Uniwersytet Wrocławski

WYDZIAŁ BIOTECHNOLOGII
prof. dr hab. **Paweł Mackiewicz**
ZAKŁAD BIOINFORMATYKI I GENOMIKI
ul. F. Joliot-Curie 14a
50-383 Wrocław
tel. +48 71 375 63 03
pamac@smorfland.uni.wroc.pl

Wrocław, 17.04.2025

Recenzja rozprawy doktorskiej Pana mgr Adriana Kani pt. "Reprezentacja gry chaosu w badaniach sekwencji nukleotydowych i aminokwasowych"

Już od czasu uzyskania pierwszych sekwencji nukleotydowych i aminokwasowych naukowcy zastanawiali się jak je analizować. Bardzo popularne stało się przyrównywanie (dopasowanie) sekwencji w celu znalezienia podobieństwa i różnic między nimi. Mimo dużej skuteczności tych metod, wyniki przyrównań bardzo zależą od przyjętego systemu punktacji i kary na wprowadzane przerwy. Optymalne wartości tych parametrów nie są łatwe do ustalenia dla porównywanego zbioru sekwencji. Ponadto, ze znacznym wzrostem liczby dostępnych sekwencji, metody te stają się zbyt czasochłonne, aby przeprowadzić analizy w rozsądnym czasie na dużym zbiorze sekwencji pomimo ulepszania algorytmów przyrównujących sekwencje. W związku z tym zaczęto proponować i stosować różne analizy nieopierające się na przyrównaniach. Jedną z nich jest reprezentacja gry chaosu (ang. Chaos Game Representation, CGR) opracowana została przez H. J. Jeffreya w 1990 roku oraz zaczerpnięta z teorii układów dynamicznych i fraktali. Metoda ta przekształca ciągu symboli (np. nukleotydów DNA) w dwuwymiarowy wykres, który oddaje zarówno lokalne, jak i globalne wzorce obecne w analizowanej sekwencji. Metoda ta może dawać interesujące wyniki i nowe spojrzenie na analizowane sekwencji. Dlatego bardzo słusznie ambitnym przedmiotem pracy doktorskiej Pana mgr Adriana Kani stało się zastosowanie tej metody w badaniach sekwencji nukleotydowych i aminokwasowych.

Rozprawa doktorska została napisana w języku polskim i ma klasyczny układ. Zawiera ona: Spis treści, Słowa kluczowe, Wykaz najważniejszych skrótów, Streszczenie, Abstract w języku angielskim, Wprowadzenie, Rozdział zawierający hipotezy badawcze i cele pracy, Metody i zasoby, Wyniki i dyskusję, Podsumowanie, Dodatki, Literaturę, Spis rysunków i Spis tabel.

We Wprowadzeniu doktorant zebrał najważniejsze informacje dotyczące analiz sekwencji biologicznych skupiając się na metodach nieopartych na klasycznych przyrównaniach sekwencji, a zwłaszcza na metodzie reprezentacji gry chaosu (CGR). Podał zastosowania tej metody w analizie genomów, badaniu genów niezbędnych, przewidywaniu odporności bakterii na antybiotyki oraz celów mikroRNA.

Sądzę, że w tej części można byłoby podać przykładową graficzną reprezentację sekwencji metodą CGR i uwypuklić zalety tej metody w porównaniu do innych. Można byłoby też wyjaśnić przynajmniej ogólnie wspominane w tej części parametry służące do opisu sekwencji badanych tą metodą, tj. wykładnik Hursta i dyskretną transformatę Fouriera. Przy stwierdzeniu, że obliczenia przeprowadzono dla genu neuraminidazy wypadałoby podać, że pochodził on z ptasich wirusów grypy. Przy zdaniu opisującym, że znacząca część wyników tej pracy została już opublikowana dobrze byłoby zacytować te prace.

Hipotezy badawcze i cele pracy doktorskiej zostały jasno sformułowane oraz właściwie zweryfikowane i zrealizowane. Warto podkreślić, że w tej części poza ich przedstawieniem podano także ich uzasadnienie i potencjalne aplikacyjne zastosowanie opracowanych metod i uzyskanych wyników. Ogólnie celem pracy było zbadanie stosowalności i ograniczeń CGR w analizach sekwencji nukleotydowych oraz aminokwasowych. Poza zastosowaniem tej metody doktorant zaproponował także nowe rozwiązania metodyczne. W szczególności mgr Kania zastosował metodę CGR do scharakteryzowania motywów sekwencyjnych mikroRNA człowieka, cech i klasyfikacji sekwencji genów niezbędnych u bakterii oraz rekonstrukcji relacji ewolucyjnych genów, genomów oraz białek. Ponadto doktorant zbadał własności kodowania sekwencji nukleotydowych DNA i RNA oraz aminokwasowych tą metodą oraz zaproponował użyteczne deskryptory.

Mam pewne uwagi do tej części. Stwierdzenie, że celem jest potwierdzenie zapewne znanych z literatury wzorców nie jest fortunne, ponieważ doktorant powinien sprawdzić czy te wzorce są już znane. Słowo zapewne jest tutaj niewłaściwe. Dobrze byłoby także uzasadnić dlaczego do rekonstrukcji relacji ewolucyjnych wybrano gen neuraminidazy z ptasich wirusów grypy oraz gen SPARC z organizmów wyższych.

W rozdziale Metody i zasoby w przejrzysty sposób przedstawiono wykorzystane bazy danych oraz zastosowaną metodykę. Umieszczono tam także pewne opisy pełniące funkcję wprowadzenia lub wstępu do poszczególnych zagadnień. Sądzę, że jest to dobre rozwiązanie niż umieszczanie wszystkiego w jednej części będącej Wstępem.

W tym rozdziale tym scharakteryzowano biologiczne bazy danych dotyczące sekwencji biologicznych oraz podano podstawowe informacje o przyrównywaniu sekwencji oraz analizach filogenetycznych. Opisano także wybrane specyficzne cechy i właściwości rzeczywistych sekwencji biologicznych. Szczegółowiej scharakteryzowano metody opisu i reprezentacji sekwencji nukleotydowych i białkowych: entropię i logo, k-mery, złożoność Lempel-Ziva, krzywą Z, reprezentację gry chaosu, genomacierz, spacer DNA i deskryptory PseAAC. Osobne rozdziały poświęcono na charakterystykę dyskretnej transformaty Fouriera (DFT), wykładnika Hursta, ilościowej analizy rekurencji (RQA), momentu bezwładności oraz indeksu podobieństwa strukturalnego (SSIM).

Generalnie opisy są jasne z interesującymi przykładami, jednakże mam również uwagi do tej części. Podana w opisie baza Pfam nie jest już wspierana, a jej dane znajdą się w bazie InterPro. Można było podać przykład tej drugiej bazy jako wtórnej zawierającej motywy i domeny. Chciałbym się dowiedzieć skąd doktorant zaczerpnął podział baz na pierwszo-, drugo- i zwłaszcza trzeciorzędowe. Jakie są kryteria ich klasyfikacji? Wypadałoby podać adresy internetowe do baz miRBase i miRTarBase. Dobrze byłoby podać jakie metody filogenetyczne oraz cechy i parametry sekwencji (opisane w 3.5. Specyfika sekwencji nukleotydowych oraz prawa biologiczne oraz 3.6. Metody opisu i reprezentacji sekwencji nukleotydowych i białkowych) zastosowano w pracy doktorskiej. Wspominając o różnej reprezentacji kodów genetycznych dobrze byłoby wyjaśnić na czym ona polega. W opisie podano, że Wacław Szybalski zauważył, że u wirusów występuje średnio więcej puryn niż pirymidyn i ma to zapobiegać tworzeniu struktur dwuniciowych pomiędzy fragmentami RNA. Jednak nie jest jasno napisane czy tworzenie tych struktur czy ich zapobieganie ma spowalniać replikację RNA i wpływać na formowanie się białek. Jeśli chodzi o drugą regułę parowania Chargaffa, to warto byłoby wspomnieć o odchyleniach od niej, które wynikają z występowania sekwencji kodujących (ich nici sensownych) o specyficznym składzie na danej nici DNA oraz tzw. asymetrii DNA związanej z różnymi substytucjami nukleotydowymi na niciach wiodącej i opóźniającej podczas replikacji. Dobrze byłoby wyjaśnić, co to są sekwencje o zwykłej i odwrotnej komplementarności. Można byłoby też podać jak interpretować wartości złożoność Lempel-Ziva. Przy opisie metody CGR podano, że autor zaproponował przyporządkowanie aminokwasów do wierzchołków z uwzględnieniem ich właściwości fizykochemicznych, jednak dobrze byłoby szczegółowiej to opisać. Niektóre wykresy gry w chaos przypominają trójkąty Sierpińskiego o czym można byłoby też wspomnieć. We wzorze 27 dobrze byłoby wyjaśnić f_i. Można byłoby też podać jawne wyliczenia niektórych wartości w Tabeli 3 i ich interpretację oraz wyjaśnić jaką skalę wybrano do liczenia hydrofobowości i hydrofilowości.

Wspomniano także o kryterium Nyquista, ale nie opisano go szczegółowiej. Dziwi mnie wzór 36 na średnią cząstkową, ponieważ jest on zwykłą sumą wartości a nie ilorazem. Dla jasności dobrze byłoby podać pewne przykłady oraz jak interpretować wartości z ilościowej analizy rekurencji oraz indeksu podobieństwa strukturalnego.

Analizy oparte na metodzie spacerów DNA oraz jej modyfikacjach były powszechnie stosowane też przez polskich naukowców z grupy prof. Stanisława Cebrata i prof. Mirosława Dudka, w których miałem zaszczyt brać udział. Poniżej przedstawiam przykładowe prace, w których były one stosowane. Można byłoby je zacytować we Wprowadzeniu i części metodycznej:

Stanisław Cebrat, Mirosław Roman Dudek (1998) The effect of DNA phase structure on DNA walks. The European Physical Journal B 3: 271-276

Stanisław Cebrat, Mirosław Roman Dudek, Agnieszka Gierlik, Maria Kowalczuk, Paweł Mackiewicz (1999) Effect of replication on the third base of codons. Physica A 265 (1-2): 78-84

Paweł Mackiewicz, Agnieszka Gierlik, Maria Kowalczuk, Mirosław Roman Dudek, Stanisław Cebrat (1999) How does replication-associated mutational pressure influence amino acid composition of proteins? Genome Research 9 (5): 409-416

Paweł Mackiewicz, Agnieszka Gierlik, Maria Kowalczuk, Dorota Szczepanik, Mirosław Roman Dudek, Stanisław Cebrat (1999) Mechanisms generating correlation in nucleotide composition in Borrelia burgdorferi genome. Physica A 273 (): 103-115

Paweł Mackiewicz, Maria Kowalczuk, Dorota Mackiewicz, Aleksandra Nowicka, Małgorzata Dudkiewicz, Agnieszka Łaszkiewicz, Mirosław Roman Dudek, Stanisław Cebrat (2002) How many protein-coding genes are there in the Saccharomyces cerevisiae genome? Yeast 19 (7): 619-629

Paweł Mackiewicz, Maria Kowalczuk, Dorota Mackiewicz, Aleksandra Nowicka, Małgorzata Dudkiewicz, Agnieszka Łaszkiewicz, Mirosław Roman Dudek, Stanisław Cebrat (2002) Replication associated mutational pressure generating long-range correlation in DNA. Physica A 314 (1-4): 646-654

Paweł Mackiewicz, Jolanata Zakrzewska-Czerwińska, Anna Zawilak, Mirosław Roman Dudek, Stanisław Cebrat (2004) Where does bacterial replication start? Rules for predicting the oriC region. Nucleic Acids Res. 32 (13): 3781-3791

Wyniki i ich dyskusja ostały połączone w jeden rozdział, co można uznać za uzasadnione, ponieważ ułatwiło to komentowanie na bieżąco uzyskanych wyników i uniknięcie zbędnych powtórzeń. Warto podkreślić, ze doktorant jest współautorem 4 publikacji przedstawiających wyniki związane z tematem pracy doktorskiej. We wszystkich jest autorem korespondencyjnym.

Wyniki analiz motywów sekwencyjnych w miRNA wydają się bardzo ważne z punktu widzenia zrozumienia mechanizmów ich oddziaływania z mRNA. Doktorant znalazł wyraźne różnice między grupami miRNA. Jednakże bardziej zrozumiałe i informatywne są zwykłe analizy częstości dinukleotydów niż skomplikowane obrazy CGR. Co nowego one wnoszą niż analizy częstościowe? Dobrze byłoby podać także jaka wartość została zastosowana do podziału miRNA na nisko- oraz wysokoenergetyczne. Nie rozumem stwierdzenia, że analogiczne wzorce w grupie miRNA wysokoenergetycznych i o największej wiarygodności mają pośrednio wskazywać na istotną rolę innych czynników w procesie degradacji, w tym kompleksu RISC. Prosiłbym o wyjaśnienie. Mam również pytanie dlaczego dinukleotydy z zasadami tworzącymi podwójne wiązania wodorowe występują częściej w grupie miRNA o wysokiej energii, a tworzące potrójne w grupie o niskiej energii? Czy ta energia to jest energia oddziaływań między miRNA i mRNA? Jeśli tak, to czy nie powinna być odwrotna zależność?

Więcej interesujących analiz przeprowadzono dla genów niezbędnych i opcjonalnych u bakterii. Nie wykazano istotnych różnic w ich długościach opierając się na analizach korelacji. Mam jednak pytanie dlaczego nie zastosowano do oceny różnic w długościach genów innych testów statystycznych jak t-Studenta lub Manna-Whitneya bezpośrednio oceniających te różnice? Doktorant znalazł także, że geny niezbędne mają istotne odchylenie od drugiej reguły Chargaffa liczonej na różne sposoby a niektóre z nich wykazały większą entropię. W przypadku analiz korelacyjnych w tym przypadku dobrze byłoby wyjaśnić jakiego testu dotyczyły wartość p. Zastosowano też w tych analizach dywergencję Kullbacka-Leiblera, o której wypadałoby wspomnieć w Metodach i zasobach. Pan mgr Kania wykazał także, że geny niezbędne są reprezentatywne ze względu na skład do innych genów tego samego genomu, a mniej w porównaniu do innych genomów, co może wydawać się oczywiste ze względu na różnice w składzie między genomami. Zastosowano także połączenie metody CGR ze spacerem DNA, które wykazało różnice między sekwencjami rzeczywistych genów i sekwencji losowych. Dobrze byłoby wyjaśnić jak te sekwencje losowe wygenerowano. Na Rysunku 42 uzyskano dwa zbiory punktów dla zależności momentów bezwładności, a niektóre genomy wykazały duże wartości tych parametrów. Jest to interesujące i dobrze byłoby te wyniki wytłumaczyć.

Wspominano także o zastosowaniu uczenia maszynowego do rozpoznawania genów niezbędnych. Dobrze byłoby wyjaśnić jaki był udział doktoranta w tych badaniach.

CGR w kombinacji z innymi parametrami oraz metodę Lempel–Ziva zastosowano także w analizie filogenetycznej dla sekwencji wygenerowanych i rzeczywistych oraz z AFproject. Wykazano, że najlepiej oczekiwane drzewa odtworzyły metoda CGR z DFT i SSIM. Dokonano też porównania czasów działania różnych metod. Dobrze byłoby pokazać ile czasu zajęłyby klasyczne analizy z przyrównaniem sekwencji i określeniem drzewa metodą odległościową lub inną. Mam też pytanie dlaczego w niektórych przypadkach pokazywano nieznormalizowane a w innych znormalizowane odległości między drzewami?

Ważne i interesujące wyniki dotyczą rozszerzenia metody CGR dla DNA, RNA oraz białek, ponieważ zaproponowano w nim i przetestowane różne sposoby reprezentacji i nadawania wag dla punktów dodawanych w metodzie CGR. Ciekawe jest zastosowanie nowego sposobu umieszczania punktów dla nukleotydów sparowanych w drugorzędowych strukturach RNA oraz wybór cech fizykochemicznych aminokwasów do CGR z wykorzystaniem sieci neuronowych. Zabrakło mi jednak szczegółowszego wyjaśnienia w jaki sposób dane z tych sieci posłużyły do przypisania współrzędnych wierzchołków dla kolejnych aminokwasów. Przetestowano także różne funkcje wagowe, aby sprawdzić, które są najbardziej odporne na zmiany w położeniach punktów w CGR po wprowadzonych mutacjach w jednej z sekwencji. Nie uważam jednak, że funkcja wagowa odporna na mutacje jest zawsze lepsza, jeśli interesują nas niewielkie różnice między sekwencjami. Nie jest dla mnie jasny opis rankingu i stwierdzenie, że im wyższa wartość S_c, tym bardziej odporna na mutacje jest dana reprezentacja, skoro wartość 0 oznacza to samo położenie w rankingu porównywanych punktów. A zatem wartość bliska zeru powinna być najlepsza. Co to znaczy, że ranking, czyli s_2 może być niższy? Według mnie albo dany punkt jest na początku rankingu i odpowiada s_1, albo może być dalej czyli wyższy w rankingu. Zastosowanie różnych funkcji wagowych i właściwości aminokwasów pokazało, ze jest możliwe uzyskanie wiarygodnych relacji filogenetycznych w oparciu o CGR i DFT. Jest to ważne osiągnięcie, które może mieć praktyczne zastosowanie.

Podsumowanie jest poprawnie napisane i zawiera najważniejsze wnioski wynikające z przeprowadzonych analiz. Wskazuje, że doktorant potrafi wybrać najważniejsze rezultaty ze swoich badań. Do pracy dołączono też adres internetowy do programów napisanych w ramach tej rozprawy.

Generalnie praca jest napisana poprawnym językiem. Znalazłem następujące błędy z propozycją poprawek: dinukloetydów -> dinukleotydów; scharakreryzować -> scharakteryzować; orgazmami -> organizmami; organizmów dwuniciowych -> organizmów z dwuniciowym DNA; występujące w sekwencji korelacji -> występujące w sekwencji korelacje; nie obserwowane -> nieobserwowane; ilość bakterii -> liczba bakterii; *tuberculosis* -> *tuberculosis*; kilu -> kilku; ilością sparowanych nukleotydów -> liczbą sparowanych nukleotydów. Ponadto w wielu miejscach brakuje przecinków przed: aby, który, które, gdy, co, a zatem; a.

Skrót DEG powinien być wyjaśniony przy pierwszym użyciu w Streszczeniu, a ND5 w Hipotezach badawczych i celach pracy. Zamiast terminu dopasowanie sekwencji stosowałbym przyrównanie sekwencji. Termin Alignment-Free Methods przetłumaczyłbym jaki metody nieoparte na przyrównaniu, a Unweighted Pair Group Method With Arithmetic Mean jako metoda średnich połączeń nieważonych. Bazy danych pierwszorzędowe zmieniłbym na bazy danych pierwotne albo podstawowe, natomiast bazy danych drugorzędowe na bazy danych wtórne albo pochodne.

Praca jest dobrze sformatowana. Jednak często nie było powołań w tekście na rysunki i tabel, co utrudniało analizowanie treści. Rysunki powinny być większe.

Reasumując mogę stwierdzić, że moje powyższe uwagi nie umniejszają znaczenia pracy doktorskiej. Doktorant wykazał się dużą wiedzą teoretyczną oraz włożył dużo trudu w opracowanie nowatorskich metod i przeprowadzenie analiz, a przedstawione opisy wyników i ich interpretacje świadczą o jego dużej dojrzałości naukowej, samodzielności, wnikliwości i głębokim zrozumieniu zagadnienia. Należy podkreślić, że analizy zostały przeprowadzone skrupulatnie, a zastosowana metodyka okazała się skuteczna.

Tematyka pracy doktorskiej jest ważna dla rozwoju nauk biologicznych, ponieważ istnieje potrzeba tworzenia nowych metod i narzędzi analizujących sekwencje nowymi metodami, co stało się przedmiotem rozprawy. Warto zaznaczyć, że doktorant nie tylko wykorzystywał dotychczasowe metody, ale po ich dogłębnym zrozumienia opracował i zastosował nowe. Uważam, że recenzowana rozprawa jest oryginalna i stanowi istotny wkład w metodykę i analizę sekwencji w oparciu o metody nieoparte o klasyczne przyrównania.

Uważam, więc, że przedstawiona do recenzji rozprawa doktorska Pana mgr Adriana Kani spełnia wszystkie wymogi określone w art. 187 Ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (Dz. U. z 2018, poz. 1668 z późn. zm.). Zgłaszam, zatem wniosek do Rady Dyscypliny Nauki Biologiczne Uniwersytetu Jagiellońskiego w Krakowie o

uznanie rozprawy Pana mgr Adriana Kani za odpowiadającą wymogom stawianym rozprawom doktorskim i o dopuszczenie doktoranta do dalszych etapów postępowania ws. nadania stopnia doktora w dziedzinie nauk ścisłych i przyrodniczych w dyscyplinie nauki biologiczne.

Prof. dr hab. Paweł Mackiewicz

Paweł Mackiewicz