Wrocław, 9.12.2024

dr hab. inż. Witold Dyrka
Politechnika Wrocławska
Wydział Podstawowych Problemów Techniki
Katedra Inżynierii Biomedycznej
witold.dyrka@pwr.edu.pl

Recenzja rozprawy doktorskiej

mgr. inż. Michała Żurkowskiego

zatytułowanej:

Algorithms for feature exploration and modeling of quadruplex structures

1. Problem badawczy i jego znaczenie

Przedmiotem rozprawy doktorskiej jest opracowanie metod oraz narzędzi informatycznych do analizy kwadrupleksów i innych złożonych motywów strukturalnych cząsteczek kwasów nukleinowych. Zasadniczym problemem naukowym podjętym przez autora jest identyfikacja oraz parametryczny opis wzorców strukturalnych w cząsteczkach DNA i RNA zdeponowanych w repozytoriach cyfrowych. Zagadnienie jest aktywnym polem badań z zakresu bioinformatyki, co wynika z pełnionych przez te motywy funkcji, m.in. stabilizacji struktury przestrzennej cząsteczek oraz regulacji ekspresji genów. Tym samym kwadrupleksy (i inne złożone motywy strukturalne) stanowią potencjalny cel dla zastosowań terapeutycznych. Walor naukowy przejawia się także w zaprojektowaniu i realizacji oprogramowania mającego umożliwić lub ułatwić analizę motywów strukturalnych. Ten aspekt pracy ma bardzo duże znaczenie praktyczne, ponieważ tworzy warunki do rozwoju badań strukturalnych nad kwasami nukleinowymi.

Na rozprawę doktorską mgr. inż. Michała Żurkowskiego składają się 1) zbiór opublikowanych i powiązanych tematycznie pięciu artykułów naukowych, w których zostały wskazane samodzielne i wyodrębnione udziały autorskie Doktoranta, 2) dysertacja w języku angielskim stanowiąca pisemną prezentacją ogólnej wiedzy teoretycznej, będącej podstawą pracy, oraz omówienie otrzymanych wyników, w tym niepublikowanych (razem 61 stron, łącznie z bibliografią), oraz 3) oprogramowanie naukowe, którego kod źródłowy udostępniono w repozytoriach, stanowiące element projektowo-wdrożeniowy osiągnięcia. Promotorem rozprawy jest prof. dr hab. inż. Marta Szachniuk.

2. Wkład autora

Autor wskazał jako cel pracy *„przyczynienie się do szerszego zrozumienia związku między strukturą a funkcją kwadrupleksów poprzez stworzenie solidnego ekosystemu narzędzi i algorytmów wielokrotnego użytku, które ułatwią dalsze odkrycia w dziedzinie badań nad kwasami nukleinowymi obejmujących te specyficzne motywy*". To ogólne założenie Doktorant precyzuje, do pewnego stopnia,

poprzez cele i kryteria szczegółowe, z których następnie „rozlicza" się w konkluzjach omówienia wyników przedstawiając swoje osiągnięcia. Do deklarowanych celów szczegółowych należały:

1. *Opracowanie efektywnych algorytmów do identyfikacji cech kwadrupleksów*

   Zdaniem Autora osiągnięcia w tym zakresie obejmują zaproponowane algorytmy:

   a. adnotacji par zasad niekanonicznych na potrzeby klasyfikacji topologii tetrad,
   b. agregacji danych o kwadrupleksach na potrzeby analizy statystycznej,
   c. DrawTetrado, do wizualizacji kwadrupleksów w formie diagramów warstwowych,
   d. LinkTetrado, do identyfikacji i klasyfikacji złożonych motywów multimerycznych.

   W mojej ocenie istotną wartość naukową posiadają przede wszystkim osiągnięcia (c) i (d). **DrawTetrado** jest autorską metodą tworzenia wizualizacji kwadrupleksów w postaci diagramów warstwowych. Algorytm został zaimplementowany w językach Python i C++ w postaci narzędzia o tej samej nazwie, którego p. Michał Żurkowski jest wyłącznym projektantem i głównym deweloperem (¾ linii kodu, ponad ⅗ aktualizacji). Algorytm **LinkTetrado** rozszerza tetrady o wchodzące z nimi w kontakt nukleotydy leżące w tej samej płaszczyźnie w oparciu o kryteria heurystyczne. Metoda, która stanowi pierwsze systematyczne podejście do problemu, pozwoliła zidentyfikować łącznie kilkadziesiąt tego typu motywów. LinkTetrado został zaimplementowany przez Doktoranta w języku Python. W przypadku osiągnięć (a) i (b), niestety nie znalazłem – ani w odnośnej publikacji, ani w omówieniu – opisu pozwalającego ocenić ich oryginalność i walory naukowe.

2. *Opracowanie algorytmów ułatwiających przewidywanie trójwymiarowych struktur kwadrupleksów na podstawie sekwencji nukleotydowych,*

   Autor przyznaje, że zrezygnował z realizacji celu automatycznego przewidywania struktury kwadrupleksów na podstawie sekwencji ze względu na zbyt małą podaż danych, co dobrze świadczy o jego uczciwości intelektualnej. Jednocześnie deklaruje jako osiągnięcie w ramach doktoratu opracowanie dwóch algorytmów lokalnego dopasowania strukturalnego kwasów nukleinowych, GEOS i GENS, które mogą być wykorzystany m.in. do oceny predykcji struktury. Włączenie tego osiągnięcia do rozprawy uważam za kontrowersyjne, ponieważ Doktorant oba algorytmy opisał w podobny sposób już wcześniej, w ramach swojej pracy magisterskiej pt. *New algorithm for RNA 3D structure alignment* z 2019 r., zrealizowanej również pod opieką prof. Marta Szachniuk. Praca magisterska jest publicznie dostępna poprzez odnośnik w repozytorium https://github.com/michal-zurkowski/rnahugs. Analiza chronologii wprowadzanych do projektu zmian kodu wskazuje na pewien rozwój projektu w trakcie doktoratu, zapewne związany z przygotowanie publikacji, jednak w mojej ocenie nie pozwala to na uznanie algorytmów GEOS i GENS za oryginalne wyniki doktoratu.

3. N*apisanie programów umożliwiających skuteczne i przyjazne dla użytkownika korzystanie z tych algorytmów*

   Jako osiągnięcia w tym zakresie, Autor przedstawia serwisy internetowe RNAhugs, ONQUADRO oraz WebTetrado. **RNAhugs** umożliwia porównanie i wizualizację struktur RNA z wykorzystaniem algorytmów GEOS i GENS. Deklaracja autorska wskazuje na wiodący wkład Doktoranta w projektowaniu serwisu oraz wyłączny – w napisaniu backendu. **WebTetrado** pozwala na wizualizację oraz opis parametryczny struktur RNA i DNA,

oferując przy tym możliwość zapisu wyników (wysokiej jakości grafika wektorowa, tabele CSV) oraz czasowe przechowywanie wyników. Deklaracja autorska wskazuje na udział Doktoranta w projektowaniu systemu, kluczowy wkład w napisaniu backendu oraz wsparcie w tworzeniu frontendu. **ONQUADRO** jest bazą tetrad (obecnie 1946 elementów) i kwadrupleksów (615 elementów), udostępniającą także niektóre analizy WebTetrado. Co bardzo istotne, baza jest automatycznie aktualizowana poprzez cotygodniową synchronizację z PDB. Zwraca uwagę zintegrowanie bazy w serwisie Nucleic Acids Knowledge Base (NAKB), świadczące o docenieniu zasobu w środowisku. Deklaracja autorska wskazuje na udział Doktoranta w projektowaniu systemu oraz kluczowy wkład w implementację backendu (w tym algorytmy (a)-(c)). Z perspektywy oceny pracy doktorskiej największe znaczenie ma wykazany istotny wkład p. mgr inż. Michała Żurkowskiego w projektowanie oprogramowania, jako mający walor rozwiązania problemu naukowego. Z perspektywy realizacji celu zdefiniowanego jako „program przyjazny dla użytkownika" zastanawia niewielki udział Doktoranta w tworzenie frontendu serwisów. Niemniej jednak, przedstawiony w rozprawie dorobek projektowy jest bardzo znaczący i charakteryzuje się wysoką jakością.

Ponadto Autor określił następujące trzy wymagania jakościowe realizacji projektu doktorskiego:

1. *Zapewnienie wysokiej jakości i miarodajności wyników poprzez rygorystyczne testy obliczeniowe*

   We wszystkich pięciu projektach Doktorant brał udział w testowaniu (w tym przygotowaniu zbiorów danych) algorytmów i oprogramowania pod kątem prawidłowości i wydajności; w czterech przypadkach wkład ten był wyłączny lub wiodący. O realizacji kryteriów jakości pośrednio świadczy publikacja wyników w renomowanych czasopismach. W szczególności dwie publikacje w wydaniu serwerowym czasopisma *Nucleic Acids Research* wiązały się, wedle deklaracji redakcji, z przejściem przez serwisy dogłębnych testów technicznych.

2. *Upublicznienie kodów źródłowych opracowanych algorytmów, umożliwiające ich swobodne użycie, modyfikację i dystrybucję*

   Kod źródłowy oprogramowania DrawTetrado, LinkTetrado i RNAhugs został udostępniony w repozytoriach autora na platformie GitHub. Wybrana została permisywna licencja MIT, która pozwala m.in. na komercyjne wykorzystanie kodu; taki wybór uważam za zdecydowanie trafny, zwiększający potencjalne grono beneficjentów projektowych osiągnięć Autora. Jednocześnie brak informacji o udostępnieniu kodu źródłowego serwisów internetowych. W zakresie jakości udostępnionego kodu źródłowego, charakteryzuje się on dobrym ustrukturyzowaniem, a kluczowe miejsca są opatrzone komentarzami. W repozytoriach brakuje natomiast komentarzy dokumentacyjnych (tzw. docstringów) oraz testów jednostkowych, które pomogłyby w utrzymaniu kodu przy aktualizacjach.

3. *Rozpowszechnianie uzyskanych wyników poprzez prezentacje konferencyjne oraz publikacje w czołowych czasopismach z zakresu biologii i bioinformatyki.*

   Dorobek publikacyjny Autora jest imponujący: do doktoratu zostało włączonych pięć publikacji w najbardziej cenionych czasopismach bioinformatycznych: *Nucleic Acids Research* (IF=16, punktacja MNiSW=200) oraz *Bioinformatics* (IF=7, MNiSW=200); oba wydawane przez Oxford University Press. Prace opublikowane w *NAR* są prezentacjami bazy

danych ONQUADRO (2021, 6. autor z 7, 17 cytowań obcych wg Google Scholar) oraz serwisów internetowych WebTetrado (2023, 2. autor z 4, 5 cytowań obcych) i RNAhugs (2024, 1. autor z 5, 1 cytowanie obce) w corocznych wydaniach dedykowanych tym formom zasobów; podobnie jeden z artykułów w *Bioinformatics* ma charakter noty aplikacyjnej DrawTetrado (2022, 1. autor z 3, 1 cytowanie obce). Co prawda przygotowanie tego typu manuskryptów wymaga zwykle mnie nakładu pracy niż typowy artykuł naukowy, to jednak są to dokładnie te łamy, na których bioinformatyk tworzący oprogramowanie naukowe chce przedstawić swoje narzędzia; pozostaje zatem pogratulować publikacji. Piąty artykuł, opublikowany także w *Bioinformatics* (2023, 1. autor z 3, 2 cytowania obce), ma charakter badawczy i prezentuje algorytmy lokalnego dopasowania struktur RNA. Cytowania obce wskazują na zauważenie badań i narzędzi Doktoranta przez środowisko. Kwerenda w niektórych z nich pozwala zauważyć kilka prac używających bazy ONQUADRO jako podstawowego źródła danych oraz dwie korzystające z WebTetrado do analizy parametrów oraz wizualizacji kwadrupleksów. Ponadto, Doktorant prezentował swoje osiągnięcia w formie wystąpień ustnych na siedmiu konferencjach międzynarodowych.

3. Poprawność

Należy zauważyć, że problem badawczy nie został sformułowany poprzez klasyczne tezy pracy. Zamiast tego tego Autor postawił sobie cel opracowania wartościowych narzędzi informatycznych na potrzeby analizy i modelowania złożonych motywów strukturalnych w kwasach nukleinowych, przy zachowaniu pewnych ogólnych standardów tworzenia oprogramowania przyjętych w bioinformatyce. Nie będąc zwolennikiem sztywnego trzymania się czasem sztucznej formuły „tez pracy", nie mogę jednak nie zauważyć, że pozwalają one syntetyczne i precyzyjnie spojrzeć osiągnięcia, czego moim zdaniem w recenzowanej rozprawie nieco brakuje. Uważam, że warto pokusić się o takie spojrzenie podczas obrony.

Na podkreślenie zasługuje istotny udział Doktoranta, w tym projektowy, w tworzeniu oprogramowania, które zostało włączone w międzynarodowe zasoby na rzecz badań kwasów nukleinowych, i jest rzeczywiście wykorzystywane. Trzy narzędzia jego współautorstwa zostały zaprezentowane w dedykowanych wydaniach czasopisma *Nucleic Acids Research*, prezentujących znaczące zasoby internetowe.

Wysoką jakością charakteryzuje się publikacja pierwszoautorska w *Bioinformatics*, przedstawiająca oryginalne algorytmy lokalnego dopasowania struktur RNA. Zwraca uwagę szeroka gama zaprojektowanych przez Autora eksperymentów obliczeniowych, obejmujących m.in. testy porównawcze z dostępnymi metodami (ilościowo i jakościowo), wydajności oraz powtarzalności. Niemniej jednak, w mojej opinii, w samej dysertacji Autor powinien wyróżnić wyniki uzyskane w zakresie tego projektu przed i po napisaniu pracy magisterskiej, a ona sama powinna być cytowana.

W moim odczuciu rozprawa mogłaby zyskać na charakteryzacji tworzonego zaimplementowanego oprogramowania naukowego i procesu jego powstawania z perspektywy inżynierii oprogramowania. Byłoby interesujące, aby podczas obrony, choćby hasłowo, Doktorant przedstawił zastosowane podejścia z zakresu DevOps/DataOps itp., które doprowadziły do powstania jakościowych narzędzi. Czy i w jaki sposób podejścia te są stosowane w procesie rozwoju narzędzia LinkTetrado?

Praca została napisana poprawnym językiem i stoi na wysokim poziomie redakcyjnym. Pozostałe uwagi, co do jakości rozprawy zostały przedstawione w punktach dot. wkładu autorskiego oraz wiedzy kandydata.

4. Wiedza kandydata

Ogólny stan wiedzy został przedstawiony w rozdziale 1. Podrozdział 1.1 prezentuje podstawowe informacji o kwasach nukleinowych oraz stosowanych w bioinformatyce formatach ich opisu. Podrozdział 1.2 przedstawia aktualną wiedzę o motywach strukturalnych, przede wszystkim kwadrupleksach, przedstawiając również dedykowane im notacje i sposoby wizualizacji. Z mojej perspektywy, jako bioinformatyka specjalizującego się analizie białek, zarówno zakres jak i sposób prezentacji stanowią właściwe wprowadzenie do przedmiotu badań. Podrozdział 1.3 wprowadza w historię modelowania struktur przestrzennych kwasów nukleinowych oraz metody oceny jakości struktur, w tym oparte na dopasowaniu, pozostawiając jednak pewien niedosyt w zakresie rozwoju omawianego pola badań w ciągu kilku ostatnich kilku lat. W moim odczuciu brakuje m.in. szerszego poruszenia kwestii wpływu rozwoju metod i zastosowań uczenia głębokiego na dziedzinę modelowania i analizy kwasów nukleinowych. Podrozdział 1.4 przedstawia, na dużym poziomie ogólności, wybrane podstawowe pojęcia z zakresu teorii algorytmów, najwięcej uwagi poświęcając metodom heurystycznym i metaheurystycznym. Dwa ostatnie podrozdziały niewątpliwie bezpośrednio odnoszą się do dorobku z dziedziny informatyki. Mając na uwadze, że ważnym osiągnięciem doktoratu jest tworzenie oprogramowania naukowego, w tym serwisów internetowych, zabrakło podrozdziału obejmującego teoretyczne podstawy inżynierii tego procesu, pozwalającego lepiej zrozumieć (i bardziej docenić) decyzje projektowe Autora. Bibliografia samej dysertacji liczy niespełna 80 pozycji, w zdecydowanej większości opublikowanych w renomowanych czasopismach, z czego jedynie 12 z ostatnich 5 lat. Podsumowując, pomimo wskazanych wyżej mankamentów, nie mam wątpliwości, że Kandydat posiada ogólną wiedzę w dyscyplinie Informatyka techniczna i telekomunikacja. Język tekstu dysertacji oraz publikacji świadczy o umiejętności komunikowania tej wiedzy, a przedstawione oprogramowanie – o umiejętności jej praktycznego zastosowania.

6. Podsumowanie

Biorąc pod uwagę opinie zaprezentowane w poprzednich punktach i wymagania zdefiniowane przez art. 187 Ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (z późniejszymi zmianami) moja ocena rozprawy pod względem trzech podstawowych kryteriów jest następująca:

A. Czy rozprawa zawiera oryginalne rozwiązanie problemu naukowego? (wybierz jedną opcję stawiając znak **X**)

[ ]	[X]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

B. Czy po przeczytaniu rozprawy zgadzasz się, że kandydat posiada ogólną wiedzę teoretyczną w dyscyplinie Informatyka techniczna i telekomunikacja?

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

C. Czy kandydat posiada umiejętność samodzielnego prowadzenia pracy naukowej?

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*