Review of Doctoral Dissertation

Title: Simulations of ligand binding processes in proteins

Candidate: Dheeraj Kumar Sarkar, Adam Mickiewicz University, Poznań, Poland

Dheeraj Kumar Sarkar presents a doctoral thesis summarising his pioneering scientific achievements by applying molecular simulations to understand how ligands enter protein binding pockets. Many proteins contain the active site buried in deep regions, gating binding to only specific ligands. Thus, understanding these processes is essential in understanding proteins' selectivity, promiscuity, or stability. Such knowledge is highly important in modern biotechnological research as it allows the rational design of novel proteins with altered functions.

Note: Since some thesis pages were not numbered, here, I will solely use raw page numbers from the electronic PDF version of the thesis.

The thesis is a synopsis of the published research articles accompanied by a short introduction and research summary. The candidate published results in two papers in impacted journals (Green Chemistry, Bioinformatics) and one submitted as a reprint to bioRxiv. The candidate's contribution to two papers is significant as he is the first or shared first author. The contribution to the second paper seems to be minor from the authors' order. Still, from the co-author's contribution statement (page 72), his contribution seems large enough to be eligible for the thesis. My only problem is that the use case 8 he wrote is in the supplementary files, which are not included directly in the thesis. He also contributed to two other papers (page 9). However, they do not seem to be affiliated with Adam Mickiewicz University. Thus, these articles should probably not be mentioned in the thesis and eventually be left for the candidate's *Curriculum Vitae* only.

The thesis is in English except for the Polish abstract and other bibliographical parts. The main text consists of an introduction to ligand transport processes in proteins (3 and a half pages), a summary of the doctoral research (6 and a half pages), conclusions and future perspectives (1 page), and references (5 pages, 57 items). Then, the three papers are included but without supplementary materials, followed by co-author contribution statements.

The introductory part is, on one hand, illustrative, providing a key summary of past knowledge related to ligand transports. Still, it is also too short, considering the complexity of the studied phenomena. While the thesis is mainly biologically oriented, it is also based on rather sophisticated computational simulations. An elaborate summary of these approaches and a critical evaluation of their abilities and limitations would be helpful in the thesis reading. Namely, in his research, the candidate employed Markov State Models (MSMs) to describe an intricate network of thermodynamic states and their connection. MSMs are quite complicated, and some general introductions similar to those provided in the review "Markov State Models: From an Art to a Science." (ref 29) would be highly beneficial. Followed by more specialized topics, such as how the bias introduced during the adaptive sampling is handled. I will return to this topic again in the question and comments section.

In the following section, the summary of three articles is presented. The summary is concise. However, the candidate focused more on describing what was done without further discussing the work in a broader scope. Thus, he missed the opportunity to explain many aspects which could not be discussed in the individual articles. Moreover, a broader discussion of data presented in the first and second articles would be more appropriate as they seem very similar. The third article studies the effect of hydrotrope agents (ATP and choline salicylate ionic liquid [IL]) on the activity of Cyt c. This work combines experimental and computer simulations. Computer simulations helped identify the protein's essential parts interacting with two hydrotrope agents.

Even though the thesis and articles were carefully prepared, I have found several errors and incorrect statements in the thesis. Some are rather cosmetic. Such as the RAMD abbreviation, used only once (pages 10 and 12), and CytC defined (page 10) but not used anymore. Missing comma in the list of methods (page 12). Some are, however, more severe. For example, the incorrect use of the terms period and frequency. If the sampling period increases, the sampling frequency decreases and vice versa. However, he seems to have interchanged these two terms in discussion on Page 18. Also, he incorrectly calls ATP and IL as hydrotrope solvents. At least for ATP, this is a rather exaggerated statement regarding its low concentrations in the setup (5 mM). Using terms such as hydrotrope agents or additives would be much better.

In the final chapter (page 21), he summarizes the achieved results and provides possible future perspectives. The candidate showed that molecular simulations are valuable assets to decipher interactions of ligands with proteins at the atomistic details. I appreciated his approach, which did not use only standard computer simulations but also sophisticated high-throughput molecular dynamics simulations. These simulations require non-trivial planning and skill with data and job organizations. On the other hand, he also used sophisticated post-processing approaches (MSM), which could be presented much better than in the current form.

In conclusion, Dheeraj Kumar Sarkar demonstrates that he can independently and successfully pursue modern academic research employing molecular modelling of biomolecular systems. The published work represents an original solution to scientific problems. Dheeraj Kumar Sarkar's doctoral thesis meets the requirements for doctoral theses in biological sciences. Therefore, I unequivocally recommend this scientific work for the defence.

In Brno, 13th December 2023

RNDr. Petr Kulhánek, Ph.D.
National Centre for Biomolecular Research
Masaryk University
Kamenice 5
625 00 Brno
Czech Republic

Additional comments and questions:

1) I am curious about two additional papers mentioned in the thesis (page 9). Can the candidate briefly comment on the current status of the submitted paper? In addition, the papers seem quite large. Thus, I assume that working on them significantly (negatively) impacted the work on the main thesis papers. Can the candidate also briefly comment on this?

2) While several computational methods suitable to describe ligand transport are summarized in the introduction, no clear selection criteria would show that the MSM method is the best. Can the candidate rationalize the selection he made?

The first paper:

3) What is the current state of the paper published in bioRxiv? Was it submitted to any peer-reviewed journal? If yes, which one, and what is its current status?

4) And here a tricky question comes. It follows from "Materials and Methods" that the Markov states were classified based on distances between protein Cα atoms and heavy atoms of the DBE ligand (1,2-dibromoethane). If this is the only input, then MSM requires all snapshots to be from the same unperturbed distribution. In my opinion, this is not the case in the presented work. First, each generated protein/ligand complex was solvated individually, leading to systems with different numbers of solvent molecules and, thus, different reference states. While each MD simulation can be considered unperturbed after equilibration, their joint distribution is not unperturbed anymore. However, this will probably be a minor problem. The major issue is that the ligand is positioned along tunnels. This is equivalent to biasing the systems. Since the bias is not compensated (reweighed), the joint distribution from many simulations differing in ligand positions cannot be unbiased anymore. I would ask the candidate to comment on these problems, how they affect the presented data and if they can be somehow compensated/corrected?

5) The studied process is formally an association/dissociation. Since the simulation setup is rarely close to the standard state definition, a non-negligible correction must be applied when results from MD simulations (non-standard states) and experiments (standard states) are compared. Can you evaluate this correction for your system and apply it to the comparison shown in Figure 4, panel C (page 41).

6) Can you explain why you use the NVT ensemble instead of the NPT one in the production simulations? Ligand binding into the protein cavity certainly leads to volume change, normally compensated in the NPT simulations. However, this will not happen under the NVT conditions, leading to distortions in the system and incorrect distributions.

7) DBE is a small and flexible ligand. Do you think the developed methodology is general enough to apply to more complex ligands? What could be the possible pitfalls?

The second paper:

8) What is the difference in the data sets used in the first and second (use case 8) articles?

The third paper:

9) What are the molar concentrations of protein, ATP, and IL in the individual simulation setups shown in Fig. 3 (page 58)?

10) Can a standard AMBER force field reasonably describe anionic species such as ATP or IL at high concentrations?