UC RIVERSIDE
School of Medicine

Adam Godzik., Ph.D.
Professor of Biomedical Sciences
Bruce D. and Nancy B. Varner Presidential
Endowed Chair in Cancer Research
University of California Riverside
School of Medicine
adam.godzik@medsch.ucr.edu

Riverside, Aug 6, 2023

REVIEW of PhD dissertation of Kaustav Sengupta “The meta-network model of the complete human biomolecular interactome at the whole cell scale”.

This dissertation presents series of computational tools developed for several different problems (modeling of the 3D genome structure, adding directional information to the protein-protein interactions), bringing it together in Pfp-go, an integrated tool for predicting function of proteins. The dissertation has an impressive breadth, each of the problems being addressed is on the forefront of current biological research and in each of them Mr. Sengupta developed new computational tools, showing their usefulness and presented interesting, novel results obtained with their help. The most impressive aspect of the thesis is its broad scope and author’s appreciation of the need for integrating various levels of descriptions of biological systems to arrive at a full understanding of its function and misfunction in diseases. Each of the individual components of the dissertation, both the analysis and the associated tools, could be sufficient for a Ph.D. thesis!

While impressive, the dissertation as presented here is not without some issues:

- This dissertation is based on five out of about 13 papers co-authored by Mr. Sengupta. Almost half of them were published in various conference proceedings.

It is not clear if these were peer reviewed. While I assume that they are, making it clear in the dissertation would help to better evaluate Mr. Sengupta achievements.

- The authors claim novelty for all of the tools described in this thesis. While technically correct (these are newly developed algorithms), this sounds a bit like grandstanding. Honest comparison of the specific results to that achieved by other tools, for instance on the examples discussed in detail would provide a more balanced view of the authors accomplishments.
- The literature review and background is solid overall, but many choices of references appear to be a bit random and biased toward older literature. For instance (Wright and Dyson, 1999) is cited several times in various context (three-dimensional genomics has become a significant challenge in molecular biology, protein structure is encoded in its sequence), none of which corresponds to its real subject (intrinsically disordered proteins).
- Naming conventions for proteins are very inconsistent. For instance, the meta-network analysis of SNPs pointing to a narrow genetic location identifying specific proteins that are discussed in details, one which (page 49) is referred to by a UniProt name (P05538) – without any explanation nor citation to UniProt, but in a second example on a next page (page 50) the protein there is referred to by an Ensemble gene ID (ENSG00000174373) – again without explanation nor a reference to the Ensemble database. None of these proteins are described by their proper name.
- This type of confusion also shows up in other places. The author compares 3D spatial graph structures of two cancers: “Chronic Lymphatic Leukemia(CLL) and Brest Cancer (MCF7)”. CLL is a correct acronym for the Chronic Lymphatic Leukemia, but MCF7 is a cell line derived from the ER+ breast cancer. Confusion continues - MCF10A is cell line that doesn’t represent a normal breast tissue, it was derived from a highly fibrotic breast tissue. It is not clear where normal blood data was taken from. No references are given in this section.
- Evaluation of networks models by comparing their statistical properties to that of random models shows only that – that Excat models are not random, but it doesn’t

prove they are correct. Two examples given strongly suggests that they might be, but again these examples were chosen because they worked.

- The Pfp-go protein function prediction algorithm puts together several tools developed by the author. It shows impressive accuracy as compared to several other methods, as described in the authors paper – but this section of the dissertation looks like an introduction to the subject. I had to reach to a published paper to see the conclusion and discussion of the results. It is a disappointing, as it is sort of a grand finale. It would be nice to at least show the results from the paper, but even more, examples of some specific proteins where this tool really provided significant insights into their function. Such examples were presented for other tools, why not for the ultimate one?

Overall, the scientific accomplishments of Mr. Sengupta, as illustrated both by this dissertation and by his publication list, is impressive for somebody at his stage of the research carrier and clearly demonstrate that he is a well-trained scientist, ready to assume more independent role. The dissertation, while showing some problems discussed above, is one of the best I have reviewed in a long time.

Comparing Mr. Sengupta's dissertation to that at the institutions I routinely participate in the reviews of Ph.D. dissertations and Ph.D. committees, such as UC Riverside, UC San Diego, The Scripps Research Institute or Sanford Burnham Prebys, and I would classify this work as very good and meeting the requirements for doctoral dissertation. In the US doctoral degrees are not graded, they are pass or fail, but I would nominate it to be classified as "passed with distinction". I strongly encourage the committee to seek a analogous distinction at Politechnika Warszawska.

Sincerely,

Ad. Godi