Recenzja rozprawy doktorskiej

„Models of cancer genome evolution used to evaluate the role of selection and occurrence of new mutations"

Doktorant: mgr inż. Paweł Kuś

Przedłożona do recenzji rozprawa doktorska dotyczy problemu inferencji statystycznej parametrów w modelach ewolucji biologicznej nowotworów, na podstawie spektrum częstości mutacji uzyskanych metodą sekwencjonowania DNA.

Jestem pod bardzo dobrym wrażeniem układu pracy, użytego języka, i poziomu objaśnień niezakładającego dogłębnej znajomości tematu badań. Cała dysertacja napisana jest bardzo dobrym językiem angielskim. Dysertacja ma bardzo przejrzysty układ, dzięki czemu przyjemnie się ją czyta.

W rozdziale 1 doktorant zwięźle przedstawia motywację, cel i hipotezę badawczą pracy, oraz ogólny plan rozprawy. W rozdziale 2 opisane zostały podstawy biologii raka oraz tych aspektów ewolucji biologicznej, które są istotne w rozwoju nowotworów. Rozdział 3 w przystępny sposób przedstawia metody sekwencjonowania DNA i analizy danych wykorzystywane w biologii raka, ich zalety i wady, wprowadza też pojęcie „Variant Allele Frequency (VAF) Spectrum" – podstawowy obiekt badań w dalszej części pracy, oraz sposób jego teoretycznego modelowania. Część „historyczna" jest w mojej opinii zbyt długa, zwłaszcza że większość opisanych zagadnień nie jest istotna dla dalszych rozdziałów, ale to może kwestia gustu. W rozdziale 4 autor opisuje zbiory danych które będzie analizować, oraz narzędzia do ich analizy. Rozdział 5 przedstawia wyniki zastosowania tych narzędzi do danych pochodzących z kilku różnych rodzajów nowotworów. Część 5.1 skupia się na analizie danych wygenerowanych w projekcie NCN, w którym doktorant uczestniczył, część 5.2 dotyczy analizy danych z projektu ze współpracy z grupą z Dr. B. Czerniaka z MD Anderson CC. Końcowy rozdział 6 podsumowuje wyniki badań. Praca zawiera jeszcze dwa suplementy: listę genów typu „driver" (onkogeny, suppressor genes) istotnych dla pracy, i zwięzły opis modelu matematycznego istotnego dla Rozdz. 5.2.

Dysertacja zawiera też listę publikacji doktoranta oraz abstrakty konferencyjne; jego wkład do publikacji których wyniki zostały wykorzystane w dysertacji jest jasno opisany. Większość skrótów jest wyjaśniona w tekście i osobnej tabeli.

Literatura cytowana w pracy jest dobrze dobrana tematycznie i reprezentatywna dla omawianych zagadnień. Bibliografia jest dobrze sformatowana, w jednolitym stylu, bez oczywistych błędów, choć muszę przyznać, że nie sprawdzałem poprawności danych wszystkich 136 pozycji.

Głównym celem pracy jest zbadanie hipotezy, czy zmiany w dynamice ewolucyjnej raka podczas przerzutów i nawrotów choroby mogą być wykryte sekwencjonowaniem DNA typu „bulk" tzn. makroskopowych próbek nowotworu, w których DNA pochodzi z wielu różnych komórek raka. Tak postawiona hipoteza jest bardzo ogólna, i wydaje mi się, że już została w pewnym sensie potwierdzona przez prace A. Sottoriva, T. Graham, C. Curtis, T. Reiter, i inne, wiele z nich cytowane w dysertacji. Jednakże w dalszej części doktorant precyzuje, że chodzi mu o analizę spektrum VAF w specyficznych rodzajach nowotworów: określenie, które mutacje są neutralne a które przystosowawcze w sensie ewolucyjnym, jak ich proporcja zależy od rodzaju raka, oraz znalezienie odchyleń w neutralnym „ogonie" VAF (mutacje o małej częstości) od przewidywań prostego modelu matematycznego typu „well-mixed" bez selekcji.

Zastosowane metody to dość standardowe metody bioinformatyczne, polegające na użyciu (ze zrozumieniem) dostępnych pakietów narzędzi typu command-line oraz R i Python, oraz proste modelowanie matematyczne. Nowym wkładem doktoranta jest program *cevomod* – alternatywa dla istniejących programów np. MOBSTER pozwalających wyznaczać współczynniki selekcji mutacji na podstawie VAF.

Głównym wynikiem pracy jest zastosowanie *cevomod* do analizy danych z sekwencjonowania DNA nowotworów BRCA i LSCC oraz AML (guz pierwotny i przerzuty). Doktorant rozwinął też analizę z pracy [10] której jest współautorem, dotyczącej procesu powstawania raka pęcherza moczowego. Wyników jest dość dużo: liczba mutacji w różnych próbkach z podziałem na „driver mutations" i neutralne, korelacje między próbkami z pierwotnego guza vs przerzuty, liczba sub-klonalnych mutacji, i wreszcie oszacowanie na częstość mutacji i wartości współczynników selekcji. Wyniki są generalnie w zgodzie z podobnymi danymi z literatury dla innych nowotworów. Interesujące jest, że MOBSTER nie zadziałał dla wielu z próbek, dla których można było otrzymać parametry ewolucyjne z *cevomod*. Autor uznaje to za dowód na przewagę jego podejścia, ale ja nie jestem do końca przekonany (zob. moje uwagi poniżej). Doktorant pokazuje też, że wykładnik α w prawie potęgowym $\sim f^{-\alpha}$ opisującym dolny ogon VAF często nie jest równy 2 jak zakłada prosty model matematyczny, i sugeruje selekcję oraz zmieniającą się szybkość mutacji jako możliwe wyjaśnienia. W sekcji 5.2 dot. raka pęcherza moczowego jest przedstawiona podobna analiza, dodatkowo z analizą rozkładu przestrzennego mutacji.

Przedstawione rezultaty są nowe, ich prezentacja i interpretacja w świetle literatury przedmiotu jest na poziomie oczekiwanym od kandydata na doktora. Wyniki uzyskane przy pomocy *cevomod* stanowią interesujący dodatek do podobnych wyników otrzymanych wcześniej dla innych nowotworów. Najciekawsze wydają mi się wyniki uzyskane z projektu raka pęcherza moczowego – jeśli rzeczywiście udałoby się wykryć stan przed-nowotworowy za pomocą podobnej analizy u żywego pacjenta (a nie na wyciętym pęcherzu), byłby to duży postęp. Jeśli chodzi o praktyczne zastosowanie otrzymanych wyników to z jednej strony, pakiet *cevomod* zostanie zapewne użyty do dalszych badań, generujących publikowalne wyniki. Z drugiej strony, ograniczenia modelu użytego w *cevomod* powodują, że ciężko użyć go do uzyskania zupełnie nowego wglądu w procesy powstawania nowotworów tak jak np. zrobili to A. Sottoria *et. al.* ze swoim modelem „Big Bang".

Podczas lektury zwróciłem uwagę na kilka problemów, niezręcznych sformuowań, oraz brakujących informacji:

- Formatowanie wyrażeń matematycznych jest niezbyt nieeleganckie, np. logarytm $\ln$, funkcja $\exp$, są pisane kursywą jakby to były zmienne, subskrypty zmiennych też (niepoprawnie) są pisane kursywą np. N_{ref} (str. 24). Wybór symboli reprezentujących niektóre zmienne jest mylący np. CN_{mut} jest jedną zmienną, ale zapis sugeruje dwie: C i N_{mut}, podobnie $N\mu$ – liczba nowych wariantów. Nie ma puktuacji po wyrażeniach matematycznych. Na szczęście dla autora, praca nie jest bardzo zmatematyzowana i takie problemy ograniczają się do kilku stron.
- Minimalne legendy (opisy symboli) w figurach czasami utrudnia zrozumienie (np. Fig. 5.10, 5.18)
- Mały font na osiach wykresów, i legend utrudnia czytanie.
- „Mutation rate" jest użyte na oznaczenie liczby mutacji, zamiast prawdziwej „rate" (wielkości o wymiarze 1/czas)
- Fig. 3.4 – skąd wzięto przedstawiony VAF?
- Fig. 5.20 – nie jest wyjaśnione co to są „maps of analyzed specimens". Generalnie trudno zrozumieć tę figurę z powodu minimalnej ilości objaśnień.

- „Darwinian vs non-Darwinian evolution" – poprawniej było pisać „non-neutral vs neutral evolution" gdyż to zwykle autorzy prac używających te sformuowania mają na myśli.

Nie wymieniam drobnych problemów takich jak (rzadkie) błędy językowe czy nieprezycyjne stwierdzenia nieistotne dla zrozumienia wyników pracy.

Mam też kilka uwag krytycznych natury merytorycznej:

(1) MOBSTER pozwala na dopasowanie zależności potęgowej („power law", Pareto Distribution type I) z wykładnikiem różnym od 2. Co zatem innego robi *cevomod*, oprócz braku testowania neutralności modelu (co akurat byłoby korzystne, patrz niżej)? Dlaczego MOBSTER działa gorzej, skoro robi fundamentalnie tę samą rzecz? Zapewne ma to związek z optymalizacją przedstawioną w 4.2.7, ale nie do końca rozumiem jej ideę. W szczególności, nie przedstawiono uzasadnienia tego podejścia na gruncie statystyki.

(2) Związany z tym problem: brakuje testów analizy VAF na danych syntetycznych. Pozwoliłoby to ustalić wiarygodność otrzymanych rezultatów np. współczynników selekcji s. Naiwnie oczekiwałbym, że pewna liczba mutacji da niezerowe s z powodów fluktuacji statystycznych, pomimo że naprawdę są neutralne, i *vice versa*, neutralne mutacje mogą okazjonalnie dać sygnatury mutacji typu „driver". Skąd wiemy, że otrzymane wyniki to nie szum statystyczny? Taki test byłby tym bardziej pożądany, że brakuje słupków błędów przy podawanych wartościach liczbowych, nie można więc ocenić ich dokładności.

(3) Wiele z przedstawionych konkluzji brzmi prawdopodobnie, ale wydaje mi się, że często można zaproponować inne wyjaśnienie niż podane przez kandydata. Np. jeśli chodzi o wykładnik $\alpha \neq 2$ w potęgowym ogonie VAF, to może on powstać również w wyniku ograniczeń przestrzennych (zob. np. D. Fusco, et al., Nature Communications 7, 12760 (2016)). Dysertacja w ogóle nie dotyka tego problemu tzn. że guzy nowotworowe nie są „well mixed" ale mają przestrzenną strukturę i rosną w niejednorodny sposób, co wpływa na dynamikę ewolucyjną. Prawdziwość założenia „well mixed" jest bardzo ważna dla wartości wyznaczonych współczynników selekcji.

(4) „Copy number changes" – jest wspomniane że mogą się pojawiać wcześnie w rozwoju raka, a więc teoretycznie mogłyby mieć duży wpływ na VAF, ale w modelu nie są uwzględnione. Czy da się oszacować wielkość efektu takich zmian na wyniki pracy?

(5) „Emergence times" mutacji muszą być skorelowane ze współczynnikami selekcji z powodu fundamentalnej zależności wynikającej z założonej dynamiki populacyjnej (eksponencjalny wzrost klonów). Czy jest zatem zaskakujące znalezienie takiej korelacji na Fig. 5.15? Czy korelacja jest silniejsza czy słabsza niż oczekiwana na podstawie założonego modelu selekcji?

Na podstawie informacji zawartych w dysertacji, poziom wiedzy ogólnej kandydata w dziedzinie biologii nowotworów i bioinformatyki, oraz umiejętność pracy samodzielnej wydają się być na poziomie oczekiwanym od osoby kończącej doktorat. Kandydat jest współautorem 9 prac, ale tylko jednej z nich (Wydawnictwo Politechniki Śląskiej) jako pierwszy autor. Ponieważ wkład doktoranta w prace nie związane z dysertacją nie jest opisany, sprawdziłem „author contributions" publikacji, do których miałem dostęp: wkład doktoranta polegał głównie na stworzeniu sofware'u i analizie danych, chociaż zdarzyło się mu mieć wkład do koncepcji projektu. Może to świadczyć o zainteresowaniu bardziej techniczną stroną badań niż ich interpretacją, ale jest to według mnie normalne i nie odbiega od moich doświadczeń z osobami o podobnym profilu badawczym.

Podsumowując, **zaprezentowana dysertacja posiada, mimo drobnych mankamentów przedstawionych powyżej, wystarczającą wartość poznawczą do zaklasyfikowania jej jako oryginalne rozwiązanie problemu badawczego**. Stwierdzam zatem, że praca doktorska spełnia wymogi określone w art. 187 ustawy z dnia 20 lipca 2018 r. – Prawo o

szkolnictwie wyższym i nauce, i **wnioskuję o dopuszczenie mgr. inż. Pawła Kusia do dalszych etapów przewodu doktorskiego.**

Podczas obrony doktorskiej będę oczekiwać od kandydata odniesienia się do mojej krytyki (punkty 1-5).

Bartłomiej Wacław 04.08.2023