



The review of Ph.D. thesis  
„High-performance exploration and querying of selected multi-  
dimensional spaces in life sciences“  
by Miroslav Kratochvíl

The topic of the thesis is rather broad and, accordingly, the solved problems are rather diverse. While somebody may consider this as a weakness, we all know how it works with PhD topics in reality and, actually, I see this as candidate's advantage. The reason is that during his work the candidate successfully mastered a wide variety of computational techniques and applied them successfully to solve real problems from the biomedical domain. His scientific competence is demonstrated by rather long list of publications with clearly defined candidate contributions.

The thesis is written with a minimum number of mistakes and typos. The Commentary part is clear and it represents very understandable introduction to candidate scientific projects. The figures are well chosen and they sufficiently illustrate main thesis concepts. I have only a few points for the discussion:

#### **Sachem/IDSM**

1. While the substructure search is an important part of any cheminformatics service, even more critical is the similarity search. If I understand it well, Sachem does not contain similarity search capability. Or does it? How is the Similarity search implemented in the Sachem GUI that is available from the IDSM website? Would the proposed fingerprint lead to any improvement also in similarity search? Please, could the candidate comment if there are any plans to improve also similarity search virtual screening and would he proceed?
2. Sachem utilizes Apache Lucy, the library for full-text search. Considering that Lucy was not developed to index chemical data, I wonder what are the advantages and disadvantages of using Lucy for this type of data. Were there any serious problems in the “bending” of Lucy for chemical data?
3. If understand it well, the whole database must be re-indexed upon the addition of new compounds. For large databases, such as PubChem, this must be rather time-consuming step. Can the candidate comment on the problem of adding new compounds into the database?

#### **SOM visualization**

4. I wonder what was the reason for the rejection of EmbedSOM paper in “a biology-oriented journal”?
5. For the analysis of single-cell cytometry data, there apparently exist many visualization approaches. Could the candidate briefly summarize advantages and disadvantages of EmbedSOM, its main selling points?



6. I don't fully understand the relation between Folsom and EmbedSOM. Is the EmbedSOM the improved re-implementation of FlowSOM or does it call FlowSOM and processes further its outputs?
7. In single-cell cytometry data analysis, does the out-of-sample problem apply? I mean, if new data are added, does the SOM map has to be re-calculated or is it possible to project new data into the existing map without its update? I think the second is correct, but I would like to hear more detailed analysis on this question.

Summarizing, the candidate has performed a large amount of insightful research and obtained new original results. The dissertation work has been performed at a high scientific level and the candidate demonstrated his capability of critical thinking and of independent scientific work in this specific field. Judging by the thesis, the candidate merits the PhD degree and I clearly recommend its acceptance.

Doc. Daniel Svozil, Ph.D.

30. 11. 2020