

Recenzja rozprawy doktorskiej pt.

Generowanie długich otwartych ramek odczytu w genomach prokariotycznych

Krystiana Bączkowskiego.

Przedstawiona do recenzji praca doktorska dotyczy zagadnienia z zakresu genomiki, gdzie doktorant pokazuje mechanizmy pojawiania się w genomach nowej jakościowo informacji genetycznej w postaci otwartych ramek odczytu (ORF) generowanych w alternatywnych fazach genów kodujących białko. Są to zagadnienia fundamentalne dla rozumienia mechanizmów ewolucji genetycznej i w przypadku tej rozprawy istotnie uzupełniają standardowe traktowanie zmian w informacji genetycznej wyłącznie jako prostej modyfikacji informacji już istniejącej np. poprzez mechanizmy fuzji i fizji genów, modyfikacji genów w wyniku procesów duplikacji czy tasowania fragmentów genu. Wydawałoby się, że obecnie przy przeogromnej ilości danych sekwencji różnych organizmów w bazach danych, które wciąż zwiększają się, praca badawcza w genomice to głównie statystyczne analizy porównawcze pokrewieństw organizmów. Tymczasem to właśnie dzięki ilości i różnorodności wyspecjalizowanych genetycznych baz danych możliwe było pokazanie przez doktoranta nowych statystycznie istotnych wielkoskalowych własności kodu genetycznego, tj. generowania nowej informacji genetycznej poprzez istniejące geny organizmu co z kolei umożliwia szybszą ewolucję tego organizmu. Te wielkoskalowe własności informacji genetycznej związane są z funkcją genu i korelacjami pomiędzy funkcjami różnych genów a tym samym zdeterminowane są przez kolektywny wkład bardzo wielu nukleotydów. Użyłem określenia kolektywny bo cecha kolektywności to trochę jak z opisem drgań atomów w kryształach przez fizyków, albo drgania pojedynczych atomów zostaną wygaszone albo zamienią się w kolektywne drgania wszystkich atomów. Obecność funkcji genów i korelacji pomiędzy nimi zaznacza się w genomie między innymi poprzez jego specyficzny skład nukleotydowy. Szczególnym przykładem może być procentowa zawartość guaniny i cytozyny w genomie z której wzrostem rośnie udział ORFów dla których można znaleźć homologi po przeczytaniu w jednej z faz alternatywnych. Tego typu

korelacje zostały przedstawione na Ryc. 64 w rozprawie. Podobnie, jest z zależnością dla znalezionych przez doktoranta domen, motywów i struktur białkowych w fazach alternatywnych ORFów (Ryc. 65 w rozprawie). W rozprawie pokazano, że ze względu na skład GC największe możliwości generowania ORFów w fazach alternatywnych mają Acidobacteria, δ -Proteobacteria, Deinococcus-Thermus, β -Proteobacteria i Actinobacteria. Analizy statystyczne genomów przeprowadzone przez Autora rozprawy udowadniają, że ponad 18% badanych ORFów jest podobna do innych ramek odczytu w alternatywnych fazach. Przetłumaczone przy pomocy programów komputerowych na białka alternatywne odczyty adnotowanych ORFów wykazują największe podobieństwo do rzeczywistych białek w przypadku genomów mających 35-50% zawartości GC (Ryc. 63). Dane te dotyczą genomów bakteryjnych i analizom poddano 457 genomów a w nich około 1,5 mln ORFów.

Rozprawa doktorska składa się z ładnie napisanego wstępu dotyczącego historycznego rozwoju biologii molekularnej w kontekście stosowanych technik informatycznych i organizacji informacji genetycznej w genomach oraz czterech rozdziałów zasadniczych pracy, a następnie wniosków, aneksu i bogatej bibliografii. W Rozdziale 2. określono cele pracy, dokładnie osiem punktów i wszystkie zostały zrealizowane. Za najważniejsze uważam zbadanie czy w dostępnych kompletnie zsekwencjonowanych genomach mikroorganizmów możliwe jest pojawienie się nowej informacji genetycznej poprzez mechanizm generowania ORFów. Drugim bardzo ważnym zadaniem było moim zdaniem zbadanie czy elementy funkcjonalne i strukturalne białek (domeny i motywy białkowe) mają wpływ na generowanie nowej informacji genetycznej w fazach alternatywnych. Jako trzecie co do ważności zrealizowanych zadań uważam sprawdzenie czy odpowiedzi na zagadkę sierocych ORFów można szukać w generowaniu ORFów w alternatywnych odczytach genów. Kolejny ważny cel pracy, bardziej o charakterze logistycznym, to utworzenie zintegrowanego narzędzia badań do analiz możliwości generowania ORFów w genomach i odnajdowania błędnie adnotowanych sekwencji.

Rozdział 3. Materiały i metody pokazuje warsztat badawczy Autora rozprawy. Sekwencje genomów i dodatkowe informacje dotyczące badanych organizmów takie jak dane taksonomiczne, dane o funkcji kodowanych białek itd. pobrano z bazy GenBank. Do badań porównawczych wykorzystano szereg pakietów programowych takich jak Emboss do tłumaczenia sekwencji nukleotydowych na aminokwasowe, Blast do poszukiwania homologów na poziomie sekwencji aminokwasowych oraz do przyporządkowywania sekwencji do odpowiednich grup ortologicznych COG, GeneMark do określania potencjału kodowania oraz pakiet InterProScan z 14 aplikacjami klienckimi do

poszukiwań motywów, domen i odpowiednich rodzin białkowych. Dodatkowo, Autor rozprawy użył ponad 40 własnych programów pomocniczych w języku Python i w językach skryptowych Bash i Perl automatyzujących poszukiwania i umożliwiających eksport wyników analiz do baz danych MySQL. Analizy statystyczne były robione przy pomocy pakietu R. Informacje te dostępne są w Tabeli 10.

Dla porównań na poziomie aminokwasowym wszystkim aminokwasom przyporządkowane zostały wartości liczbowe (indeksy) skorelowane z ich własnościami fizykochemicznymi. Wprowadzono sześć rodzajów indeksów: indeksy hydropatyczności, indeksy polarności, indeksy α helis, indeksy β struktur, indeksy transbłonowości i indeksy mutabilności. Odległość w składzie aminokwasowym produktów ORFów określano przy pomocy metryki euklidesowej w 4-wymiarowej przestrzeni zadanej przez parametry: częstość reszt kwasowych, częstość reszt zasadowych, częstość reszt polarnych i częstość reszt niepolarnych. Tutaj mam pytanie: Jak były liczone te częstości? Czy to średnia arytmetyczna czy też częstości z wagą uwzględniającą różnice w długościach porównywanych sekwencji?

Do badań substytucji aminokwasowych w pakiecie BLAST wybrano macierze podstawień PAM30, PAM70, BLOSUM80 i BLOSUM62 dla odpowiednich czterech zakresów długości sekwencji liczonej w aminokwasach (Tabela 6) tj. do 35 aminokwasów, od 35 do 50 aminokwasów, od 50 do 85 aminokwasów i powyżej aminokwasów. Uważam, że w Rozdz. 3.3.2 powinno się znaleźć dodatkowe uzasadnienie podziału sekwencji na te cztery zakresy długości oraz uzasadnienie dla użycia wybranych wersji tablic PAM i BLOSUM, dlaczego takie a nie inne?

Podsumowując Materiały i Metody przyznam, że jestem pełen podziwu dla bardzo wszechstronnego doboru baz genomowych, doboru pakietów programowych i przygotowania interfejsu programowego umożliwiającego późniejsze wykorzystanie przez innych.

Z wielkim zaciekawieniem czytałem Rozdział 4, w którym przedstawione zostały wyniki analiz statystycznych dla generowania ORFów, generowania genu przez gen i sugerowanych mechanizmów ewolucji informacji genetycznej. Pokazano, że największy potencjał generowania ORFów mają fazy czwarta i piąta sekwencji kodujących białko.

Wnioski doktoranta oparte są na danych statystycznie istotnych. Dane statystyczne uzupełnione zostały wybranymi przykładami co pozwala lepiej zrozumieć proponowane mechanizmy ewolucyjne. Wyniki niektórych szczegółowych analiz, na przykład dla rozkładu punktu

izoelektrycznego – Ryc. 33 na pewno czekają na dalsze badania, na przykład w kontekście diskutowanych mechanizmów generowania informacji genetycznej.

Rozdział 5. to skondensowane omówienie uzyskanych wyników.

Dodam, że w całej pracy znalazłem tylko 4 literówki na 200 stron. Praca zawiera 200 pozycji literaturowych.

Na zakończenie dodam, że Autor rozprawy proponuje wprowadzenie nazwy alterlogi dla sekwencji otrzymanych w wyniku alternatywnego odczytu innych sekwencji. Życzę doktorantowi i zespołowi w którym prowadzone są badania aby to pojęcie przyjęło się analogicznie do pojęć takich jak ortologi, paralogi i ksenologi,

Stwierdzam, że merytoryczna zawartość pracy spełnia wymagania stawiane pracom doktorskim.

Proszę o dopuszczenie jej do dalszych etapów przewodu doktorskiego. Uważam, że praca zasługuje na wyróżnienie.

