COVER LETTER

Submission ID: 129

Title: Comparative Visualization of Protein Secondary Structures

Date: July 17th, 2016

We would like to thank the anonymous reviewers for their helpful comments. We attempted to address all remarks in the revised version of the paper. In this letter, we give detailed responses to all reviewer comments and describe the corresponding changes in our manuscript.

Reviewer 3 - primary

1. *The case studies need to be significantly strengthened and better described. It needs to be clarified how the participants were selected and what their background was, along with the specific questions and tasks they were asked to address and carry out. Statements about the merit of the current tool relative to other tools need to mention the specific tools that are being compared. In addition, the authors should provide further clarification about the biological/structural insights claimed in the case study, as specifically requested by Reviewer 1. The authors should also address the question of whether and how the angles of features shown in the 2D view are useful to experts.*

   The presented case studies (now named usage scenarios) now contain more thorough introduction describing the importance of studying these cases. We also added the information about the group of biochemists participating on this research. Also all the remaining requirements are addressed and details about our improvements can be found in the corresponding responses to reviewer's comments below.

2. *The evaluation of the gap insertion algorithm used in the alignment needs to be better explained. This should include explanation of 1) what it means for the method to fail, 2) what test set was used, 3) the failure rate.*

   We added an example of the situation when the greedy algorithm fails (along with an image illustrating such a case). The test set used for the evaluation was selected in close cooperation with the biochemists and covered several examples of different proteins of the same family and several simulations of molecular dynamics (i.e., more examples of the scenarios we describe in the paper). The problem of determining the failure rate is discussed in detail in our reply to comment no. 36 below.

3. *The method of 2D encoding used in the juxtaposed and superposed views needs to be better described along with a discussion about scalability of the approach.*

   To describe the principle of the used 2D encoding, we added the following explanation:

"The basic element of both superposed and juxtaposed views is depicted in Figure 6. It demonstrates the case when two proteins chains are aligned. It consists of two main parts. The first part represents the sequential information about protein chain along with its secondary structures. Here we use three types of glyphs to determine between individual types of secondary structures. Arrows represent beta-sheets, spirals stand for alpha-helices, and lines represent coils. The length of the glyph corresponds to the size of the secondary structure (i.e., the number of amino acids forming the secondary structure). The reference chain is completely straightened. Then we take the information about the mutual spatial position between the secondary structures of the reference chain and the aligned chains. This determines the positioning of the glyphs representing the secondary structures of the aligned chains along the reference chain."

As for the scalability, we added the following text to the paper:

"The scalability of our approach highly depends on the input data and the similarity between the scrutinized chains. Theoretically there is no limit for the number of displayed chains, the only problem can be the readability of the resulting appearance. If the differences in the constitution and spatial orientation are small the approach can be used for dozens of solutions. On the other hand, when comparing significantly different solutions, the visualization will suffer from the occlusion problems even for a very small number of chains. This can be partially suppressed by the ability to interactively select only a desired subset of proteins and thus remove, e.g., those with the most significant differences."

4. *The motivation for the specific visualization tasks that are required for the comparative analysis of proteins envisioned in this work needs to be elaborated and these tasks should inform the discussion of related work, which is currently not well focused and does not provide a clear case for novelty.*

   We tried to improve this section by better describing the motivation and better targeting the description of related approaches.

5. *The paper is reasonably clear overall, but there are a number of small errors and confusing statements that could be clarified. The Conclusions section in particular needs some further editing.*

   The Conclusion section was reviewed and corrected.

6. *Lack of strong motivation for specific visualization tasks. The "requirements" for the design of the tool are introduced on page 4 with little motivation for the specific tasks. It would be better if these tasks were introduced earlier, before the Related Work section, and were more clearly justified.*

   We moved the tasks to the Background section and tried to strengthen the motivation behind our design rationale.

7. *Description of Related Work is too narrow and doesn't adequately address the issue of novelty of the current work. The lack of a clear statement of the visualization tasks to be performed seems to lead here*

to a Related Work section that isn't well focused on the issue of how the current work differs from previous approaches. It is difficult to assess from the description provided whether the proposed approach is truly novel, and if so, how specifically?

We tried to improve the Related work section by discussing more the limitations of the mentioned approaches. Basically there are two main problems of the existing approaches to the unfolding of protein structure to a 2D representation. The first problem is that many of the existing approaches are not suitable for comparison because the similarity between chains is not maintained by the unfolded representation. The second problem is that the existing 2D representations do not take into account the orientation of secondary structures. This information can play a crucial role, e.g., in assessing the protein reactivity with other molecules. Therefore, our approach is novel in the possibility to see the spatial differences between corresponding secondary structures and therefore compare the protein chains similarly to 3D but without the occlusion problems of 3D representation.

8. *The description of the gap insertion method is validated and the limitations of the approach lacks important details. On p. 6 it is stated that "The correctness was tested on many protein structures and only in some cases our greedy approach inserts a few unnecessary gaps into the chains." Later in the Conclusions: "Insertion algorithm which, due to its simplicity, can in some specific cases insert too much gaps." I think it would be helpful to show at least one example of how the method can fail and to give more specific information about the size of the test set used to evaluate the performance and the rate of failure. Even if these aren't rigorous tests, more specific information would be helpful.*

   The description was corrected to be more specific. We also added an example when the method fails. The test set consisted of approximately five cases for each of the two scenarios presented in the paper. The problem of determining the failure rate is addressed in our reply to comment no. 36.

9. *Methods for testing visualization are not adequately explained. Only biology background is given, but there are no specifics about what tasks users were asked to perform and how it was determine whether the objectives were attained. Furthermore, no information is given about the number or background of the "domain experts." Rather than "case studies," which implies a substantial degree of rigor, I think it is more accurate to describe what has been presented as "examples of use." Finally, although it is claimed that the tool enables users to perform visual comparisons "more quickly and intuitively than before," no information is given about what other tools are being referenced here. How does the information given by the present tool compare with that given by the Aquaria tool referred to in Figure 1?*

   We described the testing and evaluation phase more specifically. Also the information about the collaborating group of biochemists was added. The case studies were renamed to usage scenarios. Finally, we added more specific information about the limitations of the existing tools and the benefits of our proposed representation. When comparing our approach with the

Aquaria tool, our benefit lies in the possibility to encode the mutual spatial orientation of the secondary structures. This information can be highly beneficial (see reply to comment no. 28).

10. *It is not clear whether the 3D view can be zoomed to highlight differences in the regions being compared. This was perhaps just not mentioned in the examples shown, but it would seem to be an important feature, especially for performing comparisons of more than two structures. For example, in the example shown in Figure 10, the different colors used to distinguish among the structures being shown would seem to make it very difficult to see any highlighting of particular regions of the proteins. This would seem to limit the capability of the comparative analysis.*

    The 3D view can be zoomed and also the selected parts can be highlighted (as can be seen in Figure 8). These features are definitely very important for the comparison, however, as also the reviewer points out, when comparing more structures with different coloring the highlights have a limited benefit.
    So we added the information about the zooming possibility as well as the reference to Figure 8.

Minor points:

11. *Figure 1, right panel, page 2*
    Should label these structures and give some indication about what features are represented.

    We labeled the structures with their PDB identifiers and added more explanation about their selection and appearance.

12. *"It informs the user about the length and overall alignment of the compared structures", page 4*
    I don't see how the length information is conveyed - at least there are no numbers indicating position along the sequence.

    We tried to clarify this by correcting the sentence in the following way:
    "This part also gives the user the information about the relative length and global alignment of the compared chains. In other words, it aims to show the mutual positioning of the aligned chains which can be of different lengths."

13. *Highlight, page 5*
    This could be better explained. What is the meaning of the offset in the start positions of the elements?

    The offset corresponds to the shift between the carbon atoms of the first amino acids of the corresponding secondary structures. We clarified this in the text (see reply to comment no. 25).

14. *"The orientation of the secondary structures in the remaining aligned chains is adjusted according to the difference between the position and rotation of the corresponding secondary structures in the reference chain (see Figure 7)", page 5*

More detail about how the offset and rotation are determined would be helpful.

We added this information to the text (see reply to comment no. 25).

15. *"Secondary structures", page 6*
What algorithms are used to determine these? I guess this is a standard part of protein visualization software, but it would seem helpful to have a reference.

We use the DSSP algorithm to determine the secondary structures. We added the notion about this as well as the reference to the algorithm to the beginning of the Methodology section.

16. *Figure 8, caption. "…until both proteins are not processed", page 6*
Check the wording. I think "not" could be removed.

Corrected.

17. *In "Algorithm for Processing Molecular Dynamics, page 6*
It seems strange to insert gaps in aligning proteins that have the same sequence and it seems strange to align features that appear at different places in the sequence. I think it would help to at least explain the reasoning here.

It is true that the aligned proteins have the same sequence of amino acids but the problem we have to address is that these amino acids can form slightly different secondary structures. As mentioned in the paper, when the secondary structure, e.g., a helix, is too small (consisting of two or three amino acids), it can happen that it disappears in some time step. Therefore, we have to insert a gap to this chain to maintain the correspondence between amino acids further in the chains.

18. *lenghts -> lengths, page 6*

Corrected.

19. *End of Methodology Section, page 7*
Logical place to put reference to supplementary video, which is otherwise not referred to.

We added the reference.

20. *"...can merge when the protein structures change. This merge is largely caused by movements...", page 7*
I this refers to shifts in position in the structure because no dynamics are involved, but the language is a bit confusing.

We tried to clarify this and we also added more information about the importance of presence of channels in the P450 proteins.

21. *"…using the juxtaposed views illustrated in the paper", page 7*
It would help to refer to a specific figure here for comparison.

We added the reference to the corresponding Figure 2 in the paper by Cojocaru et al.

22. *"the the", page 7*

    Corrected.

23. *"The representation uses combines the advantages …", page 7*
    Sentence has numerous typos.

    Corrected.

24. *"When comparing many proteins or many time steps, the visualization starts to be too complex", page 8*
    Possible to be more specific about how many structures it is practical to compare?

    We tried to clarify this by adding the following text:
    "Another issue occurs when comparing many proteins or many time steps. Even when the compared chains are very similar, the superposed visualization will be too complex at some point (i.e., there will be so many overlapped chains that the visualization becomes unreadable)."

    The number of structures which is practical to compare is highly dependent on the differences between the compared chains. Therefore, we added the discussion about this issue to the end of the Results and Discussion section (see reply to comment no. 3).

Reviewer 1

---

25. *The approach presented in this paper first aligns multiple sequences with a greedy gap insertion algorithm and then visualizes the alignment with secondary structure ribbons and "adjusted orientations". This allows one to overlay multiple sequence representations with encoding regions of differing secondary structure composition. While the resulting figures look promising, I would have liked to see more information on how exactly angles for each of the arrows and ribbons are computed.*

    To clarify this, we added the following text:
    "To be more specific, the mutual position of two glyphs representing the corresponding secondary structures is calculated in the following way. It consists of two parts, the angle and the shift. Both are derived from the mutual position of the secondary structures in 3D space. To calculate the angle between two glyphs, we take two direction vectors of the secondary structures in 3D and compute the angle between them. This value is then projected to the angle between the glyphs in 2D. To determine the shift between glyphs, we calculate the shift between the direction vectors. In our solution we simply ignore the Z coordinate but in the future we could extend this by calculating the best viewing position to minimize the skew. The length of the glyphs is taken from 3D as well by simply computing the length from the start position (first carbon atom of a secondary structure) to the end position (last carbon atom of a secondary structure)."

26. *Also, I would like to see some discussion on the scalability of the approach: In the introduction, as the case studies show a maximum of 5 aligned structures.*

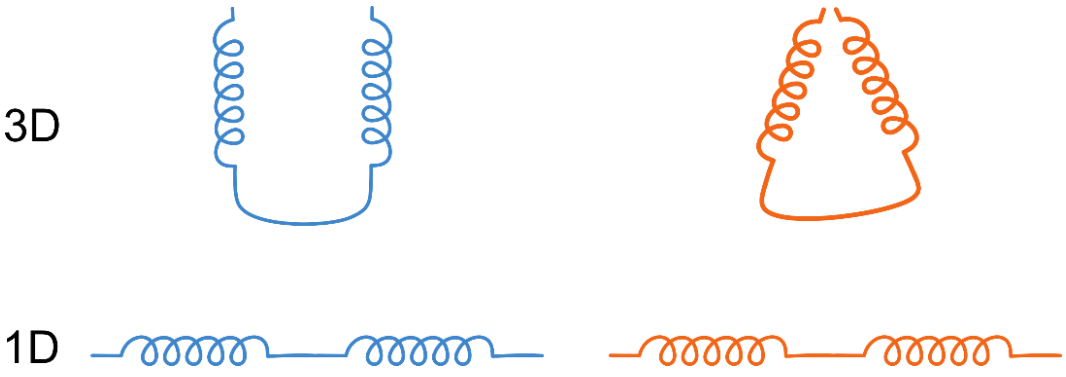
    The discussion about the scalability of our approach was added to the end of the Results and Discussion section (see reply to comment no. 3).

27. *The supplementary video is well done and is very useful to demonstrate interactivity and that the system was successfully implemented. I would love this to be available online with source code made available to the public.*

    The supplementary video is added to the final submission as well so we hope it will appear as a supplementary material at the conference web page. As for the release of the source code, we are prepared to send it on demand. And of course, we plan to release the online version as soon as possible.

28. *The first study is based on proteins with channels that 'can merge when the protein structures change,... largely caused by movements of specific secondary structures'. Is this merge visible in Figure 10? Are there any insights about the underlying structural changes in the protein family that can be inferred from the Figure, apart from the fact that some are well aligned (red box) and others are not (blue box)? Are angles useful to experts? If so, how are they using the angular information?*

    The information about the merging of channels in the cytochrome P450 family was introduced namely in order to stress the importance of studying these proteins and their structure. The channels detected in the selected proteins are not visible in our representation. Our solution is novel in that sense that the user can get the information about the changes in the void space of the proteins even prior to the computation and comparison of channels. The changes of channels are in fact a side effect of changes in positions of the secondary structures.
    The angular information is highly interesting namely because it determines the void space between the secondary structures which can be followed by a ligand. To demonstrate this, we added the following image to the paper. It shows the situation when the sequential information for the blue and orange chain is identical but the spatial position between the two helices determines the size of the entrance gorge to the structure.



Minor issues:

29. *As far as I know, the first work mentioning ribbon representations is: [1] J. S. Richardson, "The Anatomy and Taxonomy of Protein Structure," in Advances in Protein Chemistry, vol. 34, J. T. E. and F. M. R. C.B. Anfinsen, Ed. Academic Press, 1981, pp. 167–339.*

    We want to thank the reviewer for this valuable reference, we added it to the Background section where we introduce the cartoon (ribbon) representation.

Reviewer 2

Negatives:

30. *poor English*

    We tried to do our best to improve the readability and the language.

31. *better case studies or the details of how rhe case studies were chosen, and how the work was evaluated would help.*

    The case studies (now renamed to usage scenarios to follow the comment of reviewer no. 4) were selected by our collaborators from the field of biochemistry, namely protein engineering. We added the information about this group to the paper. Then we added also the information that the scenarios address the most typical problems of this group and the impact of the selected P450 family of proteins on living organisms.
    We believe that the selected scenarios, even being specific, are selected appropriately and target two significant problems in biochemistry. Moreover, the principle used in them can be applicable to other similar cases as well.

    We also added the information about the evaluation process and strengthened the Results and Discussion section by adding more details from the evaluation by the domain experts.

32. *addressing soem of the issues central to the approach like the gap issue*

    We added more details about the problems related to the gap insertion algorithm and how it differentiates from the optimal solution.

33. *it wasn;t clear in the paper if the algorithm creates too many gaps or gaps that are too big - one of the issues with the English.*

    We tried to clarify this by improving the description of the limitations of the greedy approach and its comparison with the optimal solution. Basically the problem is in the number of inserted gaps, not their size.

Reviewer 4

34. *I missed a clear description and discussion of the 2D encoding used in the juxtaposed and superimposed views as well as a discussion of the scalability of the approach.*

The discussion about the scalability of our approach was added to the end of the Results and Discussion section (see reply to comment no. 3).

35. *The manuscript frequently mentions domain experts, but it is not clear if they are co-authors or a different group of experts. Related to that, it might be more appropriate to refer to the "Case Studies" as "Usage Scenarios" if they were conducted by the authors themselves rather than by end users.*

We clarified this in the text by adding the following information:
"Our proposed visualization was tested on several usage scenarios proposed by the domain experts in biochemistry, namely in protein engineering. The group of experts consisted of one professor (head of the research group), two post-doc researchers, and two PhD students. All of them are active in designing modifications of protein structures."

36. *In the Gap Insertion Algorithm section, the authors say that they "propose a greedy algorithm with produces a sufficiently correct solution" but they don't define what they mean by "sufficiently". Also, further down they argue that "The correctness was tested on many protein structures and only in some cases our greedy approach inserts a few unnecessary gaps into the chains.". I assume this is correct, but the authors need to provide more information about this evaluation: 1. How was is done? 2. How are unnecessary gaps defined? 3. What was the gold standard? 4. How many are "some" cases?*

We added the argumentation about the sufficiency:
"To overcome this, we propose a greedy algorithm which produces sufficiently correct solution in a fraction of time of the optimal solution. The sufficiency means that the number of inserted gaps does not influence the understandability of the visualization. Our greedy solution was tested and evaluated by the biochemists who agreed that the sufficiency condition was met."

As for the correctness of the gap insertion algorithm and the insertion of unnecessary gaps, we added the following explanation:
"The correctness was tested on dozens of protein structures and in several cases our greedy approach inserted a few unnecessary gaps into the chains. The algorithm can insert these unnecessary (i.e., redundant) gaps because it has no prior knowledge about the secondary structures following the currently processed position. The optimal solution would create a hierarchical structure of all possible solutions and select that one with the smallest number of inserted gaps. The gap insertion problem is also tightly related to the definition of the correspondence between the compared secondary structures. In other words, we have to define when two secondary structures from different chains correspond to each other. In case when the secondary structures have the same constitution, the solution is trivial. However, in many cases only a portion of the secondary structures is the same. Then it is a complex problem which has to be solved in tight cooperation with the domain experts. Their expertise should help to

define a set of parameters which play a role in the similarity definition and these parameters should be incorporated into the gap insertion algorithm."

The gold standard would be defined by the optimal solution which we suggest in the paper. It would create a hierarchical representation of all possible solutions and then by traversing this representation we could find the solution with the smallest number of inserted gaps. However, this would be very time and computational intensive and the complexity would increase with the increasing number of compared chains. As we decided not to implement this robust solution, it is also hard to measure the number of unnecessary gaps inserted by our greedy approach. By simple observing the results we revealed on several places that some inserted gaps were redundant and other shift of the secondary structures would not insert them. But from discussions with the domain experts we concluded that this is not a serious limitation of our approach and that it does not decrease the understandability of the representation.

37. *In the last paragraph of the Methodology section contains a grammatical error: "except" needs to be replaced with "in addition".*

Corrected.