

Opponent's Assessment of the Bachelor's Thesis

Title: **Sequence features indicate a positive net charge is influential for the sub-cellular localization of proteins in organisms with complex plastids.**

Author: **Fuad Sarker Jeem**

Oponent: Mgr. Jan Michálek, Ph.D.

The bachelor thesis of Fuad Sarker Jeem focuses on searching for patterns characteristic of proteins targeted to complex plastids. Using sophisticated bioinformatic tools the author is trying to extend our knowledge of mechanisms of intracellular protein trafficking in a group of protozoans with complicated evolutionary history. Our present *in silico* predictions of targeting into complex plastids which are not always precise rely on N-terminal bi-partite presequences. Any addition of parameters that could increase the accuracy of predictions is desired. Therefore, as a user of these tools, I consider the topic as meaningfully chosen and appreciate the invested effort.

The introductory part explains the theoretical background behind mechanisms responsible for protein targeting, and its consequences for the following bioinformatic analysis and statistical testing. In this part, some chapters seem a bit oversimplified. For example, the evolutionary relationship of selected organisms could be explained more closely and supported with fresher references. The presented knowledge is sometimes tangled with logical flaws. Nevertheless, its general meaning and significance for the experimental part remain understandable.

The methodology represents the strong side of the thesis. The datasets used in the study are compiled of genomes of diatoms and chromerids from public databases. The logical strategy follows the protein-trafficking flow to plastids. Individual parts of protein sequences are virtually dissected and subjected to measurements of parameters that may influence their targeting signal. Outputs of pre-existing software tools are further analyzed using custom scripts and statistics. An extensive load of bioinformatic work is apparent. The developed analysis pipeline represents a decent foundation for future analyses of wider datasets comprising other groups of organisms with complex plastids. Presented results are well summarized and suitably displayed in comprehensive graphs. Finally, the brief discussion logically suggests the meaning of the results and puts them in the context of previously published studies. The theory of how charge differences in compartments attract peptides is very interesting.

The imbalance between slightly shortcoming theoretical and thorough experimental parts gives a feeling that the author could give the introduction a bit more time and attention. However, my overall impression is quite good and I believe this thesis meets the requirements for the defense.

Here, I would like to list some remarks and questions:

- Some formulations or words in the introductory part are inconveniently chosen which often disturbs the logic of given facts (sometimes the message is not completely clear) or can be misleading. But I understand that it is just a matter of practice.

e.g.:

“A secondary endosymbiosis occurs when a living cell engulfs a previously endosymbiotic eukaryote.”

“The diverse metabolic pathways of *T. pseudonana* and a small (32MB) genome size represent a large and predominant genus”

“...because it is enriched with genetic resources...”

“These organisms include non-photosynthetic, also photosynthetic organisms such as apicomplexan parasites.”

- Some of the used acronyms are not properly explained or linked to full words (NTT, BLS, SPase, SPP). Common terms such as ER are clear to most readers but you should always explain acronyms with their first use in the text (especially the less common ones).

- You extrapolate your results to “organisms with complex plastids” but the taxon sampling used in the study is composed strictly of diatoms and the two chromerids as an outgroup. As you mention in the introduction, organisms with complex plastids include many more groups. Some of them bearing also non-photosynthetic plastids. Be careful when you generalize your findings.

- The first paragraph in the chapter “2.5 Sequence features – Sequence length” does not tell much about sequence lengths.

Question 1. – Please, could you explain the paragraph?

- The statement “However, diatoms neither synthesize the ATP in plastids nor export the energy to other organelles.” is not true. What would be the purpose of photosynthetically active plastids in diatoms then?

Question 2. – Please, could you clarify this information?

- In the chapter about data cleansing you state: “The NCBI database is a great data source, however, the sequences are not screened and not quality controlled therefore, I have found multiple invalid sequences which are also deleted from the dataset.”

There is no doubt that some public databases contain junk, but this can mean many things (wrong annotation, taxonomy, the sequence itself, etc.).

Question 3. – Why did you discard the mentioned ‘invalid’ sequences? What was the particular problem in this case?