Instytut Biocybernetyki i Inżynierii Biomedycznej
im. Macieja Nałęcza
Polskiej Akademii Nauk

Dr hab.
Jan Poleszczuk

Warszawa, 25 sierpnia 2023

Recenzja rozprawy doktorskiej

mgr inż. Pawła Kusia pt. *Models of cancer genome evolution used to evaluate the role of selection and occurrence of new mutations*

Promotor: Prof. dr hab. inż. Marek Kimmel

1. Zagadnienia naukowe rozpatrywane w pracy

Przedmiotem rozprawy jest zaproponowanie, implementacja i wstępne zweryfikowanie metody pozwalającej na analizę dynamiki ewolucyjnej nowotworu na podstawie danych pochodzących z sekwencjonowania genomu, które nie osiągnęło odpowiedniej głębokości odczytu. W rozprawie Autor stara się również odpowiedzieć na pytanie, czy sekwencjonowanie próbki nowotworu całościowo, tj. bez izolacji samych komórek nowotworowych, może być wykorzystane do analizy jego dynamiki ewolucyjnej. Postawione pytanie, jak i próba zaproponowania nowej metody analitycznej, jest niezwykle istotne dla zrozumienia dynamiki ewolucyjnej nowotworu, co może być w dłuższej perspektywie wykorzystane do celów medycznych. Od dłuższego bowiem czasu badacze starają się zaproponować nowe schematy terapeutyczne, które opierają się na pryncypiach ewolucji i mają na celu np. zminimalizowanie ryzyka wytworzenia się oporności nowotworu na dany rodzaj leczenia. Co więcej, w dobie terapii kombinowanych, które celują jednocześnie w różne podpopulacje komórkowe, dogłębne zrozumienie mechanizmów ewolucyjnych może niewątpliwie pozwolić na zwiększenie ich skuteczności poprzez odpowiednie dobieranie dawek. Biorąc powyższe pod uwagę, należy stwierdzić, że rozpatrywane w pracy problemy są aktualne i dobrze wpisują się w nurt obecnie prowadzonych prac naukowych. Tematyka rozprawy niewątpliwie wpisuje się w dyscyplinę Informatyka Techniczna i Telekomunikacja.

W rozprawie wyróżniona została jedna hipoteza badawcza: *Zmiany w dynamice ewolucyjnej raka w ogniskach przerzutowych oraz z momentu nawrotu choroby można określić ilościowo na podstawie danych zsekwencjonowania DNA przeprowadzonego na nieoczyszczonych próbkach, czyli takich, w których nie wyizolowano samych komórek nowotworowych*. W celu weryfikacji powyższej hipotezy zebrano dane z sekwencjonowania różnych typów nowotworów, a następnie je przeanalizowano przy wykorzystaniu metody zaproponowanej w pracy. Powyższa hipoteza

została wstępnie zweryfikowana, acz dyskusyjne pozostaje to, czy wszystkie założenia i wnioski nie wymagają dalszych analiz – na to wskazują uwagi zawarte w dalszej części recenzji.

2. Struktura pracy

Rozprawa składa się ze streszczenia, listy publikacji Autora, opisu wkładu Autora do publikacji przedstawionej w jednej z części rozprawy, listy abstraktów konferencyjnych Autora, spisu treści, listy skrótów, sześciu numerowanych części razem z podsumowaniem, spisu rysunków i tabel, trzech dodatków oraz bibliografii. Pierwszy numerowany rozdział wprowadza krótko w tematykę rozprawy i przedstawia motywację podjęcia tematu wraz z badaną hipotezą. Rozdziały drugi i trzeci poświęcone są omówieniu zagadnień związanych z tematyką pracy. W rozdziale drugim opisane są zagadnienia związane z biologią nowotworów, genami i ich mutacjami, a także kluczowe zagadnienia opisujące zjawisko ewolucji nowotworów, w szczególności modele spełniające założenia ewolucji darwinowskiej, jak i niespełniające, model Big Bang, czy też model macierzystych komórek nowotworu. W rozdziale trzecim Autor przedstawił historię postępu w sekwencjonowaniu (od metody Sangera, do sekwencjonowania trzeciej generacji), opisał ograniczenia w sekwencjonowaniu, a także przedstawił istniejące modele i metody do analizy wyników sekwencjonowania pod kątem badania dynamiki ewolucyjnej nowotworu.

Rozdział 4 przedstawia dane wykorzystane w pracy oraz zaproponowane metody analityczne. Rozdział ten stanowi dokładny opis nowej metodologii analitycznej zaproponowane przez Autora i udostępnionej w postaci pakietu do programu R. Rozdział 5 przedstawia wyniki zastosowania zaproponowanej metody do zebranych danych klinicznych i ich porównanie z wynikami otrzymanymi przy wykorzystaniu innej szeroko wykorzystywanej metody wcześniej opisanej w literaturze. W Rozdziale 5 przedstawione zostały również wyniki prac wykorzystujących inną metodologię badawczą, w których Autor brał udział (publikacja [10]). Całość została posumowana w Rozdziale 6.

Układ pracy oceniam jako prawidłowy, choć może bardziej wskazane byłoby stawianie hipotez badawczych po dokładniejszym wprowadzeniu czytelnika w rozważane zagadnienia, czyli np. na koniec Rozdziału 3.

3. Analiza źródeł

Spis literatury zawiera 136 pozycji, z czego cztery ([10], [57], [68] i [69]) są współautorstwa Autora. W zdecydowanej większości są to prace bezpośrednio związane z tematyką rozprawy. Dobór bibliografii świadczy o dobrym rozeznaniu Doktoranta w literaturze światowej w tematyce, którą się zajmuje. Pewne zastrzeżenia może rodzić odnoszenia się Autora do stron w Internecie (np. [16], [42], [102], [108], [133]), Wikipedii (np. [17]), czy też repozytoriów niezrecenzowanych preprintów (np. [25], [34], [77], [78]). Co ważne, odnośniki te w znakomitej większości mają za zadanie przybliżyć ugruntowaną wiedzę, którą można znaleźć w klasycznych pozycjach literaturowych, lub opisują źródło, które powinno być zacytowane bezpośrednio (np.

[42]). Pewne pozycje bibliografii są również niekompletne (np.[5], [69]), czy też niepotrzebne z punktu widzenia merytoryki pracy (np. [54], [60], [81], [86]).

4. Oryginalność i silne strony rozprawy

W pracy wykorzystane zostały metody dość standardowe, jednak sposób ich wykorzystania oraz uzyskane wyniki są niewątpliwie oryginalne. Uzyskane wyniki nie zostały jeszcze opublikowane, ale wydaje się, że po pewnych uzupełnieniach powinny znaleźć miejsce w znaczących międzynarodowych czasopismach naukowych. Do szczególnie wartościowych, oryginalnych elementów rozprawy, istotnych z naukowego punku widzenia można zaliczyć:

- Zaproponowanie metody, która pozwala na analizowanie dynamiki ewolucyjnej w próbkach pochodzących z sekwencjonowania genomu o mniejszej głębokości odczytu, niż ta wymagana przez inne istniejące metody analityczne.
- Wskazanie, że możliwe wydaje się analizowanie dynamiki ewolucyjnej na podstawie nieoczyszczonych próbek tkanki nowotworowej, tj. takich które nie zawierają jedynie komórek nowotworowych.

Silną stroną rozprawy jest niewątpliwie całościowe podejście do problemu, czyli zaproponowanie nowej metody analitycznej, wstępne zweryfikowanie jej na różnorodnym zbiorze danych, a następnie przygotowanie publicznie dostępnego pakietu do szeroko wykorzystywanego języka programowania R, który każdy badacz może pobrać i wykorzystać do swoich analiz. W dzisiejszym świecie, w którym nowe metody są opisywane w literaturze niemal każdego dnia, szczególnie ten ostatni element (publikacja pakietu) zasługuje na pochwałę.

5. Słabsze strony rozprawy

Do słabszych stron pracy należą:

- brak dokładanego uzasadnienia dla przyjętych parametrów poszczególnych kroków metody oraz jakiegokolwiek badania wpływu zmian parametrów na uzyskiwane wyniki – uwaga 6.1 poniżej
- bardzo swobodnie potraktowano fakt, że próbki z ognisk przerzutowych zwierają inną tkankę zdrową, niż te z ogniska pierwotnego – uwaga 6.3 poniżej.
- dość niedbale przedstawiono w pracy wzory i niektóre modele – uwaga 6.7 poniżej
- w rzeczywistości, aby potwierdzić, że można wykorzystywać próbki nieoczyszczone do analiz (bulk data), należałoby przeprowadzić analizy na wyizolowanych komórkach nowotworowych i porównać otrzymane wyniki – tego zabrakło w pracy.

Oryginalność i silne strony rozprawy przeważają jednak nad słabszymi.

6. Szczegółowe uwagi merytoryczne i redakcyjne

W pracy pojawił się braki lub niejasności w elementach istotnych dla merytoryki pracy, do których Doktorant powinien odnieść się w trakcie obrony. W szczególności:

6.1. W rozdziale *4.2.6 Fitting neutral power-law component with covemod* Autor opisuje sposób dopasowywania części modelu opisującej neutralny dryft i podaje szczegóły dotyczące procedury, np. zawężanie spektrum VAF poprzez usunięcie 5% wariantów na obu końcach, zaokrąglanie wartości f do dwóch miejsc po przecinku, dzielenie rozkładu na sekcje o długościach 0.05, czy też odfiltrowywanie wyników gdy R^2 ma wartość mniejszą niż 0.98. Argumentacja za przyjętymi założeniami jest jednak dość skąpa i nie pozwala stwierdzić, że inne możliwości zostały gruntownie przebadane. W rozprawie powinna znaleźć się wnikliwa analiza przyjętych założeń, najlepiej z symulacyjnym zbadaniem ich wpływu na uzyskiwane wyniki (rodzaj analizy wrażliwości). Podobny problem występuje w sekcji *4.2.7 Fitting the models with the best-fitting power coefficient*, gdzie Autor opisuje kolejny zestaw przyjętych wartości parametrów i założeń bez dokładnego wyjaśnienia (np. „Then we smooth the spectrum using the stats::filter() function with a vector weights ‘c(1/3,1/3,1/3)’"). Analogicznie w podrozdziale 4.2.8 czytamy, że “[…] to fit the VAF distribution […] with a mixture of 1 to 3 binomial models” – czemu nie np. 4 albo np. 5? Takiego typu założenie wymaga wyjaśnienia.

6.2. W pracy zostały wykorzystane m.in. próbki guzów litych pochodzące z guzów pierwotnych (N próbek), ognisk przerzutowych (N próbek) oraz zdrowej tkanki (N próbek). W przypadku tych nowotworów brakuje w rozprawie kluczowych informacji klinicznych dla poszczególnych pacjentów, takich jak czy występowały przerzuty odległe, jaka była klasyfikacja TNM, w ilu węzłach chłonnych były wykryte przerzuty (zwykle sprawdza się wiele). Można sobie wyobrazić, że dynamika i status choroby jest inny u osoby z małym guzem i mikroprzerzutami w jednym węźle chłonnym w porównaniu z kimś kto ma wiele przerzutów odległych, wiele zajętych węzłów chłonnych i duże ognisko pierwotne. Tego typu informacje wzbogaciłyby prace i może pozwoliłyby na dogłębniejszą analizę uzyskanych wyników. Dodatkowo, pomimo tego, że dane dotyczące białaczki pochodzą z innej pracy, w rozprawie warto byłoby przedstawić jaki był rozkład czasu pomiędzy próbkami (ile minęło od diagnozy do wznowy). Może to by wyjaśniło spadek indeksu Jaccard’a na Rys. 5.4?

6.3. Biorąc pod uwagę, że próbki pochodzące z raka piersi i krtani nie były oczyszczone (były mieszaniną komórek nowotworowych i zwykłych) wątpliwość może budzić opisany powyżej układ eksperymentalny (2N próbek zmienionych + N próbek tkanki zdrowej). Mianowicie komórki nienowotworowe składające się na węzeł chłonny są zupełnie innego typu niż tkanka, z której była pobierana próbka kontrolna. To może mieć znaczący wpływ na uzyskane wyniki, bo Autor porównuje próbkę nowotworową z węzła chłonnego do zdrowej tkanki.

6.4. Autor wykorzystuje również dane literaturowe pochodzące od pacjentów z białaczką, a dokładniej wyniki sekwencjonowania próbek z momentu diagnozy, wznowy oraz próbek

kontrolnych. Powstaje pytanie czym jest próbka kontrolna. Czy to znaczy, że były dostępne próbki krwi pacjenta zanim został zdiagnozowany?

6.5. Kompletnie niezrozumiały jest podrozdział *4.2.3 Intra-Tumor Heterogeneity (ITH) measure* i później związane z nim wyniki na str. 52. Autor poprawnie podaje, że indeks Jaccard'a może być wykorzystywany do badania podobieństwa pomiędzy zbiorami (tutaj mowa o zbiorach występujących wariantów), ale w pracy Autor nie ma do dyspozycji różnych próbek pochodzących z ogniska pierwotnego (to byłoby badanie tzw. intra-tumor heterogeneity; ITH). Porównywanie próbek z ognisk przerzutowych i ogniska pierwotnego nie ma nic wspólnego z ITH. Co więcej, w przypadku białaczki Autor dysponuje próbkami z różnych punktów czasowych.

Pozostałe uwagi ogólne i komentarze (niewymagające odnoszenia się w trakcie obrony):

6.6. Autor przedstawia we wprowadzeniu do pracy (Rozdziały 2 i 3) wiele ciekawych informacji nt. nowotworów, mutacji oraz metod ich detekcji, ale zainteresowanym czytelnikom będzie brakować odnośników do literatury. Dla przykładu we wprowadzeniu do podrozdziału *2.1.2 DNA, genome and genes*, w którym Autor opisuję budowę DNA, geny i białka, nie znajdziemy nawet jednego odnośnika do literatury. Dalej pojawia się *The Cancer Genome Atlas* beż żadnego odnośnika gdzie można go znaleźć. Autor powinien zadbać, aby elementy pracy, które zawierają fakty, czy też znane hipotezy, miały odniesienia do literatury, z którą czytelnik mógłby się zapoznać.

6.7. Autor w pracy dość niedbale przedstawia wzory matematyczne, popełniając przy tym błędy i wprowadzając nieścisłości/zamieszanie. W szczególności:

- W mianowniku równania (3.4) powinna być suma, a nie iloczyn członów $MCF*CN_{tot}$ oraz $(1-MCF)*CN_{norm}$
- Autor wykorzystuje znak tyldy (~) do określenia rozkładu, a potem wykorzystuje go jako przybliżenie np. w równaniu (3.6)
- W równaniu (3.8) jest jeden zestaw parametrów (a, b, λ), a w opisie do tego równania Autor wymienia parametry M, b oraz c
- Autor powinien zadbać, aby postać równań (3.7), (3.8) i (3.9) przedstawiała ten sam punkt początkowy, czyli tą samą wartość N(0). Teraz jest tak, że np. (3.7) ma N(0) = 1 w przeciwieństwie do pozostałych równań.
- Autor stosuje różne symbole do oznaczania mnożenia (np. w równaniu (3.4) jest ·, w (3.10) jest *, w (3.18) jest x).
- W (3.15) Autor pisze binomial(,) a w (3.3) pisze B(,) do określenia tego samego rozkładu

6.8. Na str. 51 Autor porównuje obciążenie mutacyjne pomiędzy różnymi próbkami. Wątpliwe wydaje się jednak porównywanie danych dotyczących białaczki z pozostałymi, ponieważ dla białaczki był robiony WGS, a dla pozostałych WXS.

6.9. Niepotrzebne zamieszanie wprowadza w pracy dokładne opisanie wyników uzyskanych w [10]. Doktorant miał do tej pracy relatywnie mały wkład i duża część przedstawionego

materiału nie dotyczy tego co było efektem jego prac. Wystarczyłoby przedstawienie krótkiego opisu, a potem przedstawienie dodatkowych wyników uzyskanych przez Autora.

Inne uwagi szczegółowe:

6.10. Na str. 48 Autor pisze, że rak potrójnie ujemny jest rzadki, ale jednocześnie pojawia się informacja, że to 15% przypadków raków piersi – nie można tego uznać zatem za rzadki przypadek.

6.11. W pracy pojawiają się niedociągnięcia językowe i edycyjne. Między innymi:

- Lista skrótów i symboli, którą Autor zawarł po spisie treści nie jest pełna i nie zawiera szeregu ważnych skrótów, które potem pojawiają się w tekście (np. MCF na str. 24, ISM na str. 27, MR na str. 61, EMT na str. 74)
- Autor niefortunnie wykorzystuje do określania nowotworów piersi skrótu BRCA. Skrót ten jest wykorzystywany dla jednego z najbardziej znanych genów związanych ze zwiększonym ryzykiem wystąpienia raka piersi (a dokładnie BRCA1 i BRCA2).
- drobne błędy językowe (np. ,,needed to be […] for cancer to *growth*" gdzie powinno być *grow* na str. 8; "Variant calling algorithms try to [..] distinguish true algorithms" na str. 22); step -> steep na str. 66
- do Rysunku 2.3 nie ma odniesienia w tekście
- na Rys. 5.6 byłoby lepiej przedstawić procent, a nie liczbę pacjentów, ponieważ rozważane kohorty nie są równoliczne.

7. Ocena końcowa rozprawy

Podsumowując, pomimo uwag szczegółowych przedstawionych powyżej, uważam, że silne strony pracy przeważają nad słabszymi. Stwierdzam, że mgr inż. Paweł Kuś wykazał się wiedzą i umiejętnościami uprawniającymi go do ubiegania się o stopień doktora nauk technicznych w dyscyplinie Informatyka Techniczna i Telekomunikacja. Przedstawiona praca doktorska spełnia wymagania stawiane pracom doktorskim przez ustawę Prawo o szkolnictwie wyższym i nauce z dnia 20 lipca 2018 r. W szczególności, jej przedmiotem jest oryginalne rozwiązanie problemu naukowego, zdefiniowanego przez postawioną hipotezę badawczą. Autor posiada tytuł magistra inżyniera oraz jest współautorem siedmiu artykułów opublikowanych w czasopismach naukowych, ujętych w odpowiednim wykazie. Wnioskuję o dopuszczenie mgra inż. Pawła Kusia do publicznej obrony rozprawy doktorskiej.

Signed by / Podpisano przez:

JAN POLESZCZUK

Date / Data: 2023-08-27 21:12