Enhanced conformational exploration of protein loops using a global parameterization of the backbone geometry

Prabal Ghosh[1 [0009−0004−3449−5811]]

Universite Cote d'Azur, Sophia Antipolis, France
prabal5ghosh@gmail.com

1 Summary

Protein loops are critical structural components that play diverse roles in protein function, including enzyme activity, membrane transport, immune response, and signal transduction. These loops often undergo conformational changes to facilitate substrate entry or product exit in enzymes, modulate binding affinity in antibodies, or trigger intracellular signaling in G-protein-coupled receptors. However, the structural diversity and flexibility of loops pose significant challenges for experimental determination, as evidenced by the high prevalence of missing regions in protein structures solved at low resolution. Approximately 83% of structures resolved at 2.0 A or worse exhibit missing regions, with 90% of these located in loops or unstructured regions.

Theoretical approaches to loop modeling are grounded in the concept of energy landscapes, which encompass structure, thermodynamics, and dynamics. While all-atom simulations can explore loop conformations, their computational cost has led to the development of simplified strategies. These strategies include continuous geometric transformations such as Crankshaft and Backrub, which deform loops by rotating rigid backbone segments. These methods effectively reproduce hinge-like motions observed in crystal structures but are limited in their ability to model more complex conformational changes. Loop closure techniques solve inverse problems to ensure that loop endpoints satisfy geometric constraints. These methods, inspired by robotics, involve pre- and post-rotation stages to sample loop conformations. Early approaches relied solely on dihedral angles, while modern methods incorporate both valence and dihedral angles for finer control over atomic displacements. An example is Tripeptide Loop Closure (TLC), which uses the six dihedral angles of three C_α carbons to achieve loop closure. Fragment-based methods stitch together high-resolution structures from the Protein Data Bank (PDB) to sample loop conformations. These methods suffer from combinatorial explosion and bias toward metastable conformations found in the PDB. Hybrid approaches combine structural data with loop closure techniques to improve prediction accuracy. Despite these advancements, predicting large-amplitude conformational changes or thermodynamic properties for long loops beyond 12 amino acids remains a significant challenge due to

the high dimensionality of loop conformational space and the subtle biophysical constraints involved.

The study introduces an innovative paradigm inspired by the Hit-and-Run (HAR) Markov Chain Monte Carlo technique [4] [1] . The approach decomposes a protein loop into tripeptide segments and employs a new mathematical characterization to define the necessary conditions for Tripeptide Loop Closure (TLC) to have valid solutions. The method operates in an angular space with a dimensionality proportional to the number of tripeptides in the loop, where constraints are imposed on specific regions of this space to ensure feasible loop conformations.

Traditional loop sampling methods fall into several categories, including continuous deformations, inverse kinematic closure techniques, fragment-based approaches, and hybrid strategies. While these methods have shown success in predicting loop conformations, they often struggle with large amplitude motions, high computational costs, or limited sampling diversity. In contrast, the proposed method leverages a global parameterization of the backbone geometry, allowing it to explore loop conformations more effectively without relying on precomputed databases or statistical biases. The approach leverages a global continuous parameterization of the loop's conformational space, based on the rigidity of peptide bodies comprising four atoms: C_α, C, N, and C_α [3]. This parameterization enables the coupling of individual TLC problems and provides initial conditions for solving them. The method is the first to exploit such a global parameterization, allowing it to handle loop lengths previously considered out of reach.

The proposed algorithm models a protein loop consisting of multiple tripeptides while adhering to a rigid geometry model, where bond lengths, valence angles, and peptide bond dihedral angles remain fixed. Each tripeptide is characterized by 12 internal angles, leading to an overall angular space of dimension 12m for m tripeptides. The approach leverages a Hit-and-Run (HAR) inspired sampling strategy to explore this high-dimensional conformational space efficiently [2] [5]. The region of interest in the angular space is defined by necessary conditions for Tripeptide Loop Closure (TLC) to admit solutions, ensuring that valid loop configurations are generated. The geometric model is based on a decomposition of the loop into peptide bodies and tripeptide cores, where each peptide bond is treated as a rigid unit. Loop conformations are generated by independently moving peptide bodies in three-dimensional space through rigid motions, followed by solving TLC problems for each tripeptide. TLC computes all valid geometries of a tripeptide while maintaining fixed endpoints and allowing only dihedral angles to vary. It provides up to 16 geometrically diverse solutions, ensuring substantial flexibility in sampling loop conformations.

The algorithm efficiently samples loop conformations by iterating through random trajectories in the angular space. Intersections between these trajectories and hyper-surfaces defining the necessary conditions for TLC determine valid configurations. The process follows a HAR-like scheme where new conformations are generated by selecting solutions to individual TLC problems and combining

them. The sampled configurations are further filtered to remove steric clashes, ensuring biologically relevant structures. Two algorithmic variations are introduced: the unmixed loop sampler (ULS) and the mixed loop sampler (MLS). The unmixed approach maintains fixed peptide body positions throughout the sampling process, while the mixed approach introduces an additional step that temporarily shortens the loop to improve flexibility. The ULS method involves sequentially moving peptide bodies, identifying valid configurations in the angular space, and solving TLC problems to generate new conformations. The MLS method introduces periodic modifications in loop length to alleviate constraints imposed by fixed peptide bodies, enabling a more flexible exploration of conformational space. Both approaches allow efficient sampling of large-amplitude loop conformational changes, addressing challenges associated with high-dimensional sampling and biophysical constraints. The algorithm provides a novel method for generating diverse and low-energy loop conformations, making it a promising tool for studying flexible protein loops beyond previously established length limitations.

The study examines various loop datasets, including PTPN9-MEG2, CCP-W191G, and CDR-H3-HIV, to assess conformational diversity, exploration capabilities, failure rates, and computational efficiency. The study highlights key differences between the proposed method and MoMA-LS, noting that the latter samples specific angles that influence loop closure and conformation diversity. This distinction impacts the space explored, potentially limiting or expanding the range of conformations generated. Three loop structures were analyzed. PTPN9-MEG2, a 12-amino acid loop, exhibited a clustering pattern with one outlier conformation, showing that the proposed method produced greater fluctuations in root mean square fluctuation (RMSF) values than MoMA-LS, suggesting improved conformational diversity. The CCP-W191G loop, slightly longer at 15 amino acids, displayed a similar trend, with the proposed method achieving higher RMSF values, indicating a broader conformational space exploration. For the longest loop, the 30-amino acid CDR-H3-HIV, the RMSF results demonstrated that the method could generate substantial variations, particularly in specific loop regions, while maintaining stability in others. The findings indicate that the proposed algorithm effectively explores diverse loop conformations and provides an alternative to MoMA-LS, particularly for long loops where traditional methods struggle. While MoMA-LS benefits from additional angular sampling, the new approach achieves broader structural variation, demonstrating its potential for structural biology and protein modeling applications.

Keywords: Protein Loops, Conformational Sampling, Global Parameterization, Tripeptide Loop Closure (TLC), Hit-and-Run (HAR) Sampling, Markov Chain Monte Carlo (MCMC); Molecular Dynamics (MD), Fragment-Based Sampling, Parameter-Free Algorithm, Loop Modeling

2 Methodological Strengths and Weaknesses

The proposed method for protein loop modeling represents a substantial improvement in structural biology by enabling efficient sampling of diverse loop conformations through a global parameterization of the backbone geometry.

One of the strongest aspects of the method is its ability to explore a broad conformational space without being biased toward pre-existing PDB-derived structures. Many loop modeling methods depend on statistical information from databases, limiting their applicability to novel protein sequences. The proposed approach, however, constructs loop conformations from first principles, allowing it to be used in a wide range of structural contexts. Additionally, the method is computationally more efficient than molecular dynamics (MD)-based simulations, which require extensive simulation time to generate meaningful conformational ensembles. The elimination of predefined energy functions also prevents the algorithm from being trapped in local minima, facilitating a more thorough exploration of possible loop geometries. Moreover, the built-in steric clash detection mechanism enhances physical realism by filtering out sterically infeasible conformations, ensuring that the generated loops are biophysically plausible.

Despite these advantages, the method has some notable limitations. The most significant drawback is the absence of explicit energy-based refinement. While the method generates geometrically valid loop conformations, it does not incorporate molecular mechanics force fields to assess energetic stability, meaning the generated structures may not always correspond to the lowest-energy states in a biological context. Integrating molecular mechanics-based scoring functions or refinement techniques such as Rosetta or MD relaxation could significantly enhance the accuracy of the predicted loops. Another limitation is the computational complexity associated with sampling in a high-dimensional space. Although the approach is more scalable than brute-force methods, computing intersections with hyper-surfaces in a 12m-dimensional space (where m is the number of tripeptides) can still be computationally demanding, particularly for very long loops. The efficiency of the method depends on the choice of initial conditions and parameter settings, which can influence the convergence speed and diversity of sampled conformations.

Another limitation is the lack of explicit thermodynamic ensemble sampling. While MD-based methods generate Boltzmann-weighted ensembles that provide insights into loop flexibility and free energy landscapes, this approach focuses solely on geometric feasibility without assigning probabilities to sampled conformations. As a result, it may not be as effective for thermodynamic studies of loop dynamics. Additionally, the method does not explicitly handle side-chain flexibility, which is an important factor in loop stability and function. Many biologically relevant interactions, such as hydrogen bonding and salt bridges, are influenced by side-chain dynamics, and their absence could limit the accuracy of predicted loop structures. Incorporating side-chain repacking techniques could improve the biological relevance of the sampled conformations.

Experimental validation remains another challenge. The method has been evaluated by comparing generated loops to crystallized structures, but direct

experimental confirmation using NMR spectroscopy, X-ray crystallography, or cryo-EM would be required to fully assess the accuracy of the predicted ensembles. This is particularly relevant for flexible loops, where static crystallographic data may not fully capture dynamic conformational variations. In terms of computational efficiency, the method is significantly faster than MD-based approaches. For short loops (12-15 residues), conformations can be generated within seconds to minutes, while medium-length loops (15-30 residues) require additional sampling iterations. However, for very long loops exceeding 30 residues, performance may degrade due to the increased number of tripeptide units, which increases the dimensionality of the sampling space. While the steric clash detection mechanism is computationally efficient, it introduces additional processing overhead when handling large loop ensembles.

3 Proposed Modifications and Enhancements

3.1 Enhancing Sampling in High-Dimensional Space

The proposed solution is to implement advanced sampling strategies like adaptive sampling, Replica Exchange Monte Carlo (REMC), or Metropolis-adjusted Langevin algorithms. These methods can improve convergence and reduce the computational overhead associated with high-dimensional conformational spaces. This approach increases efficiency for very long loops while maintaining extensive conformational diversity.

3.2 Incorporating Side-Chain Flexibility

The current method only models backbone motion and does not account for side-chain flexibility, which can significantly affect loop stability and function. To address this limitation, the solution involves integrating side-chain packing and optimization. By implementing side-chain repacking algorithms and using rotamer libraries, biophysically realistic side-chain conformations can be achieved. Combining backbone sampling with side-chain minimization ensures steric clashes are resolved at both backbone and side-chain levels. This approach improves the biological relevance of sampled loops and avoids steric clashes caused by backbone-side-chain interactions.

3.3 Handling Non-Canonical Loops and Post-Translational Modifications (PTMs)

The proposed method is primarily designed for standard protein loops and does not account for post-translational modifications (PTMs) such as disulfide bonds, glycosylation, or phosphorylation. To overcome this, the solution extends to non-canonical backbone constraints. Modifying the TLC constraints to accommodate non-standard backbone geometries and introducing additional steric and geometric constraints for chemically modified amino acids expands the method's applicability beyond standard proteins. This enhancement makes the method more versatile for real-world biological systems.

3.4 Hybrid Approach: Combining HAR Sampling with Deep Learning

Deep learning has shown promising results in protein modeling, but these models still struggle with flexible loops. A hybrid approach is proposed, combining HAR sampling with deep learning. The solution involves training a deep learning model, such as a Variational Autoencoder (VAE), to predict high-likelihood loop conformations. The HAR-based sampling method is then used as an exploration step, biased toward loop geometries predicted by the deep learning model. This hybrid pipeline reduces the number of sampling iterations, combines the flexibility of physics-based methods with the predictive power of deep learning, and makes the approach faster and more generalizable.

References

1. HCP Berbee, CGE Boender, AHG Rinnooy Ran, CL Scheffer, Robert L Smith, and Jan Telgen. Hit-and-run algorithms for the identification of nonredundant linear inequalities. *Mathematical Programming*, 37:184–207, 1987.
2. Marcos R Betancourt. Efficient monte carlo trial moves for polypeptide simulations. *The Journal of chemical physics*, 123(17), 2005.
3. Frederic Cazals. https://www.youtube.com/watch?v=3clmak-0so4ab$_{c}hannel$ = $centreinternationalderencontresmath$
4. Timothée O'Donnell and Frédéric Cazals. Enhanced conformational exploration of protein loops using a global parameterization of the backbone geometry. *Journal of Computational Chemistry*, 44(11):1094–1104, 2023.
5. Timothée O'Donnell, Charles H Robert, and Frédéric Cazals. Tripeptide loop closure: a detailed study of reconstructions based on ramachandran distributions. *Proteins: structure, function, and bioinformatics*, 90(3):858–868, 2022.