Efficient conformational space exploration in *ab initio* protein folding simulation

Ahammed Ullah, Nasif Ahmed, Subrata Dey Pappu, Swakkhar Shatabda, A. Z. M. Dayem Ullah and M. Sohel Rahman

**Article citation details**

*R. Soc. open sci.* 2: 150238.  
http://dx.doi.org/10.1098/rsos.150238

**Review timeline**

- Original submission: 30 July 2014  
- 1st revised submission: 2 June 2015  
- 2nd revised submission: 18 July 2015  
- Final acceptance: 27 July 2015

Note: This article was transferred from another Royal Society journal without peer review.

**Review History**

RSOS-140204.R0 (Original submission)

**Review form: Reviewer 1 (Kathleen Steinhofel)**

Is the manuscript scientifically sound in its present form?  
Yes

Are the interpretations and conclusions justified by the results?  
Yes

Is the language acceptable?  
Yes

Is it clear how to access all supporting data?  
The tables and figure provided are clear.

Do you have any ethical concerns with this paper?  
No

© 2015 The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted use, provided the original author and source are credited.
Have you any concerns about statistical analyses in this paper?
No

Recommendation?
Accept with minor revision (please list in comments)

Comments to the Author(s)
The paper proposes an interesting way of evaluating protein conformations derived in ab initio folding simulations. The proposed energy function has been tested with a genetic algorithms using a discrete lattice model. The results presented outperform state-of-the-art algorithms on the tested sequences, and only missed the best RMSD on 3 out of the 12 instances. It is well written, with an easy to follow structure and clearly labelled tables and figures. Results are thoroughly compared to published methods with regard to free energy, however no comparisons appear to have been made with regard to the efficiency of the conformational space exploration, or computation time. Either a relevant experimental comparison is included or the title adapted to highlight the properties of the adapted energy function.

There are some very minor grammatical issues, for example - line 40 in the abstract should read “On the other hand, the HP model considers...”. Also, I would consider adding some form of shape notation to the lines on Figures 2 through 5. Printing the paper in black and white makes it difficult to determine which line is which. Due to the ordered groupings used, Figure 1 is easier to read when printed in black and white.

In the introduction, the authors state that most state of the art results are on small (less than 75 residue) proteins. Recent literature (including Rashid et al, 2013) use sequences from the same CASP9 competition the authors obtain sequences from, but of longer length (most common is up to length 279, or protein 3on7). Given that the authors suggest their algorithm should perform better on longer sequences, it is surprising they limited the length of sequences they tested.

The title of the paper suggests that the goal of their algorithm is to achieve “efficient” exploration. With respect to this, however, there are a number of issues. Firstly, the analysis of the algorithm appears to be focused on the structures obtained (in terms of free energy and similarity to native structures) as opposed to the efficiency of the algorithm. No comparisons are made to the efficiency of other state-of-the-art algorithms, even though published works have used “Energy Evaluations” - the number of times the energy of a structure is evaluated during the run, as a metric. Since there is no analysis, there is no comparison, efficiency-wise, to published material.

Furthermore, in section 2.7, the authors state that generic operators are used exhaustively, which comes with a “penalty of increased execution time”. There is no mention of the severity of this penalty, which, given the title of the paper, would be expected. Does this dramatically affect the runtime of the algorithm, or does it cause a minor decrease in efficiency.

In Section 3.1, the authors state that the ratio of probabilities for move selection is 0.4:0.24:0.2:0.15 and that this was crudely computed in pilot runs. There is no mention of how these values were chosen, or what property the authors were attempting to optimise through changing these values in the pilot runs.

The paper can be published with the above points addressed. This could be achieved through including an in-depth analysis of the efficiency of their proposed method in comparison to published literature. It would also be good to see experimental results on longer sequences (as used in other literature) to corroborate the idea that their algorithm works better with longer sequences.
Review form: Reviewer 2

Is the manuscript scientifically sound in its present form? Yes

Are the interpretations and conclusions justified by the results? No

Is the language acceptable? Yes

Is it clear how to access all supporting data? Yes

Do you have any ethical concerns with this paper? No

Have you any concerns about statistical analyses in this paper? Yes

Recommendation? Major revision is needed (please make suggestions in comments)

Comments to the Author(s)
Please refer to the attached file for my comments. (see Appendix A).

Decision letter (RSOS-140204)

11-Dec-2014

Dear Dr Rahman:

Manuscript ID RSOS-140204 entitled "Efficient Conformational Space Exploration in Ab Initio Protein Folding Simulation" which you submitted to Royal Society Open Science, has been reviewed. The comments from reviewers are included at the bottom of this letter.

In view of the criticisms of the reviewers, the manuscript has been rejected in its current form. However, a new manuscript may be submitted which takes into consideration these comments.

Please note that resubmitting your manuscript does not guarantee eventual acceptance, and that your resubmission will be subject to peer review before a decision is made.

You will be unable to make your revisions on the originally submitted version of your manuscript. Instead, revise your manuscript and upload the files via your author centre.

Once you have revised your manuscript, go to https://mc.manuscriptcentral.com/rsos and login to your Author Center. Click on "Manuscripts with Decisions," and then click on "Create a Resubmission" located next to the manuscript number. Then, follow the steps for resubmitting your manuscript.
Your resubmitted manuscript should be submitted by 10-Jun-2015. If you are unable to submit by this date please contact the Editorial Office.

I look forward to a resubmission.

Sincerely,
Emilie Aime
Senior Publishing Editor, Royal Society Open Science
openscience@royalsociety.org

Author's Response to Decision Letter for (RSOS-150238)

See Appendix B.

RSOS-150238.R1 (Revision)

Review form: Reviewer 1

Is the manuscript scientifically sound in its present form?
Yes

Are the interpretations and conclusions justified by the results?
Yes

Is the language acceptable?
Yes

Is it clear how to access all supporting data?
Yes

Do you have any ethical concerns with this paper?
No

Have you any concerns about statistical analyses in this paper?
No

Recommendation?
Accept with minor revision (please list in comments)

Comments to the Author(s)
See Attached Document (Appendix C).
Decision letter (RSOS-150238)

07-Jul-2015

Dear Dr Rahman

On behalf of the Editor, I am pleased to inform you that your Manuscript RSOS-150238 entitled "Efficient Conformational Space Exploration in Ab Initio Protein Folding Simulation" has been accepted for publication in Royal Society Open Science subject to minor revision in accordance with the referee suggestions. Please find the referees' comments at the end of this email.

The reviewers and Subject Editor have recommended publication, but also suggest some minor revisions to your manuscript. Therefore, I invite you to respond to the comments and revise your manuscript.

• Ethics statement
If your study uses humans or animals please include details of the ethical approval received, including the name of the committee that granted approval. For human studies please also detail whether informed consent was obtained. For field studies on animals please include details of all permissions, licences and/or approvals granted to carry out the fieldwork.

• Data accessibility
It is a condition of publication that all supporting data are made available either as supplementary information or preferably in a suitable permanent repository. The data accessibility section should state where the article's supporting data can be accessed. This section should also include details, where possible of where to access other relevant research materials such as statistical tools, protocols, software etc can be accessed. If the data has been deposited in an external repository this section should list the database, accession number and link to the DOI for all data from the article that has been made publicly available. Data sets that have been deposited in an external repository and have a DOI should also be appropriately cited in the manuscript and included in the reference list.

• Competing interests
Please declare any financial or non-financial competing interests, or state that you have no competing interests.

• Authors' contributions
All submissions, other than those with a single author, must include an Authors’ Contributions section which individually lists the specific contribution of each author. The list of Authors should meet all of the following criteria: 1) substantial contributions to conception and design, or acquisition of data, or analysis and interpretation of data; 2) drafting the article or revising it critically for important intellectual content; and 3) final approval of the version to be published.

All contributors who do not meet all of these criteria should be included in the acknowledgements.

We suggest the following format:
AB carried out the molecular lab work, participated in data analysis, carried out sequence alignments, participated in the design of the study and drafted the manuscript; CD carried out the statistical analyses; EF collected field data; GH conceived of the study, designed the study, coordinated the study and helped draft the manuscript. All authors gave final approval for publication.

• Acknowledgements
Please acknowledge anyone who contributed to the study but did not meet the authorship criteria.
Funding statement
Please list the source of funding for each author.

Because the schedule for publication is very tight, it is a condition of publication that you submit the revised version of your manuscript within 7 days (i.e. by the 16-Jul-2015). If you do not think you will be able to meet this date please let me know immediately.

To revise your manuscript, log into https://mc.manuscriptcentral.com/rsos and enter your Author Centre, where you will find your manuscript title listed under "Manuscripts with Decisions". Under "Actions," click on "Create a Revision." You will be unable to make your revisions on the originally submitted version of the manuscript. Instead, revise your manuscript and upload a new version through your Author Centre.

When submitting your revised manuscript, you will be able to respond to the comments made by the referees and upload a file "Response to Referees" in "Section 6 - File Upload". You can use this to document any changes you make to the original manuscript. In order to expedite the processing of the revised manuscript, please be as specific as possible in your response to the referees.

When uploading your revised files please make sure that you have:

1) A text file of the manuscript (tex, txt, rtf, docx or doc), references, tables (including captions) and figure captions. Do not upload a PDF as your "Main Document".
2) A separate electronic file of each figure (EPS or print-quality PDF preferred (either format should be produced directly from original creation package), or original software format)
3) Included a 100 word media summary of your paper when requested at submission. Please ensure you have entered correct contact details (email, institution and telephone) in your user account
4) Included the raw data to support the claims made in your paper. You can either include your data as electronic supplementary material or upload to a repository and include the relevant doi within your manuscript
5) Included your supplementary files in a format you are happy with (no line numbers, vancouver referencing, track changes removed etc) as these files will NOT be edited in production

Once again, thank you for submitting your manuscript to Royal Society Open Science and I look forward to receiving your revision. If you have any questions at all, please do not hesitate to get in touch.

Best wishes

Emilie Aime
Senior Publishing Editor
openscience@royalsociety.org

Author's Response to Decision Letter for (RSOS-150238)

See Appendix D.
Dear Editor,

In the manuscript titled “Efficient Conformational Space Exploration in Ab Initio Protein Folding Simulation”, the authors Ahammed, U. et al. proposed an improved strategy for ab initio protein folding based on non-uniform scaling version of 20x20 pairwise energy function (which is different from standard HP model). From their benchmark studies of twelve proteins, they claim that their newly developed method produces structures that are closer to the native structures compared to other state-of-the-art methodologies. They also claim that this energy function explores regions of conformational space that are unreachable using other methods.

In general, the manuscript is written in a straightforward manner but has a few grammatical/English errors. In this study, they included other interaction terms (in addition to hydrophobic-polar interaction included in HP model) that is crucial for the formation of protein tertiary structures. It is a technically sound study, in which systematic approaches were adopted and conclusive results were obtained. In my opinion, however, it will strengthen the manuscript further if the authors address the following concerns. The questions are the following:

1. In their manuscript (lines 15-23, page 19) they state “We notice that there is not much improvement in the RMSD values of the best structures from the initial random structures. This indicates the absence of strong positive correlation between the value of the energy function and RMSD measure. Similar findings are also reported in [35]. Details of the energy function values and corresponding RMSD values are given in supplementary Table S4.” If there is no strong correlation between the RMSD and energy, how do they explain the significant drop in energy between the initial and best structures?

2. Are the RMSDs reported in the units of Angstroms? Also, there is no noticeable improvement between Rashid et al. and their new findings. Especially for the larger proteins, Rashid et al. produces structures with better RMSDs (e.g., 3NBM, 3MQO, 3MRO). In that case, authors should point out how is their new strategy better than previous models?

3. In Table S3, they report the GW and BM energies of the system for runs with different time lengths. For smaller proteins, the GW energies obtained are not so different (comparing 30 mins, 1 hr and 2hrs runs), however, the differences are prominent not only in GW energies with 30 mins, 1hr and 2hrs but also much lower than the ones obtained from BM energy model.

-My first question is what is the unit of the energy reported here. Is it in reduced energy units? If so, how is it related to the standard units of energy? Without this comparison, it was difficult to interpret the meaning of these differences. And the authors did not explain it.

4. Authors mention that, they perform 50 independent runs. I am assuming all the folding simulations were performed several times. In that case, they should report standard deviation values in the relevant tables and figures for reference. Statistical data would provide more credibility in their proposed methodology.

5. Did they perform any direct comparison with other state-of-the-art techniques, e.g., I-Tasser, for these particular proteins? If so, they should add a comparative study to the manuscript for demonstrating how their methodology is superior.
6. As the authors acknowledged in the introduction, several other ab initio folding methods that are based on discrete molecular dynamics (Shirvanyants D. et al. J. Phys. Chem., 2012, 116, 8375-8382) and replica exchange molecular dynamics simulations (proposed by Karplus, Levitt and Warshel) are present in the literature. They should discuss (at least qualitatively) how are their energy function model better than these approaches both in terms of structure prediction and folding time efficiency. Why should the scientific community adopt this approach instead?

7. In the introduction the authors raise an important point that “because of the complex nature of the folding process and unknown contributing factors of the energy function, why and how do proteins adopt a specific structure remains one of the top outstanding issues in modern science”. Does their study shed any light to protein folding process? If so, it needs to be written explicitly.

8. In addition to hydrophobic-polar interactions, what other interaction terms (e.g., solvation energy etc.) were added to this energy function? It was not clear to me.

Minor corrections:

9. Only in the figure legend of Figure 4, BM is defined as real energy function. Please define it at the beginning of the text.

10. X and Y axis labels of the figures can be larger in size – difficult to read on printed version.

11. X-label of Figure 1 should be placed on the outside of the plot. That will make it more visible. Same comment is valid for Figure S1.

12. The conclusions and interpretation of the results need further work. In it’s current state the relevance/impact of the work is buried.

13. Authors need to pay close attention to the proper sentence constructions of the manuscript before next submission. Examples are the following:

   Line 18-20, page 11. “pull moves were first proposed as a complete moveset for lattices in [22] and later proved not to be completely reversible in [14].” Mentioning the name of the first author or group is preferable.

   Line 38-39, page15. “Among 99 types of potential contacts present in 1TCF …”. Should it be 1CTF?
Response to Review Comments:

Reviewers' Comments to Author:
Reviewer: 1

Comments to the Author(s)
The paper proposes an interesting way of evaluating protein conformations derived in ab initio folding simulations. The proposed energy function has been tested with a genetic algorithms using a discrete lattice model. The results presented outperform state-of-the-art algorithms on the tested sequences, and only missed the best RMSD on 3 out of the 12 instances. It is well written, with an easy to follow structure and clearly labelled tables and figures. Results are thoroughly compared to published methods with regard to free energy, however no comparisons appear to have been made with regard to the efficiency of the conformational space exploration, or computation time. Either a relevant experimental comparison is included or the title adapted to highlight the properties of the adapted energy function.

There are some very minor grammatical issues, for example - line 40 in the abstract should read "On the other hand, the HP model considers....". Also, I would consider adding some form of shape notation to the lines on Figures 2 through 5. Printing the paper in black and white makes it difficult to determine which line is which. Due to the ordered groupings used, Figure 1 is easier to read when printed in black and white.

In the introduction, the authors state that most state of the art results are on small (less than 75 residue) proteins. Recent literature (including Rashid et al, 2013) use sequences from the same CASP9 competition the authors obtain sequences from, but of longer length (most common is up to length 279, or protein 3on7). Given that the authors suggest their algorithm should perform better on longer sequences, it is surprising they limited the length of sequences they tested.

R.- The reviewer is right pointing out that recent literature including Rashid et al. 2014 have used longer sequences but those studies are on HP model. Our model uses 20x20 pairwise energy interaction matrix provided by Berrera et al. [5] and the state-of-the-art results on this model are of Shatabda et al [36] and Rashid et al [31]. Both Shatabda et al [36] and Rashid et al [31] used protein sequence
having length up to 160 amino acid residues. In our experiments we have also used sequence ranging 54-160 residues and have provided comparison with Rashid et al, 2013 as it manifests state-of-the-art results in terms of energy and RMSD. Nevertheless, in Sections 3.3 to 3.7 of the revised manuscript we have provided results of 5 longer sequences including protein 3on7. However we have only compared those results with different versions of our algorithm since there are no state-of-the-art results on these sequences in our equivalent model.

The title of the paper suggests that the goal of their algorithm is to achieve "efficient" exploration. With respect to this, however, there are a number of issues. Firstly, the analysis of the algorithm appears to be focused on the structures obtained (in terms of free energy and similarity to native structures) as opposed to the efficiency of the algorithm. No comparisons are made to the efficiency of other state-of-the-art algorithms, even though published works have used "Energy Evaluations" - the number of times the energy of a structure is evaluated during the run, as a metric. Since there is no analysis, there is no comparison, efficiency-wise, to published material.

R.- The reviewer is absolutely right pointing out that our focus was on structure obtained in terms of energy values and similarity to native structures. In the revised manuscript we have provided comparison to our proposed algorithm with Maher et al., 2014 [24] who have used "Energy Evaluations" metric (the number of times the energy of a structure is evaluated during a run) to demonstrate efficiency of algorithm. We thank the reviewer for stressing on the use of this metric since this is a machine independent metric and extremely useful for providing hardware and implementation independent comparison. But Maher et al. [24] have used several energy models i.e. HP, MJ, BM and several lattice models. They tested only 4 small sequences on the model we have used (FCC lattice and BM energy). Therefore, in Section 3.6 of the revised manuscript we have provided comparison considering these 4 sequences (out of 17) using this metric. Since there are no results on this metric for the other (and larger) 13 sequences in the literature, for those we have provided comparison of different versions of our algorithm using this metric (Section 3.6 of the revised manuscript).

Furthermore, in section 2.7, the authors state that generic operators are used exhaustively, which comes with a "penalty of increased execution time". There
is no mention of the severity of this penalty, which, given the title of the paper, would be expected. Does this dramatically affect the runtime of the algorithm, or does it cause a minor decrease in efficiency.

R.- Exhaustive operator usage is not a distinctive element of our algorithm. Several papers including state-of-the-art genetic algorithmic framework of Rashid et al, 2013 [31] used this approach. In fact all of the versions of our algorithm and the algorithm we compared to, i.e., Rashid et al. 2013 [31] have used this approach. Exhaustive operator usage intensifies the exploration and reduces randomness whereas traditional operator usage makes more diversification and less exploration. In effect there is a penalty of increased execution time in a single step of the algorithm but the penalty is amortized in the complete run of the algorithm by the reduction of randomness. Considering that efficiency is only visible in terms of energy values achieved in a time cut-off settings or number of times the energy of a structure evaluated, exhaustive operator usage does not decrease efficiency of the algorithm. In Section 3.7 of the revised manuscript we have provided justification of using exhaustive operator usage by comparing two versions of the same algorithm. One version uses operators exhaustively and other version does not and former has come out as superior.

In Section 3.1, the authors state that the ratio of probabilities for move selection is 0.4:0.24:0.2:0.15 and that this was crudely computed in pilot runs. There is no mention of how these values were chosen, or what property the authors were attempting to optimise through changing these values in the pilot runs.

R.- In Section 2.7 we have mentioned that we have empirically set a probability distribution for operator selection at each generation. From some initial runs on smaller sequences we have observed that after some generations some operators do not contribute to energy minimization anymore or contributes negligible amount wasting a significant number of iterations of the algorithm. So instead of using a uniform probability distribution we have used a non-uniform distribution to give bias to an operator which contributes largely to energy minimization. These values are our crude estimation for the empirical probability distribution of operator selection.
The paper can be published with the above points addressed. This could be achieved through including an in-depth analysis of the efficiency of their proposed method in comparison to published literature. It would also be good to see experimental results on longer sequences (as used in other literature) to corroborate the idea that their algorithm works better with longer sequences.

R.- In-depth analysis of the efficiency of the proposed method in comparison to published literature and experimental results on longer sequences along with analyses are added in the revised manuscript.

Finally, Shape notations are added in the figures as suggested by the reviewer. Other minor corrections suggested by the reviewer are addressed accordingly.
In the manuscript titled “Efficient Conformational Space Exploration in Ab Initio Protein Folding Simulation”, the authors Ahammed, U. et al. proposed an improved strategy for ab initio protein folding based on non-uniform scaling version of 20x20 pairwise energy function (which is different from standard HP model). From their benchmark studies of twelve proteins, they claim that their newly developed method produces structures that are closer to the native structures compared to other state-of-the-art methodologies. They also claim that this energy function explores regions of conformational space that are unreachable using other methods.

In general, the manuscript is written in a straightforward manner but has a few Grammatical/English errors. In this study, they included other interaction terms (in addition to hydrophobic-polar interaction included in HP model) that is crucial for the formation of protein tertiary structures. It is a technically sound study, in which systematic approaches were adopted and conclusive results were obtained. In my opinion, however, it will strengthen the manuscript further if the authors address the following concerns. The questions are the following:

1. In their manuscript (lines 15-23, page 19) they state “We notice that there is not much improvement in the RMSD values of the best structures from the initial random structures. This indicates the absence of strong positive correlation between the value of the energy function and RMSD measure. Similar findings are also reported in [35].

Details of the energy function values and corresponding RMSD values are given in supplementary Table S4.” If there is no strong correlation between the RMSD and energy, how do they explain the significant drop in energy between the initial and best structures?
The underlying basis of ab initio modeling is the Anfinsen's postulation that goes to say that native state of the protein structure should be its minimum energy state. Our algorithm adheres to this basis by finding the structure of minimum free energy. The algorithm is an energy minimization procedure which uses knowledge based energy function of Berrera et al., 2003 [5]. It has been observed that due to the limits of model assumption, knowledge based energy function works poorly in many cases when they are tested on the sequences outside of their knowledge domain. That means best energy structure and best RMSD structure can be different if RMSD measures and energy functions are poorly correlated. The inaccuracies of the energy function proposed by Berrera et al., 2003 [5] is revealed by our algorithm and other models' verification paper such as Shatabda et al., 2013 [35].

2. Are the RMSDs reported in the units of Angstroms? Also, there is no noticeable improvement between Rashid et al. and their new findings. Especially for the larger proteins, Rashid et al. produces structures with better RMSDs (e.g., 3NBM, 3MQO, 3MRO). In that case, authors should point out how is their new strategy better than previous models?

R.- Yes, the RMSDs reported are in the units of Angstrom. We apologize for not stating the unit in our manuscript earlier; however, it is now clearly mentioned in the revised version. There are two major bottlenecks of protein folding simulation - conformational space sampling and accuracy of energy function. Our study addresses the former bottleneck and it is evident that our proposed method successfully samples such low energy decoys of the conformational space that other methods failed to do. If the energy function provided by Berrera et al., 2003 [5] had strong positive correlation with RMSD measures then this would be reflected in our reported RMSD measures too. Since our strategy is loosely coupled with energy function we used [5], it is still relevant for addressing the bottleneck of conformational space sampling.

3. In Table S3, they report the GW and BM energies of the system for runs with different time lengths. For smaller proteins, the GW energies obtained are not so different (comparing 30 mins, 1 hr and 2hrs runs), however, the
differences are prominent not only in GW energies with 30 mins, 1hr and 2hrs but also much lower than the ones obtained from BM energy model.

-My first question is what is the unit of the energy reported here. Is it in reduced energy units? If so, how is it related to the standard units of energy? Without this comparison, it was difficult to interpret the meaning of these differences. And the authors did not explain it.

R.- These are not protein energies; these are differences of protein energies reported by two algorithms. That is why these values are much smaller relative to the energy of a protein conformation. Here difference is the subtraction of the energy found using GW guidance in the search algorithm and the energy found using BM guidance in the search algorithm. In Table S3, Difference=GW-BM. Both GW and BM guided algorithms’ energy values are reported using BM energy matrix of supplementary Table S1. Detailed analysis of these differences is given in Section 3.4.

4. Authors mention that, they perform 50 independent runs. I am assuming all the folding simulations were performed several times. In that case, they should report standard deviation values in the relevant tables and figures for reference. Statistical data would provide more credibility in their proposed methodology.

R.- Statistical data are provided in the supplementary Table S3 and S6 of the revised Manuscript.

5. Did they perform any direct comparison with other state-of-the-art techniques, e.g., ITasser, for these particular proteins? If so, they should add a comparative study to the manuscript for demonstrating how their methodology is superior.

R.- ITasser is a hierarchical modeling algorithm and uses fold recognition or threading methods. Our algorithm is based on ab initio modeling and we have provided comparison to the state-of-the-art results (Rashid et al, 2013 [31], Ullah et al, 2010 [38], Maher et al., 2014 [24]) on our model in Sections 3.3, 3.5 of the revised manuscript.
6. As the authors acknowledged in the introduction, several other ab initio folding methods that are based on discrete molecular dynamics (Shirvanyants D. et al. J. Phys. Chem., 2012, 116, 8375-8382) and replica exchange molecular dynamics simulations (proposed by Karplus, Levitt and Warshel) are present in the literature. They should discuss (at least qualitatively) how are their energy function model better than these approaches both in terms of structure prediction and folding time efficiency. Why should the scientific community adopt this approach instead?

R.- Molecular dynamics simulation uses all atom representation which requires massive computing power even for the small protein sequences, i.e., sequences having 50 or less amino acid residues. For longer sequences MD simulation is still unsuitable. That is why researchers started to use reduced representation where the whole amino acid is considered as a single sphere. In our study, we have used the reduced representation. Even in the reduced representation the complexity of the algorithm is so massive that we have to resort to discrete lattice model for folding simulation. In the reduced representation many lattice models and energy models have been used so far and we have provided all the comparative study to state-of-the-art results on our equivalent model (FCC lattice and energy function of Berrer et al., 2003 [5]). We believe and hope that the scientific community should adopt this approach because it samples better decoys, in terms of energy, than state-of-the-art approaches using the equivalent model.

7. In the introduction the authors raise an important point that “because of the complex nature of the folding process and unknown contributing factors of the energy function, why and how do proteins adopt a specific structure remains one of the top outstanding issues in modern science”. Does their study shed any light to protein folding process? If so, it needs to be written explicitly.

R.- Our study does not shed any light into the natural protein folding process. There is a long standing issue of overcoming different
energy barriers in the energy landscape during the mimicking of natural protein folding in computational algorithms. Sampling conformation space efficiently is the biggest barrier in computational protein folding and our study proposes a strategy to overcome this barrier of computational protein folding.

8. In addition to hydrophobic-polar interactions, what other interaction terms (e.g., solvation energy etc.) were added to this energy function? It was not clear to me.

   R.- We have not included any interaction terms; rather we have used an existing energy function of Berrera et al., 2003 [5]. Berrera et al., 2003 [5] give a 20x20 pairwise energy function in which various interactions among alpha-carbons, beta-carbons, side chains and backbone atoms have already been considered. The energy function of Berrera et al., 2003 [5] is derived from the potential of mean force theory and protein-solvent accessibility is estimated from residue solvent accessibility. For folding simulation previous studies have shown that only use of this matrix provided by Berrera et al., 2003 [5] forces search algorithm to converge frequently into local minima. Our contribution is scaling this raw energy matrix non-uniformly in a way that search algorithm can avoid local minima far efficiently than previous studies.

Minor corrections:

9. Only in the figure legend of Figure 4, BM is defined as real energy function. Please define it at the beginning of the text.

10. X and Y axis labels of the figures can be larger in size – difficult to read on printed version.

11. X-label of Figure 1 should be placed on the outside of the plot. That will make it more visible. Same comment is valid for Figure S1.
12. The conclusions and interpretation of the results need further work. In its current state the relevance/impact of the work is buried.

13. Authors need to pay close attention to the proper sentence constructions of the manuscript before next submission. Examples are the following:

Line 18-20, page 11. “pull moves were first proposed as a complete moveset for lattices in [22] and later proved not to be completely reversible in [14].”

Mentioning the name of the first author or group is preferable.

Line 38-39, page 15. “Among 99 types of potential contacts present in 1TCF ...”. Should it be 1CTF?

R.- All the minor corrections suggested by the reviewer are addressed accordingly in the revised manuscript.
Dear Editor,

In the revised version of the manuscript titled “Efficient Conformational Space Exploration in Ab Initio Protein Folding Simulation” as well as in the rebuttal letter, the authors have addressed all the questions/concerns satisfactorily.

The revised manuscript is well-written and proposes an efficient approach to study protein folding problems and is supported by the data provided in the manuscript. I have a few minor comments:

- Page 11 (line 4-8): consider sentence reconstruction. “So instead of ….largely to energy minimization”.
- Page 21 (line 10-11 and line 31-32): consider rephrasing “just better”. In my opinion, this is not scientific language.
- Page 23 (line 14): consider rephrasing “not much”.
- In some cases, commas are missing which makes the sentences hard to understand.

Overall, the research presented here will add knowledge to the protein folding literature and will contribute in further studies. Barring the minor corrections to be made, the manuscript is ready for publication in the Royal Society Open Science.
Response to Review Comments of RSOS-150238:

1. Review Comment:
Page 11 (line 4-8): consider sentence reconstruction. “So instead of...largely to energy minimization”.

Response:
We have revised that part as follows:

So instead of using a uniform probability distribution we have used a non-uniform distribution which is learned empirically from pilot runs. This is done to give bias to an operator which contributes largely to energy minimization.

=>

2. Review Comment:
Page 21 (line 10-11 and line 31-32): consider rephrasing “just better”. In my opinion, this is not scientific language.

Response:
In both cases we have replaced "just better" with "slightly lower (i.e., better)"

3. Review Comment:
Page 23 (line 14): consider rephrasing “not much”.

Response:
We have revised that part as follows:

Even though RMSD measures to native structures do not show much improvement these are not tied to our method;

=>

Even though RMSD measures to native structures do not show much improvement these are not tied to our method;

4. Review Comment:
In some cases, commas are missing which makes the sentences hard to understand.

Response:
We have updated some sentences accordingly.