Supplementary Material

MoMA-LoopSampler: A web server to exhaustively sample protein loop conformations

Amélie Barozet[1], Kevin Molloy[2,3], Marc Vaisset[1], Christophe Zanon[1], Pierre Fauret[1], Thierry Siméon[1], and Juan Cortés[1,*]

[1]LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France
[2]Computer Science, James Madison University, Harrisonburg, VA. 22181, USA
[3]School of Biology, James Madison University, Harrisonburg, VA. 22181, USA

*To whom correspondence should be addressed.

S1 Results on TP-synthase loop

In order to showcase the performance of MoMA-LoopSampler, we submitted a set of jobs to the web server to generate conformational ensemble models of the loop 6 in thiamin phosphate synthase (TP-synthase), using the settings described in the manuscript and default values for the parameters. We generated ensembles containing 100, 500, 1000 and 5000 sampled states. Since MoMA-LoopSampler performs a stochastic sampling process, we repeated the test 10 times for each sampling size, aiming to verify the reliability of the method. All of the 40 jobs run successfully, and generated the requested number of samples. We measured Cα-RMSD between each sampled state of the loop from the unliganded structure (PDB ID: 1G4E; for which the loop is missing) and experimentally-observed loop conformations, stabilized by interaction with ligands (PDB IDs: 1G4S, 1G4T, 1G67, 1G69, 1G6C). Figure S1 presents the minimum values of these distances for each structure and sampling size, considering the 10 runs in each case. Results are very consistent. They show that the method is able to sample states relatively similar to experimentally observed ones (around 2 Å RMSD) with fewer than 500 iterations, and that the quality increases with the sampling size. Note that the performance of MoMA-LoopSampler, including its ability to widely sample the conformational space, was carefully assessed and presented in previous publications (Barozet *et al.*, 2019, 2021).

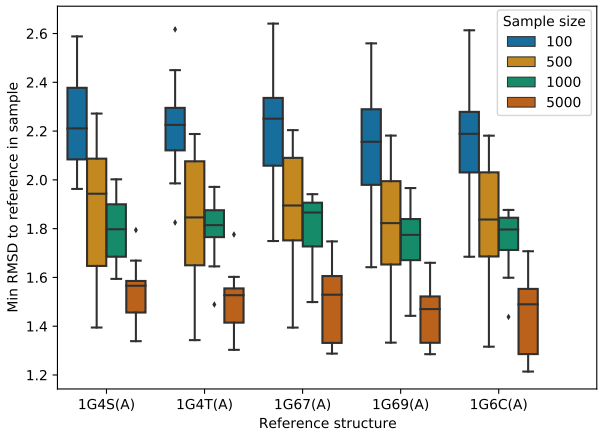

Fig. S1. Distribution of the minimum distance between sampled states of the loop 6 in TP-synthase and experimentally-determined conformations, for several sampling sizes.

S2 Another application scenario

In addition to the case of missing loops in crystallographic structures illustrated with the use-case of TP-synthase described in the manuscript, MoMA-LoopSampler can be applied to investigate flexible loops for which only a few conformations (or only one, in many cases) have been experimentally determined, but sampling is needed for a more global representation of the conformational space. We illustrate here this other application scenario using a streptavidin loop as an example. The structure of streptavidin (PDB ID: 3RY1) was determined for a tetrameric form of the protein (Le Trong *et al.*, 2011). The structural superposition of the four chains in the asymmetric unit shows some conformational variation in the loops connecting the eight β-strands that define the architecture of this protein. The most important differences are found for the loop involving residues 45-52, referred to as L3,4. This loops is supposed to act as a gate, opening and closing the access to the biotin binding site. A closed conformation of L3,4 is observed in the sub-unit A. Interestingly, this conformation is very similar to one observed in the streptavidin-biotin complex (PDB ID: 3RY2). The other three sub-units (B, C and D) contain an open conformation of the loop. L3,4 conformations in sub-units B and D are relatively similar to each other, whereas the conformation in sub-unit C is slightly different.

We applied the MoMA-LoopSampler web server to generate conformational ensemble models of L3,4 in streptavidin. This is illustrated in Figure S2. We used in this case sub-unit A as scaffold for loop sampling, but any of the four sub-units could be used. To define the anchor residues, we used the analysis presented by Le Trong *et al.* (2011) (as well as own analysis of the structures), which shows a significant fluctuation of the Cα atoms in the fragment from residue 45 to residue 52, both included. Reasoning in terms of backbone dihedral angles, this means that at least the ψ angle of residue 44 and the ϕ angle of residue 53 must be variable. Given that ϕ and ψ angles are correlated, these two residues have to be considered within the flexible fragment for loop sampling. Therefore, the anchor residues to be specified in the web interface are `43` and `54`. Since in this case the input PDB structure contains the coordinates a loop conformation, the user does not need to specify the sequence, which is directly extracted from the file. MoMA-LoopSampler also uses the bond lengths and bond angles computed from the input file, but bond torsion angles

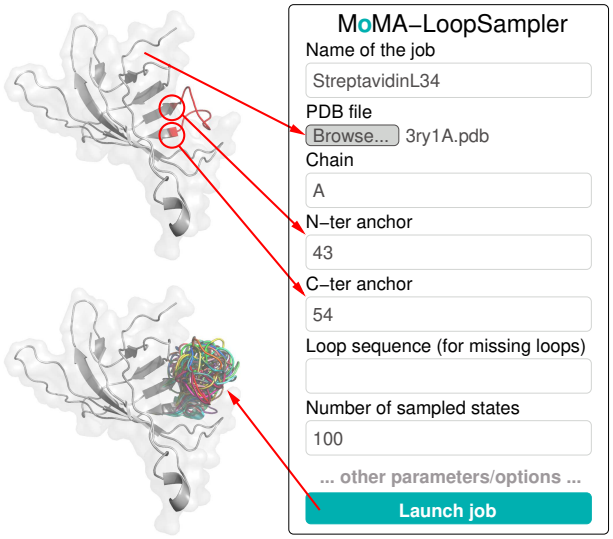

Fig. S2. Graphical explanation of the application of MoMA-LoopSampler to generate a conformational ensemble model of a flexible streptavidin loop (L3,4).

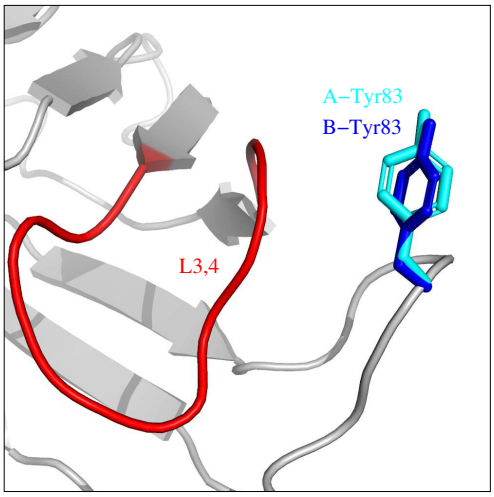

Fig. S4. Alternative locations of the side-chain of residue Try83 in sub-unit A of streptavidin. When the loop L3,4 (colored in red) if sampled from this scaffold, sampled states can not be very similar to the conformations in sub-units B and D if Try83 is placed in the first alternative location due to collisions with its side-chain.

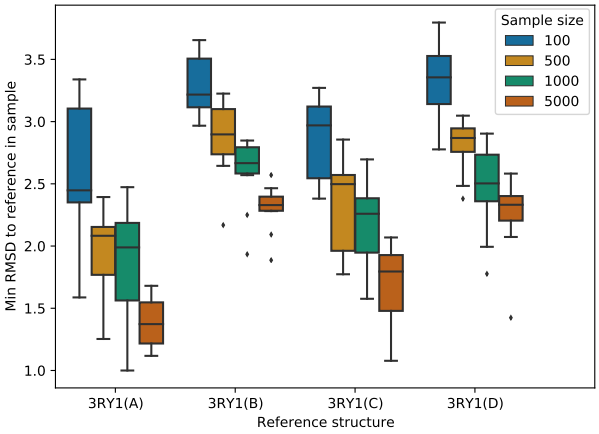

Fig. S3. Distribution of the minimum distance between sampled states of the L3,4 loop in streptavidin and experimentally-determined conformations, for several sampling sizes.

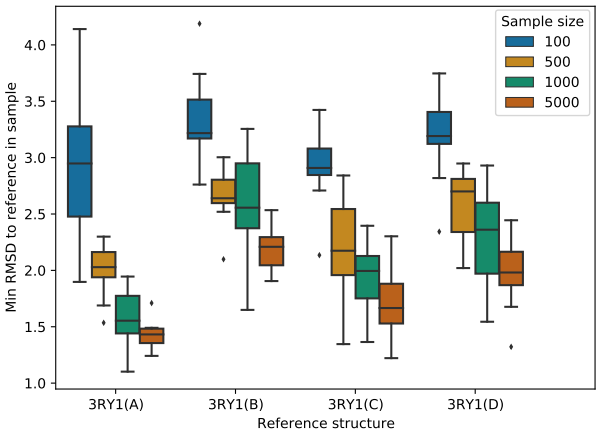

Fig. S5. Distribution of the minimum distance between sampled states of the L3,4 loop in streptavidin and experimentally-determined conformations, for several sampling sizes. In this case, Try83 was placed in the second alternative location defined in the input PDB file.

are not used (i.e. there is no bias towards the loop conformation in the crystallographic structure).

As in the case of TP-synthase presented in the previous section, we generated ensembles containing 100, 500, 1000 and 5000 sampled states, and repeated the process 10 times in each case. Results in terms of minimum Cα-RMSD between the sampled states and the loop conformations in the four sub-units are presented in Figure S3. These results show the ability of the method to sample loop conformations very similar to those of sub-units A and C. However, the results are less convincing for sub-units B and D. The analysis of the sampled states revealed that the loop could not approach the conformation in sub-chains B and D due to collision with loop L5,6, and in particular with the side-chain of residue Tyr83. Interestingly, the PDB file of streptavidin (3RY1) contains two alternative locations for this residue in sub-unit A. When this happens, MoMA-LoopSampler only considers the first one. We edited the file and repeated the calculations using the second altenative location for Tyr83. Results summarized in Figure S5 show that this minor change significantly improves the ability of the method to sample loop states similar to conformations in sub-units B and D.

TO BE MENTIONED: - loop sampling should be performed from the 4 scaffolds and considering possible alternative conformations of side-chains

- considering flexibility of the surrounding parts of the protein is important. although this is usually done in ulterior refinement stages of the sampled loop states, taking into account this flexibility within the global sampling process would be an interesting improvement. future work

References

Barozet, A., Molloy, K., Vaisset, M., Siméon, T., and Cortés, J. (2019). A reinforcement-learning-based approach to enhance exhaustive protein loop sampling. *Bioinformatics*, **36**(4), 1099–1106.

Barozet, A., Bianciotto, M., Vaisset, M., Siméon, T., Minoux, H., and Cortés, J. (2021). Protein loops with multiple metastable conformations: a challenge for sampling and scoring methods. *Proteins*, **89**(2), 218–231.

Le Trong, I., Wang, Z., Hyre, D. E., Lybrand, T. P., Stayton, P. S., and Stenkamp, R. E. (2011). Streptavidin and its biotin complex at atomic resolution. *Acta Crystallographica Section D*, **67**(9), 813–821.