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

Warszawa, 3.11.2018

Recenzja rozprawy doktorskiej

Tytuł rozprawy: MODELING AND ANALYSIS OF SPATIAL GENOME ORGANIZATION

Autor rozprawy: MGR PRZEMYSŁAW SZAŁAJ

Rozprawa doktorska dotyczy kluczowego we współczesnej biologii molekularnej zagadnienia struktury przestrzennej genomu. W ramach badań przedstawionych w rozprawie został opracowany zestaw metod obliczeniowych umożliwiających modelowanie trójwymiarowej struktury chromatyny na podstawie danych uzyskanych dzięki technice *chromosome conformation capture (3C)* oraz jej bardziej zaawansowanych odmian (Hi-C, ChIA-PET).

Wyniki przedstawione w rozprawie zostały zaprezentowane w trzech publikacjach wieloautorskich opublikowanych w bardzo dobrych czasopismach. Pierwsza publikacja w prestiżowym czasopiśmie *Genome Research* przedstawia opracowane narzędzie 3D-GNOME, a Autor rozprawy jest w niej jednym z czterech pierwszych autorów. Kolejny artykuł w *Nucleic Acids Research* prezentuje webserwis pozwalający na używanie metody modelowania 3D-GNOME, a ostatni, opublikowany wspólnie z promotorem, jest pracą przeglądową prezentującą zagadnienie organizacji przestrzennej genomu od strony biologii molekularnej i technik eksperymentalnych.

Dodatkowo Autor przedłożył obszerny autoreferat wprowadzający w tematykę rozprawy, przedstawiający główne osiągnięcia i możliwe rozszerzenia metody.

Metoda 3D-GNOME umożliwia zbudowanie trójwymiarowego modelu genomu w oparciu o dane z eksperymentu ChIA-PET. Składa się z trzech głównych modułów: (i) odszumianie danych i podsumowanie interakcji genomowych w postaci dwuwymiarowej mapy

kontaktów; (ii) budowa modelu przestrzennego w dwóch rozdzielczościach; (iii) narzędzia do wizualizacji i analizy strukturalnej. Model budowany w (ii) wykorzystuje mapę kontaktów 2D powstałą w module (i), co oznacza, że 3D-GNOME może wykorzystywać dane z technologii Hi-C do modelowania. Jednakże specyfika danych ChIA-PET jest wykorzystywana w module (ii) do budowy modelu wysokiej rozdzielczości. Z tego powodu użycie danych z popularnych eksperymentów Hi-C pozwoli uzyskać model o dość niskiej rozdzielczości. Takie rozwiązanie można uznać za pewną wadę 3D-GNOME, jednak autorzy zapewniają, że korzyści z użycia danych ChIA-PET przewyższają tę niedogodność. Kluczową cechą 3D-GNOME jest hierarchiczne podejście do budowy modelu, w którym kolejne poziomy hierarchii są zbieżne ze strukturami biologicznymi organizacji genomu.

Najciekawszą częścią narzędzia jest model 3D, który pozwala na przewidywanie, jaki efekt może mieć określona mutacja strukturalna (często takie mutacje – delecje, insercje, lub zbalansowane translokacje są przyczyną zespołów chorobowych). Do symulacji modelu zostało wykorzystanych kilka niezależnych podejść, w zależności od stopnia rozdzielczości. Najbardziej efektywne jest modelowanie niskiej rozdzielczości oparte o algorytm skalowania wielowymiarowego (MDS). Hierarchiczne modele wyższej rozdzielczości wykorzystują odpowiednio zdefiniowaną energię konfiguracji i używają metod Monte Carlo i symulowanego wyżarzania (SA) do optymalizacji.

Działanie zaproponowanej metody zostało przetestowane dla kilku zbiorów danych – zarówno cało-genomowych, jak też ograniczonych do pojedynczych chromosomów. Uzyskane rezultaty ilustrują jakość i przydatność narzędzia 3D-GNOME w analizach struktury genomu. Dodatkowym atutem jest możliwość użycia narzędzia on-line poprzez dedykowany web-serwis (artykuł w NAR wyczerpująco omawia standardowe przypadki użycia serwisu).

Rozprawa zrobiła na mnie dobre wrażenie dowodząc opanowania przez Autora warsztatu badawczego w stopniu czyniącym zadość wymaganiom stawianym rozprawom doktorskim. Rozważane w pracy zagadnienie naukowe jest ciekawe i ważne, a zaproponowane przez Autora rozwiązanie satysfakcjonujące. Szerokiej i interdyscyplinarnej wiedzy dowodzi przystępnie napisany artykuł przeglądowy. Dołączone do rozprawy zaświadczenia współautorów wskazuję, że znaczna część wyników zawartych w rozprawie stanowi oryginalny dorobek Autora.

Poniżej przedstawiam uwagi, jakie nasunęły mi się w trakcie czytania rozprawy dotyczące głównie jakości prezentacji i adekwatności użytych testów statystycznych:

- Do porównania dwóch schematów podziału genomu na segmenty (jednorodnego i wykorzystującego pętle mediowane przez CTCF) użyto dwupróbkowego testu Kołmogorowa Smirnowa. Otrzymana ekstremalnie mała p-wartość ($2.2 * 10^{-}16$) sugeruje, że badane rozkłady są statystycznie istotnie różne. Niestety taka p-wartość jest artefaktem wynikającym ze specyfiki badanych rozkładów oraz dużego rozmiaru próbki.
- Podobnie problematyczne jest użycie testu Kołmogorowa Smirnowa do testowania normalności rozkładu. Nie powinno się tego czynić zwłaszcza jeśli parametry rozkładu normalnego są wyestymowane z próbki, a nie wyspecyfikowane *a priori*.
- Prezentacja algorytmów (zwłaszcza kolejnych kroków symulacji modelu 3D) jest w moim mniemaniu zbyt lakoniczna, sformułowanie głównych procedur w postaci pseudokodów znacznie poprawiłoby ścisłość i czytelność tego fragmentu pracy.
- Ze względu na specyfikę tematyki, autoreferat zawiera wiele akronimów (TAD, ChIA-PET, MDS, CCD, etc), które nie zawsze są zdefiniowane przy pierwszym wystąpieniu.

Wymienione powyżej uwagi nie wpływają na ocenę merytorycznych wyników zaprezentowanych w rozprawie, które klasyfikuję wysoko. Podsumowując stwierdzam, że recenzowana przeze mnie praca spełnia wymagania stawiane rozprawom doktorskim przez obowiązujące przepisy i **wnoszę do Rady Wydziału Lekarskiego Uniwersytetu Medycznego w Białymstoku o dopuszczenie mgr Przemysława Szałaja do dalszych etapów przewodu doktorskiego.**

prof. dr hab. Anna Gambin

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