Response to reviewer comments on the manuscript “Multisequence algorithm for coarse-grained biomolecular simulations: exploring the sequence-structure relationship of proteins” by A. Aina and S. Wallin.

We thank the two reviewers for their careful reading of our manuscript and insightful comments. In response, we have made several changes to the text which we believe have improved the presentation of our work.

Minor note: During this revision process, we noticed that the thermalization of our simulations was quoted as 10^6 MC steps in Table II when in fact it was 10^7 MC steps. This typo has been fixed.

Below follows our detailed responses to the reviewers’ comments.

Reviewer #1 (Comments to the Author):

In this work, Aina and Wallin propose a new generalized-ensemble algorithm for coarse-grained simulations of biomolecules, inspired on simulated tempering (ST), where the sequence of a biomolecule (e.g. a protein) becomes a dynamic parameter of the simulation (instead of the temperature as in ST). The authors then test their methodology in the context of protein folding, more precisely, to study the phenomenon of protein fold switching.

I enjoyed reading the paper which, despite providing a nice set of results, is well-written and presented. I recommend its publication in JCP after the following minor points are addressed.

Our response: Thank you.

1. This method is clearly suited to study the fundamental aspects of protein fold switching. However, this phenomenon is only addressed en passant in the article's introduction. In order to broaden the readership of the paper it would be nice if the authors could provide a more detailed description of this interesting process.

Our response: We agree. Two additional paragraphs on starting on page 1 has been added, focusing on the experimental evidence and evolutionary consequences of fold switching.

References added:
[29] Y. G. Chang, S. E. Cohen, C. Phong, W. K. Myers, Y. I. Kim, R. Tseng, J. Lin, L. Zhang, J. S. Boyd, Y. Lee, S. Kang, D. Lee, S. Li, R. D. Britt, M. J. Rust, S. S. Golden, and A. LiWang, Science 349, 324 (2015).
[30] S. Meier, P. R. Jensen, C. N. David, J. Chapman, T. W. Holstein, S. Grzesiek, and S. Ozbek, Curr Biol 17, 173 (2007).
[31] E. S. Kuloglu, D. R. McCaslin, J. L. Markley, and B. F. Volk- man, J Biol Chem 277, 17863 (2002).
[32] B. M. Burmann, S. H. Knauer, A. Sevostyanova, K. Schweimer, R. A. Mooney, R. Landick, I. Artsimovitch, and P. Ro¨sch, Cell 150, 291 (2012).
[33] X. Luo, Z. Tang, G. Xia, K. Wassmann, T. Matsumoto, J. Rizo, and H. Yu, Nat Struct Mol Biol 11, 338 (2004).
[39] C. A. Ramirez-Sarmiento, J. K. Noel, S. L. Valenzuela, and I. Artsimovitch, PLOS Comput Biol 11, e1004379 (2015).
[41] N. Hansen, J. R. Allison, F. H. Hodel, and W. F. van Gunsteren, Biochemistry 52, 4962 (2013).
[42] P. N. Bryan and J. Orban, Curr Opin Struct Biol 23, 314 (2013).

[44] L. L. Porter, Y. He, Y. Chen, J. Orban, and P. N. Bryan, Biophys J 108, 154 (2015).

Also, I was wondering if the authors could discuss in more detail the applicability/scope of this method. Can the authors provide concrete examples of other biomolecular processes where the method can be useful? Could it be used, e.g., to study protein folding and evolution?

Our reponse: Some potential application were already discussed in the last paragraph of the Discussion (page 8). We inserted the following passage to provide 3 additional examples:

"This opens up for various applications, such as exploring sequence effects on the conformational properties of disordered [57] or denatured [58] proteins, or aid in efforts to combine population genetics and molecular simulations. [59] While applied to proteins in this work, we note that the MS algorithm is equally valid for other bio-macromolecules, including DNA and RNA."

References added:
[59] R. K. Das and R. V. Pappu, Proc Natl Acad Sci USA 110, 13392 (2013).
[60] W. Meng, B. Luan, N. Lyle, R. V. Pappu, and D. P. Raleigh, Biochemistry 52, 2662 (2013).
[61] A. W. Serohijos and E. I. Shakhnovich, Curr Opin Struct Biol 26, 84 (2014).

Finally, could the authors provide a comparison of their method with that based on Replica-exchange as reported in . Chem. Theory Comput., 2016, 12 (11), pp 5656-5666?

Our reponse: We thank Reviewer #1 for bringing this article to our attention. A comparison between the two methods is now included in a short paragraph in the Discussion on page 8.

References added:
[56] J. H. Meinke and U. H. E. Hansmann, J Phys Condens Matter 19, 285215 (2007).
[57] W. Zhang and J. Chen, J Chem Theory Comput 10, 918 (2014).
[58] N. A. Bernhardt, W. Xi, W. Wang, and U. H. Hansmann, J Chem Theory Comput 12, 5656 (2016).

2. How exactly is the sequence update performed? The authors mention in page 2 that the sequence update is a mutational update? Does it mean that only one residue is changed or is it the whole sequence that is changed to a new one in the sequence ensemble?

Our reponse: The sequence update is now better explained in Section III.C. on page 3. The reference in section II.A. to a "mutational update" has been changed to the more general "sequence update". See also our response to Reviewer #2, second point.

3. What is the criteria used by the authors to update a sequence every 1000MC steps? (and the temperature in ST every 100 steps?)

Our reponse: Temperature updates are less computationally demanding than sequence updates, which require a full energy calculation and other "administrative" tasks related to changing the sequence on the fly in the simulation. To compensate, we made sequence update less frequent than sequence update. To better explain our choice, we added the following two sentences in section III.C on page 4:

"The computational cost for sequence updates is somewhat higher than for temperature updates, which does not require any energy calculations and is thus extremely rapid. For the purpose of comparing computational efficiencies between ST and MS, we therefore chose sequence updates to be slightly less frequent than temperature updates."

4. I don't understand the argument that "MS simulations escape local minima through visits to sequences are not as deep or even altogether absent" (Page 7). How did the authors arrive at this conclusion? Why is the process of sequence change more efficient at low T than that of temperature change?

Our reponse: This is an important point. To better explain how we view the process by which MS simulations escape local energy minima, we have extended and re-formulated the second paragraph in the Discussion on pages 7-8. The new passage reads:

“Therefore, rather than promoting escape from local minima by visits to higher T, as in ST [40,41] or temperature replica-exchange, [44] MS simulations must escape the minima occurring for one sequence through visits to other sequences. How this process works can be envisioned by considering an MS simulation that is visiting a sequence s and is trapped in a local minimum, requiring that a high free-energy barrier is overcome for escape. The trapped state could be, e.g., a compact β-sheet rich state with a particular register. Eventually, sequence updates will carry the simulation to other sequences sˆ while still remaining in the trapped state. However, the free energy barrier of this trapped state might be substantially lower for sˆ than for s, or the barrier may even be altogether absent if the trapped state is unstable for sˆ, leading to rapid escape from the minimum.”

Reviewer #2 (Comments to the Author):

The authors discuss a generalized-ensemble algorithm that realizes a walk through sequence space. The method is a variant of Simulated Tempering akin to the way that Hamilton Replica Exchange is a variant of Replica Exchange Sampling, with the Hamilitonians here differing by the sequences. For an all-atom model such an approach would be impractical as it would involve addition or subtraction of side chain degrees of freedom, leading low acceptance rates and/or loss of detailed balance (as in Resolution Exchange Sampling). In the present paper this problem is avoided by using a backbone-only coarse-grained model, and in this context the new method allows for simultaneous sampling of a range of sequences and , as an add-on, enhanced sampling. This is used by the authors to study conformation switching (helix-strand) by mutations.

Our reponse: We thank the reviewer for the nice summary and for putting our method in context Indeed, it can be seen as an extension of simulated tempering (as should be clear from section III. Theory) and a change in sequence amounts in some way to a change in the Hamiltonian.

This is an interesting idea, but I find its implementation a bit underwhelming. It is not clear to me how relevant the obtained results on paths in sequence space and (bi)stability of "transition sequences" are. The chosen coarse-grained model is extremely simplified with its three-letter code (glycine,hydrophobic, polar), and dos not allow comparison with other work. This drawback could have been avoided by using a more sophisticated backbone-only model that can describe all 20 amino acids. This would have allowed to look into real switching proteins, as a first step to study protein evolution.

Our reponse: The model used is simplified in the sense that it only has 3 amino acid types. However, its detailed backbone geometry allows it to spontaneously and robustly fold model sequences into realistic albeit idealized folded structures, which is a unique feature. Moreover, it is not clear which 20-letter model could have been used to study fold switching of real proteins in the large-scale manner carried out here. The subtle effect of fold switching is difficult to capture in physical models; as shown by van Gunsteren et al (Biochemistry 50, 10965-10973 2011), even state-of-the-art molecular dynamics force-fields may not correctly capture the mutation-triggered switch in global free energy between of the A and B domains of protein G. We are reasonably convinced that the present work represent the first time the folding thermodynamics of >1,000 sequences in an explicit-chain model have been simulated and compared, outside of simplified lattice models where conformations can be exactly enumerated.

My second concern is the move set. If I understand the authors correctly, the walk in sequence space is not by mutating a (randomly) chosen residue (which would be easy in a three-letter model), but by randomly jumping within an ensemble of pre-selected sequences. This move makes it difficult to interpret pathways and observed stabilities, and makes the system even more artificial.

Our reponse: Indeed, a new sequence is picked randomly from the set of (pre-selected) sequences after which the new state is subject to the appropriate accept/reject condition. We chose this way because it is general and guarantees that ergodicity is satisfied. An alternative would have been, as you point out, to carry out mutational updates in which (randomly) one position is selected and given a new amino acid type. In any case, a restriction on allowed sequences would be necessary to impose, as the full sequence space for even a 16-amino acid chain would be impossibly large to sample. In our opinion, there is therefore no major difference between the two ways of updating the sequence.

The native state stabilities, i.e., the probabilities P_IA, P_IB, etc, are equilibrium observables and totally independent of the move set as long as detailed balance is satisfied (which holds for moves). We therefore disagree that the choice of sequence or mutational move would make sequence pathways or stabilities difficult to interpret, or make our system artificial. See also our response to Reviwer #1, point 2.

The authors probably consider this paper as a proof-of-concept, and that is okay, but I think they should address the above points in a revision.

Our reponse: We hope and believe that both of your main points have been appropriately addressed.

As a minor point: I suggest to extend a bit the literature on generalized-ensemble. The idea and the term itself were introduced to protein science five years before Ref. 32.

Our reponse: We agree. The following sentence was added to section II.A on page 2:

"Generalized-ensemble methods that have been frequently used for biomolecular sim- ulations include simulated tempering (ST), [40,41] replica exchange [44] or parallel tempering, [45] and metadynamics. [46] "

In addition, we point the reader to a recent historical account of generalized ensemble methods by Bernd A Berg.

References added:

[48] B. A. Berg, Eur Phys J Spec Top 226, 551-565 (2017).
[50] U. H. E. Hansmann, Chem Phys Lett 281, 140 (1997).
[51] A. Laio and M. Parrinello, Proc Natl Acad Sci USA 99, 12562 (2002)

With such a revision the manuscript would in my opinion be suitable for publication in JCP.
Our reponse: Thank you.