

Laboratory of Bioinformatics and Computational Genomics LB:GO
Faculty of Mathematics and Information Science, Warsaw University of Technology
ul. Koszykowa 75, 00-662 Warsaw, Poland

Warszawa,
27.07.2023

Laboratory of Functional and Structural Genomics LFSG
Centre of New Technologies, University of Warsaw
Banacha 2c Street, 02-097 Warsaw, Poland

e-mail: Dariusz.Plewczynski@pw.edu.pl, www: <https://plewczynski-lab.org>

Warszawa, 27/07/2023

Prof. dr hab. Dariusz Plewczyński
Laboratorium Bioinformatyki i Genomiki Obliczeniowej,
Wydział Matematyki i Nauk Informatycznych,
Politechnika Warszawska
Laboratorium Genomiki Funkcjonalnej i Strukturalnej
Centrum Nowych Technologii
Uniwersytet Warszawski

RECENZJA rozprawy doktorskiej mgr inż. Macieja Migdała

*Modeling of transcription factors influence on gene expression based on data
obtained using next-generation sequencing methods*

Ukończonej w Laboratorium Genomiki Rozwoju Danio Pręgowanego,
Międzynarodowego Instytutu Biologii Molekularnej i Komórkowej w Warszawie

pod opieką
promotor dr hab. Cecylia Winaty

zgłoszonej do
Rady Dyscypliny Nauk Medycznych
na stopień doktora nauk medycznych i nauk o zdrowiu
w dyscyplinie nauki medyczne

Praca doktorska magistra Macieja Migdała dotyczy szerokiego spektrum zagadnień związanych z regulacją ekspresji genów, koncentrując się na różnych aspektach biologii komórkowej i molekularnej. Kluczowe tematy zawarte w tej pracy obejmują zagadnienia **bioinformatyki**, tj. stosowania narzędzi i metod bioinformatycznych do analizy danych genetycznych i epigenetycznych, w tym rozwój nowych narzędzi i modeli matematycznych; **genomiki**, tj. ogólne badanie sekwencji, struktury, funkcji i interakcji genów; oraz **epigenomiki**, tj. analizy modyfikacji chromatyny i innych zmian epigenetycznych, które wpływają na ekspresję genów.

Kandydat skupia swoją uwagę na zrozumieniu procesów *ekspresji i regulacji genów*, tj. analizie sposobów, w jaki geny są włączane i wyłączane, oraz mechanizmów kontrolujących te procesy. Jest to jedno z centralnych pytań **transkryptomiki**, obejmując studiowanie kompletnego zestawu RNA w komórce, co pomaga zrozumieć, jak i jakie geny są aktywowane w różnych warunkach. Dodatkowo Doktoranta interesuje rola czynników transkrypcyjnych, badanie białek, które kontrolują proces transkrypcji DNA na RNA, co stanowi kluczowy aspekt w regulacji ekspresji genów. Doktorant szczegółowo analizuje dane **RNA-seq** wykorzystując technologię sekwencjonowania następnej generacji RNA do analizy profilu ekspresji genów, oraz dane **ATAC-seq** również wykorzystujące technologię sekwencjonowania do identyfikacji otwartych regionów chromatyny, co jest kluczowe dla zrozumienia, jak geny są regulowane na poziomie struktury DNA.

Prace funkcjonalne obejmują badanie mechanizmów genetycznych i epigenetycznych, które kierują rozwojem serca. Badanie rozwoju komórek serca (kardiomiocytów) oraz roli różnych elementów regulacyjnych w ekspresji genów, w szczególności używając *Danio Pręgowanego* jako organizmu modelowego. Doktoranta interesuje przejście nabłonkowo-mezenchymalne (Epithelial-mesenchymal transition, EMT), badanie procesu, w którym komórki nabłonkowe tracą swoje właściwości i przekształcają się w komórki mezenchymalne, co jest kluczowe w wielu procesach biologicznych, w tym w naturalnym rozwoju, oraz w wielu chorobach. W szczególności w przypadku uszkodzenia wątroby, charakteryzacji odpowiedzi na uszkodzenie wątroby na poziomie transkryptomicznym i epigenomicznym, w tym identyfikacji kluczowych czynników transkrypcyjnych.

Praca doktorska Macieja Migdała stanowi kompleksowe i wieloaspektowe podejście do zrozumienia regulacji genów, łącząc różne technologie doświadczalne i metody bioinformatyczne, aby zbadać te procesy na różnych poziomach organizacji biologicznej. Dzięki temu może przyczynić się do głębszego zrozumienia, jak geny są kontrolowane i jak ich aktywność jest koordynowana w przypadku różnych procesów rozwojowych i chorobowych.

Przedstawiona mi do recenzji praca jest więc wynikiem analizy teoretycznej w paradygmacie bioinformatyki. Autorowi udało się twórczo połączyć metodologię teoretyczną z biologicznie istotnym problemem badawczym skupiającym się na głębszym zrozumieniu roli regulacji genów oraz rozwijaniu metod analizy danych transkryptomicznych i epigenomicznych. Podejście bazujące na metodach obliczeniowych i starannie przygotowanych danych doświadczalnych stanowi fundament nowoczesnej genomiki, czy ogólniej biologii molekularnej, a także wspomaga rozwój nowych algorytmów komputerowych, umożliwiających analizę wysoce zasumionych danych doświadczalnych, tworząc podstawowe instrumentarium badawcze umożliwiające udzielenie odpowiedzi na podstawowe problemy i wyzwania współczesnej genomiki.

W myśl wymagań Ustawy o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki z dnia 14 marca 2003 r. (Dz.U. 2017 poz. 1789, z późn. zm.) oraz Rozporządzenia Ministra Nauki i Szkolnictwa Wyższego z dnia 19 stycznia 2018 r. w sprawie szczegółowego trybu i warunków przeprowadzania czynności w przewodzie doktorskim, w postępowaniu habilitacyjnym oraz w postępowaniu o nadanie tytułu profesora (Dz.U. 2018 poz. 261), jak również art. 187 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (Dz. U. z 2021 r. poz. 478, 619, 1630), przedmiotem mojej oceny jest oryginalność rozwiązanego problemu naukowego, ogólna wiedza teoretyczna Kandydata w dziedzinie nauk medycznych i bioinformatyki, a także umiejętność samodzielnego prowadzenia pracy naukowej.

Przedmiot mojej oceny, czyli rozprawa doktorska, jest w pełni zgodna z warunkami określonymi w powyższych Ustawach, zaś całkowity dorobek publikacyjny kandydata moim zdaniem przekracza zwyczajowe wymagania. Autor prezentuje oryginalność sformułowania unikalnego problemu naukowego oraz jego

rozwiązania przy użyciu zaawansowanych metod informatycznych, szeroką wiedzę teoretyczną z zakresu biologii molekularnej i informatyki, a także umiejętność prowadzenia wytrwałej i twórczej pracy naukowej. Praca Doktoranta oraz szereg publikacji powstałych we współpracy zewnętrznej wykazuje również dużą samodzielność badawczą, mimo istotnego wsparcia ze strony Pani promotor. Nie mam żadnych poważnych uwag do rozprawy doktorskiej, jakoś przedstawienia metodologii i wyników badań, dorobku publikacyjnego Kandydata. Całość rozprawy w pełni zasługuje na zaakceptowanie przez Radę Dyscypliny.

Rozprawa doktorska Pana mgr Macieja Migdała została przygotowana w Laboratorium Genomiki Rozwoju Danio Pręgowanego, Międzynarodowego Instytutu Biologii Molekularnej i Komórkowej w Warszawie pod kierownictwem dr hab. Cecylii Winaty.

Praca liczy sobie 66 stron (pomijając strony 67-92 zawierające deklaracje współautorów co do wkładu Doktoranta), w tym załączone trzy artykuły składające się na rozprawę doktorską, oraz odniesienia do artykułów istotnych dla zrozumienia dziedziny badawczej, jak również rozwiązywanego problemu. Brakuje listy innych prac opublikowanych przez Kandydata ze współautorami w punktowanych czasopismach naukowych, ich krótkiego opisu, a także listy prezentacji ustnych i posterów wygłoszonych na konferencjach polskich i międzynarodowych. Warto byłoby też przytoczyć listę grantów, w których Doktorant uczestniczył lub być może kierował.

Praca doktorska obejmuje trzy pozycje, w tym dwa pierwszo-autorskie Kandydata:

- [D1] Pawlak M, Kedzierska KZ, **Migdal M**, Nahia KA, Ramilowski JA, Bugajski L, Hashimoto K, Marconi A, Piwocka K, Carninci P, Winata CL. *Dynamics of cardiomyocyte transcriptome and chromatin landscape demarcates key events of heart development*. Genome Res. 2019 Mar;29(3):506-519. doi: 10.1101/gr.244491.118. Impact Factor: 9.438, Punkty ministerialne: 200;
- [D2] **Migdal M.**, Tralle E, Nahia KA, Bugajski L, Kedzierska KZ, Garbicz F, Piwocka K, Winata CL, Pawlak M. *Multi-omics analyses of early liver injury reveals celltype-specific transcriptional and epigenomic shift*. BMC Genomics. 2021 Dec18;22(1):904. doi: 10.1186/s12864-021-08173-1. Impact Factor: 4.558, Punkty ministerialne: 140;
- [D3] **Migdal M**, Arakawa T, Takizawa S, Furuno M, Suzuki H, Arner E, Winata CL, Kaczkowski B. *xcore: an R package for inference of gene*

expression regulators. BMC Bioinformatics. 2023 Jan 11;24(1):14. doi: 10.1186/s12859-022-05084-0. Impact Factor: 3.327, Punkty ministerialne: 100;

Poniżej lista wszystkich publikacji z bazy pubmed opublikowanych przez doktoranta:

- [P1] MIDAS-Meaningful Immunogenetic Data at Scale. **Migdal M**, Ruan DF, Forrest WF, Horowitz A, Hammer C. PLoS Comput Biol. 2021 Jul 6;17(7):e1009131. doi: 10.1371/journal.pcbi.1009131. eCollection 2021 Jul. PMID: 34228721
- [P2] Profiling subcellular localization of nuclear-encoded mitochondrial gene products in zebrafish. Uszczyńska-Ratajczak B, Sugunan S, Kwiatkowska M, **Migdal M**, Carbonell-Sala S, Sokol A, Winata CL, Chacinska A. Life Sci Alliance. 2022 Oct 25;6(1):e202201514. doi: 10.26508/lsa.202201514. Print 2023 Jan. PMID: 36283702
- [P3] xcORE: an R package for inference of gene expression regulators. **Migdal M**, Arakawa T, Takizawa S, Furuno M, Suzuki H, Arner E, Winata CL, Kaczkowski B. BMC Bioinformatics. 2023 Jan 11;24(1):14. doi: 10.1186/s12859-022-05084-0. PMID: 36631751
- [P4] Multi-omics analyses of early liver injury reveals cell-type-specific transcriptional and epigenomic shift.
- **Migdal M**, Tralle E, Abu Nahia K, Bugajski Ł, Kędzierska KZ, Garbicz F, Piwocka K, Winata CL, Pawlak M. BMC Genomics. 2021 Dec 18;22(1):904. doi: 10.1186/s12864-021-08173-1. PMID: 34920711
- [P5] Dynamics of cardiomyocyte transcriptome and chromatin landscape demarcates key events of heart development. Pawlak M, Kędzierska KZ, **Migdal M**, Nahia KA, Ramilowski JA, Bugajski Ł, Hashimoto K, Marconi A, Piwocka K, Carninci P, Winata CL. Genome Res. 2019 Mar;29(3):506-519. doi: 10.1101/gr.244491.118. Epub 2019 Feb 13. PMID: 30760547
- [P6] In silico tools for accurate HLA and KIR inference from clinical sequencing data empower immunogenetics on individual-patient and population scales. Chen J, Madireddi S, Nagarkar D, **Migdal M**, Vander Heiden J, Chang D, Mukhyala K, Selvaraj S, Kadel EE, Brauer MJ, Mariathasan S, Hunkapiller J, Jhunjhunwala S, Albert ML, Hammer C. Brief Bioinform. 2021 May 20;22(3):bbaa223. doi: 10.1093/bib/bbaa223. PMID: 32940337
- [P7] Genomic and physiological analyses of the zebrafish atrioventricular canal reveal molecular building blocks of the secondary pacemaker region. Abu Nahia K, **Migdal M**, Quinn TA, Poon KL, Lapiński M, Sulej A, Liu J, Mondal SS, Pawlak M, Bugajski Ł, Piwocka K, Brand T, Kohl P, Korzh V, Winata C. Cell Mol Life Sci. 2021 Oct;78(19-20):6669-6687. doi: 10.1007/s00018-021-03939-y. Epub 2021 Sep 23. PMID: 34557935

Doktorant uczestniczył w co najmniej trzech grantach badawczych NCN. Pierwszy to Sonata 8 kierowany przez dr Michała Pawlaka pt. *„Rola niekolagenowych białek macierzy pozakomórkowej w patogenezie i nieinwazyjnej diagnostyce zwłóknienia wątroby”* (2014/15/D/NZ5/03421). Drugi to grant OPUS 7 promotor dr hab. Cecylii Winaty pt. *„Poznanie sieci genów regulujących rozwój serca u danio przegowanego przy użyciu genomiki”* (2014/13/B/NZ2/03863). W kolejnym grantcie Kandydat był współautorem publikacji - jest to grant OPUS 10 dr hab. Wojciecha Wiszniewskiego z Instytutu Matki i Dziecka pt. *„Próba identyfikacji genów odpowiedzialnych za rozwój mózgu w badaniach genomowych pacjentów z małowłóknem”* (2015/19/B/NZ2/01824).

Dodatkowo Doktorant współ-organizował kilka edycji prestiżowej szkoły NGSchool poświęconej zagadnieniom sekwencjonowania następnej generacji, w tym metodom bioinformatycznym analizy danych otrzymanych z tego typu doświadczeń biologii molekularnej. W roku 2017 Letnia Szkoła Bioinformatyki i Analizy Danych NGS odbyła się Jachrance i zebrała 43 uczestników i 20 wykładowców obejmując tematykę analizy danych z doświadczeń przeprowadzonych na pojedynczych komórkach. Zajęcia umożliwiły poznanie różnych aspektów biologii obliczeniowej, w tym technik związanych z sekwencjonowaniem nowej generacji (NGS) i ich zastosowaniu w badaniach naukowych i przemyśle. W roku 2018 Letnia Szkoła NGSchool odbyła się na Uniwersytecie Medycznym w Lublinie, zgromadziła 31 uczestników, 11 prelegentów, 17 mentorów hackatonu, oraz 10 organizatorów. Obejmowała zagadnienia sekwencjonowania długich odczytów i zastosowania tej techniki w medycynie spersonalizowanej i precyzyjnej. Umożliwiło to Doktorantowi zmierzenie się zarówno z wyzwaniami organizacyjnymi tak dużych wydarzeń, ale też uzyskanie unikalnej wiedzy naukowej na wczesnym etapie jej rozwoju od uznanych specjalistów w tej tematyce.

W swojej pracy doktorskiej mgr Maciej Migdał koncentruje się na modelowaniu wpływu czynników transkrypcyjnych na ekspresję genów, wykorzystując dane uzyskane za pomocą metod sekwencjonowania nowej generacji (NGS). Wykorzystując zaawansowane techniki bioinformatyczne i statystyczne, autor analizuje złożone interakcje między czynnikami transkrypcyjnymi a genami, które kontrolują. Poprzez integrację danych transkryptomicznych, epigenomicznych i sekwencjonowania otwartych regionów chromatyny (ATAC-seq), mgr. Maciej Migdał

dąży do zrozumienia, jak te czynniki transkrypcyjne wiążą się z konkretnymi miejscami w DNA, aby regulować ekspresję genów. Praca obejmuje również rozwój i zastosowanie narzędzi bioinformatycznych, takich jak pakiet xcore R, umożliwiających elastyczne modelowanie ekspresji genów na podstawie dużych baz danych ChIP-seq. Ta interdyscyplinarna analiza łączy biologię molekularną, matematykę i inżynierię, mając na celu nie tylko zrozumienie podstawowych mechanizmów regulacji genów, ale także zastosowania w medycynie i badaniach chorób, takich jak rozwój serca i uszkodzenia wątroby.

Praca doktorska składa się z kilku krótkich sekcji wstępnych obejmujących podziękowania, listę uwzględnionych w rozprawie artykułów, wykaz skrótów oraz streszczenia w języku polskim i angielskim. Następnie Doktorant zamieścił Wstęp oraz listę piśmiennictwa istotnego dla pracy. Następnie przedstawione są założenia i cel pracy, pliki pdf trzech publikacji: dynamika transkryptomu kardiomiocytów i krajobrazu chromatyny w rozwoju serca; analiza omiczna wczesnego uszkodzenia wątroby ujawniająca specyficzne dla typu komórki zmiany transkrypcyjne i epigenomiczne; oraz prezentacja oprogramowania xcore: pakiet R do charakteryzacji regulatorów ekspresji genów. Na koniec znajduje się krótkie Podsumowanie i główne wnioski pracy, oraz Oświadczenia autorów.

Wnioski końcowe pracy doktorskiej obejmują:

- Rozwinięto pipeline przetwarzania danych ATAC-seq, który pozwala na identyfikację otwartych regionów chromatyny, z uwzględnieniem zduplikowanych danych, który został zaimplementowany przy użyciu Nextflow i jest dostępny w repozytorium gitLab laboratorium.
- Opisano transkryptom i epigenom kardiomiocytów na wczesnych etapach rozwoju serca przy użyciu ryby Danio Pręgowanego jako organizmu modelowego. Wykonano analizę skupień profili ekspresji genów RNA-seq i analizę wzbogacenia motywów w otwartych regionach chromatyny. Przeanalizowano mutanty *gata5*, *hand2* i *tbx5*, u których rozwój serca jest zaburzony, identyfikując odległe elementy regulacyjne w rozwoju kardiomiocytów.
- W części aplikacyjnej opisano odpowiedź transkryptomiczną i epigenomiczną na hepatotoksyczne uszkodzenie wątroby dla wybranych typów komórek wątrobowych: komórek

śródbłonka zatokowego wątroby, hepatocytów i komórek gwiaździstych wątroby, przy użyciu Danio Pręgowanego. Ponownie przeanalizowano dane z RNA-seq i ATAC-seq dokonując analizy wzbogacenia motywów DNA w regionach otwartych. Pokazano za pomocą grup ko-ekspresji, że to właśnie komórki śródbłonka stanowią pierwszą populację komórek odpowiadającą na uszkodzenie wątroby.

- Zidentyfikowana odpowiedź molekularna obejmuje aktywację genów związanych z procesami metabolicznymi i redox, w tym otwarcie chromatyny przy ich promotorach. Analiza wzbogacenia motywów znalazła dwa czynniki transkrypcyjne FOXA1 i FOXA3 jako potencjalne regulatory aktywacji komórek śródbłonka.
- Algorytmy i metody opracowane w rozprawie doktorskiej zostały opublikowane w trzech manuskryptach naukowych o istotnym impact factor i punktach ministerialnych. W szczególności rozwinięto pakiet xcore w języku R, implementujący metodę przewidywania ekspresji genów, przydatny w modelowaniu regulatorów ekspresji genów na podstawie dużych zestawów danych ChIP-seq.

Pytania do doktoranta:

- Jak można dalej rozwijać zaproponowane metody analizy transkryptomu (RNA-seq) oraz epigenomu (ATAC-seq), szczególnie w przypadku danych z pojedynczych komórek?
- Brakuje bardziej szczegółowego porównania zaproponowanej metodologii do istniejących i rozwijanych przez różne laboratoria algorytmów bioinformatycznych o podobnych celach.
- Czy i jak można wytrenować różne metody uczenia maszynowego do podobnych zadań identyfikacji klastrów ko-ekspresji? Czy użycie tych metod miałoby dodatkowe zalety?
- Jakie są przyszłe zastosowania zaproponowanej metodologii, zwłaszcza w kontekście potencjalnych aplikacji w naukach medycznych i klinicznych?
- Jakie są ograniczenia wykorzystania modeli zwierzęcych, takich jak Danio Pręgowane, w badaniu regulacji genów u ludzki, jakie są zalety stosowania tych modeli?
- Jakie są według Doktoranta kluczowe wyzwania i pytania, które pozostają niewyjaśnione w przypadku badania

mechanizmów regulacji transkrypcyjnej przy użyciu doświadczeń RNA-seq i ATAC-seq?

- Jak pozostałe publikacje Doktoranta nie zamieszczone w rozprawie pokazują różnorodne aspekty analizy danych omicznych? Czy i jeśli tak to w jakim aspekcie wsparłyby przedstawione w rozprawie tezy badawcze?

W ramach uwag do pracy doktorskiej:

- Zbyt skrótowe potraktowanie Wstępu oraz brak głębszego wprowadzenia w realizowaną tematykę badawczą.
- Brak autorskiej Figury czy Figur reprezentujących zrealizowaną tematykę badawczą i główne wyniki, chociażby w sposób skrótowy czy schematyczny.
- Zbyt pośpieszne przejście ze wstępu ogólnego do opisu (streszczenia) wyników zmieszczonych w załączonych publikacjach.
- Niski udział procentowy Doktoranta w jednej ze składających się na rozprawę doktorską publikacjach (*„Dynamics of cardiomyocyte transcriptome and chromatin landscape demarcates key events of heart development”* – 10%)
- W przypadku drugiej pracy jest to już solidne 40% *„Multi-omics analyses of early liver injury reveals cell-type-specific transcriptional and epigenomic shift”*, podobnie w trzeciej to 45% *„xcore: an Rpackage for inference of gene expression regulators using publicly available human ChIP-seq experiments”*
- Drobne błędy interpunkcyjne i językowe.
- Warto zamieścić kluczowe elementy związane z życiorysem badawczym doktoranta – listę publikacji, listę prezentacji ustnych oraz posterów na konferencjach krajowych i międzynarodowych itp.

Wnioski końcowe

W podsumowaniu oceny rozprawy doktorskiej pana mgr Macieja Migdala, pragnę wyrazić moją akceptację dla przedstawionej pracy. Suma wskaźników impact factor wszystkich publikacji wynosi 17.323, a suma punktów ministerialnych w Polsce wynosi 440. Jest to wynik moim zdaniem w pełni wystarczający, nawet jeśli uwzględnimy, że są to publikacje współautorskie, w tym jedna z niskim udziałem Doktoranta.

Biorąc pod uwagę wartość naukową rozprawy doktorskiej, udane połączenie starannie opisanych w publikacjach narzędzi obliczeniowych, a także biologicznie istotnych eksperymentów i istotnych pytań badawczych związanych z otwartymi regionami chromatyny, uważam rozprawę doktorską pana mgr Macieja Migdala za znaczący wkład w dziedzinę bioinformatyki w zakresie Nauk Medycznych.

Rozprawa doktorska spełnia warunki określone w Art. 187 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (Dz. U. 2018 poz. 1668, Dz. U. z 2021 r. poz. 478, 619, 1630). Uważam, że rozprawa ta adresuje wszystkie powszechne i ustawowe wymagania stawiane rozprawom doktorskim, stanowi oryginalne rozwiązanie problemu naukowego, wykazuje ogólną wiedzę medyczną i biologiczną kandydata w zakresie metod bioinformatycznych, oraz demonstruje zdolność do samodzielnego prowadzenia pracy naukowej.

W związku z powyższym, mam przyjemność przedłożyć Radzie Dyscypliny Nauk Medycznych Warszawskiego Uniwersytetu Medycznego moją pozytywną recenzję i wnioskować za dopuszczeniem Pana mgr Macieja Migdala do dalszych etapów przewodu doktorskiego.

Prof. Dariusz Plewczyński

Dariusz Plewczynski, PhD, Professor of Exact and Natural Sciences; Principal Investigator

Phone: +48 22 554 36 54 or +48 22 234 7219

e-mail: d.plewczynski@cent.uw.edu.pl or Dariusz.Plewczynski@pw.edu.pl www: <https://plewczynski-lab.org>

Laboratory of Functional and Structural Genomics LFSG

Centre of New Technologies, University of Warsaw; Banacha 2c Street, 02-097 Warsaw, Poland

Laboratory of Bioinformatics and Computational Genomics LB!GO

Faculty of Mathematics and Information Science, Warsaw University of Technology; ul.

Koszykowa 75, 00-662 Warsaw, Poland