INSTYTUT CHEMII BIOORGANICZNEJ
Polskiej Akademii Nauk

ul. Noskowskiego 12/14, 61-704 Poznań
tel.: centrala 61 852 85 03, sekretariat 61 852 89 19
faks: 61 852 05 32, e-mail: ibch@ibch.poznan.pl
REGON 000849327 NIP 777-00-02-062
http://www.ibch.poznan.pl

Poznań, 06.12.2022

dr hab. Barbara Uszczyńska-Ratajczak
Zakład Biologii Obliczeniowej Niekodującego RNA
Instytut Chemii Bioorganicznej
Polskiej Akademii Nauk
Noskowskiego 12/14
61-704 Poznań

Recenzja
rozprawy doktorskiej mgr Jakuba Wiedemanna, zatytułowanej: „*Combinatorial Analysis of RNA Tertiary Structures with the Use of Angular Representations*”

Przedstawiona dysertacja została wykonana w Instytucie Informatyki na Wydziale Informatyki i Telekomunikacji Politechniki Poznańskiej pod kierunkiem dr hab. Macieja Antczaka oraz dr Macieja Miłostana. Rozprawa doktorska została napisana w języku angielskim, a jej podstawą są cztery prace naukowe opublikowane w czasopismach o zasięgu międzynarodowym, włączając prestiżowe *Bioinformatics* oraz *Nucleic Acid Research*. Doktorant jest pierwszym autorem w trzech z tych prac (publikacje A1, A2 oraz A4). Natomiast w jednej pracy (A3) zajmuje czwarte miejsce na liście współautorów. Jest to wieloautorska praca będąca wynikiem realizacji ważnego i cenionego projektu RNA-Puzzles – zbiorowego eksperymentu, którego celem jest obiektywna i wiarygodna ocena jakości komputerowych predykcji struktur 3D RNA. Udział Doktoranta w tym przedsięwzięciu polegał na zaimplementowaniu nowej metody do oceny modeli trójwymiarowych struktur RNA: LCS-TA (*Longest Continuous Segments in Torsion Angle space*), która pozwala porównywać struktury bardziej w ujęciu lokalnym (analiza najdłuższych ciągów segmentów), niż globalnym (całe struktury). Dlatego też, wkład w powstanie także i tej pracy uznaję za znaczący.

Na pracę doktorską składają się streszczenie w języku angielskim i polskim, lista publikacji wraz z danymi bibliograficznymi, wstęp, skrótowe omówienie publikacji w sekcji wyniki, bibliografia, kopie publikacji stanowiących podstawę tej rozprawy, deklaracje określające wkład pozostałych współautorów w przygotowanie przedstawionych prac oraz załączniki opisujące aktywność naukową Doktoranta. Rozprawa jest przemyślana, przejrzysta i bardzo rzeczowa. Wstęp opisuje RNA od strony bardziej biologicznej, skupiając się na jego budowie oraz poziomie złożoności struktur jakie tworzy, aby następnie przedstawić ten problem ze strony informatycznej poprzez wyjaśnienie formatów danych umożliwiających rzetelny i czytelny sposób zapisu tych struktur. Połączenie przedstawionych publikacji pokazuje także rozwój naukowy Doktoranta i ewolucję zagadnień badawczych. Praca wywodzi się od relatywnie prostego problemu jakim jest porównywanie różnic w strukturach trójwymiarowych cząsteczek RNA, wykazujących

duży poziom homologii na poziomie sekwencji (publikacja A1), poprzez opracowanie nowej metody oceny podobieństwa struktur przestrzennych RNA w oparciu o identyfikację najdłuższych ciągłych segmentów o określonym współczynniku podobieństwa (metoda LCS-TA, publikacja A2), zastosowania jej w praktyce w ramach projektu RNA-puzzles (publikacja A3), aż po utworzenie bazy danych *RNAloops*, która w sposób w pełni zautomatyzowany pozwala gromadzić informacje o pętlach wieloramiennych (ang. *N-way junctions*) zidentyfikowanych w eksperymentalnie potwierdzonych strukturach przestrzennych cząsteczek RNA (publikacja A4). Multipętle jako jednoniciowe fragmenty RNA łączące ze sobą struktury 2D są jednym z trudniejszych motywów do identyfikacji w ramach predykcji obliczeniowych, ponieważ ich obecność oraz zmienność znacząco wpływa na układ przestrzenny całej cząsteczki RNA. Zgromadzone informacje na temat pętli wieloramiennych mogą być bezpośrednio wykorzystane w procesie modelowania struktur przestrzennych RNA, co pozwoli nie tylko usprawnić ten proces, ale także znacząco zwiększyć jakość otrzymywanych modeli.

Przygotowanie rozprawy doktorskiej na bazie publikacji naukowych nie jest rzeczą łatwą, gdyż wymaga systematyczności, dużego zaangażowania oraz produktywności na etapie trwania całych studiów doktoranckich. Prace naukowe w procesie recenzji oceniane są przez grono uznanych i niezależnych ekspertów, w efekcie czego przechodzą szereg, niekiedy nawet drastycznych edytorskich zmian, których należy dopełnić w relatywnie krótkim czasie. Po opublikowaniu prace te są następnie oceniane przez czytelników, a miarą ich poczytności jest liczba cytowań. Recenzja pracy doktorskiej przygotowanej na bazie publikacji, które przeszły już proces recenzji oraz cieszą się uznaniem czytelników (łączna liczba cytowań wg. Google Scholar na dzień przygotowania recenzji wynosi 54, A1=6 (2016), A2=11 (2017), A3=36 (2020), A4=1 (2022)) daje mi przywilej skoncentrowania niniejszej recenzji głównie na dyskusji przedstawionych wyników w szerszym kontekście biologicznym i bioinformatycznym.

W pierwszej publikacji (A1) przedstawiony został problem różnych struktur przestrzennych, które mogą tworzyć cząsteczki RNA o dość wysokiej homologii (>90%). Do oceny tych struktur opracowane zostało nowe narzędzie – *StructAnalyzer*, które pozwala określić globalne podobieństwo analizowanych modeli przestrzennych na podstawie odchylenia wartości średniokwadratowej (ang. *Root Mean Square Deviation*, RMSD). Jak zostało wspomniane we wstępie RMSD nie jest najlepszą metodą oceny struktur, gdyż jego wartości bardzo mocno zależą od długości sekwencji RNA. *Czy rozważa Pan zatem aktualizację oprogramowania StructAnalyzer poprzez implementację algorytmu LCS-TA, pozwalającego na analizę struktur z perspektywy bardziej lokalnej?* Byłabym wdzięczna także za komentarz na ile innowacyjne byłoby to narzędzie w kontekście obecnie dostępnych rozwiązań pozwalających na analizę podobieństwa struktur przestrzennych RNA.

Doktorant w swojej rozprawie przedstawia RNA z perspektywy bardzo ogólnej, jako liniowy polimer złożony z rybonukleotydów połączonych wiązaniem fosfodiestrowym, posiadający zdolność

tworzenia złożonych struktur na czterech różnych poziomach organizacji strukturalnej. Jednakże, w biologii wyróżnia się różne typy RNA, które pełnią określone funkcje w komórce. Grupy te obejmują m.in. rybosomalny RNA (rRNA), transferowy RNA (tRNA), matrycowy RNA (mRNA), długi niekodujący RNA (lncRNA), mikroRNA (miRNA), mały jądrowy RNA (snRNA), czy mały jąderkowy RNA (snoRNA). Nie wspominając już o kolistych typach RNA. Każdy z tych rodzajów RNA ma ściśle określone właściwości na poziomie sekwencji i struktury, które determinują jego specyficzne funkcje biologiczne. Jedną z podstawowych różnic jest jednak długość sekwencji nukleotydowej, która pozwala podzielić te cząsteczki na dwie grupy, obejmujące krótkie (< 200 nt) i długie (>200 nt) RNA. W praktyce krótkie cząsteczki jak np. miRNA osiągają długość 21-23 nukleotydów. Natomiast średnia długość matrycowego RNA, kodującego białko w genomie ludzkim wynosi ok. 2,200 nukleotydów. *W związku z tym chciałam zapytać czy istnieje jakiś optymalny zakres długość sekwencji nukleotydowej, która pozwala z dużą dokładnością modelować struktury przestrzenne cząsteczek RNA? Czy modelowanie krótszych cząsteczek jest tak samo wydajne jak tych dłuższych? A może istnieje pewien zakres długości sekwencji po przekroczeniu którego modelowanie struktur przestrzennych staje się niezwykle trudne ze względu na liczbę potencjalnych połączeń oraz złożoność struktury?* Przy okazji dyskusji tego zagadnienia chciałabym również prosić o Pański komentarz dotyczący wpływu długości badanej sekwencji RNA na zdolność metody LCS-TA do oceny podobieństwa struktur przestrzennych. *Czy ta metoda tak samo wydajnie poradzi sobie z oceną krótkich oraz długich typów RNA, czy raczej głównym czynnikiem determinującym trudność zagadnienia będzie sam skład nukleotydowy?* Jestem ciekawa jakich różnic możemy spodziewać się dla obu trybów LCS-TA: zależnego i niezależnego od sekwencji.

W świecie RNA sekwencja, struktura oraz funkcja stanowią nierozerwalne połączenie w kontekście roli biologicznej tych cząsteczek. W przypadku matrycowego RNA, które koduje białka, związek sekwencja-struktura-funkcja jest jasno określony. Obecność ramek odczytu kodujących białko pozwala przewidzieć sekwencję powstającego białka i na podstawie tej informacji z dużą dokładnością określić sposób w jaki dany gen koduje jego funkcje molekularne. Jednakże geny kodujące białko stanowią jedynie niewielki procent (ok. 1%) naszego DNA. Większość rejonów w naszym genomie nie produkuje białek, chociaż wciąż posiada zdolność wytwarzania RNA, tzw. niekodującego RNA. Jedną z liczniejszych klas niekodującego RNA są długie niekodujące RNA (ang. *long noncoding RNAs*) – cząsteczki RNA o długości >200 nt posiadające ograniczone zdolności kodowania białka. Dla lncRNA związek pomiędzy ich sekwencją, strukturą oraz funkcją wciąż pozostaje nieznany, co znacząco ogranicza ich identyfikację oraz funkcjonalną charakterystykę. Ponadto sekwencje długich niekodujących RNA są słabo zachowane na drodze ewolucji, co uniemożliwia detekcję ich ortologów lub paralogów na zasadzie podobieństwa sekwencji. Jednocześnie wielokrotnie udowodniono, że lncRNA wykazują zdolność do zachowywania funkcji biologicznych, pomimo dużych rozbieżności na poziomie sekwencji i struktury drugorzędowej,

np. lncRNA JPX. *Zatem na ile realne, Pańskim zdaniem, jest oczekiwanie, że pomimo zmienności na poziomie sekwencji i struktur drugorzędowych dwie cząsteczki RNA mogą utworzyć zbliżone struktury przestrzenne? Czy proponowane przez Pana narzędzie informatyczne StructAnalyzer umożliwia identyfikację takich przypadków? Czy zaobserwował Pan podobne przykłady podczas swojej pracy m.in. z wirusowymi RNA oraz rybozymami?*

Tworzenie repozytoriów zawierających elementy pozwalające usprawnić proces modelowania cząsteczek RNA oraz poprawę jakości tych predykcji są szczególnie ważne. Rozumiem i w pełni popieram ideę projektu *RNAloops*. Zakładam także, że aspekt modelowania funkcjonalnych elementów może być kluczowy w kontekście przewidywania struktur i ich biologicznej wartości. W ostatnim czasie dość modnym podejściem stało się klasyfikowanie cząsteczek RNA ze względu na zawartość k-merów. Porównania oparte na k-merach – ciągach sekwencji o zdefiniowanej długości, są idealnym uzupełnieniem dla algorytmów porównujących sekwencje cząsteczek, np. BLAST, zwłaszcza jeżeli te cząsteczki nie wykazują bezpośrednich związków ewolucyjnych. Metoda analizy zawartości k-merów zlicza k-mery danego typu niezależnie od ich lokalizacji w sekwencji i pozwala na wykrywanie powtarzających się relacji między sekwencją, a funkcją cząsteczek RNA. Podejście to bazuje na założeniu, iż cząsteczki RNA o podobnej funkcji mogą zawierać zbliżone motywy sekwencji, tzw. funkcjonalne domeny, nawet jeśli nie wykazują homologii liniowej. *W kontekście przewidywania struktur przestrzennych i łączenia ich z biologicznymi funkcjami RNA, na ile pożyteczna wydaje się Panu detekcja, katalogowanie i funkcjonalna adnotacja tych domen? Czy wydaje się Panu możliwe i potrzebne bezpośrednie uwzględnianie tych regionów na etapie procesu modelowania oraz porównywania struktur przestrzennych RNA?*

Identyfikacja niekodujących RNA wciąż stanowi wyzwanie, gdyż ze względu na w/w właściwości ich detekcja głównie sprowadza się do wykrywania dowodów transkrypcji – wytwarzanych cząsteczek RNA. Niestety takie podejście charakteryzuje się ograniczonym poziomem detekcji, który w dużej mierze zależy od użytej metody identyfikacji RNA oraz co ważniejsze uniemożliwia określenie funkcjonalnego potencjału wykrywanych cząsteczek RNA. Jednym ze sposobów na usprawnienie identyfikacji potencjalnie funkcjonalnych RNA jest komputerowa analiza całych genomów w poszukiwaniu zachowanych struktur RNA (ang. *conserved RNA structures*). Do tej pory analiza ta była wykonywana na poziomie struktur drugorzędowych. *Na ile realne (w kontekście zasobów i czasu obliczeń) będzie wykonanie podobnej analizy struktur przestrzennych dla np. trzech wybranych genomów kręgowców?*

Nieznany związek pomiędzy sekwencją-strukturą-funkcją lncRNA sprawia, iż znakomita większość (>97%) długich niekodujących RNA wciąż nie ma jasno określonej funkcji biologicznej. Fakt ten nastręcza wiele trudności w kontekście chociażby samej ich klasyfikacji. Ze względu na brak jasno określonych ról w komórce lncRNA są klasyfikowane względem swojej lokalizacji do najbliższego genu kodującego białko. Ta genomiczna klasyfikacja obejmująca pozycję intronową, eksonową, sensowną,

antysensowną, międzygenową, etc. jest mało informatywna i często myląca. Dużo bardziej logiczna i intuicyjna wydaje się klasyfikacja RNA względem ich struktur. Strukturalne grupowanie cząsteczek RNA pozwoli także wydajniej klasyfikować te cząsteczki w kontekście ich potencjalnych funkcji. *W związku z tym chciałabym zapytać jak wyobraża sobie Pan potencjalny podział cząsteczek RNA względem ich struktur? Czy na podstawie swojego doświadczenia może Pan zaproponować główne grupy struktur, które Pańskim zdaniem należałoby szczególnie wyróżnić?*

Na sam koniec chciałabym także poznać Pańskie dotyczące przyszłości i kierunku rozwoju bioinformatyki strukturalnej, w kontekście identyfikacji struktur przestrzennych i ich eksperymentalnej weryfikacji. *Na ile te dwa aspekty będą współistnieć i wzajemnie się wspierać? Czy oczekiwanie, iż bioinformatyka strukturalna będzie czerpać z metod biologii molekularnej jest realne, czy raczej usprawnienie procesu modelowania struktur przestrzennych będzie wynikać z głównie z poprawy precyzji i czułości metod obliczeniowych?*

Pan mgr Jakub Wiedemann posiada imponujący dorobek naukowy obejmujący pięć publikacji w czasopismach międzynarodowych, z których cztery stanowią podstawę prezentowanej rozprawy doktorskiej. Na uwagę zasługuje również wysoka aktywność naukowa Doktoranta obejmująca 27 wystąpień w formie prezentacji oraz plakatów naukowych. Autor rozprawy był także laureatem czterech stypendiów w tym stypendium Rektora dla najlepszych doktorantów na Wydziale Informatyki i Telekomunikacji. Ponadto wykazywał się aktywnością organizacyjną, obejmującą przygotowanie spotkań oraz konferencji naukowych. W skład działalności naukowej Doktoranta wchodzi także udział w czterech projektach badawczych, wymienionych w załączniku A.

Rozprawa zawiera szereg oryginalnych rozwiązań problemów naukowych z zakresu analizy, porównywania i ulepszania modeli struktur przestrzennych RNA. Doktorant posiada imponującą wiedzę teoretyczną i praktyczną w dyscyplinie Informatyka Techniczna i Telekomunikacja. Nie mam także żadnych wątpliwości, że Pan mgr Jakub Wiedemann posiada zdolność samodzielnego prowadzenia pracy naukowej. Podsumowując, z pełnym przekonaniem stwierdzam, że przedstawiona mi do recenzji praca doktorska Pana mgr Jakuba Wiedemanna spełnia wymogi stawiane rozprawom doktorskim zdefiniowane przez artykuł 13 Ustawy z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym (z późniejszymi zmianami). Stąd też, wnioskuję o dopuszczenie Doktoranta do dalszych etapów przewodu doktorskiego.

Jednocześnie mając na uwadze wysoki poziom naukowy przedstawionej rozprawy oraz jej znaczenie dla rozwoju dziedziny biologii RNA oraz bioinformatyki strukturalnej w zakresie analizy struktur przestrzennych RNA, wnioskuję o wyróżnienie niniejszej dysertacji.

B. Rotajczak