Efficient Algorithms to Explore Conformation Spaces of Flexible Protein Loops

Ankur Dhanik[1], Peggy Yao[1], Nathan Marz[1], Ryan Propper[1], Charles Kou[1], Guanfeng Liu[1], Henry van den Bedem[2], Jean-Claude Latombe[1]

[1] Computer Science Department, Stanford University, Stanford, CA 94305, USA
[2] Joint Center for Structural Genomics, SLAC, Menlo Park, CA 94025, USA

Abstract. Two efficient and complementary sampling algorithms are presented to explore the space of closed clash-free conformations of a flexible protein loop. The "seed sampling" algorithm samples conformations broadly distributed over this space, while the "deformation sampling" algorithm uses these conformations as starting points to explore more finely selected regions of the space. Computational results are shown for loops ranging from 5 to 25 residues. The algorithms are implemented in a toolkit, LoopTK, available at `https://simtk.org/home/looptk`.

1 Introduction

Several applications in biology require *exploring* the conformation space of a flexible fragment (usually, a loop) of a protein. For example, upon binding with a small ligand, a fragment may undergo large deformations to rearrange non-local contacts [14]. Incorporating such flexibility in docking algorithms is a major challenge [17]. In X-ray crystallography experiments, electron-density maps often contain noisy regions caused by disorder in the crystalline sample, resulting in an initial model with missing fragments between resolved termini [19]. Similarly, in homology modeling [15], only parts of a protein structure can be reliably inferred from known structures with similar sequences.

This problem requires satisfying two constraints concurrently: closing a kinematic loop and avoiding steric clashes. Each constraint is relatively easy to satisfy alone, but the combination is hard because the two constraints are conflicting. The subset of closed conformations with no steric clash has a relatively small volume, especially for long loops; conversely, an arbitrary collision-free conformation of a loop has null probability to be closed. So, sampling techniques proposed so far have a high rejection ratio.

Here, we present two new techniques, *seed* and *deformation* sampling, to solve this problem. Each deformation sampling operation starts from a given closed clash-free conformation and deforms this conformation without breaking closure or introducing clashes by modifying the loop's degrees of freedom (DOFs) in a coordinated way. In contrast, seed sampling generates new conformations from scratch, by prioritizing the treatment of the two constraints, so that the most limiting one is enforced first. In both techniques, detection and prevention of

steric clashes is done using the grid-indexing method described in [11]. Seed and deformation sampling complement each other very well. Seed sampling produces conformations that are broadly distributed over the loop's conformation space and provides conformations (seeds) later used by deformation sampling to explore more finely certain regions of this space. These algorithms are implemented into a toolkit, **LoopTK**, available at `https://simtk.org/home/looptk`. They have been tested on loops ranging from 5 to 25 residues.

2 Previous Work

The problem considered in this paper is a version of the "loop closure" problem studied in [2, 6, 8, 12, 13, 21]. Several works have focused on kinematic closure. Analytical Inverse Kinematics (IK) methods are described in [6, 21] to close a fragment of 3 residues. For longer fragments, iterative techniques have been proposed, like CCD (Cyclic Coordinate Descent) [2] and the "null space" technique [19]. Here we use analytical IK in a new way to close longer fragments. We use the null space technique to deform fragments without breaking closure.

Other works also consider clash avoidance. Most (e.g., [5, 8, 12]) successively sample closed conformations and next test them for steric clashes. Because of its high rejection ratio, this approach is slow when clash-free conformations span a small subset of the closed conformation space, which is the case for most long loops. This observation motivated the prioritized constraint-satisfaction approach embedded in our seed sampling procedure.

Some works try to sample conformations that locally minimize an energy function. Some use libraries of fragments obtained from previously solved structures [7, 13, 18, 20]. Others sample conformations at random and refine them through energy minimization [8, 9, 12, 16] or molecular dynamics [1]. But in the case of a truly deformable fragment, it is often more useful to explore the entire closed clash-free conformation space. For example, a fuzzy electron density map can be better explained by a distribution of conformations than by a single one, no matter how well it fits the density. Our goal in this paper is to present such exploration tools. Nevertheless, our deformation sampling technique also allows energy minimization, when this is desirable.

3 Loop Model

A loop L is a sequence of $p > 3$ consecutive residues in a protein P, such that none of the two termini of L is also a terminus of P. We number the residues of L from 1 to p, starting from the N terminus. We model the backbone of L as a serial linkage whose DOFs are the $n = 2p$ dihedral angles ϕ_i and ψ_i around the bonds N–Cα and Cα–C, in residues $i = 1, ..., p$. The rest of the protein, denoted by $P \backslash L$, is assumed rigid. We let L_B denote the backbone of L. It includes the Cβ and O atoms respectively bonded to the Cα and C atoms in the backbone.

We attach a Cartesian coordinate frame Ω_1 to the N terminus of L and another frame Ω_2 to its C terminus. When L_B is connected to the rest of the

protein, i.e., when it adopts a *closed* conformation, the pose (position and orientation) of Ω_2 relative to Ω_1 is fixed. We denote this pose by Π_g. However, if we arbitrarily pick the values of ϕ_i and ψ_i, $i = 1$ to p, then in general we get an *open* conformation of L_B, where the pose of Ω_2 differs from Π_g. The set $\mathbf{Q}$ of all open and closed conformations of L_B is a space of dimensionality $n = 2p$. The subset $\mathbf{Q}_{\text{closed}}$ of closed conformations is a subspace of $\mathbf{Q}$ of dimensionality $n-6$. Let $\Pi(q)$ denote the pose of Ω_2 relative to Ω_1 when the conformation of L_B is $q \in \mathbf{Q}$. The function Π and its inverse Π^{-1} are the "forward" and "inverse" kinematics map of L_B, respectively.

A conformation of L_B is *clash-free* if and only if no two atoms, one in L_B, the other in L_B or $P \backslash L$, are such that their centers are closer than ε times the sum of their van der Waals radii, where ε is a constant in $(0, 1)$. In our software, ε is an adjustable parameter, usually set to 0.75, which approximately corresponds to the distance where the van der Waals potential associated with two atoms begins increasing steeply. We denote the set of closed clash-free conformations of L_B by $\mathbf{Q}^{\text{free}}_{\text{closed}}$. It has the same dimensionality as $\mathbf{Q}_{\text{closed}}$, but its volume is usually a small fraction of that of $\mathbf{Q}_{\text{closed}}$.

4 Seed Sampling

Overview The goal of seed sampling is to generate conformations of L_B broadly distributed over $\mathbf{Q}^{\text{free}}_{\text{closed}}$. The challenge comes from the interaction between the kinematic closure and clash avoidance constraints. Computational tests (see Section 6) show that the approach that first samples conformations from $\mathbf{Q}_{\text{closed}}$ and next rejects those with steric clashes is often too time consuming, due to its huge rejection ratio. The reverse approach – sampling the angles ϕ_i and ψ_i of L_B to avoid clashes – will inevitably end up with open conformations, since $\mathbf{Q}_{\text{closed}}$ has lower dimensionality than $\mathbf{Q}$.

These observations led us to develop a prioritized constraint-satisfaction approach. We partition L_B into three segments, the front-end F, the mid-portion M, and the back-end B. F starts at the N terminus of L_B and B ends at its C terminus. M is the segment between them. The feasible conformations of F and B are more limited by the clash avoidance constraint than by the closure constraint; so, we sample the dihedral angles in F and B to avoid clashes, ignoring the closure constraint. Then, for any pair of conformations of F and B, the possible conformations of M are mainly limited by the closure constraint; so, we sample conformations of M using an IK procedure to close the gap between F and B and test the clash avoidance constraint afterward. The length of M must be large enough for the IK procedure to succeed with high probability, but not too large since clash avoidance is only tested afterward. In our software, the number of residues in M is set to half of that of L_B or to 4, whichever of these two numbers is larger. The number of residues of F and B are then selected equal ($\pm$ 1). Tests show that these choices are close to optimal on average.

Sampling front/back-end conformations Consider the front-end F. The angles ϕ and ψ closest to the fixed terminus of F are the most constrained by possible

clashes with the rest of the protein $P\backslash L$. So, the angles are sampled in the order in which they appear in F, that is ϕ_1, ψ_1, ϕ_2, etc. In this order, each angle ϕ_i (resp., ψ_i) determines the positions of the next two atoms $C_{\beta i}$ and C_i (resp., the next three atoms O_i, N_{i+1} and $C\alpha_{i+1}$). The angle is sampled so that these atoms do not clash with any atom in $P\backslash L$ or any preceding atom in F. Its value is picked at random, either uniformly or according to a user-input probabilistic distribution (e.g., one based on Ramachandran tables). If no value of the angle prevents the two or three atoms it governs from clashing with other atoms, the algorithm backtracks and re-samples a previously sampled angle. Clash-free conformations of the back-end B are sampled in the same way, by starting from its fixed C terminus and proceeding backward.

Sampling mid-portion conformations Given two non-clashing conformations of F and B such that the gap between them does not exceed the maximal length that M can achieve, a conformation of M is sampled as follows.

The values of the ϕ and ψ angles in M are picked at random, uniformly or according to a given distribution. This leads to a conformation q of M that is connected to F at one end and open at the other end. To close the gap between M and B, we use the IK method described in [6]. This method solves the IK problem analytically, for any sequence of residues in which exactly three pairs of (ϕ, ψ) dihedral angles are allowed to vary. These pairs need not be consecutive. Our experiments show that, on average, the IK method is the most likely to succeed in closing the gap when one pair is the last one in M and the other two are distributed in M. Let r and s denote the numbers identifying the first and last residue of M in L_B. As the IK method is extremely fast, ANALYTICAL-IK(q, i, j, s) is called for all $i = r, ..., s-2$ and $j = i+1, ..., s-1$, in a random order, until a closed conformation of M has been generated. If this conformation tests clash-free, then the seed sampling procedure constructs a closed clash-free conformation of L_B by concatenating the conformations of F, M, and B.

If the above operations fail to generate a closed clash-free conformation of M, then they are repeated (with new values of the ϕ and ψ angles in M) until a predefined maximal number of iterations has been performed.

Placing side-chains For each conformation of L_B sampled from $\mathbf{Q}^{\text{free}}_{\text{closed}}$, we use SCWRL3 [3] to place the side-chains. We may only compute the placements of the side-chains in L_B given the placements of the side-chains in $P\backslash L$. Alternatively, we may (re-)compute the placements of all the side-chains in the protein. In each case, SCWRL3 does not guarantee a clash-free conformation.

5 Deformation Sampling

Overview The deformation sampling procedure is given a "seed" conformation q in $\mathbf{Q}^{\text{free}}_{\text{closed}}$. It first selects a vector in the tangent space $T\mathbf{Q}_{\text{closed}}(q)$ of $\mathbf{Q}_{\text{closed}}$ at q. By definition, any vector in this space is a velocity vector $[\dot{\phi}_1, ..., \dot{\psi}_n]^T$ that maps to the null velocity of Ω_2 (relative to Ω_1); hence, it defines a direction of

motion that does not instantaneously break loop closure. A new conformation of L_B is then computed as $q' = q + \delta q$ where δq is a short vector in $T\mathbf{Q}_{\text{closed}}(q)$. Since the tangent space is only a local linear approximation of $\mathbf{Q}_{\text{closed}}$ at q, the closure constraint is in fact slightly broken at q'. So, ANALYTICAL-IK(q', $p-2$, $p-1$, p) is called to bring back the frame Ω_2 to its goal pose Π_g. Finally, the atoms in L_B are tested for clashes among themselves and with the rest of the protein. If a clash is detected, the procedure exits with failure.

The deformation sampling procedure may be run several times with the same seed conformation q to explore the subset of $\mathbf{Q}^{\text{free}}_{\text{closed}}$ around q. Alternatively, each run may use the conformation generated at the previous run as the new seed to generate a "pathway" in the set $\mathbf{Q}^{\text{free}}_{\text{closed}}$.

Computation of a basis of the tangent space To select a direction in $T\mathbf{Q}_{\text{closed}}(q)$, we must first compute a basis for this space. This can be done as follows [19]. Let $J(q)$ be the $6 \times n$ Jacobian matrix that maps the velocity $\dot{q} = [\dot{\phi}_1, ..., \dot{\psi}_p]^T$ of the dihedral angles in L_B at q to the velocity $[\dot{x}, \dot{y}, \dot{z}, \dot{\alpha}, \dot{\beta}, \dot{\gamma}]^T$ of Ω_2, i.e.: $[\dot{x}, \dot{y}, \dot{z}, \dot{\alpha}, \dot{\beta}, \dot{\gamma}]^T = J(q)\dot{q}$. $J(q)$ can be computed analytically using techniques presented in [4]. For simplicity, assume that J has full rank (i.e., 6) at q. A basis of $T\mathbf{Q}_{\text{closed}}(q)$ is built by first computing the Singular Value Decomposition $U\Sigma V^T$ of $J(q)$ where U is a 6×6 unitary matrix, Σ is a $6 \times n$ matrix with non-negative numbers on the diagonal and zeros off the diagonal, and V is an $n \times n$ unitary matrix [10]. Since the rows $6, ..., n$ of V do not affect the product $J(q)\dot{q}$, their transposes form an orthogonal basis $N(q)$ of $T\mathbf{Q}_{\text{closed}}(q)$.

Selection of a direction in the tangent space The deformation sampling procedure may select a direction in $T\mathbf{Q}_{\text{closed}}(q)$ at random. However, in most cases, it is preferable to minimize an objective function $E(q)$. Let $y = -\nabla E(q)$ be the negated gradient of E at q and $y_N = NN^T y$ the projection of y into $T\mathbf{Q}_{\text{closed}}(q)$. The deformation sampling procedure selects the increment δq along y_N. In this way, all the DOFs left available in L_B by the closure constraints are used to move the conformation in the direction that most reduces E.

$E(q)$ may be a function of the distances between the closest pairs of atoms at conformation q (where each pair consists of one atom in L_B and one atom in either $L \backslash B$ or L_B). Minimizing E then leads deformation sampling to increase the distances between these pairs of atoms, if this goal does not conflict with the closure constraint. In this way, deformation sampling picks increments δq that have small risk of causing steric clashes.

Another interesting objective function leads to moving a designated atom A in L_B toward a desired position x_d. This objective function can be defined as:

$$E(q) = \|x_A(q) - x_d\|^2.$$

where $x_A(q)$ is the position of A when L_B's conformation is q. This function can be used to iteratively move an atom as far as possible along selected directions to explore the boundary of $\mathbf{Q}^{\text{free}}_{\text{closed}}$. E can also be an energy function or any weighted combination of functions, each designed to achieve a distinct purpose.

Protein id	Protein size	Loop start	Loop size	Seed sampling	Naive sampling
1XNB	185	SER 31	5	0.22	0.21
1TYS	264	THR 103	5	0.06	0.06
1GPR	158	SER 74	6	0.38	0.38
1K8U	89	GLU 23	7	0.21	0.20
2DRI	271	GLN 130	7	0.42	0.46
1TIB	269	GLY 172	8	2.49	13.03
1PRN	289	ASN 215	8	0.33	0.66
1MPP	325	ILE 214	9	0.53	99.85
4ENL	436	LEU 136	9	1.46	19.35
135L	129	ASN 65	9	0.77	1.54
3SEB	238	HIS 121	10	0.50	3.80
1NLS	237	ASN 216	11	1.30	5.51
1ONC	103	MET 23	11	2.26	5.66
1COA	64	VAL 53	12	19.02	67.49
1TFE	142	GLU 158	12	0.48	8.14
8DFR	186	SER 59	13	2.02	39.36
1THW	207	CYS 177	14	1.48	9.84
1BYI	224	GLU 115	16	2.52	>800
1G5A	628	GLY 433	17	3.28	>800
1HML	123	GLY 51	25	17.74	>800

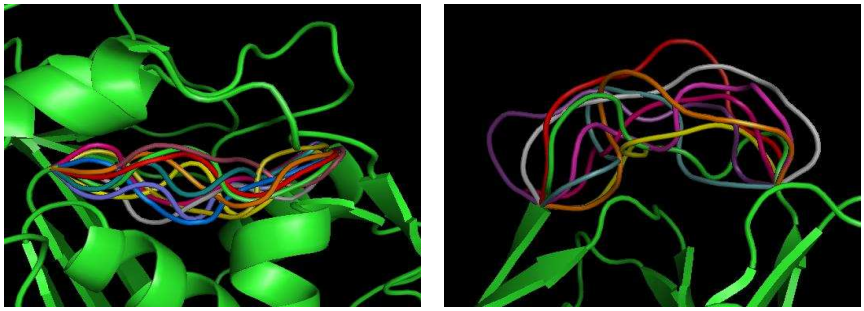
Table 1. Testset of 20 loops (see main text for comments).

Placing side-chains For each new conformation of L_B, side-chains can be placed using SCWRL3, as described in Section 4. Another possibility is to provide an initial seed conformation that already contains the loop's side-chains to the deformation sampling procedure. These side-chains are then considered rigid and the procedure deforms L_B so that the produced conformation remains clash-free.

6 Results

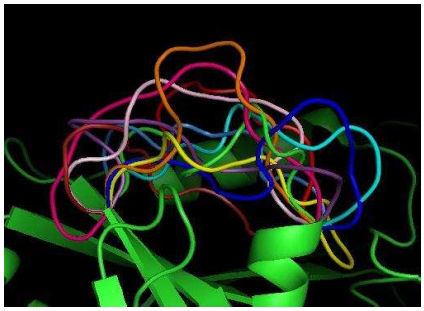
Seed sampling Table 1 lists 20 loops, whose sizes range from 5 to 25 residues, which we used to perform computational tests. Each row lists the PDB id of the protein, the number of residues in the protein, the number identifying the first residue in the loop, the number of residues in the loop, and the average times to sample one closed clash-free of the loop using two distinct procedures. Some loops protrude from the proteins and have much empty space in which they can deform without clash (e.g., 3SEB), while others are very constrained by the other protein residues (e.g., 1TIB). The loop in 1MPP is constrained in the middle by side-chains protruding from the rest of the protein. In the results presented below, all ϕ and ψ angles were picked uniformly at random (i.e., no biased distributions, like the Ramachandran's ones, were used).

Each picture in Figure 1 displays a subset of backbone conformations generated by seed sampling for the loops in 1TIB, 3SEB, 8DFR, and 1THW. The loop

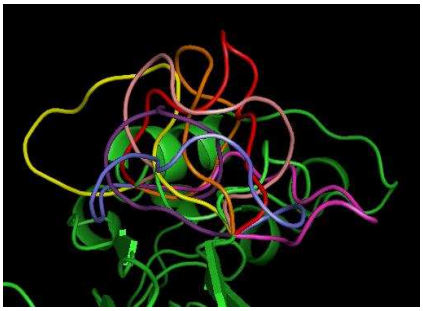


(a) 1TIB 8-residue loop

(b) 3SEB 10-residue loop



(c) 8DFR 13-residue loop



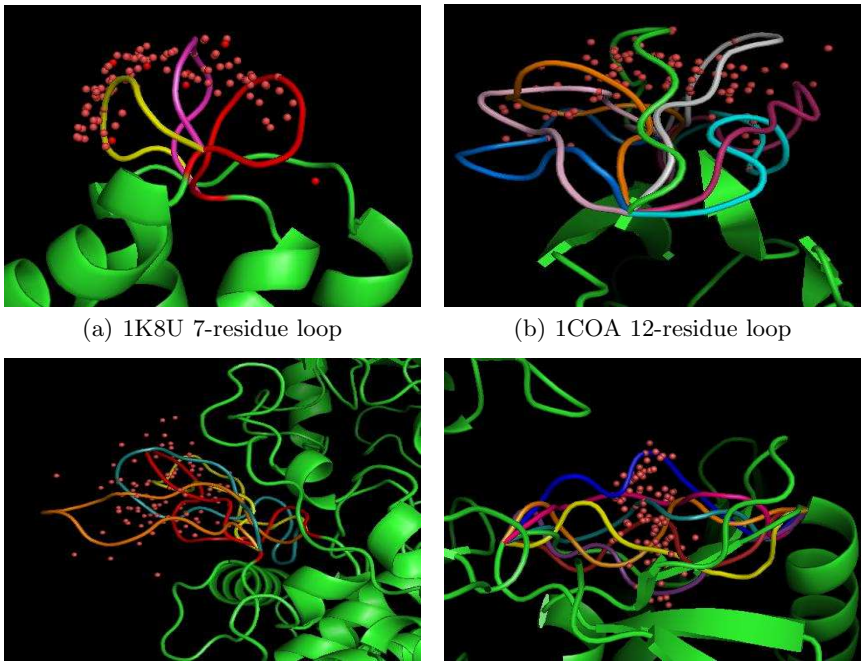
(d) 1THW 14-residue loop

Fig. 1. Some backbone conformations generated by seed sampling for the loops in 1TIB, 3SEB, 8DFR, and 1THW.

in 1TIB, which resides at the middle of the protein, has very small empty space to move in. The PDB conformation of the loop in 1THW (shown green in the picture) bends to the right, but our method also found clash-free conformations that are very different. Each picture in Figure 2 shows the distributions of the middle Cα atom in 100 sampled conformations of the loops in proteins 1K8U, 1COA, 1G5A, and 1MPP along with a few backbone conformations. The loops in 1K8U and 1COA have relatively large empty space to move in, whereas the loops in 1G5A and 1MPP are restricted by the surrounding protein residues. These figures illustrate the ability of seed sampling to generate conformations broadly distributed across the closed clash-free conformation space of a loop.

The average running time (in seconds) to compute one closed clash-free conformation of each loop is shown in Table 1 (column 5). Each average was obtained by running the procedure until it generated 100 conformations of the given loop and dividing the total running time by 100.[3] The last column of Table 1 gives the average running time of the "naive" procedure that first samples closed conformations of the loop backbone and next rejects those which are not clash-free. In both procedures, the factor ε used to define steric clashes (see Section 3) was set

[3] The algorithms are written in C++ and runs under Linux. Running times were obtained on a 3GHz Intel Pentium processor with 1GB of RAM.



(a) 1K8U 7-residue loop (b) 1COA 12-residue loop

(c) 1G5A 17-residue loop (d) 1MPP 9-residue loop

Fig. 2. Positions of the middle Cα atom (red dots) in 100 loop conformations computed by seed sampling for four proteins: 1K8U, 1COA, 1G5A, and 1MPP.

to 0.75. Our seed sampling procedure does not break a loop into 3 segments if it has fewer than 8 residues. So, the running times of both procedures for the first 5 proteins are essentially the same. For all other proteins, our procedure is faster than the naive procedure, sometimes by a large factor (188 times faster for the highly constrained loop in 1MPP). For the last 3 proteins, the naive procedure failed to sample 100 conformations after running for more than 80,000 seconds.

Not surprisingly, the running times vary significantly across loops. Short loops with much empty space around them take a few 1/10 seconds to sample, while long loops with little empty space can take a few seconds to sample. The loops in 1COA and 1HML take significantly more time to sample than the others. In the case of 1COA, it is difficult to connect the loop's front-end and back-end (3 residues each) with its mid-portion (6 residues). As Figure 5 shows, the termini of the loop are far apart and the protein constrains the loop all along. Due to the local shape of the protein at the two termini of the loop, many sampled front-ends and back-ends tend to point in opposite directions, which then makes it often impossible to close the mid-portion without clashes. In this case, we got a better average running time (4 seconds, instead of 19) by setting the length of the mid-portion to 8 (instead of 6). The loop in 1HML is inherently difficult to sample. Not only is it long, but there is also little empty space available for it.

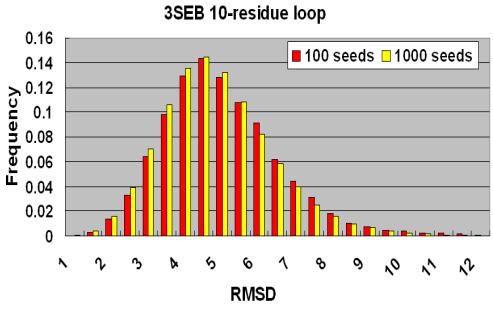


Fig. 3. RMSD histograms for one loop.

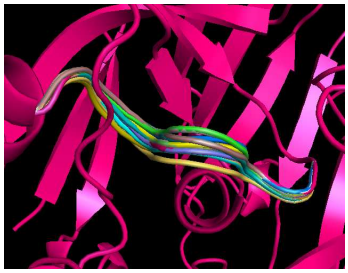


Fig. 4. Twenty conformations of the loop in 1MPP generated by deforming a given seed conformation along randomly picked directions.

Figure 3 displays two RMSD histograms generated for the loop in 3SEB. The red (resp., yellow) histogram was obtained by sampling 100 (resp. 1000) conformations of the corresponding loop and plotting the frequency of the RMSDs between all pairs of conformations. The almost identity of the two histograms indicates that the sampled conformations spread quickly in $\mathbf{Q}_{\text{closed}}^{\text{free}}$. Similar histograms were generated for other loops.

For rather long loops, any seed sampling procedure that samples broadly $\mathbf{Q}_{\text{closed}}^{\text{free}}$ can only produce a coarse distribution of samples. Indeed, for a loop with n dihedral angles, a set of N evenly distributed conformations defines a grid with $N^{1/n-6}$ discretized values for each of the $n-6$ dimensions of $\mathbf{Q}_{\text{closed}}^{\text{free}}$. If $n = 18$ (9-residue loop), a grid with 3 discretized values per dimension requires sampling 531,441 conformations. However, deformation sampling makes it possible to sample more densely "interesting" regions of $\mathbf{Q}_{\text{closed}}^{\text{free}}$.

Deformation sampling Figure 4 shows 20 conformations of the loop in 1MPP generated by deformation sampling around a conformation computed by seed sampling. To produce each conformation, the deformation sampling procedure started from the same seed conformation and selected a short vector δq in $T\mathbf{Q}_{\text{closed}}(q)$ at random. This figure illustrates the ability of deformation sampling to explore $\mathbf{Q}_{\text{closed}}^{\text{free}}$ around a given conformation.

Figure 5 shows a series of closed clash-free conformations of the loop in 1COA successively sampled by pulling the N atom (shown as a white dot) of THR 58 away from its initial position along a given direction until a steric clash occurs (white circle). The initial conformation shown in red was generated by seed sampling and the side-chains were placed without clashes using SCWRL3. Each other conformation was sampled by deformation sampling starting at the previously sampled conformation and using the objective function E defined in Section 5. Only the backbone was deformed, and each side-chain remained rigid. Steric clashes were tested for all atoms in the loop.

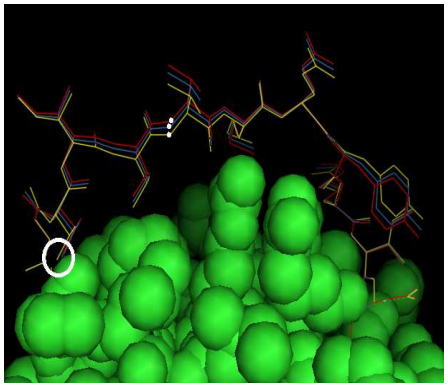


Fig. 5. Deformation of the loop in 1COA by pulling the N atom (white dot) of THR 58 along a specified direction.

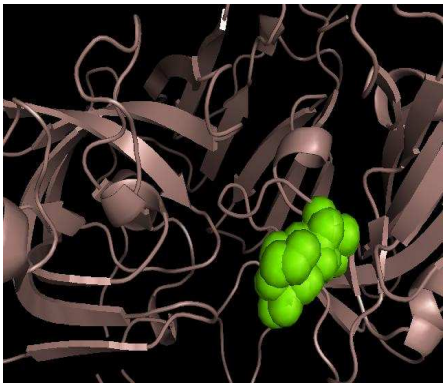
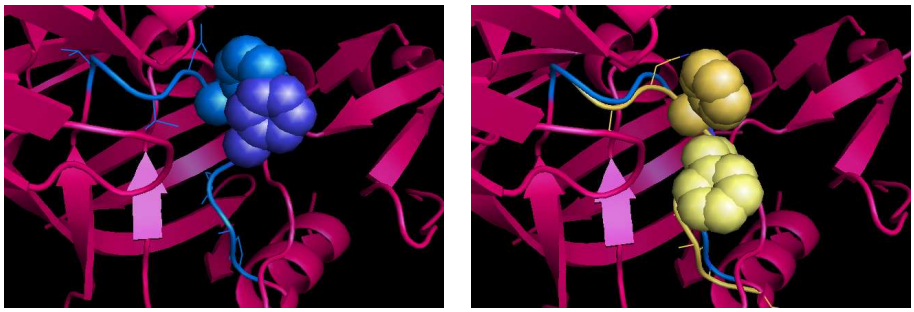


Fig. 6. Volume reachable by the 5th Cα atom in the loop of 1MPP.



(a) (b)

Fig. 7. Use of deformation sampling to remove steric clashes involving side chains.

Figure 6 displays the volume (shown green) reachable by the 5th Cα atom in the loop of 1MPP. This volume was obtained by sampling 20 seed conformations of the loop and, for each of these conformations, pulling the 5th Cα atom along several randomly picked direction until a clash occurs. The volume shown green was obtained by rendering the atom at all the positions it reached.

The running time of deformation sampling depends on the objective function. In the above experiments, it is less than 0.5 seconds per sample on average.

Placements of side-chains Our software calls SCWRL3 to place side chains. The result, however, is not guaranteed to be clash-free. We ran the seed sampling procedure to sample conformations of the backbones of the loops in 1K8U, 2DRI, 1TIB, 1MPP, and 135L, with the uniform and Ramachandran sampling distributions for the dihedral angles (see Section 4). For each loop, we sampled 50 conformations with the uniform distribution and 50 with the Ramachandran

Protein	1K8U	2DRI	1TIB	1MPP	135L
Uniform	7	9	1	0	9
Ramachandran plots	18	14	6	4	13

Table 2. Number of clash-free placements of side chains for five loops.

distribution. We then checked each conformation for steric clashes. Table 2 reports the number of clash-free conformations for each loop and each of the two distributions. As expected, the backbone conformations generated using the Ramachandran distribution facilitate the clash-free placement of the side-chains.

When seed sampling generates a conformation q of a loop backbone, such that SCWRL3 computes a side chain placement that is not clash-free, deformation sampling can be used to sample more conformations around q, to produce one where side chains are placed without clashes. In Figure 7(a) a conformation (shown blue) of the backbone of the loop in 1MPP was generated using seed sampling and the side chains were placed by SCWRL3. However, there are clashes between two side chains. In (b) a conformation (shown yellow) was generated by the deformation sampling procedure using the conformation shown in (a) as the start conformation. The new placement of the side chains computed by SCWRL3 is free of clashes. Once such a clash-free conformation has been obtained, many other clash-free conformations can be quickly generated around it, again using deformation sampling, as shown in Figure 4.

7 Conclusion

We have described two algorithms to sample the space of closed clash-free conformations of a flexible loop. The seed sampling algorithm produces broadly distributed conformations. It is based on a novel prioritized constraint-satisfaction approach that interweaves the treatment of the clash avoidance and closure constraints. The deformation sampling algorithm uses these conformations as starting points to explore more finely certain regions of the space. It is based on the computation of the null space of the loop backbone at its current conformation. Tests show that these algorithms can handle efficiently loops ranging from 5 to 25 residues in length. We have successfully used early versions of these algorithms to interpret fuzzy regions in electron-density maps obtained from X-ray crystallography [19]. Our current and future work is aimed at applying them to other applications, in particular function-driven homology (where available functional information is used to limit the search for adequate loop conformations) and ligand-protein binding.

Acknowledgements: This work has been partially supported by NSF grant DMS-0443939. Peggy Yao was supported by a Bio-X graduate fellowship.

References

1. Bruccoleri, R.E. and Karplus, M. Conformational sampling using high temperature molecular dynamics. *Biopolymers* **29** (1990) 1847–1862.
2. Canutescu, A. and Dunbrack Jr., R. Cyclic coordinate descent: A robotics algorithm for protein loop closure, *Protein Sci.* **12** (2003) 963–972.
3. Canutescu, A., Shelenkov, A., and Dunbrack Jr., R. A graph theory algorithm for protein side-chain prediction, *Protein Sci.*,**12** (2003) 2001–2014.
4. Chang, K.S. and Khatib, O. Operational space dynamics: Efficient algorithm for modeling and control of branching mechanisms. *Proc. IEEE Int. Conf. on Robotics and Automation*, San Francisco, CA, (2000) pp. 850–856.
5. Cortes, J., Simeon, T., Renaud-Simeon, M., and Tran, V. Geometric algorithms for the conformational analysis of long protein loops, *J. Comp. Chem.*, **25** (2004) 956–967.
6. Coutsias, E.A., Soek, C., Jacobson, M.P., and Dill, K.A. A kinematic view of loop closure, *J. Comp. Chem.*, **25** (2004) 510–528.
7. Deane C.M. and Blundell T.L. A novel exhaustive search algorithm for predicting the conformation of polypeptide segments in proteins. *Proteins: Struc., Func., and Gene.* **40** (2000) 135–144.
8. DePristo, M.A., de Bakker, P.I.W., Lovell, S.C., and Blundell, T.L. Ab initio construction of polypeptide fragments: efficient generation of accurate, representative ensembles. *Proteins: Struc., Func., and Gene.* **51** (2003) 41–55.
9. Fiser, A., Do R.K.G., and Sali, A. Modeling of loops in protein structures. *Protein Sci.* **9** (2000) 1753–1773.
10. Golub, G. and van Loan, C. *Matrix Computations*, John Hopkins University Press, 3rd edition, 1996.
11. Halperin, D. and Overmars, M.H. Spheres, molecules and hidden surface removal. *Comp. Geom. Theory and App.*, **11** (1998) 83-102.
12. Jacobson, M.P., Pincus, D.L., Rapp, C.S., Day, T.J.F., Honig, B., Shaw, D.E., and Friesner, R.A. A hierarchical approach to all-atom protein loop prediction. *Proteins: Struc., Func., and Bioinf.*, **55** (2004) 351–367.
13. Kolodny, R., Guibas, L., Levitt, M., and Koehl, P. Inverse kinematics in biology: the protein loop closure problem. *Int. J. Robotics Research* **24** (2005) 151–163.
14. Okazaki, K., Koga, N., Takada, S., Onuchic, J.N., and Wolynes, P.G. Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: Structure-based molecular dynamics simulations. *PNAS* **103** (2006) 11844–11849.
15. Sauder, J.M. and Dunbrack Jr., R. Beyond genomic fold assignment: rational modeling of proteins in bilological systems, *J. Mol. Biol.*, **8** (2000) 296–306.
16. Shehu, A., Clementi, C., and Kavraki, L.E. Modeling Protein Conformational Ensembles: From Missing Loops to Equilibrium Fluctuations. *Proteins: Struc., Func., and Bioinf.* **65** (2006) 164–179.
17. Sousa, S.F., Fernandes, P.A., and Ramos, M.J. Protein-ligand docking: Current status and future challenges. *Proteins: Struc., Func., and Bioinf.*, **65** (2006) 15–26.
18. Tossato, C.E., Bindewald, E., Hesser, J., and Manner, R. A divide and conquer approach to fast loop modeling. *Protein Eng.* **15** (2002) 279–286.
19. van den Bedem, H., Lotan, I., Latombe, J.C., and Deacon, A. Real-space protein-model completion: an inverse-kinematic approach, *Acta Cryst.*, **D61** (2005) 2–13.
20. van Vlijmen, H.W.T. and Karplus, M. PDB-based protein loop prediction: parameters for selection and methods for optimization. *J. Mol. Biol.* **267** (1997) 975–1001.
21. Wedemeyer, W.J. and Scheraga, H.A. Exact analytical loop closure in proteins using polynomial equations. *J. Comp. Chem.*, **20** (1999) 819–844.