Review of:
Ensembler: Enabling high-throughput molecular simulations at the superfamily scale due to
PLOS Computational Biology

Recommendation: Major Revisions

Summary:
In this manuscript, the authors detail the construction of a molecular modeling pipeline that combines previous and new tools, such as Modeller, together with open source bioinformatics modules, such as biopython, and available open source molecular dynamics algorithms to create a streamlined approach for creating a 3D structural model (as well multiple configurations of the same protein) from the amino acid sequence. Similar to standard homology modeling techniques, a FASTA sequence is compared to existing templates (through PDB) to provide the raw starting points for the modeling of the target protein structure and its ensemble of structures. As far as I can see, the majority of the initial part of this pipeline (i.e. before the molecular simulations) is very similar to other existing methods, such as MODELLER and ROSETTA (which are modeling tools used directly in this pipeline), as well as I-TASSER, which is another homology modeling approach that also has model refinement procedures built in. In my opinion, the main novelty of this work is that it provides the capability to create a target protein structure (or multiple) and directly parameterize, solvate, minimize and perform either explicit or implicit MD all within the same program. This will surely be an important tool that may help to extend the use of MD to fields outside of computational chemistry and push the limits of what can be simulated. That said, while there is a lot of potential for this tool, I believe there are a number of potential pitfalls that must be properly addressed by the authors before it can be published. While I think the overall idea of creating a streamlined tool to take a user through the construction of a model through simulation is commendable and important, there is potential danger in trying to automate and scale-up the parameterization and generation of protein structures and these pitfalls should at least be mentioned and explained in greater clarity in the text so that readers can be aware. The authors have provided an example application to illustrate the utility of this software by constructing an ensemble of structures for a family of protein kinases. I think that there are several points in this part of the text that also require further clarity and certain aspects of the chosen application that, if discussed in more detail, could perhaps better demonstrate the novelty and utility of this software. In other words, what this software actually brings to the table is still a bit unclear after reading through the provided example application. The overall intent and content of the article is explained clearly and succinctly. It is interesting and important to note that the actual tool is available on Git and the manuscript has been downloaded already by a number of people via the non-reviewed biorxiv online resource.

My main concern is that people without a strong background in molecular modeling and MD might be misled by certain parts of this manuscript and the potential pitfalls of automating the construction of simulation-ready models in this way are major. Before this manuscript can be accepted for publication, these points need to be very clearly laid out and addressed.

Major comments:

1) “However, it remains difficult for researchers to exploit the full variety of available protein sequence and structural data in simulation studies, largely due to limitations in software architecture.”

The way I understood this statement (and the one that follows it) is that the main limitation in exploring a number of confirmations of a single protein as well as confirmations of a large number of proteins is mainly due to the initial set-up of the model and its parameterization. If this was not the main point of this paragraph, I would recommend making the author’s point more clear here. If it is the main point, I don’t think that this is entirely true. While I agree that model creation, set-up and initial minimization for a single protein can be laborious (e.g. especially if there are non-standard amino acids or complex cofactors) and that current methods are not optimal for scaling up this effort to a large number of proteins, I do not think that it is impossible to do nor is it the main aspect limiting scale up. For example, there are already existing examples where previous efforts have simulated long-time MD on entire families of proteins, such as work done by Valarie Daggett’s lab, and I think that this work should be referenced at this point in the text in order to not be misleading.

Secondly, the parts of initial model set-up and parameterization that are the most laborious are not captured in the method presented here. The parameterization of non-standard residues and long-term MD simulations to stabilize and equilibrate unresolved regions of the protein (e.g. loops) are really the most time-consuming steps of this process. While the authors remediate, in part, the unresolved regions of proteins (as other homology modeling programs do) they do not really clarify whether this software is capable of addressing this level of complexity. They briefly describe these limitations at a later point in the text near the end, “Future directions,” but I believe that this statement quoted above is misleading, not clear and should be rewritten to reflect these issues. On the same lines, the statement “Furthermore, the automation of simulation setup provides an excellent opportunity to make concrete certain "best practices", such as the choice of simulation parameters.” is a bit too strong and should be removed or reworded, as this method does not appear to address some of the most variable parts of this practice, such as the parameterization and simulation of non-standard residues, metal ions, ligand complexes, etc.

According to most, what is more of an issue is the computational time required for not only a true relaxation of the structure (which is usually much longer than 100ps) as well as the actual sampling of phase space. GPU computing is accelerating and we can now reach much longer timescales at affordable rates, however this remains a commendable effort for even a subset of proteins. I understand that the authors are not recommending long-time MD due to the risk of inaccurate force fields, however, I think that these two issues should be clearly separated.

2) “The ability to fully exploit the large quantity of available protein sequence and structural data in biomolecular simulation studies could open up many interesting avenues for research, enabling the study of entire protein families or superfamilies within a single organism or across multiple organisms.”

This statement is also fairly misleading because I interpret this as saying we are not already exploiting the large quantity of available protein sequence and structural data. What I think is needed at some point in this manuscript is a clear definition of what kinds of tools already exist that do similar things and how this method takes our current status one step further. As I read through the text, I feel that fairly large claims are made which make it seem like this is the only tool available that takes an amino acid sequence, compares it to existing templates to construct models for molecular modeling applications. If the authors could put at least a paragraph in the text that compares the utility of this method to existing methods that are also doing similar modeling of target proteins, such as I-TASSER, which is the highest ranked protein structure prediction framework based on the last CASP experiments, that would be very helpful to the readers. If this method offers something entirely different than that of I-TASSER (e.g. up to model refinement) then that should also be made clear. After reading through the manuscript, I have understood that the defining point of this software is mainly the ability to perform explicit and implicit MD. Is this correct?

For the first stages of this pipeline (i.e. template searching, model construction and refinement), is this analogous to that of I-TASSER and similar methods? Using Ensembler, I understand that we can get a series of multiple structures that all may have different configurations representative of multiple conformations of the target protein, but can’t you also get this from other homology modeling tools, especially if one uses various templates as starting structures? Low-sequence identity templates are typically thrown out of most homology modeling frameworks.

3) “Here, we present the first steps toward bridging the gap between biomolecular simulation software and omics scale sequence and structural data: a fully automated open source framework for building simulation-ready protein models in multiple conformational substates scalable from single sequences to entire superfamilies.”

As mentioned above, I think it would be better to leave out “first steps” as other efforts have previously been made to scale up molecular simulations on the superfamily scale (through Valarie Daggett’s work). This should be changed and mentioning of her work and others similar should be put into the manuscript.

4) “This results in a total of almost 400,000 models, and we demonstrate that these provide wide-ranging coverage of known functionally relevant conformations”

After reading the application section which discusses the use of Ensembler in more detail, I think that there are a lot of details missing that would make this illustration more clear. Here

are the points that would be most helpful to me as a reader to address further:

- how many available templates are there for this family?
- how many of the templates already show multiple configurations of these proteins?
- to what extent does your method find relevant configurations **that are not already experimentally resolved**
- the main intent of this article seems to be the scale-up and generation of model simulations for entire families, but this application only appears to address two individual proteins of this family. Shouldn't this method be demonstrated for the entire family to show its utility?
- "This analysis indicates that higher sequence identity templates result in models with lower RMSDs,while templates with remote sequence identities result in larger RMSDs on average."

  This seems like a fairly obvious statement. What would be more useful to the reader would be a comparison of the "low sequence" set to existing structures and a further analysis that presents how likely these models are to be relevant. The authors could compare this set of proteins to existing proteins using MOLPROBITY or other programs. In other words, can you provide evidence that these higher RMSD conformations are useful? One possibility could be to run the same implicit MD protocol on a random subset of crystal structures and compare the distribution of energies to the lower RMSD group. It seems odd to me that the higher sequence identity subset is bimodal. Is this a convergence issue?
- " High sequence identity models (in opaque blue) tend to be quite structurally similar, with some variation in loops or changes in domain orientation"

  Isn't this also fairly obvious? Ensembler constructed templates from the available crystallographic models for these kinases and very little (100ps) MD was run on the models. Why should there be a difference in structure?
- The example shown here is not very convincing, because the two conformations picked up by Ensembler are already crystallized. What would be more convincing is if Ensember can pick up dual conformations in other proteins whose structures are only found through template-based homologies, thus, providing a clear objective for using a multi-structure strategy over other current methods. Can Ensembler do this?

5) "We further suggest that models with high target-template sequence identity are the most likely to represent native metastable states, while lower sequence identity models would aid in sampling of more distant regions of accessible phase space."

I have a number of concerns with a few statements similar to this one in different parts of the manuscript. In general, I agree with this statement to an extent- higher sequence identity in an experimental template would mean less modeling/modifications in the target structure. Therefore, the target structure would be much closer to the native fold of the original template than one that has lower sequence identity. The part I do not agree with, or at least is easily misunderstood when it is put in this way is that lower sequence identity may not lead

to biologically relevant regions of conformational phase space. As it is stated here, it seems certain that having low sequence identity will also be beneficial to the method so that you can add more structures to an ensemble for a given target protein and enhance the amount of phase space that is "sampled." However, this is one of the most dangerous pitfalls of this method- that if not done or analyzed in a correct way, conformations of a target structure taken from templates with low sequence identity may not be at all accurate representations of your protein and someone without a strong background may not be able to tell which conformations are relevant.

What could be a way to address this would be to implement some kind of warning to suggest potentially meaningless protein structures taken from templates with low sequence identity. At least if there is an explicit warning built into the software the users would be forewarned and could choose which confirmations to go forward with. Another suggestion would be to provide a benchmarking analysis that clearly shows the relationship between sequence identity and biologically relevant conformations presented in Ensembler. If such a benchmarking set would be available, I would be much more likely to use this tool, knowing what threshold I should set to only analyze conformations that are most relevant and meaningful. The authors have implemented an argument that takes in a threshold for sequence identity, however it is well known that the relationship between structure and sequence identity is not always clear. Some proteins with very low sequence identities still have very similar structures and so choosing a threshold is often not always a clear choice. I think that this feature will not necessarily be enough to address the concerns voiced above.

Furthermore, when taking conformations from templates with low sequence identity, it should be clearly stated that 100 ps of MD will more than likely not be enough to "correct" any non-native like folds or parts of the protein that have not been modeled correctly. The authors do mention that MSM methods would be capable of removing outliers that have non-native conformations, but to what degree is this successful? Isn't there a possibility of false positives that might not be properly captured by MSM method, and what are the chances that these can be identified and dealt with. If the authors could provide some statistics on how likely the MSM methods can remove outliers, and provide concrete examples, that would be extremely helpful to the reader.

Similarly, the authors claim that the models could be useful without any production simulation. But in order to be useful and relevant, the user should be fairly confident that the confirmations brought forward by this method are not irrelevent. If the authors could demonstrate this clearly, then these claims would be more believable.

6) It would be really helpful if comparisons to other well-known homology modeling techniques could be made for a subset of structures. For example, when available homology models are also available for a given target protein, can the authors simply compare the models generated by Ensembler to those generated from another program that also uses refinement procedures (markov-chain or similar). It would be helpful to see how similar or different the models generated by Ensembler are to

those that have been already ranked and scored in CASP. An RMSD alignment or a simple comparison of secondary structural content would be enough to say whether the structures generated are consistent with other well-known methods. Furthermore, the authors could use the PSQS score, PROCHECK, MOLPROBITY or something similiar to show how different these metrics are compared to other refined homology models. I understand that Ensembler is using the framework of MODELLER, which has also been ranked in CASP experiments, but how the overall choice of starting structures is similar or different to that of MODELLER, is not clear. In other words, how does choosing low sequence identity templates affect the predictability of MODELLER compared to other programs and its performance in CASP?

7) "While long-timescale unrestrained MD simulations (on the order of 100 us) have been found to be ineffective for recapitulating native-like conformations, possibly due to forcefield issues [34], even relatively short simulations can be useful for relaxing structural elements such as side chain orientation [33].

This is again another statement that should be changed/taken out in order to not mislead readers. When reading this, I interpreted this to say long-term, unconstrained MD is not effective in enhancing the predictive scope of these modeling frameworks, but short simulations can be useful. In some ways this is true, and the authors cite a DE Shaw paper that does show that unconstrained MD for long time scales show a good deal of drift from the native template structure. However, what is misleading about this statement is that this same paper points out that restrained long-time scale MD does dramatically improve model quality and provides substantial refinement to homology models, especially to regions of the model that are very different to the native fold. This statement, as is, is technically correct but does not really fully represent the paper that it cites and therefore should be changed to reflect this. The authors should clearly state this point that restrained MD can substantially improve model quality and state correctly the use of long-term MD in homology model refinement.

8) "Ensembler could exploit structural data from a set of homologous proteins to model in these molecules, although there will likely be a number of challenges to overcome in the design and implementation of such functionality."

What do the authors mean by this? Further explanation and discussion should be given here if this statement is left in. This, in my opinion is one of the hardest parts to parameterizing proteins and it should be made clear earlier on that Ensembler does not address this.

Minor comments:

1) line 57 - counterions is two words
2) See first comment above- please change line 63 to reflect points discussed in "major comments" to address: "Due to the laborious and manual nature of this process, simulation studies typically consider only one or a few proteins and starting

configurations."

3) For the PDBs that are filtered out based on number of resolved residues, does the method allow for selection of these PDBs if the target protein is only a portion of a protein chain (i.e. a domain)? Does the filtering take into account protein length? If not, can this be implemented?
4) For model refinement, wouldn't a monte-carlo based MD algorithm make more sense than molecular dynamics at such an early phase of target structure development?
5) For cases that fail in your pipeline due to the modeling of unresolved loops (from interatomic distance constraints), would it be helpful to implement a local minimization around the site that will be modified in order to attempt modeling on a relaxed structure?
6) Most people use the needle algorithm to align sequences, I am curious if there is a reason to use the pairwise2 algorithm in this case?
7) For the quickmodel feature, are the templates ranked? how are they ranked? how would the user know which is most representative?
8) The authors give the total times in CPU hours for using this method- Can the authors compare the timing of this method to others similar, such as I-TASSER? How different is the generation in a refined model before molecular simulation?