

**Name of the student:** Mr. Nikolas Tolar

**Thesis:** Algorithmic prediction of functional changes within protein sequence.

### Reviewer's statement

In his Bachelor thesis, Mr. Tolar developed a computational pipeline to identify the function-potential amino acid substitutions according to the protein sequence alignments.

Mr. Tolar shows a broad knowledge of the topic of his thesis. Information within the literature review is appropriate for the research topics studied but presented in an unorganized manner and with very few citations. The aims are clear and fulfilled by the presented results. The results are mostly observational and descriptive. Nevertheless, Mr. Tolar demonstrates promising expertise in bioinformatic analysis using python programming, but he did not illustrate the analysis pipeline clearly in some cases. The code was written in a very basic way without taking the advantage of using several biopython packages such as SeqIO, Bio.AlignIO; which could make the final code shorter and the computational process faster. However, I appreciate the organizing and clarity of the code.

His pipeline contributes to identifying proteins with amino acid substitutions in important positions which might reveal the unique epigenetic regulation of the Naked mole-rat (*Heterocephalus glaber*). Mr. Tolar set the basis of a computational pipeline suitable for a faster and batch analysis of Naked mole-rat (NMR) proteomic data. The output of this pipeline will broaden our understanding of NMR metabolism. Also, initiate future opportunities for experimental functional characterization of the candidate proteins with potential roles in NMR epigenetic regulation. The thesis satisfies the formal criteria. Information is correctly referenced, including the figures and tables. Figures are generally well designed but sometimes with an unclear message.

I have comments regarding the text:

Unfortunately, I have a big concern about the structure and the writing style of the thesis which put the reader in confusion to follow the workflow. The **introduction** is divided into several sections which some time unnecessary with so many non-eloquent and vague sentences while the **abstract** section is well written and informative. The **introduction** section needs to improve ignoring the division into many subtitles. The **methods** section is well written with again many non-eloquent sentences. The **results** section is well written and are correctly presented. Some results could be more understandable with better presentation. Finally, the **discussion** section is informative, and I appreciate that the pipeline future enhancement has been discussed as well.

Additionally, I have few comments regarding the formal side of the thesis, pointing out minor issues that could be improved:

1. Using more mammalian candidates in the analysis, should increase the possibility of identifying true amino acid substitutions.
2. Implement all three scripts in one pipeline, which will transform the pipeline to fully automated version.

3. Integrate a step of protein motifs search to the pipeline, which will help to identify the amino acid substitutions with potential functional impact.

3. I recommend implementing suitable Biopython and other packages such as SeqIO, Bio.AlignIO and argparse, to improve the code.

**Questions for the author:**

1. In **section 3. Materials and methods**: Why there are two different types of alignments sources? Isn't better to unify the data source since I suppose the applied aligning algorithm to the downloaded alignments from the Naked Mole-Rat Genome Resource database is unknown.
2. In **section 3. Materials and methods**: How are the alignments were downloaded from the Naked Mole-Rat Genome Resource database? since only a non-working link for the dN/dS ratios is available under the downloads sections.
3. In **section 5.2 Proteins**: Why proteins with one substitution would not show up as a top candidate in the STA output? And what the possible solution to fix this on the code level?
4. Which the best to define the functional differences between proteins? protein motifs differences or amino acid substitution scoring matrices.

Overall, I appropriate the time and effort of Mr. Tolar to develop such pipeline and analyze much data in a short time. Therefore, I recommend the thesis for defense with grade 1 (excellent).

In České Budějovice, 04.06.2021

Abdoallah Sharaf, Ph.D.