Warszawa, ...25.LIS. 2024 r.

dr hab. inż. Robert Nowak, prof. uczelni
Instytut Informatyki
Wydział Elektroniki i Technik Informacyjnych
Politechnika Warszawska
ul. Nowowiejska 15/19
00-665 Warszawa
robert.nowak@pw.edu.pl

Recenzja rozprawy doktorskiej mgr inż. Michała Pawła Żurkowskiego zatytułowanej „Algorithms for feature exploration and modeling of quadruplex structures"

Recenzja powstała na prośbę Dziekana Wydziału Informatyki i Telekomunikacji Politechniki Poznańskiej z dnia 08.10.2024, zgodnie z uchwałą Rady Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Poznańskiej z dania 24.09.2024, na podstawie:

- rozprawy doktorskiej liczącej, bez dodatków, 61 stron, z października 2024 r,
- załączonych kopii 7. publikacji powiązanych tematycznie, opublikowanych w znaczących czasopismach naukowych, w których Kandydat jest współautorem,
- dorobku naukowego Doktoranta uwzględnionego w bazach Scoupus i Google Scholar,
- kodów źródłowych prezentowanych rozwiązań udostępnionych w otwartych repozytoriach.

1 Tematyka badań

Przedstawiona rozprawa doktorska przedstawia nowe metody informatyczne do badania kwadrupleksów. Kwadrupleksy są to specyficzne struktury występujące w DNA, RNA i innych biopolimerach. Kwadrupleksy są zaliczane do niekanonicznych struktur drugorzędowych, ponieważ oddziaływania wodorowe nie obejmują analizowanych typowo oddziaływań pomiędzy komplementarnymi nukleotydami (A z T, G z C itp.) a oddziaływania pomiędzy tetradami (czwórkami nukleotydów), najczęściej guaninowymi (wtedy są nazywane G-kwadrupleksami albo G4).

Większość narzędzi i metod analiz struktur drugorzędowych skupiła się na wsparciu oddziaływań wodorowych opisanych w kanonicznym parowaniu zasad w DNA i RNA, dlatego niniejsza praca doktorska wypełnia lukę, dostarczając nowych metod i narzędzi bioinformatycznych do badania kwadrupleksów, które są istotnym czynnikiem niektórych procesów biologicznych.

W ramach pracy opracowano: algorytmy badania podobieństw struktur cząsteczek uwzględniający kwadrupleksy, metodę wizualizacji kwadrupleksów ułatwiające badanie ich relacji, w tym klasyfikację kwadrupleksów, metodę detekcji motywów multimerycznych, opartych o kwadrupleksy. Dla wszystkich nowych metod dostarczono działający program komputerowy, udostępniany jako wolne oprogramowanie, a ponadto dostarczono bazę danych, która gromadzi kwadrupleksy wraz ze skryptami do jej uzupełniania z rekordów bazy Protein Data Bank (PDB). Dostarczono także aplikację do analizy struktur nieobecnych w PDB.

Problem badawczy jest postawiony właściwie, jest on interesujący i istotny. Przedstawione rozwiązania są poprawne i potwierdzone wdrożeniem. Problem ma charakter naukowy i ma znaczenie praktyczne.

2 Główne wyniki rozprawy

W rozprawie w sposób właściwy przeprowadzono analizę problemu. Przedstawiono definicję kwadrupleksu (G-kwadrupleksu, G4), jego istotną rolę w niektórych procesach biologicznych i ubogie wsparcie w istniejących metodach bioinformatycznych do analizy sekwencji. Uzyskanie przedstawionego, poprawnego, opisu problemu badawczego wymagało wiedzy dziedzinowej (biologicznej) i informatycznej, a ponadto pogłębionego przeglądu literatury światowej. Autor w pracy cytuje kilkadziesiąt pozycji. Wnioski z przeglądu źródeł sformułowano w sposób jasny i przekonujący, pozwoliły one postawić opisane w pracy problemy badawcze.

Rozprawa, napisana w języku angielskim, ma 4 rozdziały: wprowadzenie (rozdział 1), omówienie wyników (2) oraz podsumowanie (3) i opis osiągnięć. Główne osiągnięcia Kandydata zostały zawarte w pięciu publikacjach powiązanych tematycznie oznaczonych [A1] – [A5], czyli:

A1. Tomasz Zok, Natalia Kraszewska, Joanna Miskiewicz, Paulina Pielacinska, **Michal Żurkowski**, Marta Szachniuk, ONQUADRO: a database of experimentally determined quadruplex structures. Nucleic Acids Research, 2021, 200 pkt MEiN, Q1 (WoS), IF=16.1;

A2. **Michal Żurkowski**, Tomasz Zok, Marta Szachniuk, DrawTetrado to create layer diagrams of G4 structures. Bioinformatics, 2022, 200 pkt MEiN, IF=5.8;

A3. Bartosz Adamczyk, **Michal Żurkowski**, Marta Szachniuk, Tomasz Zok, WebTetrado: a webserver to explore quadruplexes in nucleic acid 3D structures. Nucleic Acids Research, 2023, 200 pkt MEiN, IF=16.6;

A4. **Michal Żurkowski**, Maciej Antczak, Marta Szachniuk, High-quality, customizable heuristics for RNA 3D structure alignment. Bioinformatics, 2023, 200 pkt MEiN, IF=4.4;

A5. **Michal Żurkowski**, Mateusz Swiercz, Filip Wozny, Maciej Antczak, Marta Szachniuk, RNAhugs web server for customized 3D RNA structure alignment. Nucleic Acids Research, 200 pkt MEiN, IF=16.6.

Wszystkie wymienione artykuły są opublikowane w czasopismach z pierwszego kwantyla bazy Web of Science, mają 200 pkt na liście MEiN i mają wysoki współczynnik wpływu. W każdej publikacji udział Kandydata jest znaczący, w trzech artykułach jest on pierwszym (wiodącym) autorem.

Kandydat opracował następujące metody.

- Nowy algorytm znajdowania dopasowania struktur cząsteczek uwzględniający kwadrupleksy, o nazwie GEOS, który wykorzystuje metodę zachłanną (ang. greedy). Algorytm przegląda sąsiadów lokalnego rozwiązania, wybierając najlepszego kandydata, aż do uzyskania rozwiązania problemu wejściowego. Kandydat zdefiniował przestrzeń poszukiwań, znalazł obiecujące rozwiązania początkowe (lokalne) i zdefiniował relację sąsiedztwa. Udowodnił przy tym analitycznie pewne cechy problemu, pozwalające na poszukiwania lokalne.

- Nowy algorytm, który znajduje dopasowania struktur, o nazwie GENS, który ma zastosowania podobne do GEOS, ale wykorzystuje meta-heurystykę, algorytm genetyczny, do znajdowania dopasowania. Taki algorytm pozwala znajdować optimum globalne. Kandydat tutaj dostarczył reprezentację problemu dopasowania z uwzględnieniem kwadrupleksów w kontekście algorytmu genetycznego.

- Metodę detekcji motywów multimerycznych, o nazwie LinkTetrado. Motywy multimeryczne są innymi przykładami niekanonicznych struktur, podobnych do kwadrupleksów, ale obejmujący pięć (pentad, PDB ID: 2RQJ), sześć (hexad, PD ID: 10ZS) lub więcej symboli. Metoda analizuje wykryte czwórki (algorytm podobny jak poprzednio), a następnie ich sąsiednie nukleotydy za pomocą nowych warunków, m.in. przestrzennych, pozwalających utworzyć motyw multimeryczny.

- Metodę wizualizacji kwadrupleksów ułatwiające badanie ich relacji.

Kandydat opracował samodzielnie (RNAhugs, LinkTetrado, DrawTetrado, WebTetrado), lub w zespole (ONQUADRO), programy komputerowe wykorzystujące przeglądarkę jako interfejs użytkownika.

- System RNAhugs zawierającego algorytmy GEOS i GENS, do dopasowywania struktur uwzględniających kwadrupleksy. System jest udostępniony jako aplikacja internetowa na stronie uczelni `https://rnahugs.cs.put.poznan.pl`, został opisany m.in. w [A4],[A5], strona projektu: `https://github.com/RNApolis/rnahugs`;

- program LinkTetrado, strona projektu: `https://github.com/michal-zurkowski/linktetrado`;

- system DrawTetrado zawierającej autorską metodę wizualizacji kwadrupleksów, strona projektu `https://github.com/RNApolis/drawtetrado`, system opisany w [A2];
- repozytorium danych ONQUADRO, przechowująca kwadrupeksy. System jest dostępny na stronie uczelni, `https://onquadro.cs.put.poznan.pl`, został opisany w [A1];
- aplikację WebTerado, która pozwala na analizę kwadrupleksów z danych innych niż rekordy PDB, dostępną na stronie uczelni, `https://webtetrado.cs.put.poznan.pl`, opisaną w [A3].

Kandydat przeanalizował dane z istniejących baz danych za pomocą własnych narzędzi, proponując lepszą niż znana z literatury klasyfikację kwadrupleksów (DrawTetrado), czy wykrywając 25 nowych struktur multimerycznych (LinkTetrado).

Prace związane z opracowywaniem nowych algorytmów mają charakter naukowy w dyscyplinie informatyka techniczna i telekomunikacja, prace związane z dostarczaniem działających systemów mają charakter wdrożeniowy, zaś analizy danych są istotne naukowo dla biologii molekularnej.

Wkład autora w przedstawione rozwiązania jest znaczny, w niektórych elementach jest on pomysłodawcą, twórcą algorytmu oraz programistą i testerem, głównym autorem publikacji. Rozwiązania są poprawne, uzasadnione i godne zaufania. Kandydat prezentuje odpowiednią wiedzę w dyscyplinie Informatyka Techniczna i Telekomunikacja.

3 Elementy krytyczne

Znalazłem kilka elementów, które są dyskusyjne lub które należy dodatkowo wyjaśnić. Poniżej przedstawione niedociągnięcia, czy obszary dyskusyjne, nie zmieniają pozytywnej konkluzji końcowej przedstawionej pracy. Stanowią podstawę do dyskusji na publicznej obronie rozprawy.

3.1 Opis wkładu

Opis wkładu w pracy jest nieprecyzyjny dla niektórych przedstawionych wcześniej szczegółowych osiągnięć. Artykuły naukowe są współautorskie. Nie podano udziału procentowego, ani nie wyliczono wszystkich komponentów, co jest problemem przy ocenie tego aspektu. Wkład Kandydata jest nieco inny w punkcie 'Contribution to publications' (1) i w deklaracjach współautorów (2). Przykładowo dla [A1], mgr Żurkowski (inicjały MZ) w (1) pisze „I designed key components of the QNQUADRO database ...", podczas gdy w (2) jest „TZ, JM, MZ and MS designed the dababase system. TZ created the database schema and developed the auto-update functionality. TZ and MZ implemented the backend. ...".

Po lekturze pracy i artykułów Recenzent nie zarzuca zbyt małego udziału, tylko nieprecyzyjne określenie wkładu w każdą pracę.

3.2 Wydajność prezentowanych rozwiązań

Jednym z celów postawionych w rozprawie jest opracowanie wydajnych algorytmów. Tymczasem nie pokazano analizy złożoności obliczeniowej (czasowej, pamięciowej) opracowanych metod, co ma zazwyczaj największy wpływ na wydajność. W szczególności:

- Jaka jest złożoność metody GEOS ?
- Jaka jest złożoność metody GENS ?
- Jaka jest złożoność metody LinkTetrado ?

W publikacjach Kandydata zawarto wykresy czasu działania w funkcji wielkości danych, natomiast brak jest dyskusji i konkluzji. Przykładem pożądanego elementu, którego nie znalazłem, jest analiza które elementy algorytmu determinują złożoność, albo co uzasadnienia uzyskaną w eksperymentach numerycznych zależność czasu działania w funkcji wielkości danych.

3.3 Poprawność, wysoka jakość kodu, odporność, testowanie

Celem pracy (rozdział 1.5, str. 20) jest wysoka jakość dostarczonego oprogramowania, co jest potwierdzono wdrożeniem. Istotnym składnikiem zapewnienia jakości jest obecność automatycznych testów, m.in. testów jednostkowych i wysokich miar związanych z automatycznym testowaniem, np. wysokie pokrycie kodu testami. Takich testów nie znalazłem w dostarczonych repozytoriach github. W pracy nie ma także raportów z badania jakości testów. Wobec tego nasuwają się poniższe pytana.

- Jak zapewniono wysoką jakość dla RNAhugs, DrawTetrado, WebTetrado, ONQUADRO?
- W jaki sposób bada się, czy ta jakość jest zachowana po zmianach w kodzie.

3.4 Opis metod przedstawionych w pracy

Opisy poszczególnych metod przedstawionych w pracy nie zawierają tych samych elementów, a powinny. Brakuje opisu formalnego problemu wyszukiwania struktur, np. poprzez gramatykę formalną. Niektóre metody mają opis w pseudokodzie, np. metoda 'LinkTetrado' ma schemat blokowy, inne nie mają.

3.5 Otwartość kodu źródłowego

Kod źródłowy dla większości metod przedstawionych w pracy jest dostępny i jest dobrej jakości. Niestety, nie znalazłem repozytorium, gdzie jest kod źródłowy systemu ONQUADRO [A1]. Są tam m.in. skrypty do automatycznego uzupełniania bazy, schemat bazy danych itd.

3.6 Algorytmy

Opis wyboru metody, konstrukcja algorytmu, dobór parametrów i meta-parametrów jest bardzo powierzchowny. Z punktu widzenia użytkownika takie opisy nie są istotne, ale z punktu widzenia informatyki, powinien być dokładniejszy. W szczególności algorytm genetyczny, będący istotnym elementem metody GENS nie został opisany poprawnie.

- Jak dobrano wielkość populacji, prawdopodobieństwo mutacji i inne parametry? Czy ten dobór jest optymalny? Jeżeli tak, jak to badano?
- Dlaczego użyto tak prostej meta-heurystyki, choć od wielu lat używa się innych (nowszych) metod, np ewolucji różnicowej (Differential Evolution)? Innymi słowy stała wielkość populacji i stop algorytmu po n pokoleniach jest znacznie mniej skuteczny niż później opracowane meta-heurystyki.

4 Ocena dorobku

Dorobek naukowy Kandydata jest wyróżniający się. Mgr. Żurkowski posiada w dorobku 7 recenzowanych publikacji, publikacje te są znane w środowisku, suma cytowań wg. Google Scholar, sprawdzana 22.11.2024, przekroczyła 200. Kandydat jest aktywnym badaczem, prezentował wyniki swoich prac na 12 konferencjach, brał udział w trzech projektach badawczych, w tym w dwóch z nich (granty Politechniki Poznańskiej) był głównym badaczem. Za swoje prace uzyskał kilka nagród, w tym Best Paper Award.

Aplikacje, których twórcą lub współtwórcą jest Kandydat są dostępne na stronie uczelni, `https://rnahugs.cs.put.poznan.pl`, `https://onquadro.cs.put.poznan.pl`, `https://webtetrado.cs.put.poznan.pl`, oraz w repozytoriach `https://github.com/RNApolis/rnahugs`, `https://github.com/michal-zurkowski/linktetrado`, `https://github.com/RNApolis/drawtetrado`.

Mgr. inż. prowadzi zajęcia laboratoryjne ze studentami Informatyki na Politechnice Poznańskiej.

W mojej ocenie dorobek naukowy wystarczający, aby dopuścić Kandydata do dalszych etapów przewodu doktorskiego.

5 Podsumowanie

Temat badawczy uważam za bardzo istotny, teza jest poprawna i oryginalna, wykazana w stopniu wyczerpującym. Opracowane rozwiązanie jest nowatorskie, dostarcza wyniki na poziomie prezentowanym w literaturze światowej, a ponadto ma wysoki poziom gotowości technologicznej.

Pod względem trzech podstawowych kryteriów stwierdzam że,

- rozprawa zawiera oryginalne rozwiązanie problemu badawczego,
- zgadzam się, że kandydat posiada ogólną wiedzę teoretyczną w dyscyplinie ITT,
- zdecydowanie potwierdzam umiejętność samodzielnego prowadzenia pracy naukowej przez kandydata.

Stwierdzam, że **recenzowana rozprawa doktorska mgr inż. Michała Pawła Żurkowskiego spełnia warunki** określone w aktualnych przepisach i wnioskuję do Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Poznańskiej o dopuszczenie rozprawy doktorskiej mgr inż. Michała Pawła Żurkowskiego do publicznej obrony.

Z-ca DYREKTORA
Instytutu Informatyki ds. Nauki
dr hab. inż. Robert Nowak,
profesor uczelni