Silesian University of Technology
Faculty of Automatic Control, Electronics and Computer Science

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

Doctoral thesis

Paweł Kuś

Supervisor:
Prof. dr hab. inż. Marek Kimmel
Co-supervisor:
Dr inż. Roman Jaksik

Gliwice 2023

Streszczenie rozszerzone rozprawy doktorskiej

Modele ewolucji genomu nowotworów w ocenie roli selekcji i występowania nowych mutacji

mgr inż. Paweł Kuś

1 Wprowadzenie

Nowotwory są jedną z głównych przyczyn zgonów na świecie. Według Światowej Organizacji Zdrowia choroby nowotworowe były przyczyną 10 milionów zgonów na świecie w 2020 roku. W tym samym roku tylko wśród 5 najczęstszych typów nowotworów zdiagnozowano ponad 9 milionów nowych przypadków, a około 400 tysięcy przypadków zostało zdiagnozowanych wśród dzieci [3]. Wiele czynników ryzyka chorób nowotworowych jest związanych ze zmianami stylu życia ludzi w krajach wysoko rozwiniętych. Dlatego też nowotwory są przedmiotem szczególnej uwagi środowisk naukowych. Badania prowadzone w minionych dekadach pozwoliły odkryć szereg mechanizmów charakteryzujących komórki nowotworowe [8], oraz doprowadziły do wdrożenia licznych zaawansowanych terapii przeciwnowotworowych. Pomimo tych postępów, leczenie nowotworów zapobiegające ewolucji lekooporności i nawrotom choroby, wciąż jest wyzwaniem dla medycyny. Możliwości medycyny są w leczeniu nowotworów ograniczone, podobnie jak stan wiedzy na temat roli mechanizmów odpowiedzialnych za ewolucję nowotworów: procesów mutagenezy i selekcji.

Mutageneza jest procesem powstawania mutacji, czyli zmian (wariantów) w materiale genetycznym komórki. Zwiększa ona genetyczną różnorodność komórek i może skutkować powstaniem zmian w genach szczególnie istotnych dla utrzymania homeostazy, związanych np. z cyklem komórkowym i podziałem komórki lub naprawą uszkodzeń DNA. Mutacje w tych genach, zwanych genami *driverowymi* (ang. *driver* - kierowca, sterownik), mogą doprowadzić do selekcji komórek: pozytywnej, jeśli mutacja zwiększyła tempo proliferacji komórki i częstość występowania jej komórek potomnych rośnie, lub negatywnej, kiedy prowadzi do wymarcia populacji komórek posiadających mutacje negatywne w skutkach.

Rola mechanizmów selekcji i mutagenezy w ewolucji nowotworów jest wciąż przedmiotem dyskusji. Zaproponowane zostały modele takie jak model neutralnej ewolucji, w której nowo-występujące mutacje nie dają komórkom przewagi ewolucyjnej, sprawność wszystkich komórek w nowotworze jest podobna i wzrost nowotworu jest pojedynczą ekspansją wcześnie powstałych klonów, oraz model ewolucji klonalnej, w którym pojawiające się mutacje zapoczątkowują ekspansje nowych klonów, które ostatecznie zastępują

całkowicie wcześniejsze populacje komórek o mniejszej sprawności.

Rozwój metod sekwencjonowania następnej generacji (ang. *Next Generation Sequencing*, NGS) umożliwił badanie genomów, w tym genomów nowotworów, na niespotykaną wcześniej skalę. NGS obniżyło znacznie koszta sekwencjonowania DNA, przez co stało się ono powszechną praktyką w badaniu nowotworów i doprowadziło do powstania baz danych takich jak *The Cancer Genome Atlas*, udostępniających dane molekularne tysięcy zsekwencjonowanych próbek nowotworowych. Szczególnie powszechne stało się sekwencjonowanie typu *bulk*, w którym materiał genetyczny jest izolowany z próbki zawierającej miliony komórek, a następnie zbiorczo sekwencjonowany.

Sekwencjonowanie DNA, po odpowiednim przetworzeniu wyników obejmującym m. in. kontrolę jakości materiału, mapowanie odczytów do genomu referencyjnego, detekcję wariantów (mutacji) i filtrowanie wariantów celem odrzucenia wyników fałszywie pozytywnych, dostarcza informacji takiech jak lokalizacja wariantów i ich częstość alleliczna (ang. *Variant Allele Frequency, VAF*), która jest powiązana z częstością występowania w nowotworze komórek z danym wariantem.

Pojawienie się szeroko dostępnych danych z sekwencjonowania DNA nowotworów umożliwiło rozwój metod analizy ich ewolucji, opartych o częstości występowania mutacji. Popularne algorytmy takie jak SciClone [9] czy PyClone [10], łączą dane o mutacjach z wielu próbek by określić strukturę klonalną nowotworu. Algorytm PhyloWGS [5] oprócz detekcji klonów analizuje również pokrewieństwo między nimi, tworząc ich drzewo filogenetyczne. Algorytmy te nie analizują jednak parametrów dynamiki ewolucyjnej nowotworu takich jak tempo mutacji, czy współczynniki selekcji nowych klonów. Powstały niedawno algorytm MOBSTER [4] określa je, jednak do skutecznego działania wymaga on danych z sekwencjonowania całogenomowego o pokryciu przynajmniej 100x i nie działa poprawnie z danymi pochodzącymi z bardziej powszechnie wykonywanego sekwencjonowania całoeksomowego. Bazuje on również na pewnych założeniach, takich jak eksponencjalny wzrost populacji czy stałe tempo mutacji, które nie we wszystkich guzach są spełnione.

2 Teza

Teza, którą stawiamy i której dowodzimy w niniejszej pracy brzmi: *Zmiany w dynamice ewolucyjnej nowotworów związane z powstawaniem przerzutów i nawrotami choroby mogą być wnioskowane w oparciu o wyniki sekwencjonowania DNA typu bulk.* Podstawowe cele tej pracy obejmują:

1. analizę dynamiki ewolucyjnej podczas progresji choroby nowotworowej, jej nawrotów i powstawania przerzutów,

2. utworzenie oprogramowania zdolnego do analizy dynamiki ewolucyjnej nowotworów z użyciem danych z sekwencjonowania eksomowego i sekwencjonowania o niższym

pokryciu,

3. walidację założeń popularnych modeli takich jak założenie eksponencjalnego wzrostu populacji czy założenie stałości tempa mutacji.

3 Dane

W celu zaadresowania postawionej tezy, zebraliśmy 4 zestawy danych pochodzące z eksperymentów sekwencjonowania typu *bulk*. Zebrane dane reprezentowały 4 typy nowotworów i zawierały dane z minimum dwóch próbek nowotworowych od każdego pacjenta.

Dynamika ewolucyjna nawracających nowotworów została zbadana na przykładzie ostrej białaczki szpikowej (ang. *Acute myeloid leukemia*, AML), która jest odpowiedzialna za około 40% wszystkich białaczek wykrywanych w Polsce. Wykorzystane zostały dane z sekwencjonowania całogenomowego opublikowane w studium Shlush et al. [11]. Zbiór danych obejmował 11 pacjentów, od których zostały pobrane po dwie próbki nowotworowe reprezentujące momenty diagnozy i nawrotu choroby, oraz po jednej próbce kontrolnej.

Dynamika ewolucyjna nowotworów z przerzutami zastała zbadana na przykładzie nowotworów piersi (BRCA) i krtani (LSCC). Te dwie kohorty danych pochodzą z finansowanego przez Narodowe Centrum Nauki grantu nr 2018/29/B/ST7/02550 *Podejście systemowe do progresji i prognozy raka: Nowe modele i statystyki do analizy danych genomicznych* i zawierają nieopublikowane dotąd dane. Nowotwory piersi są najczęściej diagnozowanymi nowotworami na świecie i są wykrywane niemal wyłącznie wśród kobiet. Nowotwory krtani są typem nowotworu głowy i szyi (ang *Head and Neck Squamous Cell Cancers*, HNSC), występującymi 5-krotnie częściej wśród mężczyzn niż wśród kobiet. Zarówno nowotwory piersi jak i krtani są nowotworami nabłonkowego pochodzenia, oba również wykazują często aktywność procesu tzw. przejścia nabłonkowo-mezenchymalnego (ang. EMT), związanego z powstawaniem przerzutów. Kohorty BRCA i LSCC obejmowały odpowiednio 15 pacjentek z nowotworami piersi oraz 12 pacjentów obojga płci z nowotworami krtani. Od wszystkich pacjentów pobrano po jednej próbce z guza pierwotnego, jednej z przerzutu do lokalnego węzła chłonnego, oraz po jednej próbce kontrolnej. Wszystkie próbki zostały poddane sekwencjonowaniu całoeksomowemu z docelowym pokryciem 100x.

Zmiany dynamiki ewolucyjnej podczas progresji choroby nowotworowej zostały zbadane na przykładzie dwóch nowotworów pęcherza moczowego z użyciem danych z sekwencjonowania całoeksomowego wielu próbek o różnym stopniu zaawansowania choroby, tworzących mapy całego pęcherza. Dane te zostały pozyskane z laboratorium Dr. Bogdana Czerniaka w MD Anderson Cancer Center w Houson, TX, USA, i były uprzednio wykorzystane w opublikowanym we współpracy z autorem rozprawy studium Bondaruk et al. *The origin of bladder cancer from mucosal field effect* [1]. Dane pochodzą od dwóch pacjentów,

oznaczonych numerami 19 i 24, którzy reprezentowali dwa odmienne typy molekularne nowotworów pęcherza: luminalny i bazalny. Oba pęcherze zostały otwarte wzdłuż tylnej ściany pęcherza i podzielone na fragmenty o rozmiarze 1 x 2 cm, które zostały zaklasyfikowane do 1 z 4 grup: normalnego urothelium (ang. *normal urothelium*, NU), śródbłonkowej neoplazji o niskim stopniu złośliwości (ang. *low-grade intraurothelial neoplasia*, LGIN), śródbłonkowej neoplazji o wysokim stopniu złośliwości (ang. *high-grade intraurothelial neoplasia*, HGIN) i raka urotelialnego (ang. *urothelial carcinoma*, UC). W pracy Bondaruk et al. [1] wybrane regiony obu map zostały poddane wielo-omicznym eksperymentom obejmującym sekwencjonowanie całego eksomu, sekwencjonowanie RNA, całogenomową analizę metylacji, czy całogenomową analizę zmian w liczbie kopii. W niniejszej pracy powtórnie wykorzystane zostały dane pochodzące z sekwencjonowania całego eksomu.

4 Metody

Przetwarzanie danych NGS. Kontrola jakości surowych wyników sekwencjonowania w kohortach BRCA i LSCC została przeprowadzona z wykorzystaniem narzędzi FastQC i FastQ Screen. Następnie odczyty zostały zmapowane do genomu referencyjnego GRCh38 za pomocą programu BWA-MEM. Pliki BAM ze zmapowanymi odczytami zostały przygotowane za pomocą narzędzi z zestawu GATK do detekcji mutacji, która została przeprowadzona za pomocą programu Mutect2. Warianty zostały następnie przefiltrowane za pomocą FilterMutectCalls z zestawu GATK, i opisane za pomocą Variant Effect Predictor.

Dane z kohorty AML zostały pobrane w postaci zmapowanych do genomu GRCh37 plików BAM z repozytorium European Genome-Phenome Archive (ID: EGAD00001003234), a następnie przetworzone za pomocą protokołu opisanego dla kohort BRCA i LSCC.

Analizy rozwoju raka pęcherza bazowały na plikach VCF zawierających wyniki detekcji mutacji przeprowadzonej w ramach studium Bondaruk et al. [1]. Proces przetwarzania danych był jednak podobny: zsekwencjonowane odczyty zostały zmapowane do genomu referencyjnego GRCh38 za pomocą BWA-MEM, następnie przygotowanie plików BAM i detekcja mutacji została przeprowadzona za pomocą zestawu narzędzi GATK i programu Mutect2. Wykryte warianty opisano za pomocą programu Oncotator.

Modelowanie. Prezentowane analizy zostały wykonane w środowisku R v4.2.2.
Początkowo dane zostały zamodelowane za pomocą algorytmu MOBSTER. Algorytm ten dopasowuje do spektrum VAF mieszaninę rozkładów o kształcie potęgowym i dwumianowym:

$$N(f) \sim \frac{A}{f^{\alpha}} + \sum_{i=1}^{K} N_k \cdot binomial(n, f_k) \tag{1}$$

gdzie f jest częstością alleliczną mutacji, $N(f)$ określa liczbę mutacji o częstości f, K jest liczbą klonów i sub-klonów w nowotworze, N_k liczbą mutacji związanych z (sub)klonem k, f_k częstością alleliczną mutacji w (sub)klonie k, oraz A jest stałą proporcjonalną do częstości mutacji na efektywny podział komórki, opisany w [15]. Potęgowa komponenta tego modelu jest nazywana *neutralnym ogonem* i odpowiada pojawiającym się we wszystkich komórkach neutralnym mutacjom. Częstość alleliczna tych wariantów zależy głównie od czasu wystąpienia mutacji i jest równa odwrotności liczby komórek w nowotworze w chwili mutacji. Jak wykazali Durrett [6], Williams [15] i inni, wykładnik α krzywej potęgowej opisującej neutralny ogon jest równy 2, przy założeniu stałego tempa mutacji i eksponencjalnego wzrostu populacji. Wówczas współczynnik A pozwala określić średnie tempo mutacji na podział komórki [15]. Składowe dwumianowe odpowiadają wariantom związanym z klonami i sub-klonami, a parametry je opisujące pozwalają określić czas pojawienia się sub-klonu oraz współczynnik jego selekcji [16].

W przypadku większości analizowanych w pracy próbek, MOBSTER nie był uprzednio w stanie poprawnie dopasować modelu z powodu niewystarczającej liczby mutacji reprezentujących neutralny ogon rozkładu. Z tego powodu, w ramach rozprawy, zaimplementowany został w języku R pakiet *cevomod*, zawierający zaproponowane przez nas metody dopasowywania modelu dla danych z sekwencjonowania całoeksomowego i sekwencjonowania o niewystarczającym dla algorytmu MOBSTER pokryciu.

W *cevomod* zaproponowane zostały dwie alternatywne metody dopasowywania do danych komponenty potęgowej. Pierwsza z nich zakłada eksponencjalny wzrost populacji i stałe tempo mutacji. W tym przypadku $\alpha = 2$, a A zależy od liczby przedziałów w spektrum VAF i częstości mutacji na efektywny podział μ. μ zostaje wyznaczone z zależności:

$$M(f) \sim \mu/f \tag{2}$$

opisanej w [15], gdzie $M(f)$ jest liczbą mutacji o częstości większej od f. Druga metoda została zaimplementowana celem walidacji założeń modelu i detekcji występowania w ogonie licznych, posiadających przewagę selekcyjną mikro-klonów [14]. W metodzie tej oba parametry krzywej potęgowej A i α są dopasowywane do danych, a otrzymane wartości α różne od 2 świadczą o niespełnieniu założeń lub obecności mikro-klonów. Krzywa potęgowa dopasowywana jest w procesie optymalizacji, w którym maksymalizowana funkcja celu I składa się z dwóch elementów: nagrody za liczbę mutacji pod krzywą (ang. *mutation count reward*, MCR) i kary za odłączenie się krzywej od spektrum (ang. *spectrum detach penalty*, SDP).

$$I = MCR - SDP \tag{3}$$

W obu metodach, dwumianowe komponenty modelu są dopasowywane poprzez klasteryza-

cję wariantów nie wyjaśnionych przez komponentę potęgową. W pierwszym kroku w każdym przedziale i spektrum VAF losowanych jest n_i wariantów, gdzie n_i jest dodatnią częścią różnicy między liczbą wariantów w przedziale i i predykcją komponenty potęgowej w tym przedziale. Następnie, za pomocą pakietu BMix [4], warianty są klasteryzowane za pomocą mieszaniny od 1 do 3 rozkładów dwumianowych, z uwzględnieniem głębokości sekwencjonowania wariantów.

5 Wyniki

Wyniki naszej pracy opisaliśmy w dwóch częściach. Pierwsza część przedstawia analizy zbiorów danych AML, BRCA i LSCC, w których od każdego pacjenta pobrane były po dwie próbki nowotworowe. W drugiej części przedstawione są wyniki analizy zbioru danych BLCA, zawierającego łącznie 66 próbek od dwóch pacjentów.

5.1 Analiza ewolucji przerzutujących nowotworów piersi i krtani oraz nawracających białaczek.

Guzy z kohort AML i BRCA/LSCC wykazały różne kształty rozkładów VAF. W większość próbek AML rozpoznawalne były dwie grupy wariantów. Rozkład VAF wariantów o wysokiej częstości miał symetryczny kształt ze średnią wartością VAF oscylującą wokół 0.5, co w komórkach diploidalnych (o dwóch kopiach każdego genu) odpowiada wariantom klonalnym obecnym we wszystkich komórkach. Duża liczność mutacji w tej grupie jest oznaką przeszłych *selektywnych zastąpień* (ang. *selective sweeps*), w których wcześniejsze klony zostały całkowicie zastąpione przez bardziej agresywne klony powstałe później, posiadające większą liczbę mutacji. Rozkład VAF wariantów o niskich częstościach był asymetryczny, z wydłużonym prawym zboczem, zgodnym z modelem *neutralnego ogona*. W kohortach BRCA i LSCC rozkłady VAF składały się z pojedynczych, asymetrycznych rozkładów o niskich częstościach, bez wyraźnych grup mutacji klonalnych.

Z użyciem zaimplementowanego przez nas pakietu *cevomod* dopasowaliśmy do danych mieszaniny rozkładów potęgowych i dwumianowych. Używając modelu z wykładnikiem $\alpha = 2$ określiliśmy współczynniki mutacji w próbkach, oraz zidentyfikowaliśmy w neutralnych ogonach subklony, dla których wyznaczyliśmy czasy narodzin i współczynniki selekcji. We wszystkich kohortach zidentyfikowaliśmy częste zmiany współczynnika mutacji pomiędzy próbkami z nowotworów pierwotnych (z momentu diagnozy w AML lub z guza głównego w BRCA i LSCC) i wtórnych (z momentu nawrotu choroby w AML lub przerzutu do węzła chłonnego w BRCA i LSCC). W LSCC dominował wzrost tempa mutacji w guzie wtórnym, którego średnia wartość wynosiła 14%. W kohortach AML i BRCA zmiany tempa mutacji w obu kierunkach były równie częste, najczęściej nie większe niż 2-krotne. Tylko w u jednego pacjenta AML rozpoznaliśmy blisko 10-krotny spadek tempa mutacji w

próbce z nawrotu choroby.

W BRCA i LSCC, większość wykrytych subklonów pojawiła się wcześnie, w trakcie pierwszych 20% podwojeń objętości nowotworu. Ich współczynniki selekcji zawierały się najczęściej (za wyjątkiem dwóch próbek) pomiędzy 0% do do 20% wzrostu tempa podziałów komórek. Wczesne powstanie klonów, niskie bądź umiarkowane współczynniki ich selekcji i brak obecności wyraźnych grup mutacji klonalnych wskazują na niemal neutralną ewolucję analizowanych nowotworów BRCA i LSCC od momentu inicjacji guza do momentu sekwencjonowania. Klony o współczynnikach selekcji bliskich 0 mogą odpowiadać punktowej ewolucji nowotworów, w której wszystkie subklony powstają na początku rozwoju choroby i współistnieją w jej trakcie. Model taki został niedawno zaproponowany dla niektórych raków jelita grubego [12] i dla ewolucji zmian liczb kopii genów w nowotworach piersi [7]. Próbki z klonami o wyższych współczynnikach selekcji, bliskich 20%, mogły zostać zsekwencjonowane przed dokonaniem się pierwszego *selektywnego zastąpienia*. Niskie częstości alleliczne wykrytych w tych próbkach mutacji w genach *driverowych* są zgodne z symulacjami młodych nowotworów w studium Bozic et al. [2].

W kohorcie AML czasy narodzin i wartości współczynnika selekcji klonów były znacznie bardziej zróżnicowane. Czasy narodzin subklonów w większości próbek wynosiły do 60% podwojeń objętości nowotworu, jednak w przypadku 3 próbek estymaty te przekraczały czas życia nowotworu. W próbkach tych dopasowania modelu z wykładnikiem $\alpha = 2$ były wyjątkowo niedokładne. Używając drugiego modelu o optymalizowanym współczynniku α zidentyfikowaliśmy znaczne odstępstwa α od wartości teoretycznej, świadczące o niespełnieniu założeń modelu. Porównanie optymalnych wartości α w próbkach guzów pierwotnych i wtórnych wykazało dominujący wzrost parametru α w próbkach z nawrotu lub przerzutów nowotworu.

Jak wykazali Tung i Durrett [14], występowanie $\alpha < 2$ może być wynikiem obecności w neutralnym ogonie licznych mikroklonów podlegających pozytywnej selekcji. Rozwiązanie to nie obejmuje jednak wartości $\alpha > 2$. W rozprawie zaproponowaliśmy matematyczne wyjaśnienie tego zjawiska wykazując, że jeśli prędkość gromadzenia się w nowotworze M' nowych mutacji zależy od wielkości nowotworu:

$$M' = \mu\lambda N(t)^{\kappa} \tag{4}$$

gdzie μ jest początkowym tempem mutacji, λ współczynnikiem eksponencjalnego wzrostu nowotworu, $N(t)$ wielkością nowotworu w chwili t, a κ dodatnią stałą $\in (0, \infty)$, to liczba mutacji o częstości f może zostać wyrażona za pomocą równania:

$$X(f) = \frac{\mu}{f^{\kappa+1}} \tag{5}$$

Wówczas $\alpha > 2$ może wynikać ze wzrostu tempa mutacji w nowotworze. Jeśli szybkość mutacji nie zależy od wielkości populacji ($\kappa = 1$), wówczas równanie 5 jest zgodne z

równaniami zaproponowanymi przez Durrett [6] i Williams [15].

5.2 Analiza ewolucji nowotworów pęcherza z pola kanceryzacji

W większości próbek obu map zidentyfikowaliśmy liczne mutacje w znanych genach *driverowych*, w tym wiele mutacji w polach zaklasyfikowanych jako normalne urothelium lub neoplazja o niskim stopniu złośliwości. Większość tych mutacji była obecna w pojedynczych polach map, rozpoznaliśmy jednak również 18 mutacji rozprzestrzenionych na wiele pól w mapie 19 i 7 takich mutacji w mapie 24. Co ciekawe, w mapie 19 rozpoznaliśmy 4 wzory rozprzestrzenienia mutacji, z których tylko 2 były związane z obszarem pęcherza zajętym przez nowotwór. 2 wzory rozprzestrzenienia mutacji dotyczyły głównie pól zaklasyfikowanych jako NU lub LGIN i wskazywały na ekspansje zmutowanych komórek nie związanych z głównym nowotworem.

Z użyciem opracowanego przez nas pakietu *cevomod*, dopasowaliśmy potęgowe komponenty modelu do neutralnych ogonów w spektrach VAF większości pól obu map. Estymowane współczynniki wykazały wysokie tempo mutacji w polach nowotworowych oraz podwyższone tempo mutacji w licznych polach zaklasyfikowanych jako normalne urothelium lub neoplazja o niskim lub wysokim stopniu złośliwości. W mapie 19, rozpoznaliśmy nie-nowotworowy obszar o podwyższonym tempie mutacji pokrywający się z obszarem lokalnej ekspansji komórek z mutacjami w genach ELF3, KMT2D i RHOB, niewykrytymi w próbkach nowotworowych.

Następnie używając drugiego zaproponowanego przez nas modelu, zmierzyliśmy odchylenia wykładnika α od przewidywanej wartości 2. Wartości większe niż 2 przeważały w mapie 19 i często dotyczyły próbek zaklasyfikowanych jako normalne urothelium lub nisko-inwazyjna neoplazja. W mapie 24 częstość występowania wartości mniejszych i większych od 2 była zbliżona, jednak odchylenia α w dół były bardziej znaczące. Wartości współczynnika α były podobne w sąsiadujących ze sobą polach map, świadcząc o rozpowszechnieniu na sąsiednie pola procesów które wpłynęły na te wartości.

Odchylenia α mogą wynikać z różnych procesów, takich jak występowanie mikroklonów o przewadze selekcyjnej, czy zmiana tempa mutacji. Wykazaliśmy, że szacunki tempa mutacji w przypadku $\alpha > 2$ są niższe niż w modelach z $\alpha = 2$, wysokie wartości α mogą więc sygnalizować wyższe tempo mutacji. Obniżenie wartości α, wynikające z obecności mikro-klonów, może zatem być skompensowane wzrostem tempa mutacji. Rzeczywiście, wartości $\alpha < 2$ obserwowane były rzadziej niż $\alpha > 2$, a tempo mutacji było wyższe w próbkach o bardziej zaawansowanym stadium choroby. Niskie wartości α zostały zaobserwowane w niektórych próbkach o niskim tempie mutacji, ale wysokiej liczbie mutacji w genach *driverowych*. Ostateczna odpowiedź, próbki te rzeczywiście reprezentują przypadek mikroklonów opisanych przez Tung i Durrett [14], może być poza zasięgiem analizy danych sekwencjonowania typu *bulk*, która nie umożliwia rozróżnienia małych

klonów [13].

6 Podsumowanie

W prezentowanej rozprawie użyliśmy danych z sekwencjonowania DNA typu *bulk* przeszło 140 próbek nowotworowych. Wykazaliśmy, że parametry dynamiki ewolucyjnej nowotworów mogą zostać estymowane w oparciu o dane sekwencjonowania typu *bulk* oraz, że istnieją mierzalne różnice w parametrach tej dynamiki pomiędzy próbkami reprezentującymi guzy główne, przerzuty, nawroty nowotworu czy też różne stadia jego zaawansowania. Dowiedliśmy zatem prawdziwości postawionej tezy, mówiącej że: *Zmiany w dynamice ewolucyjnej nowotworów związane z powstawaniem przerzutów i nawrotami choroby mogą być mierzone w oparciu o wyniki sekwencjonowania DNA typu bulk*, co było pierwszym celem rozprawy.

Algorytm MOBSTER nie dopasował poprawnie modeli do użytych w rozprawie danych ze względu na niewystarczającą liczbę mutacji w neutralnym ogonie. Z tego powodu zaproponowaliśmy nowe metody dopasowania modelu do danych pochodzących z sekwencjonowania całoeksomowego lub sekwencjonowania o niewystarczającym pokryciu. Metody zostały zaimplementowane w nowej paczce w języku R, *cevomod*, spełniając drugi cel niniejszej pracy. Pakiet jest publicznie dostępny w repozytorium GitHub pod adresem `https://github.com/pawelqs/cevomod`. Pozwala on wybrać pomiędzy dwoma typami modeli: modelem *neutralnym* o wykładniku komponenty potęgowej równym 2, oraz modelem w którym wykładnik jest optymalizowany do danych. Pierwszy z modeli pozwala wyznaczyć parametry dynamiki ewolucyjnej takie jak tempo mutacji, czas pojawienia się klonów, czy współczynniki selekcji klonów przy założeniu eksponencjalnego wzrostu guza i stałego tempa mutacji. Drugi model pozwala zwalidować te założenia: odchylenia optymalnej wartości α świadczą o niespełnieniu któregoś z nich.

Używając zaproponowanych metod wykazaliśmy, że założenia na których bazują popularne modele używane do estymacji parametrów ewolucji nowotworów, mogą nie być zachowane w wielu guzach. Zaobserwowaliśmy i opisaliśmy częste odchylenia wykładnika opisującego neutralny ogon rozkładu od oczekiwanej wartości 2. Zaproponowaliśmy również matematyczne wyjaśnienie odchyleń, wiążąc je ze zmianami w tempie mutacji podczas wzrostu nowotworu. Tym samym osiągnęliśmy trzeci, ostatni cel rozprawy.

Podziękowania. Praca była współfinansowana przez Unię Europejską w ramach Europejskiego Funduszu Społecznego (grant POWR.03.02.00-00-I029) oraz przez Narodowe Centrum Nauki (grant 2018/29/B/ST7/02550). Obliczenia wykonano przy użyciu klastra obliczeniowego Ziemowit (`http://www.ziemowit.hpc.polsl.pl`) znajdującego się na wyposażeniu Laboratorium Biologii Obliczeniowej i Bioinformatyki, Centrum Biotechnologii Politechniki Śląskiej w Gliwicach, zakupionego w ramach projektu Śląska BIO-

FARMA, nr POIG.02.01.00-00-166/08.

References

[1] Jolanta Bondaruk et al. "The origin of bladder cancer from mucosal field effects". In: *iScience* 25.7 (June 2022), p. 104551. ISSN: 2589-0042. DOI: 10.1016/j.isci.2022.104551. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9209726/ (visited on 01/18/2023).

[2] Ivana Bozic, Chay Paterson, and Bartlomiej Waclaw. "On measuring selection in cancer from subclonal mutation frequencies". en. In: *PLOS Computational Biology* 15.9 (2019), e1007368. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1007368. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007368 (visited on 01/18/2023).

[3] *Cancer*. en. Feb. 2022. URL: https://www.who.int/news-room/fact-sheets/detail/cancer (visited on 12/22/2022).

[4] Giulio Caravagna et al. "Subclonal reconstruction of tumors by using machine learning and population genetics". en. In: *Nature Genetics* 52.9 (Sept. 2020), pp. 898–907. ISSN: 1546-1718. DOI: 10.1038/s41588-020-0675-5. URL: https://www.nature.com/articles/s41588-020-0675-5 (visited on 12/15/2022).

[5] Amit G. Deshwar et al. "PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors". In: *Genome Biology* 16.1 (Feb. 2015), p. 35. ISSN: 1465-6906. DOI: 10.1186/s13059-015-0602-8. URL: https://doi.org/10.1186/s13059-015-0602-8 (visited on 04/12/2023).

[6] Rick Durrett. "POPULATION GENETICS OF NEUTRAL MUTATIONS IN EXPONENTIALLY GROWING CANCER CELL POPULATIONS". In: *The annals of applied probability : an official journal of the Institute of Mathematical Statistics* 23.1 (2013), pp. 230–250. ISSN: 1050-5164. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3588108/ (visited on 12/15/2022).

[7] Ruli Gao et al. "Punctuated copy number evolution and clonal stasis in triple-negative breast cancer". en. In: *Nature Genetics* 48.10 (Oct. 2016), pp. 1119–1130. ISSN: 1546-1718. DOI: 10.1038/ng.3641. URL: https://www.nature.com/articles/ng.3641 (visited on 01/12/2023).

[8] Douglas Hanahan and Robert A. Weinberg. "Hallmarks of Cancer: The Next Generation". English. In: *Cell* 144.5 (Mar. 2011), pp. 646–674. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2011.02.013. URL: https://www.cell.com/cell/abstract/S0092-8674(11)00127-9 (visited on 12/15/2022).

[9] Christopher A. Miller et al. "SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution". en. In: *PLOS Computational Biology* 10.8 (2014), e1003665. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003665. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003665 (visited on 04/12/2023).

[10] Andrew Roth et al. "PyClone: statistical inference of clonal population structure in cancer". en. In: *Nature Methods* 11.4 (Apr. 2014), pp. 396–398. ISSN: 1548-7105. DOI: 10.1038/nmeth.2883. URL: https://www.nature.com/articles/nmeth.2883 (visited on 04/12/2023).

[11] Liran I. Shlush et al. "Tracing the origins of relapse in acute myeloid leukaemia to stem cells". en. In: *Nature* 547.7661 (July 2017), pp. 104–108. ISSN: 1476-4687. DOI: 10.1038/nature22993. URL: https://www.nature.com/articles/nature22993 (visited on 12/15/2022).

[12] Andrea Sottoriva et al. "A Big Bang model of human colorectal tumor growth". en. In: *Nature Genetics* 47.3 (Mar. 2015), pp. 209–216. ISSN: 1546-1718. DOI: 10.1038/ng.3214. URL: https://www.nature.com/articles/ng.3214 (visited on 01/18/2023).

[13] Xianbin Su et al. "Accurate tumor clonal structures require single-cell analysis". eng. In: *Annals of the New York Academy of Sciences* 1517.1 (Nov. 2022), pp. 213–224. ISSN: 1749-6632. DOI: 10.1111/nyas.14897.

[14] Hwai-Ray Tung and Rick Durrett. "Signatures of neutral evolution in exponentially growing tumors: A theoretical perspective". en. In: *PLOS Computational Biology* 17.2 (2021), e1008701. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1008701. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008701 (visited on 01/18/2023).

[15] Marc J. Williams et al. "Identification of neutral tumor evolution across cancer types". en. In: *Nature Genetics* 48.3 (Mar. 2016), pp. 238–244. ISSN: 1546-1718. DOI: 10.1038/ng.3489. URL: https://www.nature.com/articles/ng.3489 (visited on 12/15/2022).

[16] Marc J. Williams et al. "Quantification of subclonal selection in cancer from bulk sequencing data". en. In: *Nature Genetics* 50.6 (June 2018), pp. 895–903. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0128-6. URL: https://www.nature.com/articles/s41588-018-0128-6 (visited on 12/15/2022).