Institute of Pharmacology
Polish Academy of Sciences
*Department of Medicinal Chemistry*

12 Smętna Street
31-343 Kraków, Poland
tel.: (+48) 12 662 32 20
fax: (+48) 12 637 45 00
www.if-pan.krakow.pl

General Secretary
tel.: (+48) 12 662 32 05
tel.: (+48) 12 662 32 49
e-mail: ifpan@if-pan.krakow.pl

Kraków, January 28, 2019

**Assessment**

**of the PhD thesis of Przemysław Szałaj MSc, entitled:**

**"Modeling and analysis of spatial genome organization"**

Supervised by professor Dariusz Plewczyński at the Faculty of Medicine of the Medical University of Bialystok. Part of the research was carried out at the Hasselt University, Belgium, as part of the Etiuda scholarship granted by the National Science Center in 2016.

Almost 16 years have passed since the international consortium announced the results of the Human Genome Project and published the data of 3 billion base-pairs of the human DNA. The continuous development in the fields of genetics and genetic technology makes DNA sequencing a routine part of scientific research nowadays. However, the nucleotide sequence identification is just the beginning. Considering that human genetic material is encoded in a DNA strand of two-meter long and packed in the cell nucleus of an average diameter of 10 micrometers, the knowledge how the DNA is folded inside the nucleus is no less important for the regulation of biological functions than the sequence itself. A number of high-throughput experimental methods such as Hi-C and ChiA-PET provide data for mapping chromatin interactions, which processed with the use of bioinformatics tools allow generation of models of 3D genome organization. And it is modeling of the organization of the genome that the PhD thesis of Mr. Szałaj is focused on. It is composed of three published papers, of which two are original publications, and the third is a review. The dissertation is in English and it is supplemented with a short Polish summary.

A brief introduction presented the research goal, i.e. the development of the original 3D-GNOME algorithm (acronym: 3-Dimensional Genome Modeling Engine) for the reconstruction of the spatial chromatin organization. The multiscale nature of the modeling process reflects the hierarchical biological structure of the genome. The algorithm used data from ChIA-PET experiments but can also be used to model the selected region of the genome, single chromosomes or the entire genome in high resolution.

In the next chapter, experimental methods the PhD thesis is based on are outlined and different levels of genome organization (mainly chromatin loops), domains and specific protein factors (like CTCF) influencing the architecture of chromatin are described. Subsequently, the correlation between the structure of the genome and the occurring diseases is shortly discussed. In Chapter 3, the Author presented different approaches of experimental

data analysis for establishing the 3D configuration of the chromatin chain and identification of topological domains, along with the review of known genome modeling methods.

General overview of the 3D-GNOME algorithm, its advantages and limitations are described in Chapter 4. The Author pointed out that the algorithm had been first presented in the work of Tang et al. (in the Cell journal, in 2015), in which the importance of specific factors (mainly CTCF binding protein and RNAPII polymerase) for spatial organization of chromatin and regulation of the transcription process were described. Mr. Szałaj, as a co-author of this publication (though not included in the thesis), was a member of a four-person team that developed a 3D-NOME (3D NucleoOme Modeling Engine) program to visualize data obtained from ChIA-PET experiments and to model the hierarchical 3D structure of the genome. At this point, the question arises: what was the reason for changing the name of the 3D-NOME algorithm to 3D-GNOME in the next publication, which is already an essential part of the doctoral thesis? Was it just a 'cosmetic change', or was it caused by the modification of the previously developed approach?

The most important features that distinguish 3D-GNOME against other available methods, are among others: close connection with structural units of the genome organization (i.e. with chromosome territories, topological domains, chromatin loops) which facilitates the interpretation of the obtained models in terms of their biological significance; genome modeling approach based on the ChIA-PET data (at the time of the manuscript publishing, the only software of its kind); orientation on CFCT loop motifs playing the important role in organizing chromatin architecture; multi-scale approach that allows to create high resolution whole genome models as well as modeling of individual genomic regions; high quality of the data enabling predictions relating to structural changes induced by genetic mutations.

In addition to the strengths of the 3D-GNOME algorithm, its limitations were also analyzed. These are mainly the difficulties in: modeling condensed regions of chromatin, modeling domains with a very high number of loops or regions with no loops (which may even prevent the use of the method). It has to be emphasized, that critical approach to own research is seldom found in doctoral theses and proves scientific maturity, reflected in a reliable assessment of the real possibilities of the developed method.

Copies of three publications forming the basis of the dissertation were provided in Chapter 5. Mr. Szałaj is the first author in all of them, and in a short introductory part, contribution of other co-authors to individual articles are specified.

The first research paper provided a detailed description of the 3D-GNOME algorithm. The software is composed of three basic modules enabling: 1) generation of 2D heat maps; 2) construction of 3D models of both individual chromosomes and the whole genome on two levels: using multidimensional scaling to obtain low-resolution 3D structures, or building 3D models at multiple scales by simulated annealing; 3) visualization of 3D models for structural analysis. In addition to the default functions, which were originally developed by the Author, the 3D GNOM contains additional tools that can be used independently. They include: an option for normalization

of the interaction frequency matrix, a method of modeling structures at the segment level and an interactive 3D viewer.

In the second paper, the web server for the 3D-GNOME platform was developed and described. The freely available tools allow the users (even those inexperienced) to perform simulations independently using their own experimental data or those available in dedicated databases.

The review paper entitled: "Three-dimensional organization and dynamics of the genome", recently published in the Cell Biology and Toxicology, summarized the current knowledge on the complexity of the genome structure and discussed input of experimental techniques for studying chromosomal contacts.

In Chapter 6, all the presented results and potential further development of the 3D-GNOME were discussed. Mr. Szałaj once again emphasized the distinctive features of the created software, and its successful usage in different computer analyzes. It is a pity, however, that examples of its application (for example, by scientists referencing to the dissertation papers published in 2016), have not been mentioned here and briefly characterized.

Plans related to the further optimization of the 3D-GNOME tools include above all, increasing versatility of the developed approach by using experimental data other than CHiA-PET (such as data from microscopy-based studies and novel single-cell approaches to guide the generation of individual DNA structures) and introducing dynamics simulations for studying chromatin conformations in time.

The Author also mentioned that because specific bioinformatics tools for data obtained in Hi-C method (which is the standard experimental technique to study genome organization) are available, so the 3D-GNOME may be less popular. Therefore, it would be beneficial to carry out additional analyses and comparisons using Hi-C and ChIA-PET data, to encourage new users to try the 3D-GNOME algorithm for testing. Therefore, given the fact that the 3D-GNOME software has been available since 2016, have the indicated modifications and benchmarks been carried out and has user feedback (if any) helped to introduce new features into 3D-GNOME?

In the final part of the doctoral thesis there was a list of publications and conferences, statements of co-authors and a list of publications cited in the thesis. The full list of publications co-authored by Mr. Szałaj included items that were not used in the PhD.

The material presented by Mr. Szałaj is an example of a bioinformatics solution, the application of which may contribute to a better understanding of factors influencing the organization of the genome and its mechanisms of functioning. The thesis is logically organized, carefully edited, and it is difficult to find major drawbacks. The only shortcomings that I noticed are the lack of numbering of chapters in the text of the dissertation (although they are numbered in the table of contents) and the lack of a list of abbreviations used. Few typographical errors, which happen even to perfectionists, are not worth mentioning. I do not have any substantive comments either. The research presented in the publications comprising the PhD thesis has been evaluated in detail by international reviewers, thus obtaining a positive opinion. It is important to note the very high impact factor of the journals publishing the thesis papers, which is 24.417. Moreover, according to the bibliometric summary, the combined

impact factor of six papers with the co-authorship of Mr. Szałaj is 91.265 points, which emphasizes the importance of the research.

To sum up, the topic taken up in the doctoral thesis entitled: "Modelling and analysis of spatial genome organization" is an important contribution to genomics, which is currently one of the fastest growing fields of science. The software developed by Przemysław Szałaj makes it possible to approximate the three-dimensional structure of the genome and find links between selected regions and biological function, and in the long run may also help to determine the probability of the occurrence of specific diseases.

Generally, the presented PhD thesis clearly meets the requirements set out in the Act on Academic Degrees in order to obtain a PhD degree, therefore I do recommend to the Scientific Council of the Faculty of Medicine of the Medical University of Białystok accepting the PhD thesis of Mr. Szałaj.