 17.29 Å RMSD prediction (Table 10). Aside from 2ZBX, five hairpin cases remain where the predictions remain at around 2 Å or worse. These cases are highlighted in red in Table 10. For these cases, we explored the use of the RFS throughout the entire loop prediction, rather than just to rescore the final loop candidates. The results for these five cases when using the RFS are shown in Table 11.

Table 10. Energy of the 2ZBX Loop−Hairpin−Loop Predictions after Application of the Frequency-Based Penalty Term

RMSD (Å)	freq.-based score (kcal/mol)	total energy (kcal/mol)	ΔE (kcal/mol)
0.0 (native)	9.89	−15697.1	0.0
1.02	25.65	−15692.4	4.7
17.29	4387.82	−9927.01	5770.09

Table 11. Reprediction of Hairpin Cases with Initial RMSDs of around 2 Å or Worse[a]

	standard energy model		standard energy model + RFS	
PDB	RMSD (Å)	ΔE (kcal/mol)	RMSD (Å)	ΔE (kcal/mol)
2O36	3.61	−5.00	0.93	−10.71
2OKX	2.88	−6.87	3.62	6.42
3EI9	2.12	−2.04	0.36	0.24
3EJA	1.97	−32.25	1.86	−27.2
3FAU	6.21	1.33	0.51	−11.6

[a]Repredictions were performed by using the RFS throughout the prediction, rather than just to rescore the final putative loops.

The RFS appears successful at correcting the energy error and leading to a lower RMSD in three of the five cases. PDB 2OKX remains a difficult case. Although this case appears to exhibit an energy error before penalizing unlikely structures with the RFS, now a sampling error remains where we appear unable to produce the native conformation. PDB 3EJA appears to remain an energy error, and this case warrants further discussion.

PDB 3EJA contains a seven-residue hairpin within a 15-residue loop that satisfies the various criteria specified in the Materials and Methods section. In particular, the global quality criteria of having suitably high resolution and superior R factors was satisfied as well as the local criteria for B factors and real-space R factors. Inspection of the predicted loop reveals that we are able to form a reasonable hairpin (Figure 10A) and further,

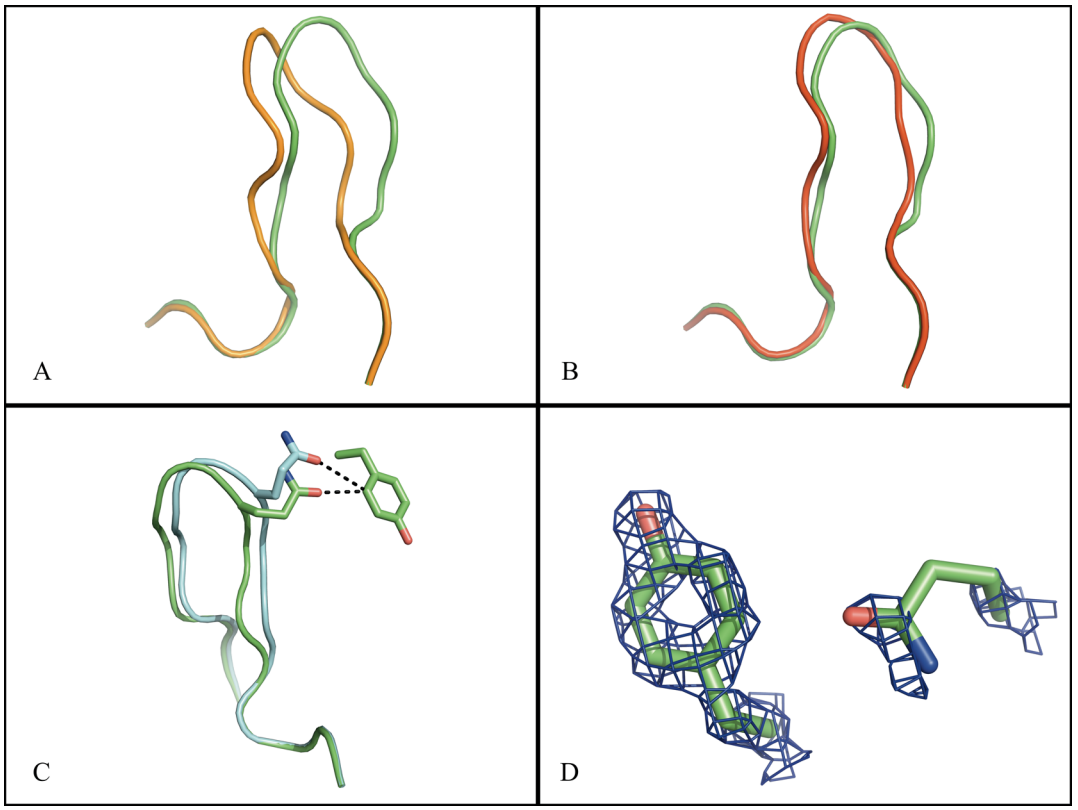


Figure 10. Loop−hairpin−loop predictions in PDB 3EJA. In all panels, the native loop is shown in green. (A) Native hairpin versus the lowest energy prediction using the RFS. (B) Native hairpin versus an intermediately ranked loop. This loop has a 0.94 Å RMSD and a ΔE of −1.16 kcal/mol. (C) Native hairpin versus minimization of the native hairpin. After minimization, the distance between Q108 and Y191 increases from 3.0 Å to 3.5 Å. (D) 2Fo-Fc map contoured at 2σ around residues Q108 and Y191. Observe that while Y191 is confidently built, Q108 has very poor density.

Table 12. Results from LHL Prediction in an Inexact Environment[a]

		native environment		perturbed native		perturbed native + addl. side-chain randomization		perturbed native + addl. side-chain randomization + RFS	
helix length	PDB	RMSD (Å)	ΔE (kcal/mol)	RMSD (Å)	ΔE (kcal/mol)	RMSD (Å)	ΔE (kcal/mol)	RMSD (Å)	ΔE (kcal/mol)
4	1BKR	0.55	1.05	1.67	24.54	2.77	2.42	0.61	−2.11
5	1LSW	0.4	−3.74	0.78	−1.39	0.98	−8.97	0.54	−15.03
6	1WOV	0.67	−2.88	1.29	5.25	1.32	−12.85	0.66	−22.55
7	3HL0	0.39	2.52	0.62	−6.97	0.6	−16.28	0.68	−17.84
8	2EX0	0.33	−4.40	2.28	23.84	0.54	10.49	0.76	7.77

[a]The RMSD is relative to the native structure. The ΔE shown is relative to the energy of the native where the loop and surrounding side chains are minimized.

that during hierarchical loop prediction we succeed in producing a near native loop with an RMSD of 0.94 Å and a ΔE of −1.16 kcal/mol, relative to the native (Figure 10B). This would seem to suggest the sampling is not an issue here. The fact that the lowest energy loop predicted (Figure 10A) was found nearly 30 kcal/mol lower in energy than the native was surprising. Inspection of the individual residues comprising the loop revealed an unusual close contact between the oxygen on the amide side chain of Q108 and an aromatic carbon on Y191. The distance between these polar and nonpolar atoms was a surprising 3.0 Å. Loop minimization perturbs the hairpin such that this distance is increased to 3.5 Å where Y191, like all surrounding residues, is held fixed (Figure 10C). The suspicion was that these residues might have been improperly built in the crystal structure and indeed inspection of the electron density showed Y191 to be confidently placed while Q108 was modeled into sparse density (Figure 10D). We see no alternative positions to place Q108; however, it is beyond the scope of this work to construct the necessary omit maps and attempt model refinement. In describing the structure, the crystallographers do describe a possible role for Y191, but no mention is made of Q108 and so perhaps this residue simply does not hold a stable conformation.[43] Difficulty in modeling an occasional residue in a high resolution crystal structure is certainly not uncommon. We attempted to exclude loops that were affected by problems such as these in using a real-space R-factor cutoff of 2.0. However, this residue has a real-space R factor of 0.185. In future studies, it appears a more stringent cutoff is required.

Predictions Performed in an Inexact Environment. Throughout all loop predictions, we have relied on the crystal structure to provide the surrounding environment of the loop. This too, like the precise knowledge of helical bounds, may not be accurately known in a homology modeling experiment. To

Table 13. Results from Hairpin Prediction in an Inexact Environment[a]

		native environment		perturbed native		perturbed native + addl. side-chain randomization		perturbed native + addl. side-chain randomization + RFS	
hairpin length	PDB	RMSD (Å)	ΔE (kcal/mol)	RMSD (Å)	ΔE (kcal/mol)	RMSD (Å)	ΔE (kcal/mol)	RMSD (Å)	ΔE (kcal/mol)
6	1F0L	0.41	−1.71	0.72	12.68	0.74	−10.01	0.73	−14.16
7*	2C0D	0.29	0.89	0.89	−14.19	2.34	−27.39	1.71	−1.54
8	2SLI	0.48	−6.85	0.54	−1.64	3.22	−8.67	0.49	−12.72
9	1GYH	0.38	−3.18	0.73	0.45	0.82	0.18	0.9	1.54
10	2CIU	0.29	−7.11	6.18	29.67	0.41	−22.61	0.57	−10.21
11	2ZWA	0.53	−10.55	0.91	−10.16	0.46	−6.17	0.77	7.68
12	3CSS	0.30	−3.06	0.57	2.05	0.4	−4.86	0.37	−3.73

[a]The RMSD is relative to the native structure. The ΔE shown is relative to the energy of the native where the loop and surrounding side chains are minimized. The hairpin of length 7, 2C0D, is shown before protonation of D136 in chain B. After protonation of this residue, the energy errors shown here are eliminated. Energy errors occur when predicted loops are reported substantially lower in energy than the native but have poor RMSD. This is discussed in greater detail in the text.

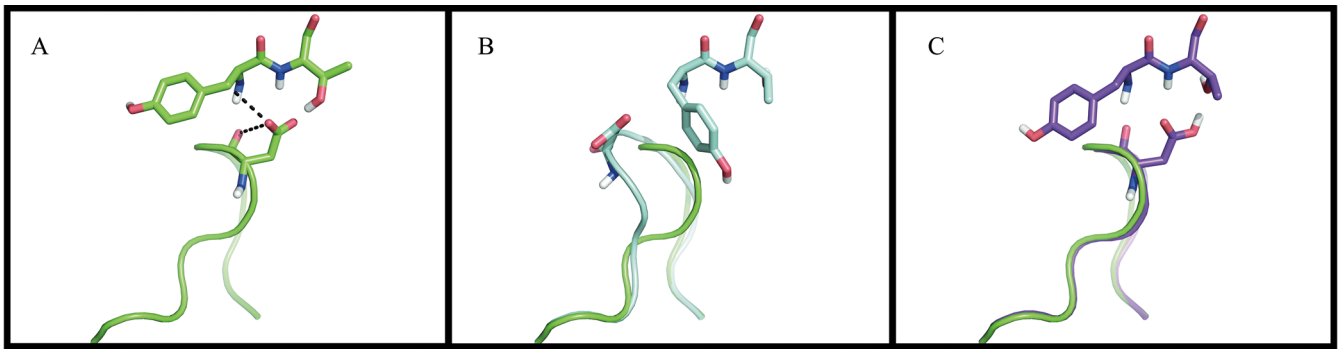


Figure 11. PDB 2C0D. In all panels, the native loop is shown in green for comparison. (A) The native loop with all atoms shown for D136 and surrounding side chains Y63 and T64. The suspicious close contacts that motivated protonation of D136 are shown dotted in this panel. (B) The coordinates of the same atoms in the RFS prediction with D136 deprotonated. (C) The coordinates of the RFS prediction with D136 protonated. Notice the similarity to the native loop in panel A.

explore the effectiveness of our sampling and energy model in a more realistic setting, we minimized the surrounding environment in the presence of a predicted, but poor, 3 Å RMSD loop. This produced a non-native but locally minimized surrounding side chain environment. However, the backbone environment is still that of the native. From here, we deleted the target loop and performed loop prediction with simultaneous refinement of all surrounding residues. This approach was for both loop−helix−loops and hairpins. We repredicted in an inexact surrounding environment one loop for each secondary-structure length. The loops selected had a sub-1-Å RMSD when predicted in the native environment. For loop−helix−loops, this selection was based on the results from predictions using the exact helical bounds. As would be expected, prediction of the loop as well as surrounding side chains increases the sampling required and computational cost of these predictions. In particular, we found it necessary to introduce additional rounds of side-chain randomization (Table 12). Hence, we used only the exact helical bounds to avoid the added complication and expense of sampling surrounding side chains with all the combinations of alternative helical bounds. We also explored the use of the rotamer frequency score (RFS), mentioned previously when describing the improvement in hairpin case 2ZBX (Figure 9 and Table 10) and others (Table 11). Here, we used the RFS throughout the loop prediction, penalizing all intermediate loops as necessary so that only structures with the lowest penalty are likely to advance on to subsequent refinement. The results of these predictions for LHLs are shown in Table 12.

In all cases, we were able to recover the loop with sub-1-Å RMSD when utilizing additional rounds of side-chain randomization and the RFS. The use of additional rounds of side-chain randomization finds in all cases a lower energy structure. In 2EX0, the effect is most pronounced where a 2.28 Å prediction is improved to 0.75 Å. Still in the cases 1BKR, 1L5W, and 1WOV, additional rounds of side-chain randomization are further improved with the addition of the RFS, which brings, in the most striking example, a 2.77 Å prediction down to 0.61 Å.

Similar results were seen for hairpins as is shown in Table 13. As before, the use of additional rounds of side-chain randomization improves results. Most notably, this additional side chain sampling takes the perturbed native prediction for 2CIU from 6.18 Å to 0.41 Å.

PDB 2C0D evidently posed a significant challenge. The lowest energy structure reported is substantially lower in energy than the native and other similar calculations on 2C0D (Table 13). This suggests a problem separate from sampling. Visual inspection of the predicted structure relative to the native illustrates the source of this energy error being due to incorrect protonation state assignment.

This situation is illustrated in Figure 11. Shown is the close contact between D136 and Y63. Both residues are part of chain B, but Y63 is interacting from a crystallographically related monomer. The distance from the carboxylic oxygen in D63 to the C_β is only 3.2 Å, while the distance from that same carboxylic oxygen to that residue's backbone carbonyl is 3.35 Å. Were D63 to be assigned as charged, as it originally was using our previously published algorithm,[34] substantial repulsion

dx.doi.org/10.1021/ct301083q | J. Chem. Theory Comput. 2013, 9, 1846−1864

Table 14. The Effect of Protonation of D136 on the Hairpin Prediction in PDB 2C0D

	perturbed native		perturbed native + addl. side-chain randomization		perturbed native + addl. side-chain randomization + RFS	
D136 protonation state	RMSD (Å)	ΔE (kcal/mol)	RMSD (Å)	ΔE (kcal/mol)	RMSD (Å)	ΔE (kcal/mol)
deprotonated	0.89	−14.19	2.34	−27.39	1.71	−1.54
protonated	1.34	19.14	0.71	−22.56	0.56	−19.02

between D136 and Y63, as shown in Figure 11B, is expected. D136 lies at the tip of the hairpin, and so a large deviation of this residue can lead to a significant RMSD for much of the hairpin. Once D136 is assigned as protonated, the successful prediction shown in Figure 11C results. Here, a 0.56 Å loop is produced with a ΔE of −19.02 kcal/mol. The effect of protonation of this residue on all three perturbed native predictions performed for PDB 2C0D is shown in Table 14.

Remarkably, the prediction of this hairpin when the surrounding environment is native is possible with D136 left as deprotonated (Table 13). As shown in Figure 11B, incorrect protonation state assignment of D136 leads to residue Y63 being perturbed from its native conformation. Evidently, leaving Y63 and all surrounding environment residues constrained to their native position removes the heavy dependence on correct protonation state assignment of D136. The fact that the removal of this constraint leaves our predictions sensitive to additional factors is not surprising. Additional perturbed native experiments such as these will be run in the future to expose more weaknesses in our algorithm; however, for the cases presented in this work, the difficulties appear isolated to this case and are tractable.

Interpretation of the Relative Energies. Throughout this work, we have reported results comparing the geometry of our predicted structure to the native coordinates via the RMSD and comparing the energy of our predicted structure to the minimized native via the ΔE. As mentioned in the Materials and Methods section, $\Delta E = E_{\text{prediction}} - E_{\text{native}}$. In any successful energy model, the minimized native structure should be reported as being lowest in energy, and yet we report negative ΔE values across various predictions. It is worth speculating on the source of this. We believe there are two general possibilities:

1. There are problems in the backbone of the crystal structure that cannot be rectified with our gradient-based minimization as our energy model places the backbone in a local minimum. This seems perfectly plausible in crystal structures, even for the high quality structures explored in this work, as hydrogen atoms positions are not experimentally known, preventing, at the least, the use of an all-atom energy model for refinement. Indeed, Bell et al. report a successful reduction in nonbonded clashes in crystal structures, introduced after consideration of explicit hydrogen atoms, through the use of an all-atom refinement procedure without any loss in adherence to the diffraction data.[44] Thus, what we may be observing instead is a slightly physically superior structure obtained during the extensive sampling performed during our *ab initio* loop prediction.

2. That negative ΔE values observed in predictions with remarkably low sub-Ångström backbone RMSD may instead be due to improper side-chain contacts being formed. For example, Table 12 includes a 0.33 Å prediction of an LHL in PDB 2EX0 with a ΔE of −4.40 kcal/mol. It may well be that these improper contacts are due to a flaw in our energy model, and although this is possible, our ability here to select the lowest energy structure and achieve sub-Ångström RMSDs appears unaffected. As such, in this paper we do not investigate in greater detail the source of these errors.

We also observe systematic differences in the ΔE across methods and secondary structure. For example, Table 1 reports the RMSD and ΔE of LHL predictions performed using just our normal dipeptide dihedral library versus the helical dihedral library presented in this work. In this table, the mean ΔE for all helix lengths predicted is lower with the helical dihedral library than without. This suggests that without the helical dihedral library, there are sampling errors which are removed by seeding the helix.

For the hairpin predictions, Tables 8 and 9 show that the vast majority of predictions conclude with a structure with a negative ΔE. Referring to the first of our two speculations on the source of these negative ΔE values, it may be that the extensive sampling performed in loop prediction is producing superior backbone hydrogen bonds that are not accessible through minimization of the crystal structure.

CONCLUSIONS

We have developed a robust algorithm to exploit secondary structure prediction of small helical segments in loops to yield routinely accurate loop−helix−loop predictions to atomic accuracy. Furthermore, we have demonstrated that our previous dipeptide-dihedral library and all-atom energy model can successfully predict loops containing hairpins. By running parallel loop predictions with a systematically generated set of putative helical bounds from two secondary structure prediction algorithms (SSPro4 and PSIPRED) as well as the normal loop prediction protocol, we have demonstrated that the native loop−helix−loop can be reliably sampled and accurately scored.

This application of a separate, helical dihedral library to a subset of loop residues is at the crux of our method. It affords us increased likelihood of the formation of the coupled hydrogen bonds that define secondary structure by performing loop buildup with the coupled dihedral angles already in place, but it has also introduced a sort of lever effect, where small changes at the base of the helix lead to significant displacement of the terminal end of the loop. For smaller helices, this is obviously less of a problem but for larger helical bounds, such as the LHLs predicted in PDBs 1OAO and 2YR5 where the helical bounds were supplied by PSIPRED, it became impossible for loop buildup to be performed—all possible helix conformations produced loop halves that were considered impossible to close.

Rather than seek to expand the size of our helical dihedral library to include more rotamers, we found it more effective to attempt loop−helix−loop prediction with shorter helical bounds, one that would be less likely to demonstrate a lever effect. This led to the use of our truncated helix sampling method. We leave it up to subsequent rounds of further minimization and sampling to form the remainder of the helix, and indeed this appears to be effective. Nonetheless, for very large helices, our limited dihedral library may fail to contain a

sufficient number of rotamers to avoid a sampling error, and the truncation method may leave too large of a subloop to correctly sample and form the remaining coupled dihedrals that are necessary to complete the helix. In practice though, this is not a very large concern for us. Such large helices are likely the well-conserved regions between homologous proteins. Knowledge of these helical bounds would likely be found with sequence-based secondary structure prediction methods, but crucially, the conformation of these large loop−helix−loops lies squarely within the purview of our previous rigid helix placement algorithm.[21]

Hairpins, somewhat surprisingly, appeared as a simpler type of secondary structure to predict. The small nonlocality of the hydrogen bonds deterred us from wanting to introduce a separate hairpin dihedral library as such a library would seem to produce a bias in the non-hydrogen bond turn-region of the hairpin between the two β strands. Rather, we attempted loop−hairpin−loop prediction using only our previous dipeptide-dihedral library.[2b] Low RMSD loops were successfully predicted to atomic accuracy with no significant change to our past algorithms, other than permitting a flexible ofac to be tried throughout all rounds of hierarchical loop prediction. For both hairpins and loop−helix−loops, it would be desirable in the future to further establish this methodology by running blind tests where the structure of a given loop is available but unknown to the researcher. However, we do not anticipate the results of such experiments to diverge from what we present here as our method is automated, using only the energy and not user input, to determine the final loop conformation.

Predictions performed in a non-native surrounding environment were successful, albeit requiring additional sampling and the use of our rotamer frequency score to accurately predict the loop. An apparent caveat is that the additional degree of freedom now present in the surrounding environment can magnify energy errors. As shown in the hairpin in PDB 2C0D, incorrect protonation state assignment of an aspartic acid is compensated by the coupled movement of a surrounding environment residue. Although only this case had such a problem, clearly more experiments need to be performed across a large set of loops, with and without secondary structure, to expose weaknesses in our algorithm and correct them. These experiments are already underway and will be discussed in a future publication.

AUTHOR INFORMATION

Corresponding Author

*Phone: 212-854-7606. E-mail: rich@chem.columbia.edu.

Notes

The authors declare the following competing financial interest(s): RAF has a significant financial stake in Schrdinger, Inc., is a consultant to Schrdinger, Inc., and is on the Scientific Advisory Board of Schrdinger, Inc..

ACKNOWLEDGMENTS

This research was supported by a National Institutes of Health grant to R.A.F. under Grant No. GM-40526. Additional support was provided by an award from the Department of Energy (DOE) Office of Science Graduate Fellowship Program (DOE SCGF) to E.B.M. The DOE SCGF Program was made possible in part by the American Recovery and Reinvestment Act of 2009. The DOE SCGF program is administered by the Oak Ridge Institute for Science and Education for the DOE. ORISE is managed by Oak Ridge Associated Universities (ORAU) under DOE contract number DE-AC05-06OR23100. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship to D.A.G. under Grant No. DGE-07-07425. C.S.M. is supported by the NIH Training Program in Molecular Biophysics T32GM008281. R.A.F. has a significant financial stake in Schrödinger, Inc., is a consultant to Schrödinger, Inc., and is on the Scientific Advisory Board of Schrödinger, Inc. All opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the policies and views of DOE, ORAU, ORISE, NIH, or NSF.

REFERENCES

(1) Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J.; Honig, B.; Shaw, D. E.; Friesner, R. A. A hierarchical approach to all-atom protein loop prediction. *Proteins: Struct., Funct., Bioinf.* **2004**, *55* (2), 351−367.

(2) (a) Li, J.; Abel, R.; Zhu, K.; Cao, Y.; Zhao, S.; Friesner, R. A. The VSGB 2.0 model: a next generation energy model for high resolution protein structure modeling. *Proteins: Struct., Funct., Bioinf.* **2011**, *79* (10), 2794−2812. (b) Zhao, S.; Zhu, K.; Li, J.; Friesner, R. A. Progress in Super Long Loop Prediction. *Proteins: Struct., Funct., Bioinf.* **2011**, *79* (10), 2920−2935.

(3) Go, N.; Scheraga, H. A. Ring Closure and Local Conformational Deformations of Chain Molecules. *Macromolecules* **1970**, *3* (2), 178−187.

(4) Palmer, K. A.; Scheraga, H. A. Standard-geometry chains fitted to X-ray derived structures: Validation of the rigid-geometry approximation. I. Chain closure through a limited search of "loop" conformations. *J. Comput. Chem.* **1991**, *12* (4), 505−526.

(5) Moult, J.; James, M. N. G. An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins: Struct., Funct., Bioinf.* **1986**, *1* (2), 146−163.

(6) Bruccoleri, R. E.; Karplus, M. Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers* **1987**, *26* (1), 137−168.

(7) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4* (2), 187−217.

(8) Bassolino-Klimas, D.; Bruccoleri, R. E. Application of a directed conformational search for generating 3-D coordinates for protein structures from α-carbon coordinates. *Proteins: Struct., Funct., Bioinf.* **1992**, *14* (4), 465−474.

(9) DePristo, M. A.; de Bakker, P. I. W.; Lovell, S. C.; Blundell, T. L. Ab initio construction of polypeptide fragments: Efficient generation of accurate, representative ensembles. *Proteins: Struct., Funct., Bioinf.* **2003**, *51* (1), 41−55.

(10) de Bakker, P. I. W.; DePristo, M. A.; Burke, D. F.; Blundell, T. L. Ab initio construction of polypeptide fragments: Accuracy of loop decoy discrimination by an all-atom statistical potential and the AMBER force field with the Generalized Born solvation model. *Proteins: Struct., Funct., Bioinf.* **2003**, *51* (1), 21−40.

(11) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117* (19), 5179−5197.

(12) (a) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii. *J. Phys. Chem. A* **1997**, *101* (16), 3005−3014. (b) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **1990**, *112* (16), 6127−6129.

(13) Kolodny, R.; Guibas, L.; Levitt, M.; Koehl, P. Inverse Kinematics in Biology: The Protein Loop Closure Problem. *Int. J. Robot. Res.* **2005**, *24* (2−3), 151−163.

(14) Petrey, D.; Honig, B. Protein Structure Prediction: Inroads to Biology. *Mol. Cell* **2005**, *20* (6), 811−819.
(15) Cherezov, V.; Rosenbaum, D. M.; Hanson, M. A.; Rasmussen, S. G.; Thian, F. S.; Kobilka, T. S.; Choi, H. J.; Kuhn, P.; Weis, W. I.; Kobilka, B. K.; Stevens, R. C. High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. *Science* **2007**, *318* (5854), 1258−1265.
(16) Warne, T.; Serrano-Vega, M. J.; Baker, J. G.; Moukhametzianov, R.; Edwards, P. C.; Henderson, R.; Leslie, A. G.; Tate, C. G.; Schertler, G. F. Structure of a beta1-adrenergic G-protein-coupled receptor. *Nature* **2008**, *454* (7203), 486−491.
(17) Knight, S.; Andersson, I.; Branden, C. I. Crystallographic analysis of ribulose 1,5-bisphosphate carboxylase from spinach at 2.4 A resolution. Subunit interactions and active site. *J. Mol. Biol.* **1990**, *215* (1), 113−160.
(18) (a) Goldfeld, D. A.; Zhu, K.; Beuming, T.; Friesner, R. A. Successful prediction of the intra- and extracellular loops of four G-protein-coupled receptors. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108* (20), 8275−8280. (b) Nikiforovich, G. V.; Taylor, C. M.; Marshall, G. R.; Baranski, T. J. Modeling the possible conformations of the extracellular loops in G-protein-coupled receptors. *Proteins: Struct., Funct., Bioinf.* **2010**, *78* (2), 271−285.
(19) Zhu, J.; Xie, L.; Honig, B. Structural refinement of protein segments containing secondary structure elements: Local sampling, knowledge-based potentials, and clustering. *Proteins: Struct., Funct., Bioinf.* **2006**, *65* (2), 463−479.
(20) Rohl, C. A.; Strauss, C. E.; Chivian, D.; Baker, D. Modeling structurally variable regions in homologous proteins with rosetta. *Proteins: Struct., Funct., Bioinf.* **2004**, *55* (3), 656−677.
(21) Li, X.; Jacobson, M. P.; Friesner, R. A. High-resolution prediction of protein helix positions and orientations. *Proteins: Struct., Funct., Bioinf.* **2004**, *55* (2), 368−382.
(22) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235−242.
(23) Wang, G. L.; Dunbrack, R. L. PISCES: a protein sequence culling server. *Bioinformatics* **2003**, *19* (12), 1589−1591.
(24) Jones, T. A.; Zou, J. Y.; Cowan, S. W.; Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1991**, *47* (Pt 2), 110−119.
(25) Kleywegt, G. J.; Harris, M. R.; Zou, J. Y.; Taylor, T. C.; Wahlby, A.; Jones, T. A. The Uppsala Electron-Density Server. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2004**, *60*, 2240−2249.
(26) Zhu, K.; Pincus, D. L.; Zhao, S. W.; Friesner, R. A. Long loop prediction using the protein local optimization program. *Proteins: Struct., Funct., Bioinf.* **2006**, *65* (2), 438−452.
(27) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure - Pattern-Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **1983**, *22* (12), 2577−2637.
(28) (a) Hartigan, J. A. *Clustering Algorithms*; John Wiley: New York, 1975. (b) Hartigan, J. A.; Wong, M. A. Algorithm AS 136: A K-Means Clustering Algorithm. *Appl. Stat.* **1979**, *28* (1), 100−108.
(29) Xiang, Z.; Honig, B. Extending the accuracy limits of prediction for side-chain conformations. *J. Mol. Biol.* **2001**, *311* (2), 421−430.
(30) Jacobson, M. P.; Friesner, R. A.; Xiang, Z.; Honig, B. On the role of the crystal environment in determining protein side-chain conformations. *J. Mol. Biol.* **2002**, *320* (3), 597−608.
(31) (a) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118* (45), 11225−11236. (b) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides. *J. Phys. Chem. B* **2001**, *105* (28), 6474−6487. (c) Jacobson, M. P.; Kaminski, G. A.; Friesner, R. A.; Rapp, C. S. Force Field Validation Using Protein Side Chain Prediction. *J. Phys. Chem. B* **2002**, *106* (44), 11673−11680.
(32) Ghosh, A.; Rapp, C. S.; Friesner, R. A. Generalized Born Model Based on a Surface Integral Formulation. *J. Phys. Chem. B* **1998**, *102* (52), 10983−10990.
(33) Zhu, K.; Shirts, M. R.; Friesner, R. A. Improved Methods for Side Chain and Loop Predictions via the Protein Local Optimization Program: Variable Dielectric Model for Implicitly Improving the Treatment of Polarization Effects. *J. Chem. Theory Comput.* **2007**, *3* (6), 2108−2119.
(34) Li, X.; Jacobson, M. P.; Zhu, K.; Zhao, S.; Friesner, R. A. Assignment of polar states for protein amino acid residues using an interaction cluster decomposition algorithm and its application to high resolution protein structure modeling. *Proteins: Struct., Funct., Bioinf.* **2007**, *66* (4), 824−837.
(35) (a) Pollastri, G.; Przybylski, D.; Rost, B.; Baldi, P. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins: Struct., Funct., Bioinf.* **2002**, *47* (2), 228−235. (b) Cheng, J.; Randall, A. Z.; Sweredoski, M. J.; Baldi, P. SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res.* **2005**, *33* (suppl 2), W72−W76.
(36) Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **1999**, *292* (2), 195−202.
(37) Sellers, B. D.; Zhu, K.; Zhao, S.; Friesner, R. A.; Jacobson, M. P. Toward better refinement of comparative models: predicting loops in inexact environments. *Proteins: Struct., Funct., Bioinf.* **2008**, *72* (3), 959−971.
(38) Tyagi, M.; Bornot, A.; Offmann, B.; de Brevern, A. G. Analysis of loop boundaries using different local structure assignment methods. *Protein Sci.* **2009**, *18* (9), 1869−1881.
(39) Koh, I. Y.; Eyrich, V. A.; Marti-Renom, M. A.; Przybylski, D.; Madhusudhan, M. S.; Eswar, N.; Grana, O.; Pazos, F.; Valencia, A.; Sali, A.; Rost, B. EVA: Evaluation of protein structure prediction servers. *Nucleic Acids Res.* **2003**, *31* (13), 3311−3315.
(40) Pirovano, W.; Heringa, J. Protein secondary structure prediction. *Methods Mol. Biol.* **2010**, *609*, 327−348.
(41) Cendron, L.; Berni, R.; Folli, C.; Ramazzina, I.; Percudani, R.; Zanotti, G. The structure of 2-oxo-4-hydroxy-4-carboxy-5-ureidoimidazoline decarboxylase provides insights into the mechanism of uric acid degradation. *J. Biol. Chem.* **2007**, *282* (25), 18182−18189.
(42) (a) Kim, K.; Park, J.; Rhee, S. Structural and functional basis for (S)-allantoin formation in the ureide pathway. *J. Biol. Chem.* **2007**, *282* (32), 23457−23464. (b) French, J. B.; Ealick, S. E. Structural and mechanistic studies on Klebsiella pneumoniae 2-Oxo-4-hydroxy-4-carboxy-5-ureidoimidazoline decarboxylase. *J. Biol. Chem.* **2010**, *285* (46), 35446−35454.
(43) Harris, P. V.; Welner, D.; McFarland, K. C.; Re, E.; Navarro Poulsen, J. C.; Brown, K.; Salbo, R.; Ding, H.; Vlasenko, E.; Merino, S.; Xu, F.; Cherry, J.; Larsen, S.; Lo Leggio, L. Stimulation of lignocellulosic biomass hydrolysis by proteins of glycoside hydrolase family 61: structure and function of a large, enigmatic family. *Biochemistry* **2010**, *49* (15), 3305−3316.
(44) Bell, J. A.; Ho, K. L.; Farid, R. Significant reduction in errors associated with nonbonded contacts in protein crystal structures: automated all-atom refinement with PrimeX. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2012**, *68* (Pt 8), 935−952.
(45) Gouet, P.; Courcelle, E.; Stuart, D. I.; Metoz, F. ESPript: analysis of multiple sequence alignments in PostScript. *Bioinformatics* **1999**, *15* (4), 305−308.
(46) *The PyMOL Molecular Graphics System*, version 1.4.1; Schrodinger, LLC: New York, 2011.