UNIWERSYTET WARSZAWSKI

Instytut Informatyki
ul. Banacha 2
02–097 Warsaw
POLAND

prof. dr hab. Anna Gambin
Phone: +(48 22) 5544 212
Fax: +(48 22) 5544 400
e-mail: aniag@mimuw.edu.pl

Warsaw, 7.11.2018

Review of the doctoral dissertation

Title: MODELING AND ANALYSIS OF SPATIAL GENOME ORGANIZATION

Author: MGR PRZEMYSŁAW SZAŁAJ

The doctoral thesis deals with the key issue in the contemporary molecular biology, namely the spatial structure of the genome. The research outcomes presented in the dissertation include the computational framework that allow modeling the three-dimensional structure of chromatin based on the data obtained by the chromosome conformation capture (3C) technology and its more advanced varieties (Hi-C, ChIA-PET).

The results presented in the dissertation were published in three multi-author publications in very good journals. The first publication in *Genome Research* presents the 3D-GNOME tool developed by a team of researchers and the author of the dissertation shares first authorships with three others. Another article in *Nucleic Acids Research* presents the webservice allowing the on-line use of the 3D-GNOME modeling method, and the last one published jointly with the supervisor is a review paper presenting the issue of spatial genome organization from the point of view of molecular biology and experimental techniques.

Moreover, Author included a comprehensive summary of the dissertation introducing the main achievements and discussion of drawbacks and possible extensions of the method.

The 3D-GNOME framework provides the set of tools to build a three-dimensional model of genome structure based on data from the ChIA-PET experiment. It consists of three main modules: (i) data denoising and generation of heatmap summarizing the genomic interactions; (ii) building a spatial model of genome structure in two resolutions (medium

and high); (iii) visualization and structural analysis. The model built in (ii) is based on the 2D contact map created in module (i). Therefore 3D-GNOME may also use data from Hi-C technology to create the 3D model. However, the specificity of the ChIA-PET data is used in module (ii) to build high resolution model. For this reason, data from popular Hi-C experiments will provide only the medium resolution model. This solution can be considered a disadvantage, but the Authors assure that the benefits of using the ChIA-PET data outweigh this drawback. A key feature of 3D-GNOME is the hierarchical approach to designing the model in which successive levels of the hierarchy correspond directly to biological structures of the genome organization.

The most interesting and potentially useful in clinical applications module is the simulation of 3D model of genome structure. In particular it allows to predict to what extend the specific structural variant influences the chromatin structure. Structural mutations (deletions, insertions, or balanced translocations) are the cause of various disease syndromes (genomic disorders). Several independent approaches have been explored to simulate the model, depending on the level of resolution. Low resolution modeling based on the multidimensional scaling algorithm (MDS) turned out to be the most efficient computationally. Hierarchical models of higher resolution apply appropriately defined configuration energy and use Monte Carlo simulated annealing (SA) for optimization, i.e. finding the configuration minimizing the energy landscape.

The proposed method has been validated on several data sets. The 3D model has been built for whole genome data and also for the smaller example limited to single chromosome. The results obtained clearly demonstrate the quality and suitability of the 3D-GNOME tool in the analysis of the genome structure. Moreover, a dedicated web-service has been developed to enable on-line usage of the tool.

The dissertation is well written and satisfies the requirements for doctoral dissertations. The scientific issue considered in the work is interesting and important, and the solution proposed by the author is satisfactory. Wide and interdisciplinary knowledge is demonstrated by an easy to read review article. The co-authors' attestations attached to the thesis argued that a significant part of the results constitute the original contribution of the Author.

Below I list few comments concerning mainly the quality of presentation and the adequacy of the statistical tests used:

- Two-sample Kolmogorov Smirnov test was used to compare two segmentations of the genome (homogeneous and based on CTCF-mediated loops). The extremely low p-value obtained ($2.2 * 10^{-}16$) suggests that the distributions are significantly different. Unfortunately, such p-value is an artefact resulting from the specificity of the distributions studied and the large sample size.
- Similarly, it is problematic to use the Kolmogorov Smirnov test to verify the normality of the distribution. This should not be done especially if parameters of normal distribution are estimated from the sample and not specified *a priori*.
- The presentation of algorithms (especially details of simulation of the 3D model) is in my opinion too laconic, the formulation of the main procedures in the form of pseudocodes would significantly improve the accuracy and legibility.
- Due to the specificity of the subject, the summary contains many acronyms (TAD, ChIA-PET, MDS, CCD, etc.), which are not always defined at the first occurrence.

The aforementioned comments do not affect the assessment of the substantive results presented in the dissertation, which I classify as satisfactory. In summary, I find that the work meets the requirements for doctoral dissertations in current regulations and **I am applying to the Council of the Faculty of Medicine of the Medical University of Bialystok for the admission of Przemysław Szałaj to the defence of doctoral dissertation.**

prof. dr hab. Anna Gambin