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, 20.08.2022

Recenzja rozprawy doktorskiej Pani mgr Anny Macioszek pt. "HMM-based method for identifying enrichment in signal from sequencing-based experiments"

Intensywnie rozwijające się techniki biologii molekularnej i metody sekwencjonowania dostarczają coraz więcej danych sekwencyjnych. Dzięki temu biolodzy mają coraz lepszy wgląd w procesy molekularne zachodzące w organizmach żywych. Bardzo cennym źródłem takich danych jest sekwencjonowanie nowej generacji (ang. NGS, Next Generation Sequencing), które staje się coraz szybsze i tańsze, a biolodzy wciąż poszukują w nich odpowiedzi na nowe pytania z różnych zagadnień biologii molekularnej. Jednakże duże ilości danych są bezwartościowe, jeśli nie można ich obrobić i wyciągnąć odpowiednich informacji. W związku z tym istnieje konieczność ciągłego tworzenia nowych narzędzi bioinformatycznych służących do tych celów.

Dlatego bardzo słusznie ambitnym przedmiotem pracy doktorskiej Pani mgr Anny Macioszek stało się opracowanie nowego narzędzia bioinformatycznego nazwanego HERON (ang. HiddEn maRkov mOdel based peakcalliNg) przeznaczonego do analizy danych sekwencyjnych pochodzących z eksperymentów precypitacji chromatyny ChiP-seq (ang. chromatin immunoprecipitation-sequencing) identyfikujących miejsca w DNA oddziałujące z białkami. Technika ta polega na uzyskiwaniu fragmentów DNA, do których przyłączyły się badane białka, a następnie sekwencjonowaniu tych fragmentów i mapowaniu na genom. Końcowe etapy tych badań wymagają odpowiednich analiz bioinformatycznych. Kluczowym etapem jest rozpoznawanie obszarów wzbogaconych w sygnały pokrywające genom, co nie jest proste. Dlatego celem opracowanego programu stało się udoskonalenie tej procedury zwanej po angielsku peakcalling.

Mimo, że istnieje wiele programów służących do takiej analizy tych danych, ich wyniki nie są satysfakcjonujące i porównywalne. Dlatego konieczne jest opracowywanie wciąż nowych podejść i algorytmów. Jednym z nich jest właśnie HERON.

Rozprawa doktorska została napisana w języku angielskim ze streszczeniem w języku polskim. Zawiera ona wstęp przedstawiający podstawowe informacje dotyczące zagadnień genetycznych i sekwencjonowania. Osobny rozdział poświęcono identyfikowaniu sygnału pochodzącego z sekwencjonowania w genomie. Sam program scharakteryzowano w kolejnym rozdziale. Odpowiada on części metodycznej. Następne rozdziały odpowiadają wynikom i dyskusji. Zawierają one porównanie wyników działania opracowanego programu z innymi narzędziami na podstawie danych symulacyjnych i dwóch typach danych eksperymentalnych. Każdy z tych trzech rozdziałów zawiera podsumowanie zawierające dyskusję i najważniejsze wnioski wynikające z tych porównań. Na końcu rozprawy zamieszczono końcowe podsumowanie zawierające ogólne wnioski i dyskusję wynikające z otrzymanych rezultatów.

Wstęp jest przejrzyście napisany i odpowiedni do wprowadzenia czytelnika do omawianych zagadnień oraz zrozumienia stawianych problemów. Przedstawiono w nim budowę DNA i RNA oraz ekspresję informacji genetycznej, transkrypcję i translację, oraz konwencję oznaczania i nazywania nici w sekwencji kodującej. Opisano także organizację chromatyny eukariotycznej uwzględniając histony tworzące nukleosomy, a także scharakteryzowano ich modyfikacje oraz rolę w regulowaniu ekspresji genów. W rozdziale dotyczącym sekwencjonowania została krótko scharakteryzowana metoda Sangera oraz szerzej metoda NGS. Wspomniano także o metodach sekwencjonowania trzeciej generacji. Doktorantka wskazała wady i zalety metod sekwencjonowania. W osobnym podrozdziale opisano przykłady eksperymentów stosujących NGS, jak ChiP-seq, ATAC-seq i RNA-seq oraz przedstawiono jedną z najczęstszych i podstawowych analiz, czyli mapowanie odczytów na genom referencyjny i normalizację liczby odczytów. Pod koniec części wstępnych uzasadniono konieczność opracowywania nowych narzędzi bioinformatycznych do analizy danych NGS.

Opisy poszczególnych zagadnień są precyzyjne i jasne. Świadczą, że doktorantka dobrze zrozumiała podstawy procesów molekularnych, które stały się przedmiotem opracowanego narzędzia. Jednakże mam kilka uwag do tej części. W opisie RNA-seq uściśliłbym, że ta technika służy do oceny ilości transkryptów, tzn. ilości RNA, bo z opisu wynika, że również białek. Podając ekstremalne wielkości genomów można było podkreślić, ze chodzi o prokarioty i eukarioty, bo są mniejsze genomy wirusowe i organellowe. Użyto termin "non-coding genes", ale go nie wytłumaczono. Zdanie: "The two strands are antiparallel to each other, i.e. where one has the 5' end, the other has the 3' end." nie jest do końca ścisłe, ponieważ każda z nici posiada koniec 3' i 5' tylko ustawione w przeciwnych kierunkach. Zdanie sugeruje, że mają tylko po jednym końcu. Zamiast "translated into ribonucleotides"

powinno być "transcribed into ribonucleotides". Uważam, że stosowanie terminu nić sensowna i antysensowna do całego genomu jest niepoprawne. Nawet jeśli ktoś je użył w tym kontekście, to jest to niezgodne z ich definicją. Powinny być one stosowane tylko w kontekście genu, jak to zostało opisane wcześniej przez doktorantkę. Natomiast terminy nić Watsona (+) i Cricka (-), a jeszcze lepiej nić podana (forward) i komplementarna (reverse) powinny być stosowane do sekwencji całych genomów deponowanych w bazie danych. Przy opisie ATAC-seq można było dodać, że po oczyszczeniu DNA otagowane sekwencje są namnażane w reakcji PCR. Jednakże warto podkreślić, że doktorantka zamieściła ładne rysunki, które są jej oryginalnym dziełem. Bardzo dobrze obrazują one opisywane zagadnienia.

Rozdział dotyczący identyfikowania sygnału pochodzącego z sekwencjonowania w genomie (ang. peakcalling) został dobrze opracowany. Uwzględniono najważniejsze problemy z tym związane, np. fałszywe sygnały, porównanie z tłem, niejednorodność i nierównomierność sygnału. W osobnych podrozdziałach zostały opisane algorytmy i zasady działania programów służących do identyfikowania sygnałów: MACS, BayesPeak i SICER oraz inne podejścia. Pod koniec tego rozdziału znalazło się także uzasadnienie konieczności opracowywania nowych programów, ponieważ skuteczność działania dotychczasowych narzędzi zależy od typu danych. Szczególny problem dotyczy identyfikowania stosunkowo długich domen słabo wzbogaconych w odczyty pochodzących np. z eksperymentów, w których metylowane są histony. Istotne jest tutaj uwzględnienie zależności pozycji w genomie, a nie traktowanie ich niezależnie jak większość programów. Dlatego celem rozprawy doktorskiej stało się opracowanie narzędzia służącego do analizy takich danych. To uzasadnienie można traktować jako cel pracy, który został w pełni zrealizowany. Jednakże dobrze byłoby podać więcej przykładów eksperymentów, które generują takie pokrycie genomu i sygnały. Rozumiem konieczność stosowania innych programów do identyfikowania domen o małym pokryciu, ale chciałbym zapytać doktorantkę, czy zmieniając odpowiednie parametry programów wyspecjalizowanych do silnych i wąskich sygnałów, np. długość oczekiwanego sygnału i próg istotności siły sygnału, można uzyskać lepszą skuteczność w identyfikowaniu takich słabo wzbogaconych domen.

Zasady działania algorytmu pakietu HERON zostały formalnie przedstawione w rozdziale 3. Zaletą tego programu jest uwzględnienie zależności między pozycjami, co ma uzasadnienie biologiczne. Takiej zależności nie zakłada popularny program MACS. Narzędzie HERON opiera się na ukrytych modelach Markowa (HMM) z ciągłymi emisjami. Kolejną zaletą programu doktorantki jest uwzględnienie trzech stanów: braku sygnału, tła i obecności

sygnału. To prowadzi do bardziej precyzyjnej identyfikacji sygnałów w sekwencji genomu. Do określania najbardziej prawdopodobnej kolejności stanów dla obserwowanej sekwencji oraz parametrów generujących obserwowaną sekwencję zastosowano algorytm Viterbiego i Bauma-Welcha. Trzecią zaletą algorytmu jest uwzględnienie dwóch typów rozkładów emisji stanów, tj. rozkładu Gaussa i ujemnego dwumianowego, które posłużyły do opisu wartości emitowanych przez stany. Rozkład ujemny dwumianowy jest również używany w programie BayesPeak także stosujący HMM, ale drugiego rzędu i inną metodę estymowania parametrów, tj. metodę Monte Carlo, niż zastosowane w narzędziu HERON.

Implementacja pakietu HERON napisanego w języku python została jasno przedstawiona. Doktorantka korzystała z pakietu hmmlearn, ale wyraźnie napisała, co było jej oryginalnym wkładem, m.in. implementacja rozkładu dwumianowego ujemnego. Jej autorstwem jest także wiele skryptów służących do analizy danych. Ważną opcją programu HERON jest możliwość analizy wielu danych na raz z tego samego typu eksperymentu, co jest szczególnie istotne w przypadku słabej jakości sygnału w porównaniu do szumu. Możliwe jest też analizowanie porównawcze wielu zbiorów (subpopulacji) pochodzących z modyfikacji danego eksperymentu. Mam w związku z tym pytanie, jak poradziłby sobie program HERON, gdyby w tym przypadku plik kontroli potraktować jako jedną subpopulację, a plik z oczekiwanymi sygnałami jako drugą subpopulację? Czy dałoby to rozsądne wyniki? W opisie opcji programu podano, że wykorzystuje on plik kontroli. Z dalszych opisów wynika, że jest on niekonieczny, dlatego można by o tym wspomnieć właśnie w tym miejscu. Chciałbym się dowiedzieć także jak w takim przypadku jest liczony znormalizowany sygnał? Do programu HERON można podać wiele plików kontroli dla każdego pliku próbki. W związku z tym mam pytanie: czy pary tych plików są rozpoznawane po kolejności ich wprowadzenia, czy po jakieś wspólnej części nazwy pliku? Proszę podać także jaką minimalną i maksymalna liczbę stanów można podać w opcji –s? Jeśli wprowadzi się inna liczbę niż trzy, to czy trzeba obowiązkowo zmieniać opcję –q? Opcja –g nie jest jasno opisana. Chciałbym się dowiedzieć, jak przypisywać próbki (pliki) do grup. Proszę o podanie przykładu.

Skuteczność programu HERON została porównana z trzema innymi narzędziami: MACS (uwzględniając opcję dla szerokich sygnałów), SICER i BayesPeak. Ich działanie zostało ocenione na wielu danych symulowanych oraz dwóch typach zbiorów eksperymentalnych, pochodzących ze zdrowych i nowotworowych tkanek człowieka. Przy porównaniu programów w oparciu o dane eksperymentalne można by wspomnieć, dlaczego nie uwzględniono programu BayesPeak badanego na danych symulacyjnych.

Aby dokonać porównań, doktorantka opracowała pakiet do symulowania danych z eksperymentów opartych na NGS. Można w nim wygodnie regulować stosunek sygnału do szumu. Doktorantka zmodyfikowała także istniejący pakiet ChiP-sim. Dane symulacyjne zostały wygenerowane przy kombinacji różnych długości sygnałów, wzbogacenia sygnału i poziomu szumu, co dało 756 zbiorów. Uwzględniono dodatkowo 7 powtórzeń oraz dwie subpopulacje. Przy zapuszczaniu programów rozpatrywano różne długości okien. Wyniki tych analiz przedstawiono na Fig. 4.4, jednak dobrze byłoby podać jakiego programu dotyczą te wyniki. Domyślam się, że był nim HERON. Przydałoby się powołać się na ten rysunek w tekście przy odpowiednim opisie. Według mnie ciekawe jest to, że szerokość sygnału nie zależy w prosty sposób od długości okna.

Doktorantka rzetelnie opisała i przedyskutowała wyniki poszczególnych programów. Przy porównaniu danych symulacyjnych opartych na trzech miarach HERON okazał się generalnie lepszy niż inne programy zwłaszcza dla sygnałów o długości ponad 3000 bp. Dobre wyniki dał też dla krótszych sygnałów o słabym wzbogaceniu. Dziwny jest nagły spadek jakości programu SICER dla bardzo szerokich sygnałów. HERON okazał się lepszy pod względem czułości kosztem specyficzności. MACS wypadł lepiej dla krótkich wzbogaconych sygnałów. W legendach niektórych rysunków porównujących programy dla kompletności opisu można by napisać, że NB to rozkład ujemny dwumianowy. HERON został również porównany pod względem stosowanych rozkładów i liczby próbek oraz subpopulacji. Zwiększenie liczby próbek do czterech zwiększało skuteczność identyfikacji sygnałów, jednak dalszy wzrost liczby prób powodował tylko niewielki wzrost skuteczności. Rozkład Gaussa okazał się lepszy tylko dla krótszych sygnałów, jak wynika to z Fig. 4.7, ale nie znalazłem tego wniosku w tekście. HERON dał także dobre rozróżnienie między subpopulacjami dla silnych sygnałów, ale nie dla przypadków o słabym sygnale i dużej liczbie próbek.

Porównanie programów dla danych eksperymentalnych dla tkanek zdrowych i nowotworowych również wykazało większą skuteczność narzędzia HERON szczególnie dla danych, w których metylowany był histon H3, gdzie oczekiwano właśnie długich domen. Ocena skuteczności programów polegała na porównaniu wyników z oczekiwanymi długościami sygnałów w danych eksperymentach i stopniem ekspresji genów. Identyfikowane sygnały były dłuższe i mniej liczne w przypadku HERON. Porównanie z danymi RNA-seq potwierdziło, że zidentyfikowane domeny dotyczyły genów charakteryzujących się niższą ekspresją, co oczekiwano dla tego typu eksperymentu. HERON dał także bardziej powtarzalne wyniki niż inne programy dla analizy danych od poszczególnych pacjentów. Testowanie

przypadków z wieloma próbkami możliwymi do analiz w programie HERON pokazało skuteczność takiego podejścia. Przy analizie wielu próbek zidentyfikowane domeny okazały się krótsze, a geny związane z tymi domenami wykazywały oczekiwaną mniejszą ekspresję. Dla danych z tkanek nowotworowych okazało się jednak, że zmienność sygnałów jest większa niż dla tkanek zdrowych. Niestety analizy dwóch subpopulacji nie dały zadowalających wyników zarówno dla tkanek zdrowych, jak i nowotworowych. Jedynie rozróżnienie dwóch eksperymentów H3K4me3 i H3K27ac okazało się satysfakcjonujące.

Doktorantka przeanalizowała skuteczność opracowanego programu również pod względem stosowanych rozkładów. Rozkład dwumianowy ujemny okazał się lepiej pasować do danych niż rozkład Gaussa. Jednakże analizy wykazały, że trudniej jest oszacować jego parametry, dlatego zaproponowała, że lepiej jest stosować rozkład Gaussa. Dlatego jest on opcją domyślną w programie.

Mam kilka uwag do części dotyczącej porównania programów. W analizach statystycznych stosowane były dwa testy. Ja przedstawiłbym wyniki tylko dla jednego testu, który jest odpowiedni do takich danych po sprawdzeniu założeń testów. Wyniki na Fig. 5.3, 5.4, 6.3 i 6.8 mogłyby zostać przedstawione jako wykresy pudełkowe, aby pokazać zmienność wyników. W legendzie do Fig. 6.1 powinny być poprawione nazwy modyfikacji histonów zgodnie z tym, co jest na rysunku. Wartości osi y na Fig. 6.6 i 6.7 można by przedstawić w skali logarytmicznej. Wtedy może lepiej byłyby widoczne różnice między programami.

Warto podkreślić, że program HERON został już opublikowany i jest bezpłatnie dostępny dla użytkowników. Ja w swojej praktyce stosowałem narzędzie Peak Finder MetaServer 1.3 (PFMS), który stosuje trzy programy: MACS, CisGenome i SISSRs. Chciałbym się dowiedzieć, jak doktorantka może ocenić to narzędzie?

Mimo, że praca nie zawiera osobnego rozdziału z dyskusją, jej elementy były uwzględniane w poszczególnych rozdziałach dotyczących poszczególnych analiz. W nich doktorantka wytłumaczyła uzyskane wyniki w oparciu o specyfikę danych oraz działanie algorytmów poszczególnych programów. Dobrze przeprowadzona dyskusja wyników świadczy o dużej dojrzałości naukowej doktorantki i umiejętności wydobywania najważniejszych informacji. Warto podkreślić, że doktorantka rzetelnie i w obiektywny sposób przedstawiła wady i zalety swojego programu pokazując, w jakich przypadkach powinien być stosowany, a w jakich nie. W końcowym rozdziale doktorantka podsumowała najważniejsze cechy opracowanego narzędzia do identyfikacji długich słabo wzbogaconych odczytów oraz

narzędzia do symulacji danych z eksperymentów NGS. Jednakże dobrze byłoby wyeksponować i podsumować cechy samego algorytmu w porównaniu z innymi programami, np. w punktach i podać co jest nowością.

Praca jest dobrze zorganizowana pod względem formalnym i należy podkreślić jej poprawne sformatowanie i obecność wielu rysunków. W niektórych przypadkach mogłaby być większa czcionka. Rozdział 5 powinien posiadać bardziej precyzyjny tytuł, tak jak rozdział 6, np. "Testing on experimental data from normal tissues". Bibliografia zawiera 87 pozycji anglojęzycznych i linków internetowych. Praca jest napisana poprawnym językiem i stylem. Dostrzegłem tylko drobne błędy, np. zamiast "separatedly" powinno być "separately". Następujący fragment był dla mnie niejasny gramatycznie: "as I did not want the differences between the samples that can be attributed to the difference in grades to influence the results and make the analysis harder." Gdzieś powinien być czasowniki po "want", może tak: "I did not want to record the differences".

Stwierdzone przeze mnie zastrzeżenia nie rzutują jednak na bardzo pozytywną ocenę pracy, a przedstawione powyżej uwagi nie zmniejszają wartości ocenianej rozprawy. Opracowanie narzędzia było bardzo zasadne, ponieważ istnieje potrzeba polepszania analizy danych pochodzących z sekwencjonowania nowej generacji. Przedstawione w pracy narzędzie będzie pomocne dla biologów molekularnych analizujących tego typu dane.

Uważam, więc, że przedstawiona do recenzji rozprawa doktorska spełnia wszystkie wymogi Ustawy o Stopniach Naukowych. Zgłaszam, zatem wniosek do Rady Naukowej Dyscyplin Matematyka i Informatyka Uniwersytetu Warszawskiego o uznanie rozprawy Pani mgr Anny Macioszek za odpowiadającą wymogom stawianym rozprawom doktorskim i o dopuszczenie doktorantki do dalszych etapów przewodu doktorskiego.

Prof. dr hab. Paweł Mackiewicz

Paweł Mackiewicz