POMORSKI UNIWERSYTET MEDYCZNY
W SZCZECINIE

lek. Tomasz Falgowski

ANALIZA METYLACJI SEKWENCJI CPG W OPARCIU O UCZENIE MASZYNOWE I SIECI NEURONOWE

Rozprawa doktorska w dziedzinie nauk medycznych i nauk o zdrowiu

Dyscyplina nauki medyczne

Promotor: prof. dr hab. n. med. Tadeusz Sulikowski

Szczecin 2024

Podziękowania

Dziękuję Profesorowi Tadeuszowi Sulikowskiemu za wsparcie i cierpliwość, które stały się dla mnie fundamentem na drodze do napisania tej rozprawy.

Dziękuję mojej Narzeczonej za to, że nie pozwoliłaś mi poddać się w najtrudniejszym momencie tej pracy i za pomoc na każdym etapie jej powstawania

Dziękuję także mojej Mamie oraz Mamie Mojej Narzeczonej - bez waszej pomocy nie udałoby mi się tego ukończyć.

I Synowi Szymonowi "Me ex immenso vacuo liberasti, novam vitae significationem mihi dedisti."

Spis treści

Wstęp

Złożoność ludzkiego genomu, będącego zestawem kompletnych instrukcji genetycznych polega nie tylko na określonej kompozycji miliardów par zasad, ale także na chemicznej modyfikacji, która może być odczytywana i interpretowana przez enzymy i inne czynniki molekularne. Te chemiczne modyfikacje zależne są od epigenetycznych mechanizmów [1]. Metylacja kwasu deoksyrybonukleinowego (DNA) jest jednym z najlepiej opisanych czynników tego złożonego systemu definiowanego jako epigenetyczny czynnik polegający na kowalencyjnym przyłączeniu grup metylowych (-CH3) w pozycji 5 pierścienia pirymidynowego cytozyny w DNA przez metylotransferazy DNA (DNMTs). Proces ten jest ważny dla prawidłowego rozwoju i odgrywa znaczącą rolę m.in. w genomowym imprintingu [2, 3], inaktywacji chromosomu X [4], regulacji transkrypcji DNA [5] oraz w patogenezie wielu chorób w tym w chorobach nowotworowych, dla których stanowi potencjalny biomarker detekcji i klasyfikacji [6, 7]. Jedną z najpopularniejszych platform do analizy metylacji jest Illumina Infinum BeadChips [8]. Wartość β jest podstawową wartością używaną do pomiaru stopnia metylacji, ale niestety przez fakt, iż nie spełnia niektórych założeń standardowych testów statystycznych, jak rozkład normalny, czy homoskedastyczność jest trudna do analizy z ich użyciem. Z tego powodu poszukiwane są nowe metody do jej analizy. W ostatnim czasie nastąpił znaczący rozwój w dziedzinie sztucznej inteligencji, uczenia maszynowego i sieci neuronowych. Zdolności tych ostatnich do uniwersalnej aproksymacji dowolnej funkcji, a więc w rzeczywistości dostosowania do dowolnego problemu sprawia, że są one narzędziami bardzo efektywnymi i wszechstronnym, chociaż ich użycie dla wielu zastosowań może wydawać się zbyt ekstremalne. Jednak już sama ich zdolności do ekstrakcji istotnych cech może pozwolić na stworzenie narzędzia do identyfikacji wzorców metylacji. Dodatkowo sieci neuronowe w przeciwieństwie do powszechnie używanych testów statystycznych są bardziej odporne na wiele podstawowych problemów dotyczących klasycznych testów statystycznych takich jak problem wysokiej wariancji i szumu, niekompletność danych, czy wspomniana wcześniej heteroskedastyczność. Atutem wydaje się też być zdolność tej technologii do adaptacji, personalizacji i ulepszania tzn., w miarę jak sieć neuronowa uczy się danych, jej właściwości adaptacyjne pozwalają na jej ukształtowanie pod konkretny problem badawczy, a sam model można udoskonalać wraz z nowymi

informacjami. Niezależnie od wybranej metody badawczej do analizy metylacji, dane uzyskane przy ich udziale powinny być reprezentatywne i wskazywać realne znaczenie biologiczne.

Nazewnictwo specjalistyczne

Ze względu na fakt, że zarówno epigenetyka, jak i uczenie maszynowe oraz sieci neuronowe są nowymi i bardzo szybko rozwijającym się dziedzinami, a dominująca większość piśmiennictwa napisana jest w języku angielskim lub pochodzi od nazwisk twórców i nie posiada swojego polskiego tłumaczenia, oraz fakt, że w środowisku osób zajmujących się wymienioną tematyką bardzo często korzysta się z anglojęzycznych zwrotów, podjęto decyzję, by nie tłumaczyć niektórych użytych zwrotów, a w sytuacji, gdy definicja posiada tłumaczenie na język polski uzupełnić ją o określenie w języku angielskim.

Od DNA do białka

W ogólnym modelu w żywych organizmach informacja genetyczna jest przenoszona od DNA przez informacyjny kwas rybonukleinowy (mRNA) do białka [9]. Niektórzy autorzy, idąc daleko w swoich rozważaniach, uważają, że przepływ informacji jest analogiczny do systemu komputerowego, a nawet są próby wykorzystania DNA jako potencjalnego nośnika danych cyfrowych [10–12]. DNA to cząsteczka, która stanowi podstawę życia przez zawarte w niej informacje dotyczące syntezy białek w żywym organizmie, stanowiąc podstawę jego rozwoju, funkcjonowania i reprodukcji. Samo DNA ulega replikacji i przenosi zawartą w sobie informację w trakcie podziału komórkowego [13]. DNA zbudowane jest z mniejszych jednostek zwanych nukleotydami. Każdy nukleotyd składa się z polisacharydu (deoksyrybozy), grupy fosforanowej i zasady azotowej. Adenina, guanina, cytozyna i tymina to cztery zasady azotowe, z których DNA jest zbudowane [13]. Dzięki komplementarnemu parowaniu zasad nukleotydy w DNA parują się ze sobą w taki sposób, że adenina łączy się z tyminą za pomocą dwóch wiązań wodorowych, a guanina z cytozyną za pomocą trzech wiązań wodorowych. Struktura DNA w postaci podwójnej helisy i jej stabilność jest związana z komplementarnością parowania zasad [13]. Podstawową jednostką organizacji DNA u eukariontów jest nukleosom, składający się z

segmentu DNA owiniętego wokół oktameru białka histonowego [14]. Histony są białkami, które nie tylko pomagają zagęszczać DNA, ale odgrywają także rolę w regulacji ekspresji genów [15]. Połączenie nukleosomów przez DNA łącznikowy nazywane jest chromatyną. Podczas podziału komórki, chromatyna w jądrze kondensuje się tworząc chromosomy [16]. Kodujący DNA to część DNA, która zawiera instrukcje dotyczące tworzenia białek, a gen jest podstawową jednostką kodującą. Niekodująca część DNA to ta, który nie zawiera instrukcji tworzenia białek. Niektóre niekodujące sekwencje DNA pełnią funkcję regulacyjne, takie jak kontrola ekspresji pobliskich genów. Inne niekodujące sekwencje DNA mogą pełnić funkcje strukturalne, takie jak pomoc w organizacji chromosomu [17, 18]. Transkrypcja to proces tworzenia kopii genu z DNA na mRNA. Jest ono używane jako matryca do syntezy białek podczas translacji [13]. Translacja to proces wykorzystania informacji zawartych w cząsteczce mRNA do syntezy białka. Proces transkrypcji rozpoczyna się od inicjacji, czyli przyłączenia czynników transkrypcyjnych i polimerazy kwasu rybonukleinowego (RNA) z promotorem, który jest sekwencją DNA sygnalizującą początek genu, niczym wielka litera na początku zdania [13]. Pierwszy czynnik transkrypcyjny jest nazywany białkiem wiążącym TATA *(ang. TATA-binding protein)* i łączy się on do sekwencji DNA zwanej sekwencją TATA (*ang. TATA box*) [19], a gdy ten pierwszy czynnik transkrypcyjny zwiąże się, inne wiążą się w grupach, aż wreszcie polimeraza RNA dołącza do kompleksu tuż przed początkiem sekwencji genu. Gdy wszystkie te czynniki zostaną złożone, rozpoczyna się proces transkrypcji. Jest wiele czynników transkrypcji, które dzielą się na rodziny białek z charakterystycznymi dla siebie domenami łączącymi DNA jak helisa-zwrot-helisa (ang. helix-turn-helix), palec cynkowy (ang. zinc finger), czy zamek leucynowy (ang. leucine zipper) [20, 21]. Czynniki transkrypcyjne to białka, które mogą stymulować lub hamować transkrypcję i mają szeroki zakres zadań od wpływania na genom w reakcji na czynniki środowiskowe [22], regulowania procesów komórkowych [23], czy regulowania odpowiedzi immunologicznej [24]. Kolejnym etapem jest elongacja, kiedy to podwójna helisa DNA ulega rozwinięciu, a polimeraza RNA rozpoczyna właściwy etap transkrypcji. By transkrypcja mogła być przeprowadzona, chromatyna musi być w stanie rozwiniętym, za co odpowiada acetylaza przez przyłączenie grup acetylowych. Proces ten jest odwracalny i wykonywany przez deacetylazę [25]. Kolejnymi etapami są: terminacja, która zakańcza transkrypcję, gdy polimeraza RNA osiągnie sekwencję terminacyjną oraz proces

modyfikacji potranskrypcyjnej, by przygotować heterogenny jądrowy RNA n(pre-mRNA, hnRNA) do translacji. Sama translacja jest procesem syntezy białka [13]. Regulacja ekspresji genu jest zarządzana m.in. przez elementy cis regulatorowe, które znajdują się w niekodujących regionach DNA i wpływają na opisany powyżej przepływ informacji genetycznej na wielu jej etapach od organizacji chromatyny (jej upakowania/rozwinięcia), przez transkrypcję, po potranskrypcyjne modyfikacje, czy typowy dla eukariotów eksport RNA. Do czynników modulacji ekspresji genów zaliczyć można zarówno elementy cis-regulacyjne, takie jak promotory, jak i odległe typy elementów cis-regulacyjnych, tj. wzmacniacze i wyciszacze transkrypcji [13]. O ile te pierwszy zazwyczaj znajdują się w odległości ok 1kb od miejsca startu transkrypcji, o tyle drugie mogą działać na dłuższy dystans nawet większy niż 1 Mb [26]. Geny, które podlegają ciągłej ekspresji i regulują podstawowe procesy w komórce nazywamy referencyjnymi (ang. house keeping genes) [27]. Tylko 2% ludzkiego DNA koduje białka, pozostała część genomu składa się m.in. z intronów, promotorów, sekwencji kontrolujących, wirusowych sekwencji czy powtarzających się. Z tych ostatnich najobszerniejsze są transpozony, czyli sekwencje, które mogą przemieszczać się w genomie tej samej komórki [28–31].

Epigenetyka

Badania dotyczące chemicznych zmian, jakim podlega DNA, a które to zmiany nie modyfikują jego sekwencji dały początek obszernej nauce zwanej epigenetyką [32].

Rys historyczny epigenetyki i genetyki

Początku historii epigenetyki i genetyki można doszukać się już w XVIII wieku, kiedy to Pierre Louis Maupertuis analizował rodzinne występowanie polidaktylii, czyli występowania dodatkowego palca u ludzi [33]. W XIX wieku Charles Darwin sformułował teorię ewolucji obserwując, jak pewne cechy przekazywane są dalszym pokoleniom [34]. W tym samym wieku Grzegorz Mendel, czeski zakonnik sformułował pierwsze zasady dziedziczenia [35]. Na początku XX wieku Walter Sutton i Theodor Boveri [36] zaproponowali chromosomalną teorię dziedziczenia, a w 1928 roku Fredrick Griffith przeprowadził eksperyment na bakteriach dowodzący istnienia biologicznego czynnika, który mógłby przenieść cechy jednego organizmu na drugi

[37], co w 1946 potwierdzili Avery, MacLeod i McCarty formułując hipotezę, ze DNA jest tym czynnikiem [38]. W 1953 roku naukowcy James Watson i Francis Crick na podstawie swoich obserwacji zaproponowali, że cząsteczka DNA przyjmuje kształt podwójnej helisy [39]. W 1975 roku Robin Holliday i John Pugh [40] oraz jednoczasowo Arthur Riggs [41] zaproponowali mechanizm regulacji, który funkcjonuje dodatkowo do genomu. Odkrycie to dało początek epigenetyce, a proces metylacji DNA był wskazywanym przez nich mechanizmem. W 1986 roku Bird [42] opublikował pracę zauważając istnienie wysp CpG i opisując ich znaczenie. XXI wiek charakteryzuje się ogromnym skokiem w rozwoju wiedzy na temat genetyki i epigenetyki włączając w to olbrzymie badania, jak Projekt Poznania Ludzkiego Genomu (ang. Human Genome Project) [43], który zakończył swoją pracę w 2003 roku oraz jego następcę projekt ENCODE [44].

Metylacja i metylom — definicja

Jednym z bardziej poznanych i będących częścią tematu tego przewodu mechanizmem epigenetycznym jest proces metylacji, czyli przyłączenia grupy metylowej do nukleotydu cytozyny w dinukleotydzie cytozyna-guanina w cząsteczce DNA [45]. Metylom to całość wszystkich miejsc metylowanych w genomie organizmu. Należy pamiętać, że różne typy komórek w organizmie mogą mieć unikalne wzorce metylacji DNA, co w konsekwencji oznacza, że różne geny są metylowane i wyciszane w różnych typach komórek [46].

Wyspy, wybrzeża i szelfy CpG

Genom ssaków w ogólnym rozrachunku jest ubogi w sekwencje CpG. Jest ich ok. 28 milionów (w całym ludzkim genomie jest ok. 3,2 miliarda par zasad [43]) Sekwencje te pod względem lokalizacji i charakterystyki możemy podzielić na wyspy (ang. CpG islands), wybrzeża (ang. CpG shores), szelfy (ang. CpG shelves) i regiony otwartego morza (ang. open sea CpG regions) [47]. Wyspy CpG to bogate w sekwencje CpG regiony długości >200 par zasad i zawierające przynajmniej 50% cytozyny. W ogólnej ocenie jest ponad 29,000 wysp CpG [48] i skojarzone są one z ponad połową ludzkich genów i wszystkich genów referencyjnych (ang. house-keeping genes) [49]. Natomiast za wybrzeża uważamy regiony oddalone o 0-2 kb od wysp, szelfy o 2-4 kb a

sekwencje CpG będące odizolowane w genomie nazywamy regionami otwartego morza [47].

Stabilność metylacji

W przeciwieństwie do komórek embrionalnych stan metylacji sekwencji CpG w komórkach somatycznych jest uważany za stabilny [50] i ok. 70% dinukleotydów CpG jest metylowanych. Regiony o niższym zagęszczeniu sekwencji CpG włączając w to elementy powtarzalne są wysoce metylowane w przeciwieństwie do wysp CpG, które powszechnie znajdowane są przy promotorach (te z kolei są hipometylowane) [51, 52]. Genomowe odległe elementy regulujące, jak wzmacniacze transkrypcji (ang. enhancers), które mają pośredni poziom CpG charakteryzują się wysoce zróżnicowanym poziomem metylacji [53, 54].

Regulacja metylacji i jej znaczenie

Metylacja DNA to fundamentalny aspekt biologii człowieka, wpływający na rozwój i funkcjonowanie naszych komórek, który stanowi istotny element wielu procesów od regulacji transkrypcji [55], przez wpływ na inaktywacje chromosomu X [56, 57], genomowy imprinting [58, 59] czy stabilność genomu [60].Mimo swojej stabilności sekwencje CpG mogą ulegać metylacji de novo i demetylacji, jak również są procesy, które metylację utrzymują.

Metylacja de novo

Metylacja de novo zachodzi z udziałem metylotransferaz DNA (ang. DNA methyltransferases, DNMTs), a dokładnie DNMT3A i DNMT3B, które wchodzą w interakcję z niemetylowanymi lub semi-metylowanymi sekwencjami CpG i dokonują addycji grup metylowych do cytozyny przy użyciu s-adenozylometioniny jako grupy donorowej [61, 62]. DNMT1 z kolei utrzymuje metylację podczas podziału komórki [63].

DNMT3

Sekwencje CpG promotorów genów aktywnie transkrybowanych są niemetylowane, ale wzbogacone o histon 3 trimetylowany na lizynie 4 (H3K4me3)

[64]. W przeciwieństwie do ich promotorów sekwencje CpG w obrębie genów są metylowane [65]. Rodzina metylotransferaz DNMT3 składa się z domeny Pro-Trp-Trp-Pro (PWWP) i domeny ATRX-DNMT3D-DNMT3L (ADD), które mają znaczenie w umieszczaniu DNMT3A i DNMT3B w ich docelowych miejscach i w ich regulacji enzymatycznej, a sama domena PWWP wymagana jest do połączenia chromatyny z DNMT3A i DNMT3B [66] i do rozpoznania i przyłączenia do H3K36me3. Interakcja pomiędzy PWWP a H3K36me3 stymuluje aktywność DNMT3A i DNMT3B do rozpoczęcia metylacji [67, 68]. Dodatkowo należy zwrócić uwagę, że modyfikacja H3K36me3 z udziałem enzymu SETD2 oraz DNMT3B mogą mieć ważną rolę w tłumieniu transkrypcji ukrytych promotorów znajdujących się w obrębie genów, aczkolwiek proces wymaga dokładniejszej analizy i sprawdzenia, czy biorą w tym udział dodatkowe mechanizmy jak hamowanie czynników transkrypcyjnych, czy rekrutacja białek odczytujących metylowane CpG [69]. Katalitycznie nieaktywna DNMT3L stymuluje DNMT3A i DNMT3B [70, 71].

DNMT1

Do utrzymania metylacji w trakcie replikacji wymagane jest by nowo syntetyzowana nić DNA była tak samo metylowana, jak pierwowzór, za co odpowiada DNMT1 [72–74]. Zaburzenia funkcjonowania DNMT1 lub jej kofaktora UHRF1 (ang. ubiquitin-like, containing PHD and RING finger domains 1) mogą prowadzić do demetylacji [75]. Białko UHRF1 przez wiązanie do częściowo-metylowanego DNA przez domenę SRA (ang. SET- and RING-finger associated domain) rekrutuje DNMT1 [76]. Mechanizm ten zależny jest od aktywności ligazy ubikwitynowej E3 (ang. E3 ubiquitin ligase) [77]. Metylotransferaza DNMT1 ma także zdolność do metylowania de novo promotorów przez interakcję z kompleksem MBD3-NuRD/Mi2 [78].

Demetylacja

Dwa procesy są odpowiedzialne za demetylację. Pierwszy proces to pasywna dylucja wskutek replikacji DNA lub w trakcie naprawy zasad inicjowanego przez glikozylazę DNA, a drugi to aktywne działanie dioksygenazy TET [79].

Dioksygenazy TET

Nazwa pochodzi od translokacji pomiędzy 10 a 11 chromosomem w ostrej białaczce szpikowej, gdzie to białko zostało zidentyfikowane jako część białka fuzyjnego [80]. Odkrycie białek TET zmieniło dotychczasowy pogląd, że metylacja jest niezmiennym czynnikiem wyciszania genów. Dioksygenaza TET stopniowo utlenia 5-metylocytozynę do 5-hydroksymetylocytozyny, 5-formylocytozyny i następnie do 5-karboksycytozyny [81, 82]. Rodzina białek TET składa się z TET1, TET2 i TET3. Wszystkie trzy DNMTs mają zdolność do katalizowania stopniowej demetylacji 5-metylocytozyny przez 5-hydroksymetylocytozynę, 5-formylocytoznę, aż do 5-karboksycytozyny [83]. Wszystkie formy oksydowane mogą ulec dalszej demetylacji w trakcie replikacji[84, 85] lub przez usunięcie zasady przez glikozylazę tymina DNA (ang. thymine DNA glycosylase, TDG), a następnie przez działanie szlaku naprawy przez wycinanie zasad (ang. base excision repair pathway, BER), jak to ma miejsce w przypadku 5-formylocytozyny i 5-karboksymetylocytozyny [86, 87]. Rodzina białek TET ma znaczenie w różnicowaniu pluripotencjalnych komórek, co potwierdzają ostatnie badania z mysimi komórkami macierzystymi. Utrata TET1 i TET2 powoduje redukcję 5-hydroksymetylocytozyny oraz zaburzenie transkrypcji genów odpowiedzialnych za ten proces [88, 89]. Wpływ na różnicowanie komórek pluripotencjalnych ma również interakcja TET z czynnikiem pluripotencji NANOG [90] i kompleksem hamującym Polycomb 2 (ang. Polycomb repressive complex 2, PRC2) [91]. Czynniki transkrypcyjne są szczególnie wrażliwe na stan metylacji sekwencji CpG a ich aktywacja współgra z hipometylacją [92]. Należy jednak zwrócić uwagę, że istnieją też czynniki transkrypcyjne, które preferują sekwencje metylowane np. czynniki pluripotencji KLF4 [93]. Metylacja DNA może hamować ekspresję genów pośrednio przez rekrutację białka domeny wiążącej sekwencję metylowaną CpG (ang. methyl-CpG-binding domain, MBD) [94]. Przykładem może być interakcja między nukleosomowym czynnikiem remodelującym NuRD, który wchodzi w interakcję z metylowanym DNA przez łączenie się z białkiem MBD2 i MBD3. Kompleks NuRD składa się z deacetylazy histonowej i histonowych podkompleksów. Efektem tej interakcji pomiędzy białkami MBD i NuRD jest jak pisze Leighton i wsp [95] fizyczny most pomiędzy trzema znaczącymi ramionami epigenetycznej regulacji genetycznej –

deacetylacja histonów, ATP-zależne przekształcenie nukleosomu i selektywne rozpoznanie metylowanego DNA.

Metylacja DNA a transpozony

Transpozony to sekwencje DNA, które mają zdolność do zmiany swojej pozycji w kodzie genetycznym, a co wydaje się interesujące stanowią one ok 43% kodu genetycznego. Wyróżniamy dwie klasy transpozonów [96]. Klasę I stanowią retrotranspozony, które rozprzestrzeniają się wskutek odwrotnej transkrypcji. Klasę tę można podzielić dodatkowo na podklasę Long terminal repeat (LTR) i non-LTR elements [97]. Do grupy pierwszej należą m. in endogenous retroviruses (ERVs), czyli sekwencje, które pierwotnie pochodzą od retrowirusów, ale zostały zintegrowane do genomu organizmu gospodarza. W obrębie non-LTR elements można dodatkowo wyróżnić long interspersed elements, (LINEs) i short interspersed elements, (SINEs) [98] Do klasy II należą transpozony, które przemieszczają się przez wycinanie się i wklejanie z genomu lub poprzez mechanizm rolling-circle [99]. Kilka epigenetycznych mechanizmów hamuje mobilizacje transpozonów takich jak modyfikacja histonów, czy właśnie metylacja DNA. Pojawiła się nawet hipoteza, która mówi, że znaczenie metylacji urosło znacząco w toku ewolucji jako obrona przed transpozonami [100, 101]. Ekspresja większości aktywnych retrotranspozonów jest kontrolowana przez promotory bogate w sekwencje CpG [102]. W komórkach nowotworowych zaobserwowano zmiany metylacji DNA pod postacią globalnej utraty metylacji w regionach odpowiedzialnych za aktywność transpozonów [103] np. zaobserwowano zwiększoną aktywność TET2 i TET3, czyli dwóch enzymów odpowiadających za demetylację w guzach, które jednoczasowo wykazywały wysoką ekspresję ERVs [104].

Imprinting genomowy

Pojęcie imprintingu wprowadziła po raz pierwszy w latach 60 XX wieku Helen Crouse opisując eliminację ojcowskich chromosomów X u much [105]. U ssaków grupa autosomalnych genów podlega preferencyjnej ekspresji na podstawie tylko jednego z dwóch rodzicielskich alleli - część ojcowskiego, a część matczynego allelu. U podstawy tego procesu leży zróżnicowanie pod względem metylacji dinukleotydów CpG genów podczas gametogenezy [106, 107].U ludzi około 100 genów może ulegać

imprintingowi, z których wiele ma ogromne znaczenie w prawidłowym rozwoju, a zmiany w ich ekspresji mogą powodować różne zaburzenia w tym choroby wrodzone, czy w niektórych przypadkach zwiększone ryzyko choroby nowotworowej [108]. Genomowy imprinting ma ogromne znaczenie w prawidłowym przebiegu embriogenezy, w rozwoju prenatalnym i postnatalnym. Kilka epigenetycznych procesów ma duże znaczenie w tym fenomenie [109, 110]. Znaczniki imprintingu zostają ustanowione w gametach podczas dojrzewania komórek prapłciowych, a kluczowe znaczenie u ssaków w tym procesie ma metylacja i modyfikacja na poziomie histonów [111]. Większość imprintowanych genów znajduje się w klastrach nazywanych domenami, co pozwala na wspólną regulację m.in przez odmiennie metylowane regiony (DMRs), czyli sekwencje DNA, gdzie stan metylacji pomiędzy allelem matczyny a ojcowskim jest różny [108]. Dodatkowo każda z domen kontrolowana jest przez centrum imprintingu zdefiniowanego przez odmiennie metylowane regiony komórek macierzystych (ang. Germline differentially methylated regions, gDMRs) [108]. W zależności od metylacji allelu w obrębie gDMRs chromatyna przyjmuje konfigurację otwartą lub zamkniętą w ten sposób regulując transkrypcję genów [112]. Zależna od gDMRs regulacja ekspresji genów w komórkach somatycznych prowadzi do powstania w trakcie rozwoju dodatkowych atrybutów epigenetycznych, na które składają się tzw. wtórne DMRs, które związane są m.in z modyfikacją struktury chromatyny, czy miejscami wiązania czynników transkrypcji lub promotorami genów [108]. Przebieg imprintingu jest również zależny od rodzaju tkanki i tak przykładowo UBE3a charakteryzuje się w wielu tkankach bialleliczną ekspresją, ale już w obrębie tkanki mózgowej zależny jest od ekspresji matczynej [113, 114]. Utrata imprintingu, która powstaje w wyniku delecji, mutacji, disomii jednorodzicielskiej lub epimutacji może być przyczyną niektórych schorzeń, co ilustrować może delecja w 15q11-q13, która w zależności od utraty ekspresji ojcowskiej lub matczynej może prowadzić do dwóch odmiennych chorób – odpowiednio zespołu Angelmana lub zespołu Prader-Willi [115]. Zaburzenia metylacji są rodzajem epimutacji, która może spowodować utratę imprintingu [116]. W niektórych przypadkach nieprawidłowości imprintingu obserwuje się zaburzenia metylacji w kilku locus genomu (ang. multi-locus imprinting disorder, MLID) tj. zarówno w miejscu typowym dla danej choroby jak i dodatkowym, w efekcie [117] może dojść do

nakładania się fenotypów różnych zaburzeń [118, 119]. Opisywane zjawisko można zaobserwować w zespole Silver-Russel lub w zespole Beckwith-Wiedeman [120, 121].

Embriogeneza

W trakcie embriogenezy metylacja DNA przechodzi dynamiczne zmiany, które są bardzo istotne dla dalszego rozwoju. Można wyróżnić dwa cykle tego procesu określanego mianem „resetowania metylacji" [122]. Pierwszy zachodzi po zapłodnieniu na etapie blastocysty. Początkowo dochodzi do procesu demetylacji, a jej poziom w kontekście całego metylomu spada z ok. 70% do ok. 25% [123]. Następnie po implantacji proces ten zostaje odwrócony i tworzy się nowy wzorzec metylacji, a komórki tracąc zdolność pluripotencji rozpoczynają stopniowe różnicowanie [123]. Nowy wzorzec metylacji następnie jest podtrzymywany głównie przez aktywność DNMT1 i utrzymuje się przez kolejne podziały komórkowe [124]. Drugi cykl metylacji zachodzi w momencie tworzenia się komórek prapłciowych (ang. primoridal germ cells, PGCs). W przeciwieństwie do pierwszego cyklu, stan metylacji związany z imprintingiem zostaje tu wymazany [122]. Anomalie na etapie opisanego powyżej przeprogramowania epigenetycznego (ang. epigenetic reprogramming) mogą skutkować np. utratą imprintingu i rozwojem różnorakich zaburzeń [108].

Znaczenie metylacji DNA w patogenezie wybranych chorób

Zaburzenia metylacji DNA obserwowane są w wielu chorobach – od przewlekłej niewydolności nerek i chorób neuropsychiatrycznych, a na chorobach nowotworowych kończąc [125, 126]. Mechanizm wpływu metylacji w patogenezie chorób może przebiegać w dość złożony, wieloczynnikowy sposób. Jest to widoczne na przykład w problemie methylation quantitive trait loci (meQTLs), gdzie polimorfizm pojedynczego nukleotydu wpływa na odmienny wzorzec metylacji i pewną kaskadę zdarzeń prowadzącą ostatecznie do wystąpienia nieprawidłowości w funkcjonowaniu organizmu [127, 128]. Xue i wsp. [129] stwierdzili, że czynnikiem ryzyka cukrzycy typu 2 jest zaobserwowana przez nich w obecności allelu T rs11257655 interakcja pomiędzy powstałym białkowym kompleksem FOXA1/FOXA2 i innymi czynnikami transkrypcyjnym a wzmacniaczem transkrypcji genu CAMK1D, która prowadzi do demetylacji cg03575602 i następowej upregulacji CMK1D. Do podobnych wniosków

doszli Fogarty i wsp. [130]. Z kolei Sanchez-Mut i wsp. zauważyli, że meQTLs przez wpływ na zależną od CTCF trójwymiarową konformację chromatyny związaną z promoter PM20D1 i zmienną ekspresję genów ostatecznie moduluje neurodegenerację [131]. Podobne zależności zaobserwowano również w chorobie Parkinsona [132], czy w innych chorobach w tym monogenowych [133]. Działanie starzenia i czynników środowiskowych jak ekspozycja na dym papierosowy również wpływa na wzorzec metylacji DNA [134, 135]. Meng i wsp. zauważyli związek pomiędzy wpływem mutacji rs6933349 na sekwencję cpg cg21325723 zlokalizowaną w genie TSBP1 i występującą tylko u aktywnie palących a przebiegiem reumatologicznego zapalenia stawów [136]. Podczas karcynogenezy nieprawidłowy wzorzec metylacji, niezależnie od tego, czy jest on spowodowany przez mutagenne czynniki i procesy zewnętrzne ,czy też wewnętrzne to często dotyczy w większości sekwencji CpG promotorów regulujących ekspresję protoonkogenów i antyonkogenów oraz transpozonów [137–139]. Generalnie komórki nowotworowe charakteryzują się redukcją poziomu metylacji w DNA w regionach o niskim zagęszczeniu sekwencji CpG, natomiast wyspy i kaniony CpG mają tendencję do hipermetylacji w procesie nowotworzenia [140–142]. Hipermetylacja obserwowana jest częściej w promotorach i wzmacniaczach ekspresji antyonkogenów, co podkreśla znaczenie epimutacji w procesie onkogenezy [140]. Przykładem jest hipermetylacja DNA wraz z downregulacją antyonkogenów takich jak BRCA1, RAS, czy BCL2 obserwowana w wielu komórkach nowotworowych [143]. Niektóre klasy nowotworu jelita grubego charakteryzują się obecnością demetylacji w obrębie genów, jak CDH3 [144], a inne odwrotnie zwiększoną metylacją, jak np. hipermetylacja promotora ludzkiego genu kodującego białko MLH1, która prowadzi do zmniejszenia jego ekspresji. Obraz drugiego przypadku widoczny jest w podgrupie raków jelita grubego z niestabilnością mikrosatelitarną (ang. microsatellite unstable cancer) [145]. Podobnie hamowanie przez hipermetylację ekpresji białka DAPK1, które bierze udział w procesie apoptozy zaobserwowano w niektórych nowotworach [146–148]. Również mutacje związane z DNMTs prowadzące do demetylacji obserwowane są w procesie nowotworzenia [149]. Zaobserwowana w niektórych nowotworach raka jelita grubego wysoka ekspresja UHRF1, która wpływa na aktywność DNMT1, będącą metylotransferazą utrzymującą metylację i w efekcie hamująca ekspresję antyonkogenów jest przykładem takiej zależności [150]. Zaburzenia imprintingu opisane wcześniej i związane z genomową utratą metylacji są bardzo często

obserwowane w chorobach nowotworowych [151]. Tak jak metylotransferazy odpowiadają za metylację DNA i jej utrzymanie, tak białka TET pełnią odmienną funkcję. Oksydacja 5-metylocytozyny zależna od TET jest wymagana do utrzymania niemetylowanych wysp CpG w zdrowych komórkach. W komórkach nowotworowych zauważane jest upośledzenie tego procesu [152]. Analiza metylomu, może też usprawnić diagnostykę i stanowić podstawę do stworzenia nowych markerów do ich wczesnego wykrywania [7]. Wykonane w ostatnim czasie studium metylomu przez Ambatipudi i wsp. [153] wykazało, że okres menopauzy powoduje akumulację metylacji DNA w konsekwencji zwiększając zapadalność na raka piersi w tym okresie. Ta przykładowa obserwacja mogłaby potencjalnie posłużyć do opracowania nowoczesnego markeru nowotworowego. Zmiany epigenetyczne w przeciwieństwie do genetycznych z reguły są odwracalne co niesie ze sobą możliwość stworzenia nowych opcji terapeutycznych. Przykładem są analogi cytydyny jak decitabina będąca inhibitorem DNMTs przez co może ona indukować hipometylację DNA [154]. Ostatecznie warto zwrócić uwagę na fakt, że występowanie zaburzeń metylacji nie zawsze jest przyczyną danej choroby. Przykładem może być nieprawidłowy wzorzec metylacji widoczny w ostrej białaczce szpikowej, który uważany jest nie za przyczynę tej choroby a efekt nadmiernej proliferacji komórek w trakcie jej progresji [155].

Pomiar metylacji

Pełne zrozumienie znaczenia metylacji DNA m.in. na zdrowie człowieka wymaga wciąż poszerzenia wiedzy na ten temat przez kolejne jej pomiary w obrębie obszernego zakresu genomu i w różnym kontekście badawczym. Badanie asocjacyjne całego epigenomu (ang. epigenome-wide association study, EWAS) będące zestawem znaczników epigenetycznych, które obejmują cały genom np. metylacja DNA i powiązanie ich z określonymi cechami fenotypowymi jest jednym ze źródeł tej wiedzy [156]. Złotym standardem do precyzyjnego mapowania metylowanej cytozyny jest sekwencjonowanie genomu z użyciem wodorosiarczynu (ang. whole genome bisulphite sequencing, WGBS) [157, 158]. Niestety problemem tej metody jest jej koszt oraz wymagany wysoki poziom technicznej ekspertyzy do jej użycia, na który składa się m. in trudność w sekwencjonowaniu niektórych obszarów genomu, złożoność protokołów badania, oraz odpowiedni dobór głębokości sekwencjonowania [158].

Illumina Infinium BeadChips

Do dziś najpopularniejszą, alternatywą dla tej metody, pomimo powstania innych jest platforma Illumina Infinium BeadChips [8]. Sama platforma Illumina Infinium BeadChips przeszła w ostatnim czasie znaczący rozwój od wersji Infinium HumanMethylation27 BeadChip (27K) [159] oceniającej ok. 27 tysięcy sekwencji CpG, czyli ok 0,1% procenta wszystkich obecnych w genomie, przez wersję Infinium HumanMethylation450 array (450K) [160] oceniającą ich ok. 480 tysięcy do współczesnej Infinium MethylationEPIC BeadChip (EPIC) [161], za pomocą której można przeanalizować aż ok. 850 tysięcy CpG, co stanowi już 3% wszystkich tych sekwencji w genomie. Wersja 27K obejmowała głównie sekwencje CpG znajdujące się w proksymalnej części promotorów genów obejmujących consensus coding sequence i dobrze opisane geny nowotworowe [159], z kolei wariant 450K został opracowany po konsultacji z konsorcjum badaczy metylacji DNA i został wzbogacony m.in o geny RefSeq, promotory FANTOM4, regiony MHC i niektóre wzmacniacze transkrypcji [162]. Opracowany w 2015 EPIC BeadChip obejmuje dodatkowo potencjalne wzmacniacze transkrypcji z projektu FANTOM5 [163] i ENCODE [44].

Mikromacierze Illumina – zarys działania

Podobnie do sekwencjonowania wodorosiarczynowego całego genomu (Whole Genome Bisulfite Sequencing, WGBS) technologia Illumina oparta jest na konwersji DNA przy użyciu wodorosiarczynu sodu (ang. sodium bisulphite), ale z następowym genotypowaniem wybranych CpG korzystając z sond mikromacierzy. W końcowym efekcie uzyskujemy dla każdego CpG wyniki pomiaru intensywności sygnału dla allelu metylowanego i niemetylowanego [160]. Dane te zapisywane są w pliku o formacie idat. Celem wyekstrahowania danych plik idat może być poddany przetworzeniu w komercyjnym programie GenomeStudio [164] lub bibliotece minfi [165] z repozytorium Bioconductor. W ten sposób możemy uzyskać wynik poziomu metylacji wyrażony wartością β, która jest stosunkiem intensywności sygnału dla alleli metylowanych (M) i niemetylowanych (U) zgodnie ze wzorem [164] przedstawionym na Rycina *1*

$$\beta = \frac{M}{(M + U + \alpha)}$$

Rycina 1 Wzór wartości β określającej poziom metylacji. M - wartość sygnału dla allelu metylowanego, U-wartość sygnału allelu niemetylowanego α - stała stabilizująca wartość β , z reguły przyjmowana jest wartość 100

Wady i zalety mikromacierzy Illumina

Użycie mikromacierzy Illumina ma trzy istotne zalety: są łatwe w użyciu, oszczędne pod względem czasu i kosztów, a uzyskane wyniki są zgodne z innymi platformami [159]. Do zgłaszanych problemów z zastosowaniem Illumina należą błędy w wyniku zdarzeń związanych z krzyżową hybrydyzacją (ang. cross-hybridization) [166], efektem serii (ang. batch effect) [167], czy efektem pozycyjnym (ang. positional effect) [168], aczkolwiek opracowywane są metody, których celem jest zminimalizowanie tych problemów [169].

Charakterystyka β-value

Wartość β stanowi proporcję pomiędzy badanymi komórkami metylowanymi a niemetylowanymi w danej sekwencji CpG i jest intuicyjna do analizy [170]. Zakres jaki przyjmuje wartość β to od 0 do 1, a jej rozkład w obrębie genomu jest bimodalny [171], natomiast rozkład w kontekście pojedynczej sekwencji CpG jest raczej unimodalny. Dodatkowo wartość β charakteryzuje się heteroskedastycznością [171] (dla zachowania pewnej spójności i przejrzystości rozprawy opis problemu heteroskedastyczności umiejscowiono w części opisującej model liniowej regresji). Wymienione cechy stanowią złamanie założeń wielu klasycznych testów statystycznych sprawiając, że ich prawidłowe działanie jest utrudnione [172]. Niektórzy badacze proponują zastosowanie transformacji logitowej zamieniając wartość β na wartość M, próbując w ten sposób rozwiązać wyżej wymieniony problem [171]. Wzory matematyczne opisujące sposób uzyskania wartości M z użyciem intensywności sygnału dla alleli oraz przez przekształcenie wartości β co opisano na Rycinie 2.

$$M_i = log_2\left(\frac{max(y_{i\sim methy'}, 0) + \alpha}{max(y_{i\sim unmethy'}, 0) + \alpha}\right); M_i = log2\left(\frac{Beta_i}{1 - Betai}\right)$$

Rycina 2 Wzory opisujące wartość M i jej zależność względem wartości β

Otrzymujemy w ten sposób wartość, która nie jest ograniczona przedziałem od 0 do 1 i jest mniej heteroskedastyczna [171], aczkolwiek wśród badaczy metylacji nie ma spójnej opinii na temat zasadności jej stosowania [171–173].

Interpretacja wartości β

Wielu badaczy podczas interpretacji wyników metylacji korzysta z absolutnych definicji klasyfikując sekwencje CpG względem wartości β. Bibikova i wsp. [159] wartość β dla sekwencji hipermetylowanych określa jako >0.75 i <0.2 dla hipometylowanych, jednocześnie wyróżniając grupę sekwencji CpG oscylujących wokół wartości 0.5 [159]. Grupę tę nazywa jako częściowo metylowaną (ang. hemi-methylated). Absolutne definicje metylacji dla przykładu zastosowano również w badaniu przeprowadzonym przez Gueant i wsp. [174], projekcie ENCODE [44, 175], bazie UALCAN [176], czy w badaniach przeprowadzonych przez Oshima i wsp [177] i Laurent i wsp [178]. Inny sposób interpretacji wartości β często spotykany w literaturze [179–182], oraz nazywany odmiennym metylowaniem sond (ang. differentially methylated probes, DMPs) opiera się na zależności względnej pomiędzy próbą badaną a kontrolną, a dokładniej na różnicy pomiędzy średnimi w obu grupach w połączeniu z istotnością statystyczną np. w teście t Studenta. Obszar genomu zawierający większą liczbę DMPs nazywamy odmiennie metylowanym regionem (ang. differentially methylated region, DMR) [183].

Biblioteka ChAMP

Jednym z rekomendowanych i najbardziej popularnych narzędzi do analizy danych metylacji jest biblioteka ChAMP [184] z repozytorium Bioconductor. ChAMP stanowi zespół funkcji, który wykorzystuje inne biblioteki i własne rozwiązania stanowiąc pewnego rodzaju protokół badawczy do analizy metylacji. Biblioteka ta została opracowana dla środowiska statystycznego R [185]. Przez implementacje

wspomnianej wcześniej biblioteki minfi pozwala na przekształcenie surowych danych z pliku idat zawierającego dane z mikromacierzy Illumina Infinium na wartości β, jak również na poszukiwanie DMPs i DMRs [184]. Do poszukiwania DMPs ChAMP wykorzystuje funkcje champ.DMP(), która wykorzystuje pakiet limma. Zgodnie z dokumentacją funkcja ta jako wynik generuje tabele składająca się m.in z sekwencji cpg, wartości p dla testu dopasowania do liniowego modelu i różnicy w średniej metylacji pomiędzy grupami badanymi [186]. Analiza kodu funkcji champ.DMP() wykonaną przez autora rozprawy pokazuje, że wartości β z adnotacją fenotypową "przypadek" i "kontrola" przekazywane są do funkcji „limma_lmfit", która dopasowuje wartości względem kilku modeli liniowych metodą ważonych lub uogólnionych najmniejszych kwadratów. Następnie za pomocą funkcji limmas.contrasts.fit() otrzymywane są współczynniki i błędy standardowe. W etapie końcowym stosowana jest dodatkowo empiryczna metoda Bayesa do szeregowania w kolejności dowodów [187].

Istotność statystyczna w porównaniach wielokrotnych

Testując hipotezę statystyczną niezależnie od testu sprawdzamy, czy hipoteza zerowa, czyli stwierdzenie, że między zmiennymi lub populacjami nie ma różnicy jest prawdziwa, w innym przypadku przyjmujemy hipotezę alternatywną, czyli w praktyce tą, która nas interesuje np. skuteczność nowej metody operacyjnej. Jedną z wartości mogącą pomóc w ocenie stopnia prawdziwości hipotezy alternatywnej jest powszechnie używana wartość p. Dokładniejsza definicja tej wartości mówi o prawdopodobieństwie, że zaobserwowana zależność w losowej próbie z populacji jest przypadkowa [188]. Powszechnie stosowanym progiem dla wartości p jest 0.05, co oznacza, że jeżeli wartość p jest mniejsza niż 0.05, hipoteza zerowa jest odrzucana, a hipoteza alternatywna jest akceptowana [188]. Próg 0.05 jest jednak progiem przyjętym konwencjonalnie i współcześnie uważa się zwłaszcza w badaniach biologicznych, że powinien być on bardziej restrykcyjny np. 0.01 [189], a w badaniach dotyczących metylacji nawet przyjmować wartość $P < 9 \times 10^{-8}$ [190]. Drugim problemem przy stosowaniu testów statystycznych jest problem wielokrotnego testowania. W dużym skrócie oznacza on, że przy wielokrotnym wykonywaniu testu jest duże prawdopodobieństwo, że uda się nam znaleźć jakąś różnicę. Jeżeli jakieś zjawisko występuje z częstość 1/1000 to wykonując 1000 testów na jego obecność jesteśmy w

stanie zaobserwować je przynajmniej raz. Jednym ze sposobów rozwiązania problemu związanego z wielokrotnym testowaniem jest przyjęcie poprawki Bonferroniego. Metoda ta polega na zastosowaniu skorygowanego poziomu istotności statystycznej dla każdego testu biorąc pod uwagę liczbę wykonanych testów. Nowa granica istotności jest ilorazem zakładanego poziomu i ilości wykonanych testów [188]. Innym sposobem na przeciwdziałanie problemu wielokrotnego testowania jest użycie wskaźnika fałszywego wykrywania (ang. false discovery rate , FDR), który jest oczekiwanym odsetkiem wyników fałszywie dodatnich w zbiorze wszystkich odrzuconych hipotez [189]. Wspomniana wyżej biblioteka ChAMP wykorzystuje metodę Bonferroniego [186].

Uczenie maszynowe

W dzisiejszych czasach uczenie maszynowe jest przedmiotem szerokiego i intensywnego rozwoju, oraz jest nieświadomie używane przez nas w codziennym życiu np. jako algorytmy do klasyfikacji obrazu [191], rekomendacji filmów [192], czy filtrowania niechcianych e-maili [193]. Uczenie maszynowe to gałąź tak zwanej sztucznej inteligencji, czyli dziedziny nauki o algorytmach, które poddają się automatycznemu samodoskonaleniu przez doświadczenie lub użycie odpowiednich danych. Ich istotą jest fakt, że do podjęcia decyzji lub predykcji nie wymagają one czasochłonnego i żmudnego programowania każdej możliwej sytuacji, a opierają się na próbkach danych, które nazywamy danymi treningowymi. Historia uczenia maszynowego sięga lat 50 XX wieku, kiedy to Alan Turing w swojej pracy „Computing Machinery and Intelligence" [194] opracował koncepcję inteligentnej maszyny oraz opracował test, który pozwoliłby na rozróżnienie maszyny od człowieka. Za kolejny krok milowy można by uznać pracę opublikowaną przez Arthur Samuel, w której to autor opracował program uznawany za pierwszy program, który zdolny był do samodzielnego uczenia się gry w warcaby [195]. W 1957 Rosenblatt stworzył model perceptronu i jednocześnie zbudował w ten sposób fundamenty dla rozwoju sieci neuronowych [196]. W latach 60-tych XX wieku E. W. Forgy i Stuart Loyd niezależnie od siebie opracowali podstawy algorytmu klasteryzacji opartego o średnie k i nazywanego z czasem algorytmem Lloyda-Forgyego. Hopfield w latach 80 dał początek sieciom rekurencyjnym prezentując swój model architektury nazywany dziś siecią

asocjacyjną Hopfielda [197]. W połowie lat 80-tych Rumelhart i Hinton opracowali algorytm propagacji wstecznej, będący do dziś podstawą uczenia sieci neuronowych [198]. W tym samym czasie Yann LeCun opracował pierwszą sieć konwolucyjną [199]. Współcześnie wraz ze znacznym wzrostem możliwości obliczeniowej komputerów i niespotykanym wcześniej tempem rozwoju algorytmów związanych z sieciami neuronowymi i sztuczną inteligencją wydaje się, że można mówić, jak to ujął Klaus Schwab na Światowym Ekonomicznym Forum o początku czwartej rewolucji przemysłowej.

Metody uczenia maszynowego

Metody uczenia maszynowego możemy podzielić na [200, 201]:

1. uczenie nadzorowane (ang. supervised learning)
2. uczenie nienadzorowane (ang. unsupervised learning)
3. uczenie częściowo nadzorowane (ang. semi-supervised learning)
4. uczenie przez wzmacnianie (ang. reinforcement learning)

Uczenie maszynowe nadzorowane

Ten sposób polega na użyciu danych treningowych wraz z prawidłowymi odpowiedziami tzw. etykietami (ang. labels). Celem procesu nauki/treningu jest znalezienie przez algorytm charakterystycznych cech tego zbioru i dostosowanie swoich zmiennych tak, by w następnym kroku po ekspozycji na nowe dane, do których algorytm nie miał wcześniej dostępu przewidzieć prawidłową odpowiedź, czy wynik [202]. Sam proces uczenia nadzorowanego można z kolei podzielić w zależności od jednego z dwóch celów. Pierwszym jest klasyfikacja, czyli atrybutem decyzyjnym jest wartość binarna lub nominalna np. rozpoznawany obiekt to kot albo pies. Drugim celem jest regresja, gdzie do czynienia mamy z wynikiem podanym jako liczba rzeczywista np. do predykcji ceny rynkowej [203].

Uczenie maszynowe nienadzorowane

Uczenie nienadzorowane polega na wykorzystaniu danych treningowych niezawierających prawidłowej odpowiedzi. Z dużej puli atrybutów i zmiennych algorytm ma prawidłowo zidentyfikować pewien trend w danych, a jego elementy

podzielić/rozpoznać i podzielić na podzbiory. Przykładem takiego uczenia maszynowego jest klasteryzacja (inaczej grupowanie, analiza skupień) (ang. data clustering) [202].

Uczenie maszynowe częściowo nadzorowane

Ten rodzaj uczenia maszynowego łączy elementy dwóch poprzednich. Zbiór treningowy zawiera zarówno elementy z etykietami (prawidłowymi odpowiedziami) oraz bez nich. Zadaniem algorytmu jest zidentyfikowanie cech charakterystycznych, pogrupowanie elementów i zaopatrzenie nieoznaczonych elementów danych w odpowiednie etykiety. Korzyścią z zastosowania takiego mechanizmu jest możliwość pominięcia przypisywania etykiet w sposób ręczny, co jest znaczącym ułatwieniem pod względem kosztów i czasu [204].

Uczenie maszynowe przez wzmacnianie

Trening w tym modelu uczenia maszynowego w przeciwieństwie do poprzednich modeli nie opiera się na zestawie danych, a na interakcji ze środowiskiem, gdzie mamy niezliczoną liczbę możliwych kombinacji. Taki typ uczenia maszynowego znajduje zastosowanie w automatyce samochodowej, robotyce, czy w grach [205], a w ostatnim czasie również stosowany był w przetwarzaniu języka naturalnego (ang. natural language processing) [206]. Ten typ uczenia maszynowego składa się z polityki (ang. policy), sygnału nagrody (ang. reward signal) i funkcji wartości (ang. value function). Agent, czyli w tym przypadku algorytm reaguje ze środowiskiem zgodnie z ustaloną polityką interakcji i w zależności od podjętych akcji otrzymuje sygnał nagrody. Funkcja wartości z kolei w przeciwieństwie do sygnału nagrody nie określa natychmiastową ocenę akcji wykonanej przez algorytm, ale w kontekście długoterminowym [205].

Analiza regresji

Analiza regresji to metoda statystyczna i uczenia maszynowego nadzorowanego, która pozwala stworzyć model na podstawie oznaczonego zestawu danych i w ten sposób przewidzieć interakcję pomiędzy zmiennymi. W analizie regresji możemy wydzielić:

1. regresję liniową – liniowa zależność pomiędzy zmiennymi
2. regresję logistyczną – zmienna zależna przyjmuje jedną z dwóch wartości (skala dychotomiczna)

Istotną metodą regresji stosowaną w uczeniu maszynowym jest metoda gradientu prostego, opisana w dalszych częściach tej rozprawy [207].

Klasteryzacja (grupowanie)

Klasteryzacją określamy metodę nauczania maszynowego nienadzorowanego, która polega na grupowania elementów, które pod względem cech są zbliżone do siebie. Przykładem takich metod jest:

1. algorytm k-średnich (inaczej algorytm centroidów) (ang. k-means)
2. klasteryzacja hierarchiczna (ang. hierarchical clustering)

Algorytm k-średnich

Algorytm k średnich jest popularnym algorytmem klasteryzacji. Działanie algorytmu opiera się na znalezieniu takiego podziału, aby dystans pomiędzy każdym z punktów a powiązanym centroidem był jak najmniejszy. Proces rozpoczyna się od losowego wybrania K centroidów, dopasowania najbliższych punktów danych i wstępnego utworzenia klastrów. W kolejnych krokach algorytm poprawia położenie centroidów na podstawie średnich dystansów do punktów w danym klastrze. Proces ten jest powtarzany aż do konwergencji [208].

Klasteryzacja hierarchiczna

Klasteryzacja hierarchiczna to typ nienadzorowanego uczenia maszynowego używany do grupowania rekordów/próbek w klastry w zależności od dystansu pomiędzy parami. Człon hierarchiczny w nazwie tego typu klasteryzacji oznacza, że wynikiem użycia metody jest hierarchia klastrów w postaci dendrogramu, gdzie mniejsze klastry należą do większych. Klasycznie metody klasteryzacji można podzielić na aglomeracyjne i deglomeracyjne (różnicujące). Metody aglomeracyjne traktują poszczególne elementy danych jako oddzielne klastry, które w kolejnych etapach działania algorytmu łączone są w większe klastry, aż do momentu, w którym powstanie

jeden globalny klaster. Metody deglomeracyjne są przeciwieństwem metod aglomeracyjnych. W tym przypadku wszystkie próbki traktowane są jako jeden klaster, a następnie dzielone są na mniejsze. Jedną z metod jest metoda Warda, której cechą charakterystyczną jest dążenie do minimalizacji wariancji dystansów pomiędzy nowo utworzonymi klastrami [209]. Metoda Warda jest powszechnie stosowaną metodą i jednym z jej istotnych zalet jest balans pomiędzy liczbą klastrów a ich podobieństwem, jednakże jest dość wymagająca pod względem użycia mocy obliczeniowej komputera [210].

Redukcja wymiaru (ang. dimensionality reduction)

Redukcja wymiaru to technika używana w uczeniu maszynowym, której celem jest redukcja puli cech (zmiennych), a więc wymiarów w danym zbiorze danych. Celem tego procesu jest uproszczenie zbioru danych przy zachowaniu jak największej ilości informacji w nim zawartych np. przez usunięcie danych zbędnych, nieróżnicujących lub tzw. szumu. Może to pozwolić dodatkowo na lepszą wizualizację danych np. w formie wektorów lub graficznej, dodatkowo mniejsza liczba zmiennych ułatwia funkcjonowanie wielu algorytmów. Jedną z metod należących do tej grupy jest analiza głównych składowych (ang. principal component analysis, PCA). Narzędzie to jest typem liniowej transformacji, którego działanie polega na przekształceniu danych w nowy układ współrzędnych, gdzie największe znaczenie mają zmienne o największej wariancji. Dzięki temu uzyskiwane są dane o jak największej reprezentatywności i jak najmniejszej liczby wymiarów (zmiennych). Inne powszechnie stosowane metody redukcji wymiarowości to [211]:

1. Stochastyczna metoda porządkowania sąsiadów w oparciu o rozkład t (ang. t-distributed stochastic neighbour embedding, t-SNE)
2. Rozkład według wartości szczególnych (ang. Singular Value Decomposition, SVD)
3. Autoenkoder (ang. autoencoder, AE)

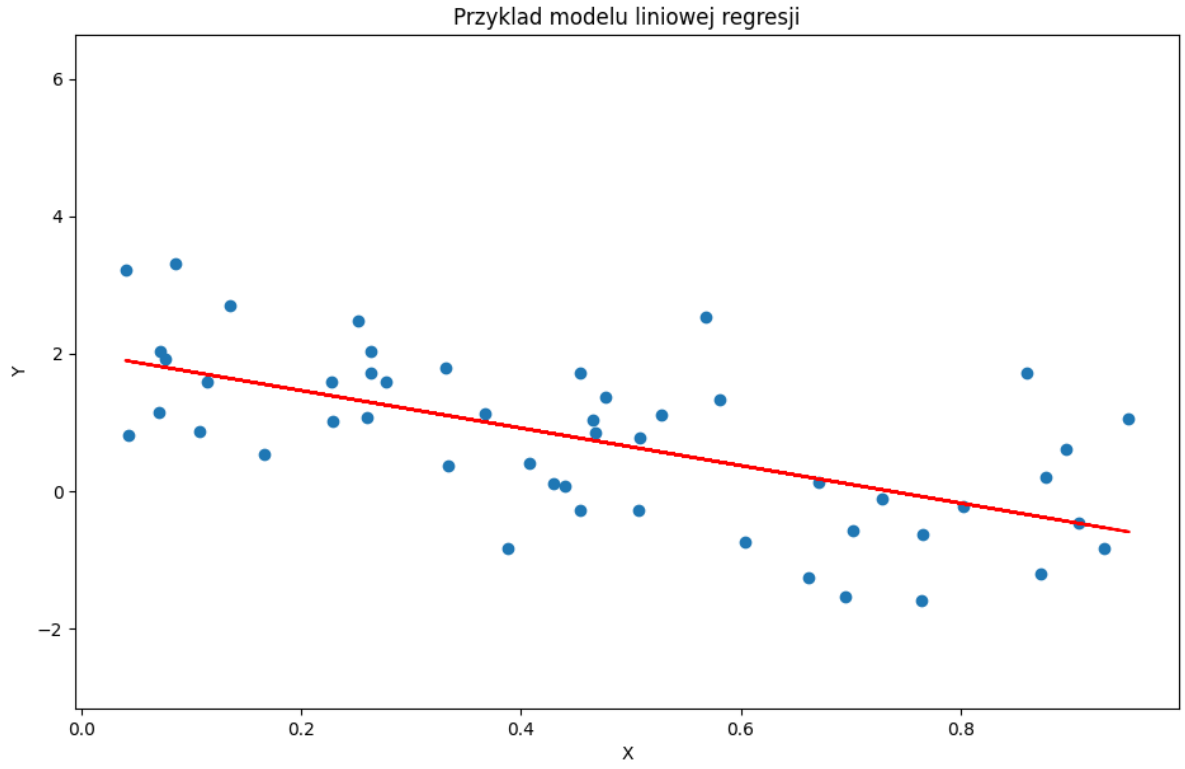
Model uczenia maszynowego oparty o regresję liniową

Model uczenia maszynowego oparty o regresję liniową jest modelem uczenia nadzorowanego. Prosty model regresji liniowej dotyczy pojedynczej zmiennej

niezależnej i jej stosunku do zmiennej zależnej (w tym kontekście przewidywanej) [212] zgodnie ze wzorem przedstawionym na Rycina 3.

$$y = \beta + \beta_1 x$$

Rycina 3 Wzór uproszczonego modelu regresji liniowej - gdzie y jest zmienna zależną, a x reprezentuje zmienna niezależną. β jest punktem przecięcia z osią y, β_1 jest stopniem nachylenia ciągłej.



Rycina 4 Przykład modelu liniowej regresji

Podstawą takiego modelu jest znalezienie ciągłej obejmującej punkty danych dla zmiennej zależnej, których średnia różnica z prawdziwymi wynikami jest najmniejsza i w ten sposób uzyskanie zdolności do predykcji innych wartości(Rycina 4).

Dla wielokrotnej liniowej regresji wzór ten wygląda następująco (Rycina 5):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

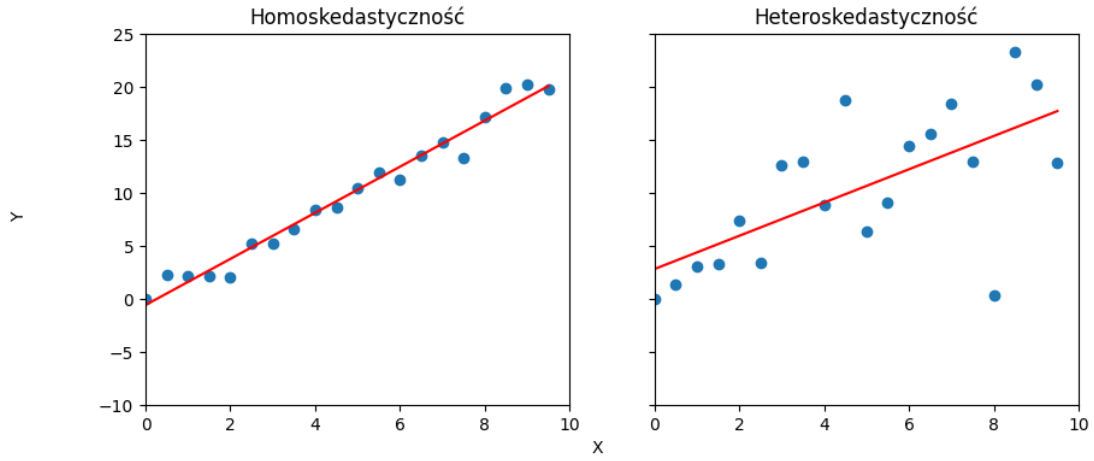
Rycina 5 Wzór dla wielokrotnej liniowej regresji. β, -parametry\wagi, x1...xn- dane wejściowe, y-wynik

Do prawidłowego funkcjonowania takiego modelu wymagane jest spełnienie przez dane pewnych założeń. [212] Są to m.in:

1. **liniowość** - zależność między zmienną zależną a niezależną jest liniowa
2. **niezależność** - obserwacje w zbiorze są niezależne od siebie
3. **homoskedastyczność** - wariancja błędu jest stała dla każdej zmiennej niezależnej
4. **normalność** - rozkład błędu jest normalny

Heteroskedastyczność

Klasyczny model liniowy zakłada homoskedastyczność wariancji błędów dla zmiennej zależnej, a gdy warunek ten nie jest spełniony to mówimy o heteroskedastyczności. Mówiąc prościej pojawiają się obserwacje, które odbiegają od przewidywanych, ale są zależne (Rycina 6). Powodem pojawienia się heteroskedastyczności mogą być wartości odstające (ang. outliers) czy błąd pomiaru. Zjawisko to nie jest jednoznaczne z brakiem zależności liniowej, jak również heteroskedastyczność może być także obserwowana w modelach nieliniowych. Zjawisko to może prowadzić do nieprawidłowych wyników w analizach regresji [213].



Rycina 6 Przykład homoskedastyczność i heteroskedastyczność

Praktycznym przykładem problemu może być zależność pomiędzy produktywnością i wykształceniem pracownika, a jego pensją. Na hipotetycznym rynku pracy można zaobserwować, że pracownik o niewielkiej produktywności i wykształceniu zarabia mniej, aczkolwiek w tej samej dziedzinie, ale w różnych miejscach pracy osoba o dużej produktywności i wykształceniu może mieć zarobki większe lub podobne do pracownika mniej efektywnego. Podobnym przykładem może być wielkość dochodu, a

poziom wydatków, który niejednokrotnie jest podobny u osób mniej majętnych, aczkolwiek wraz ze wzrostem przychodów nie zawsze obserwowany jest wzrost wydatków. Problem heteroskedastyczności dotyczy również pomiaru metylacji DNA opisywanego wartością β, o czym wspomniano we wcześniejszym miejscu tej rozprawy. [171].

Model uczenia maszynowego oparty o regresję logistyczną

Model uczenia maszynowego oparty o logistyczną regresję jest kolejnym modelem uczenia nadzorowanego, aczkolwiek znajduje ona zastosowanie w problemach klasyfikacji, niż regresji, ze względu na fakt, iż przewidywany wynik dla takiego modelu to wartość od zera do jeden. Podobnie jak w modelu opartym o liniową regresję znaczenie ma relacja pomiędzy zmienną zależną a niezależną, z tą różnicą, że dla każdej zmiennej niezależnej będącej liczbą rzeczywistą przyjmowana jest wartość od 0 do 1, co interpretować można jako prawdopodobieństwo przynależności do danej kategorii. Wzór funkcji logistycznej przedstawiony jest na Rycinie 7.

$$p = \frac{1}{1 + e^{-z}}$$

Rycina 7 Wzór funkcji logistycznej, gdzie p to prawdopodobieństwo dla danej zmiennej na przynależność do kategorii/klasy, e to podstawa z logarytmu naturalnego, z liniowe zestawienie zmiennych niezależnych i współczynników

Perceptron

Perceptron to model liniowego klasyfikatora, który został opracowany przez F. Rosenblatt w 1958 roku [196]. Perceptron to prosty, iteratywny model będący podstawą współczesnych sieci neuronowych. Jego działanie polega na przydzieleniu danej zmiennej niezależnej jednej z klas, co jest cechą wspólną z modelami opartymi o regresję logistyczną, jednakże perceptron do obliczenia używa funkcji schodkowej. Przez dodanie tej funkcji aktywacyjnej wprowadzana jest nieliniowość. Wzory opisujące perceptron znajdują się na Rycina 8

$$y = f(w \cdot x + b)$$

$$f(x) = \begin{cases} 1, & jeżeli\ (\boldsymbol{w} \cdot \boldsymbol{x} + b) > 0 \\ 0, & w\ innym\ przypadku \end{cases}$$

Rycina 8 Wzory opisujące perceptron.

Sztuczne sieci neuronowe (ang. artificial neural networks)

Sztuczne sieci neuronowe będące częścią uczenia maszynowego zainspirowane są ich biologicznymi odpowiednikami. Biologiczny neuron uzyskuje sygnał z sąsiednich neuronów, a następnie będąc pod wpływem odebranych impulsów przekazuje je odpowiednio dalej. Wraz ze złożonością takich połączeń rosną możliwości takiej sieci, co każdy może osobiście doświadczyć na samym sobie. Celem powstania sztucznych sieci neuronowych, było stworzenie matematycznego modelu ich biologicznego odpowiednika jako droga do uzyskania sztucznej inteligencji [214]. W typowej sieci neuronowej sygnał wyjściowy z jednego neuronu do drugiego stanowi zazwyczaj liczba rzeczywista, która jest wynikiem nieliniowej funkcji aktywacji, będącej funkcją sumy sygnałów wejściowych. Każdy neuron i połączenie między nimi posiada odpowiedni parametr (ang. weight/ parameter), które modulują sygnał wyjściowy i podlegają optymalizacji podczas procesu uczenia sieci neuronowej [215]. Również typowo, neurony w sztucznej sieci neuronowej agregowane są w warstwy, z których pierwsza jest warstwą wejścia, a ostatnia wyjścia. Warstwy środkowe nazywamy warstwami ukrytymi. Możliwe jest również by sygnał przez poszczególne węzły przechodził kilkukrotnie. Dokładny opis budowy podstawowej sieci neuronowej znajduje się w dalszej części rozprawy. Wraz ze zwiększeniem liczby warstw ukrytych możemy mówić o tzw. głębokim nauczaniu (ang. deep learning) [215].

Sieci neuronowe jako uniwersalny aproksymator

Sieci neuronowe wykazują możliwość tzw. uniwersalnej aproksymacji, co oznacza, że mają zdolność do optymalnego oszacowania wyniku każdej funkcji liniowej i nieliniowej. Zdolność ta wykazywana jest niezależnie od architektury sieci neuronowej [216, 217]. W praktyce widoczne jest to w ilości możliwych zastosowań

sieci neuronowych, gdzie opracowanie funkcji, czy klasycznego algorytmu komputerowego byłoby trudne i nie dające satysfakcjonującego efektu, jak np. rozpoznawanie twarzy, czy identyfikacja obiektów. Głębokie uczenie jest bardzo wszechstronne i może być zastosowane zarówno do uczenia nadzorowanego, nienadzorowanego, częściowo nadzorowanego i nauczania przez wzmocnienie [218, 219]. Dodatkowo wynikiem działania głębokiego uczenia jest nie tylko odpowiedź na zadany problem (predykcja), ale także miara poziomu ufności (ang. confidence level), co ma niebagatelne znaczenie zwłaszcza w kontekście np. medycznej diagnozy.

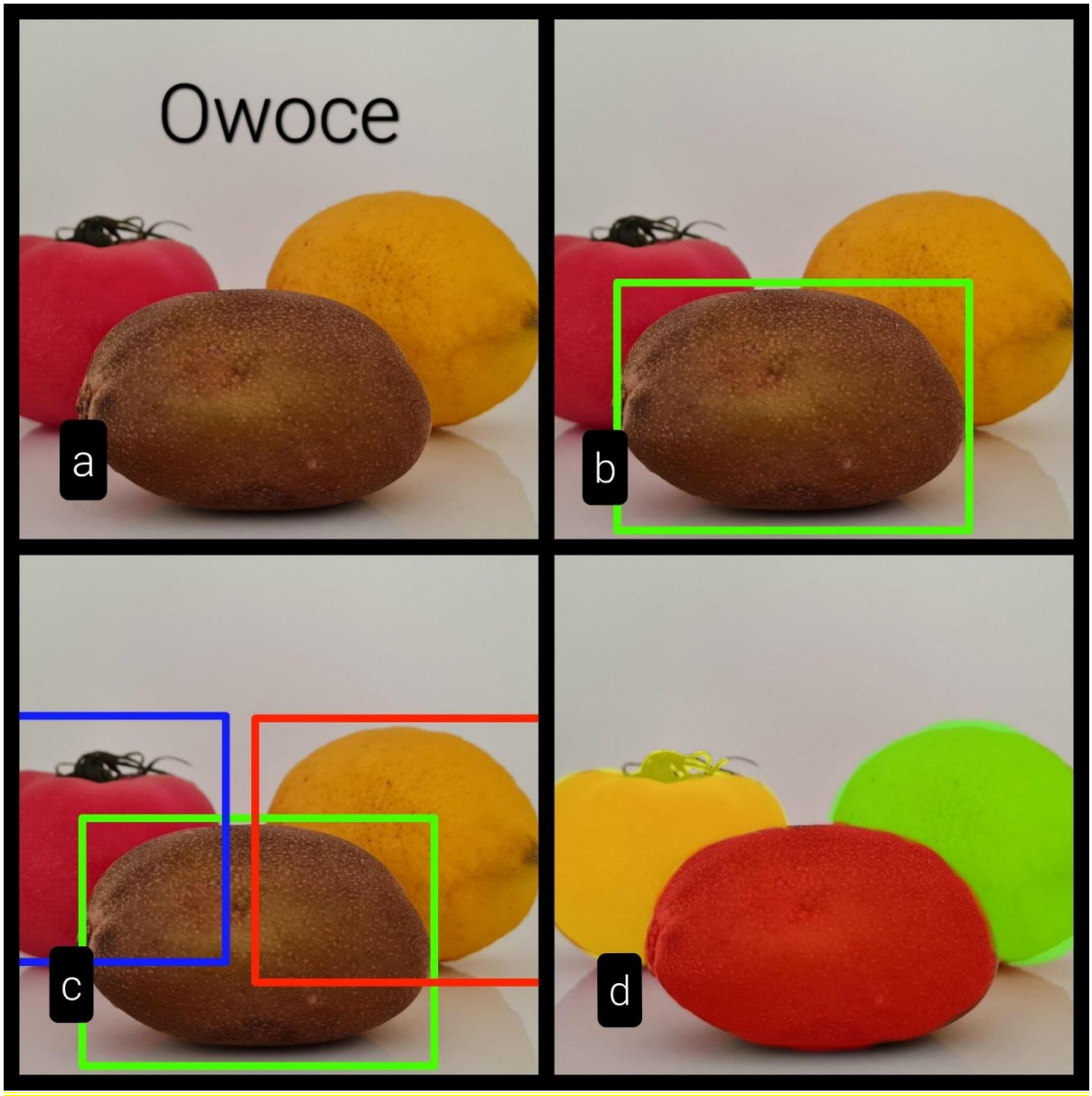
Zastosowanie sieci neuronowych i głębokiego uczenia w medycynie

Jak już wspomniano ostatnie lata charakteryzują się znaczącym rozwojem uczenia maszynowego i sieci neuronowych oraz ich rosnącym zastosowaniem w coraz to bardziej złożonych problemach, jak rozpoznawanie i klasyfikacja obrazów [220], rozpoznawanie i synteza ludzkiej mowy [221], czy system sztucznej inteligencji, który pokonał 5 do 0 europejskiego mistrza świata w grę Go [222], która jest uważana za bardziej złożoną niż gra w szachy [223].W medycynie z kolei dominują zastosowania w obrębie analizy i rozpoznawania obrazów (Rycina 9) i można je podzielić na:

1. klasyfikację
2. lokalizację
3. detekcję
4. segmentację

Chociaż te pojęcia wydają się mieć zbliżone znaczenie i używane są często naprzemiennie to jednak posiadają one różne znaczenie w tym kontekście. Za klasyfikację rozumiemy ocenę przynależności np. obrazu do jednej z kategorii, a lokalizację definiujemy jako rozpoznanie i wskazanie położenia szukanego obiektu. Detekcja pozwala na rozpoznanie i wskazanie lokalizacji kilku obiektów. Segmentacja to z kolei możliwość rozpoznania i zaznaczenia powierzchni zajmowanej przez obiekt na zdjęciu lub jego zarys [224]. W dalszej części rozprawy autor przedstawi wybrane osiągnięcia w tej dziedzinie, aby pokazać jakie potencjalne korzyści niesie ze sobą

rozwój uczenia maszynowego i sztucznej inteligencji w medycynie ogólnie, oraz w analizie metylacji.



Rycina 9 Analiza obrazu z użyciem sieci neuronowych a- klasyfikacja b- lokalizacja c-detekcja d-segmentacja

Klasyfikacja

Powstało dotychczas kilkadziesiąt sieci neuronowych do klasyfikacji zdjęć rentgenowskich (rtg). Aktualnym przykładem może być ta opracowana przez Rajpurkar i wsp. [225]. Badacze Ci wykorzystali sieć neuronową DenseNet i przetrenowali z użyciem 112 tysięcy obrazów rtg klatki piersiowej celem wykrycia 14 różnych chorób,

a jej zdolność rozpoznawania była podobna do zdolności wyszkolonych radiologów. Powstały również sieci neuronowe dające możliwość wykrywania choroby Alzheimera na podstawie radiologicznych obrazów mózgu [226, 227]. Trwają prace nad algorytmem umożliwiającym klasyfikację różnych klas nowotworów skóry [228], czy klasyfikacji biopsji barwionej hematoksyliną i eozyną pod względem występowania typów raka piersi [229].

Lokalizacja

Opracowano metodę polegającą na wykorzystaniu architektury ConvNet [230] do analizowania obrazów pozyskanych z tomografii komputerowej do lokalizacji anatomii [231]. Z kolei Zhao i wsp. [232] zaprojektowali sieć neuronową pomagającą w wykrywaniu guzów trzustki na przeglądowym zdjęciu rtg jamy brzusznej.

Detekcja

W dziedzinie rozpoznawania nowotworów skóry również poczyniono postępy. Powstały algorytmy do ich detekcji m.in na podstawie obrazów pozyskanych przy pomocy dermoskopu [233, 234]. Chouchan i wsp. [235] przetestowali w kontekście detekcji 5 innych architektur sieci neuronowych AlexNet [220], DenseNet121 [236], ResNet18 [237], InceptionV3 [238] i GoogLeNet [239]. Wszystkie te metody były wcześniej przetrenowane w innym celu, a następnie przez badaczy zoptymalizowane z użyciem transferu wiedzy [ang. transfer learning] do detekcji zapalenia płuc. W ten sposób Chouchan i wsp wskazali, najlepszą z architektur, jaką jest ResNet18 i udowodnili również, że metoda transferu wiedzy jest skuteczna w detekcji na podstawie zdjęć rentgenowskich [235].

Detekcja w obrazach histopatologicznych

Patomorfolog oceniając obrazy histopatologiczne w poszukiwaniu diagnozy ma wiele cech pojedynczej komórki do przeanalizowania od stopnia anaplazji, kształtu jąder komórkowych, czy zmian architektonicznych tkanki po cechy naciekania [240, 241]. Jest to proces żmudny i czasochłonny. Powstało wiele modeli sieci neuronowych, których celem jest pomoc w tym procesie. Ciresan i wsp. [242] opracowali postępowanie z użyciem sieci neuronowych do oceny liczby mitotycznej [243], dzięki

któremu osiągnięto najlepsze wyniki w detekcji w zawodach International Pattern Recognition Conference (ICPR) 2012 Mitosis Detection Competition. Wynik ten poprawili z użyciem tych samych danych Lei i wsp. [244]. Natomiast Sirinukunwattana i wsp. opracowali algorytm do detekcji i klasyfikacji jąder komórkowych w obrazach histopatologicznych raków jelita grubego [245], a Celik i wsp. [246] do wykrywania inwazyjnego raka przewodowego sutka.

Segmentacja

Jak wspomniano wyżej segmentacja to możliwość rozpoznania i zaznaczenia powierzchni lub zarysu obiektu. W tej metodzie wydaje się, że największy rozwój dotyczy algorytmów przydatnych w neurologii, neurochirurgii i radioterapii. Konwencjonalna metoda polega na ręcznym zaznaczaniu przez radiologa poszczególnych elementów natomiast segmentacja pozwala na automatyzacje oznaczenie poszczególnych tkanek czy zmian patologicznych celem dalszej obróbki, oceny lub przygotowania okołozabiegowego. Analizę problemu i wielu rozwiązań przeprowadzili Thaha i wsp. w swojej pracy przeglądowej [247]. Należące do guzów mózgu glejaki to agresywne nowotwory, dla których odpowiedni plan leczenia jest podstawą skuteczności terapeutycznej. Rezonans magnetyczny stanowi podstawę diagnostyki. Pereira i wsp. [248] opracowali sztuczną sieć neuronowa opartą o sieci konwolucyjne i odpowiedni rozmiar jąder konwolucji (ang. kernel) osiągając efekt pozwalający na zdobycie czołowych pozycji w Brain Tumor Segmentation Challenge 2013 (BRATS 2013) i 2015 (BRATS 2015). Xu i wsp. [249] przy użyciu wielokaskadowej konwolucyjnej sieci neuronowej (ang. multi-cascaded convolutional neural network) uzyskali uogólnioną segmentację, a następnie zwiększyli jej dokładność przy pomocy techniki fully connected conditional random fields (CRFs), by w kolejnym kroku połączyć obrazy z 3 różnych płaszczyzn (osiowej, poprzecznej i strzałkowej) finalnie uzyskując bardzo dobry wynik.

Uczenie maszynowe w epigenetyce i multiomice

Szacowanie poziomów metylacji (ang. Computational estimation of methylation levels)

Metody pomiaru metylacji dla całego genomu są drogie i często pracochłonne, oraz obarczone wieloma problemami, jak w przypadku będącego złotym standardem sekwencjonowanie genomu z użyciem wodorosiarczynu. Z kolei użycie mikromacierzy ograniczone jest do spektrum analizowanych przez nie sekwencji. Rozwiązania dla tych trudności szuka się w szacowanym obliczaniu poziomu metylacji (ang. computational estimation of methylation). Wyróżnić możemy trzy podstawowe typy rozwijanych metod - predykcja, imputacja i wzbogacania/rozszerzania (ang. expanding). Predykcja polega na przewidzeniu stanu metylacji w obrębie sekwencji CpG na podstawie innych danych np. kontekstu genetycznego. Imputacja polega na uzupełnianiu brakujących pojedynczych danych, a wzbogacanie to metoda, której celem jest poszerzenie zakresu informacji o metylacji na podstawie danych o mniejszym zakresie. Przykładem algorytmu do predykcji jest methCGI [250], który za pomocą maszyn wektorów nośnych (ang. support vector machine, SVM) i danych na temat sekwencji nukleotydów i miejsc wiązania czynników transkrypcji estymował stan metylacji wysp CpG w próbkach z tkanek ludzkiego mózgu. Kolejnym przykładem może być model MRCN oparty o sieci konwolucyjne [251]. Narzędzia do imputacji opracowali Kapourani i Sanguinetti [252], oraz Soueza [253]. Pierwsza metoda nazwana została Mellisa i oparta jest o bayesowski model hierarchiczny, a druga Epiclomal i bazuje na hierarchicznym modelu mieszanym (ang. hierarchical mixture model). Obie metody wykorzystują dane na temat innych komórek oraz sąsiadujących loci. Fan i wsp. z kolei opracowali metodę wzbogacania pozwalającą poszerzyć informację o metylacji sekwencji CpG z mikromacierzy 450K. Przez wprowadzenie do modelu oprócz danych z mikromacierzy także m.in danych na temat flankujących regionów DNA udało im się uzyskać informacje o nawet 18-krotnie większym zbiorze sekwencji CpG [254].

Klasyfikacja pełnego metylomu i danych multiomicznych

Podstawową metodą klasyfikacji metylacji DNA jest klasteryzacja hierarchiczna, pozwalająca za pomocą dendrogramu zaobserwować zależności i różnice pomiędzy grupą kontrolną a badaną [255]. Rozwinięciem tej metody jest recursive-partitioning mixture model (RPMM). Algorytm ten dzieli zbiór danych na mniejsze grupy na podstawie podobnych cech, a następnie tworzy model dla każdej podgrupy celem wyjaśnienia rozkładu danych w tej podgrupie. Niektóre analizy [256] sugerują, że ta metoda może prawidłowo rozróżnić i dostarczyć istotnych informacji przy analizie

metylacji DNA np. nowotworów. Amor i wsp. [257] użyli klastrowania k-średnich oraz samo organizujących się map (ang. self-organizing map, SOM) i mieszanej metody Gaussa, a także AE do klasyfikacji raków piersi. Si i wsp. użyli kilku warstw maszyn Boltzmanna do redukcji wymiarowości danych i wykrycia cech charakterystycznych w danych metylacji DNA w efekcie uzyskując klasyfikator do analizy nowotworów piersi [258]. Wiele prac naukowych wskazuje, że do zadań typu klasyfikacja chorób, czy ich podtypów lepiej sprawdzają się metody typu głębokiego uczenia, niż typowe modele uczenia maszynowego [259, 260]. Model MetaCancer wykorzystuje konwolucyjne wariacyjne autoenkodery do analizy wieloomicznych danych w tym metylacji celem oceny ryzyka wystąpienia przerzutów [261]. Z kolei Xia i wsp. zaproponowali model również oparty o sieć konwolucyjną do prognozowania występowania gruczolakoraków płuc, raków wątroby i raka jasno-komórkowego nerki na podstawie danych o metylacji [262]. Zespół Zhang'a. opracował algorytm do detekcji pacjentów ze schizofrenią z użyciem modelu głębokiego uczenia opartego na uwadze. Dane na temat metylacji poddawane są kilku etapowemu przetwarzaniu. W pierwszej kolejności przez sieć neuronową opartą o uwagę, następnie przez AE i ostatecznie przez SVM. Pierwszy etap ma za zadanie wskazać cechy charakterystyczne, drugi redukcję wymiarów, a ostatni pełni ostateczną funkcję klasyfikatora [263]. MethylNet z kolei jest siecią neuronową wykorzystującą metodę Shapley Additive ExPlanation (SHAP). Autorzy starali się stworzyć ławy do użycia algorytm pozwalający na wykonywanie zadań typu nienadzorowana klasteryzacja, dekonwolucja typów komórek, klasyfikacja podtypów nowotworów, czy regresja związana z wiekiem [264]. Titus i wsp. oraz Wang i wsp. opracowali metody klasyfikacji raków piersi i płuca [38,39] oparte o wariacyjne autoenkodery (ang. variational autoencoder, VAE) i redukcję wymiarowości z użycie t-SNE oraz klasyfikatora opartego o logistyczną regresję [265, 266]. Jedną ze ścieżek rozwoju jest próba stworzenia modeli do analizy bardziej złożonej np. łączącej obraz histopatologiczny z danymi genetycznym. Przykładem takiego modelu jest PAGE-net, który łączy analizę obrazów histopatologicznych z użyciem sieci konwolucyjnych oraz analizę danych genetycznych z użyciem zaadaptowanej do tego sieci neuronowej CoX-PASNet [267]. Aplikacja PathMe z kolei łączy użycie AE do wyekstrahowania istotnych cech z ekspresji genów, miRNA, zmienności liczby kopii DNA oraz metylacji DNA z klasyfikacją z użyciem SHAP (podobnie jak MethylNet) do analizy współczynnika przeżywalności [268]. Wyżej wymienione metody oparte są głównie na

głębokim nauczaniu. Należy jednak zwrócić uwagę również na te, które do swojego działania używają innych metod jak PCA [269].

Analiza różnic w metylacji

Park i wsp. opracowali instrument analityczny o nazwie methylSIG do badania różnic pomiędzy grupami kontrolnymi i badanymi w oparciu o rozkład β-binominalny. Narzędzie służy do analizy WGBS lub jego zredukowanej reprezentacji [270]. Z kolei Wang i wsp. stworzyli statystyczne narzędzie do analizy DMRs oparte o kombinacje danych z MeDIP-seq (methylated DNA immunoprecipitation followed by sequencing) i (MRE-seq) methylation-sensitive restriction enzyme sequencing o nazwie M&M. Autorzy dodatkowo zaobserwowali, że różnice w metylacji w obrębie promotorów są odmienne i charakterystyczne dla poszczególnych tkanek [271]. Su i wsp. stworzyli narzędzie CpG_MPs do analizy danych dotyczących metylacji DNA pozyskanych za pomocą sekwencjonowania z wodorosiarczynem (ang. Bisulfite sequencing). Podstawą działania tego narzędzia jest użycie entropii Shannona i odpowiedni sposób zdefiniowania interesujących regionów [272]. Entropie Shannona wykorzystuje również algorytm stworzony przez Zhang'a i wsp. o nazwie QDMRs. Aczkolwiek ten algorytm został przystosowany do znajdowania DMRs [273]. DMRMark to z kolei narzędzie stworzone przez Shen i wsp., które oparte jest na modelu Markova i mieszanym modelu Gaussa. Model ten nie wymaga określania ścisłych granic, a DMRs mogą być wykryte nawet dla pojedynczej par próbek [274]. Pomimo rozwoju wielu nowych metod analizy dotychczas powszechnie używane są metody oparte o standardowe metody statystyczne tak jakie modele liniowej regresji zastosowane w bibliotece ChAMP [187].

Dodatkowe zastosowania sieci neuronowych

Sztuczna inteligencja i sieci neuronowe wkraczają coraz szerzej do medycyny, a jej rozwój jest bardzo intensywny. Warto zwrócić uwagę, że ostatnimi czasy Amerykańska Agencja Żywności i leków (U.S. Food and Drug Administration (FDA)) przygotowała oficjalny plan dopuszczania do użycia medycznego oprogramowania opartego o uczenie maszynowe. Podobne kroki czyni Unia Europejska nadając oznaczenia Conformité Européenne, jednakże większość tych algorytmów dopuszczona jest w dziedzinie radiologii [275]. Nie zmienia to faktu, że oprócz przytoczonych

powyżej przykładów sieci neuronowe i uczenie maszynowe wdrażane są coraz szerzej m.in do analizy składania białek [276], poszukiwania nowych leków [277], a także powoli wkraczają do chirurgii [278].

Budowa i sposób uczenia się sieci neuronowych

W ogólnym zarysie sieci neuronowe niezależnie od wybranej architektury opierają swoje działanie na odpowiednio zoptymalizowanych/skalibrowanych w trakcie procesu treningowego parametrów (ang. weights, parameters), tak by w jak najlepszy sposób na podstawie danych wejściowych aproksymować prawidłowy wynik. Efektywność działania sieci neuronowej zależy od kilku czynników, którymi m.in są:

1. Treningowy i walidacyjny zbiór danych, oraz ich odpowiednie wstępne przetworzenie
2. Wybór odpowiedniej architektury sieci neuronowej
3. Sposób inicjalizacji parametrów sieci
4. Dobór funkcji straty
5. Optymalny współczynnik uczenia
6. Algorytm propagacji wstecznej
7. Odpowiedni algorytm optymalizacji parametrów sieci
8. Czas i wystarczająca moc obliczeniowa

Treningowy i walidacyjny zbiór danych

Proces trenowania sieci neuronowej oparty jest w pierwszej kolejności na odpowiednich danych wejściowych, którymi może być odpowiednie środowisko symulowane lub realne lub treningowy zbiór danych. Taki zbiór może składać z dowolnego typu rekordów, takich jak obraz, dźwięk, tekst czy wartości liczbowe. Zbiór zebranych i przetworzonych danych najczęściej dzieli się na treningowy (ang. training dataset) i walidacyjny (ang. validation dataset). Nie ma określonych zaleceń w jakim stosunku taki podział dobrać, może to być 70:30, 60:40, czy nawet 50:50 [279] [280]. Celem tego postępowania jest zoptymalizowanie sieci neuronowej przy użyciu jednego zbioru, a sprawdzenie jej rzeczywistej skuteczności na danych, do których wcześniej nie miała dostępu. Sieci neuronowe wymagają dużej ilości danych. Pierwszym problemem

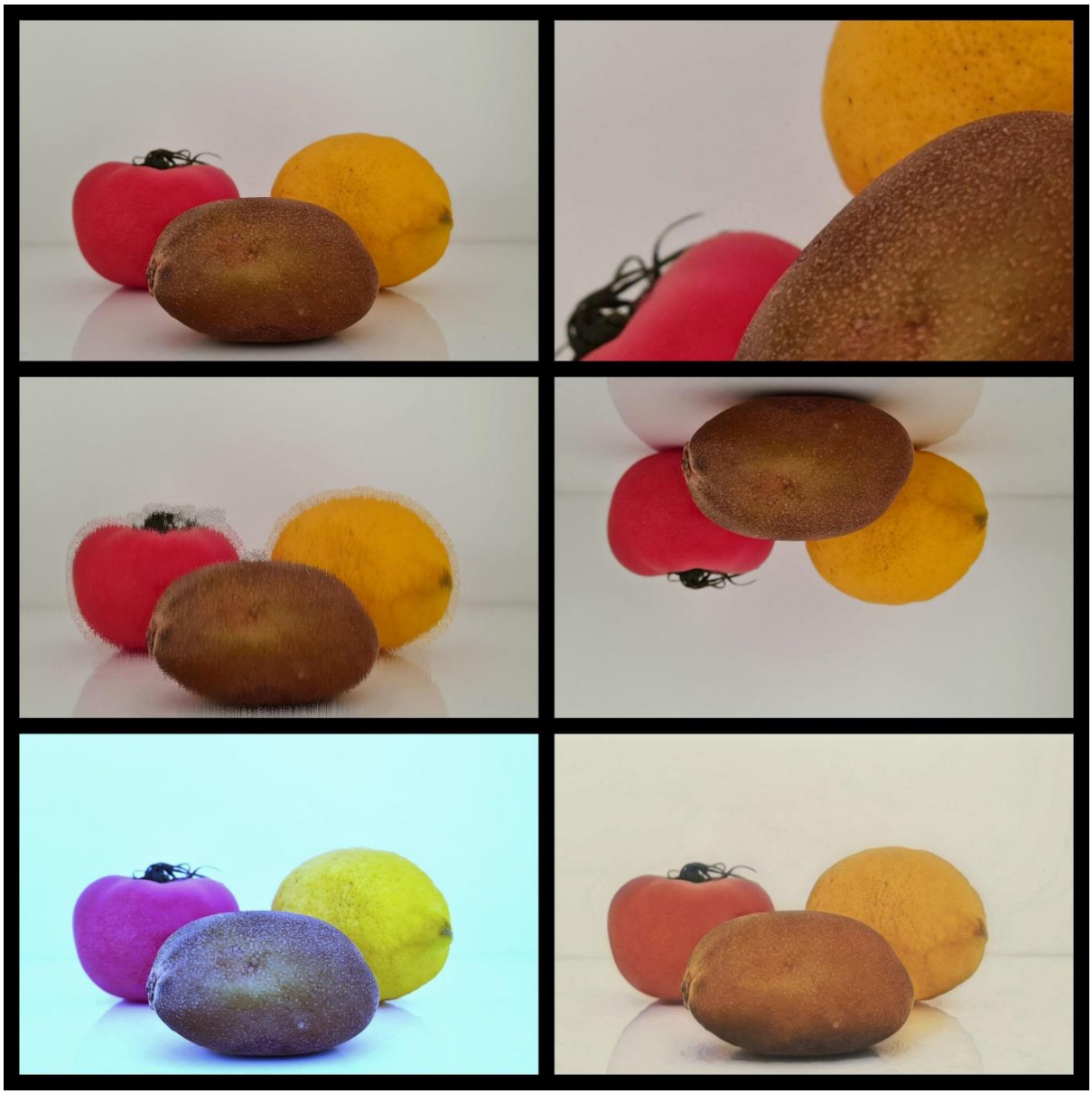
przy budowie zbioru treningowego jest odpowiednia ilość potrzebnych danych. Sieci neuronowe wymagają ich w dużej ilości, czynnik ten jest tak samo ważny, jak odpowiednia architektura sieci neuronowej. Dzięki otwartemu działaniu społeczności badaczy i naukowców dostępne są również dane publiczne, jak CIFAR, BreCaHAD, TCIA, czy Mozilla Common Voice [281–283]. Dane treningowe powinny być również reprezentatywne i odpowiednio dobrane. Informacje potrzebne sieci neuronowej do nauki powinny być odpowiednio reprezentatywne pod względem celu w jakim sieć ma działać oraz zbalansowane [284]. Dominacja jednej grupy w treningowym zbiorze może doprowadzić do nieprawidłowego klasyfikowania. Jest to szczególnie problematyczne w przypadku danych biologicznych, gdzie dane kontrolne (próbki zdrowe) są dostępne w większej ilości, niż dane badawcze [285].

Augmentacja obrazów

W celu zwiększenia różnorodności danych takich jak obraz można taki zbiór źródłowy poddać procesowi augmentacji obrazów (ang. image augmentation) (Rycina 10) [286]. Metoda ta polega m.in na

1. obrocie obrazu o określony kąt
2. przycinaniu i kadrowaniu
3. przesunięciu lub odbiciu obrazu w pionie lub poziomie
4. zmianie temperatury barw, jasności lub kontrastu obrazu
5. dodawanie szumów lub zakłóceń do obrazu
6. zniekształcaniu obrazów np. przez skalowanie czy rozciąganie
7. generowanie nowych obrazów na podstawie tych już istniejących (ekstrapolacja, wykorzystanie innych sieci neuronowych)

Niestety w przypadku danych biologicznych nie zawsze można zastosować tego typu proces, ze względu na możliwe wprowadzenie danych nieprawidłowych. Problem ten np. dotyczy wartości metylacji.



Rycina 10 Augmentacja obrazów

Dane syntetyczne i symulowane

Kolejnym sposobem na rozwiązanie problemu ilości i jakości danych jest zastosowanie danych syntetycznych lub symulowanych. Aktualnie jest coraz większy trend w kierunku stosowania tej metody. Przykładem użycia omawianych danych są prace Wang i wsp. [287] i Wood i wsp. [288]. W drugim przypadku wykorzystano wygenerowane komputerowo trójwymiarowe twarze o różnych cechach charakterystycznych, ubiorach, mimikach, oświetleniu i otoczeniu i na ich podstawie przetrenowano sieć neuronową uzyskując doskonałe wyniki.

Transfer wiedzy (ang. transfer learning) i metoda dostrajania (ang. fine tuning)

Zarówno metoda transferu wiedzy, jak i dostrajanie sieci już przetrenowanej może być bardzo pomocne w sytuacji niedoboru danych treningowych. Obie koncepcje są bardzo zbliżone do siebie i skuteczne, a ich przykłady podano we wcześniejszej części tej rozprawy. W przypadku metody dostrajania mówi się o wykorzystaniu sieci neuronowej już przetrenowanej i dotrenowaniu jej z użyciem nowych danych. Druga metoda zakłada nie tylko użycie nowych danych, ale np. dodanie kolejnych warstw do już przetrenowanej sieci czy zamrożenie parametrów w trakcie jej doszkalania [289].

Prywatność danych treningowych

Sieci neuronowe i uczenie maszynowe należy rozpatrywać również w kontekście prywatności osób, które użyczyły swoich danych do procesu uczenia. Dane takie, jeżeli są niezabezpieczone mogą zostać wykradzione. Ataki na dane wrażliwe można podzielić na te wiążące się z uzyskaniem nieautoryzowanego dostępu do bazy danych wrażliwych i te, w których próbuje się te dane odzyskać z przetrenowanego modelu. Do tych drugich należy: metoda odwrócenia modelu (ang. model inversion) [290, 291] i ataki na modele uczenia maszynowego oparte na wnioskowaniu o członkostwie (ang. Membership inference attacks) [292, 293]. Atak metodą odwrócenia modelu to bardzo efektywny sposób ataku, kiedy celem jest pozyskanie danych wrażliwych. Proces polega na stworzeniu algorytmu, który wykorzysta dane wyjściowe modelu tak by zrekonstruować dane treningowe. Atakujący pozyskuje np. ogólnodostępne dane na temat ofiary i tak manipuluje danymi wejściowymi bazując na wyniku uzyskanym z

modelu uczeniu maszynowego, by uzyskać te rzeczywiście wprowadzone podczas treningu, a w konsekwencji poznać dane wrażliwe. Oddzielną grupą ataków są ataki oparte o nieuprawniony dostęp do danych treningowych. Ich szczegółowy opis znacząco przekracza ramy tej rozprawy, ponieważ obejmuje on ogrom obszernych technik i narzędzi [294]. Ze względu na realny problem jakim jest ochrona prywatności danych opracowywane są różne metody prewencji. W przypadku nieupoważnionego dostępu do danych treningowych stosuje się m.in metody typu secure enclaves [295], gdzie dane treningowe odseparowane są od głównej jednostki na poziomie sprzętowym lub metody typu federated learning [296], w których baza danych treningowych jest zdecentralizowana, a sam proces uczenia odbywa się na oddzielnych komputerach, które po zakończeniu aktualizują parametry modelu na serwerze. W prewencji ataków, których celem jest odzyskanie danych z modelu uczenia maszynowego stosujemy:

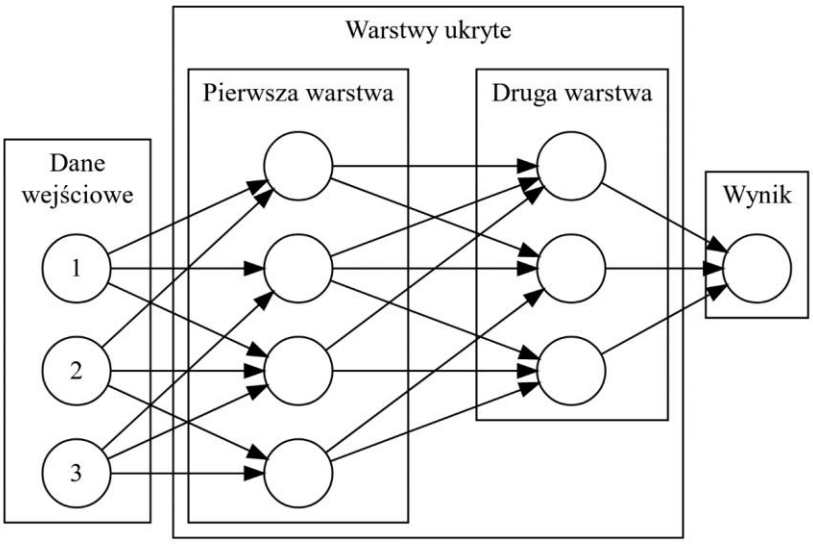
1. prywatyzację różnicującą (ang. differential privacy) [297]
2. szyfrowanie homomorficzne (ang. Homomorphic encryption) [298]
3. k-anonymity [299]

Prywatyzacja różnicująca polega na modyfikacji danych treningowych tak, by nie zaburzały one treningu, ale jednocześnie wprowadzały pewien stopień anonimizacji. Szyfrowanie homomorficzne polega z kolei na zaszyfrowaniu danych wrażliwych, ale tak, by zachować ich własności matematyczne i w efekcie można było dokonać procesu uczenia. Metoda ta jest wymagające pod względem mocy obliczeniowej i dość wolna. Wykorzystywane są zarówno metody symetrycznego, jak i asymetrycznego szyfrowania, aczkolwiek drugi typ jest powszechniejszy [297]. Metoda symetrycznego szyfrowania polega na wykorzystaniu jednego wspólnego klucza szyfrującego, a metoda asymetryczna na wykorzystaniu klucza publicznego i prywatnego. Dane szyfrowane są z użyciem klucza publicznego odbiorcy, w taki sposób by można je było odszyfrować tylko z użyciem jego klucza prywatnego. Wydzielić można homomorficzne szyfrowanie pełne, gdzie dostępne są pełne dane lub częściowe, gdzie dostęp ten jest ograniczony [298]. Innym sposobem jest metoda k-anonimizacji [299]. Można ją podzielić na generalizująca i supresującą. Metoda generalizująca zamienia pewne dane na uogólnioną kategorię np. przedział wiekowy, zamiast dokładnego wieku. Metoda supresująca z kolei pewne dane zamazuje całkowicie. Oddzielną metodą

prywatyzacji jest wykorzystanie danych syntetycznych. Jest to metoda zdobywająca coraz większą popularność i została ona wykorzystana m.in. w opracowanej przez autora metodzie [288, 300, 301]. Jej ogromną zaletą jest brak albo minimalny poziom wykorzystania danych prawdziwych osób, ale niestety równocześnie może wiązać się ona również z efektem niekorzystnym, jakim jest nieprawidłowe przeszkolenie sieci neuronowej i np. pewna stronniczość (ang. bias) objawiająca się fałszywie dodatnimi lub negatywnymi wynikami. Przyczyną tego zjawiska jest użycie nieprawidłowo stworzonego zbioru danych treningowych [302].

Budowa podstawowej sieci neuronowej

Najprostsza sieć neuronowa, której można użyć do zadania typu klasyfikacja składa się z minimum trzech typów warstw (Rycina 11). Odpowiednio będzie to warstwa wejściowa, której rozmiar odpowiada wielkości danych wejściowych np. liczbie pikseli w obrazie, warstwa wyjściowa, która opowiada liczbie klas, do których dane mają być zakwalifikowane, oraz pośrednie tzw. warstwy ukryte [303]. Taki model sieci neuronowej, gdzie wszystkie węzły jednej warstwy łącza się z wszystkimi kolejnej tworząc ciąg przesyłu informacji od wejścia do wyjścia nazywamy siecią neuronową w pełni połączoną (ang. Feed-forward neural network) [304].

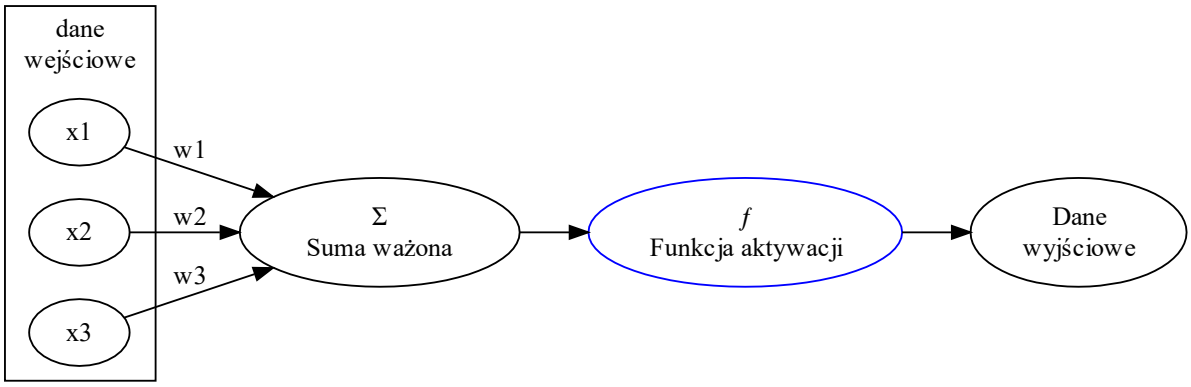


Rycina 11 Model prostej sieci neuronowej

Jest wiele algorytmów głębokiego uczenia i wciąż rozwijane są nowe. W zależności od zadania, które ma być rozwiązane wymagane jest dobranie odpowiedniej architektury. Opis wybranych architektur znajduje się w dalszej części rozprawy.

Funkcja aktywacji

Efektywność działania sieci neuronowych zależna jest od doboru funkcji aktywacji do wybranego zadania [305]. Pomimo faktu, iż sieci neuronowe składają się z więcej niż z jednej warstwy na modelu węzła przedstawionym na Rycina 12 zaznaczono kolorem niebieskim omawiany element.



Rycina 12 Model węzła z oznaczoną kolorem niebieskim funkcją aktywacji

Parametry i hiperparametry

Parametrem określamy zmienne, które podlegają modyfikacji w trakcie uczenia się sieci neuronowej. Natomiast mianem hiperparametr określamy zmienne określone zewnętrznie przez badacza. Optymalizacja parametrów podczas treningu sieci neuronowe zależna jest od funkcji kosztów, która jest różnicą pomiędzy danymi wyjściowymi obliczonymi przez sieć neuronową, a pożądanymi/prawidłowymi. Celem optymalizacji jest uzyskanie jak najmniejszej wartości dla funkcji kosztów. Przed rozpoczęciem treningu musimy nadać jakieś wartości wyjściowe poszczególnym parametrom. Proces ten nazywany jest inicjalizacją parametrów [306]. Funkcja aktywacji wprowadza nieliniowość, która jest wymagana do tego, by sieć neuronowa była zdolna do uczenia się złożonych problemów [307]. Optymalna wg. Datta [306] funkcja aktywacji powinna być:

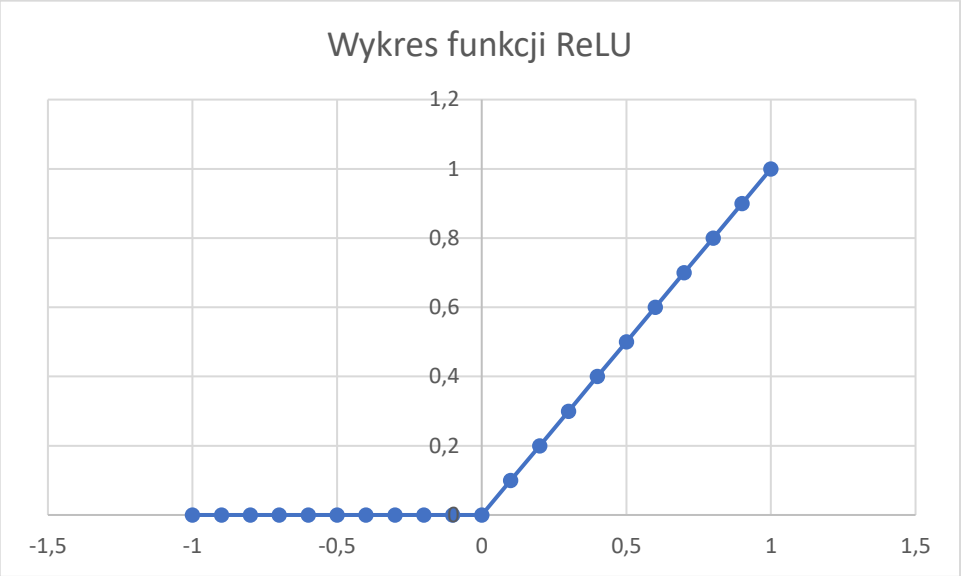
1. Nieliniowa
    1. Realne problemy często nie są liniowe. Nieliniowa funkcja może aproksymować funkcję liniową, ale nie odwrotnie
    2. Jeżeli funkcja aktywacji byłaby liniowa to warstwy ukryte mogłyby być skrócone do pojedynczego wzoru, a wtedy zwiększanie złożoności sieci neuronowej nie miałoby sensu
2. Funkcją różniczkowalną:
    1. Funkcja różniczkowa określa jak jedna zmienna zmienia się w zależności od drugiej w tym przypadku dane wejściowe od wyjściowych. Jeżeli funkcja jest różniczkowalna to zmiany nie będą miały charakteru skokowego
3. Funkcją ciągłą
    1. Funkcja nie może być funkcją różniczkowalną, jeżeli nie jest funkcją ciągłą
4. Powinna mieć granice
5. Może być poddana środkowaniu (ang. zero-centered)
    1. To znaczy, że średnia wartość ze wszystkich możliwych wyników funkcji jest równa 0. W przeciwnym wypadku warstwa wyjściowa będzie zawsze dążyć w stronę dodatnią lub ujemna w zależności od tendencji funkcji
6. Powinna wiązać się z niskim kosztem obliczeniowym

Funkcja sigmoidalna

W przeszłości najczęściej używano funkcji sigmoidalnej z powodu zdolności do ograniczenia wartości rzeczywistych w granicach 0 i 1. Miało to imitować ekscytacje biologicznych neuronów. Jednakże aktualnie nie jest ona już używana z powodu szybkiego wysycania gradientu w konsekwencji hamując uczenie się sieci neuronowej [308].

Funkcja aktywacji Rectified Linear Unit

Funkcja aktywacji Rectified Linear Unit (ReLU), która jest jedną z najbardziej popularnych w użyciu definiowana jest zgodnie ze wzorem f(x) = max (0, x). Gdzie x to wartość wejściowa, a f(x) wyjściowa. Funkcja przyjmuje wartość wejściową dla każdej wartości większej od 0 (Rycina 13) [309].



Rycina 13 Wykres przestawiający przebieg funkcji ReLU

Zaletą użycia funkcji ReLU jest znaczące przyśpieszenie uczenia się sieci neuronowej i ograniczenie potrzebnej mocy obliczeniowej [310]. Funkcja ta nie spełnia wszystkich założeń zalecanej funkcji wg. Datta [306] m.in nie jest ograniczona, ani możliwa do środkowania do zera, a dla wartości 0 nie jest różniczkowalna. Dodatkowo problemem jest wrażliwość tej funkcji na problem uśmiercania neuronów. Oznacza to, że podczas treningu parametry mogą dążyć do 0 powodując, że niektóre neurony mogą zostać nieodwracalnie wyłączone [311]. Z tego powodu powstały różne modyfikacje tej funkcji np. leaky ReLU, która miała na celu zniwelowanie powyższego problemu przez dodanie nachylenia dla wartości negatywnych. Wzór opisujący funkcje ReLU przedstawiono na Rycina 14

$$f(x) = \begin{cases} 0.01x & \text{dla } x \leq 0 \\ x & \text{w innym przypadku} \end{cases}$$

Rycina 14 Wzór funkcji aktywacji ReLu

Dodatkowo należy zwrócić uwagę, że jest bardzo dużo innych funkcji aktywacji, często zbliżonych do tej teoretycznej opisanej przez Datta [306], jak Scaled exponential linear unit (SeLU) [312], Gaussian error linear unit (GELU) [313], czy Maxout activation function [314]. Aczkolwiek mimo wielu teoretycznych rozważań nie ma ostatecznych rekomendacji na podstawie testów i badań jaką funkcję aktywacji należy w szczególności stosować [309]. Jak zauważa Dubey i wsp. [305] modyfikacje funkcji sigmoidalnej, czy funkcji tangensu hiperbolicznego mimo, że rozwiązują wiele problemów pierwowzorów to zwiększają ich złożoność, a nowe warianty funkcji ReLU pomimo bycia zaprojektowanymi do poprawienia niedoskonałości poprzedniczki to w wielu aplikacjach nie przynoszą korzyści. Dodatkowo zauważa on, że do dziś wiele sieci neuronowych korzysta właśnie z funkcji ReLU. Jest tak np. w algorytmie YOLACT opracowanym przez Bolya i wsp. [315].

Inicjowanie parametrów

Jak wspomniano powyżej kolejnym istotnym elementem wpływającym na proces uczenia się sieci neuronowych jest proces inicjowania parametrów. W trakcie treningu parametry sieci neuronowej są optymalizowane z użyciem specjalnego algorytmu optymalizacji, tak by wartość funkcji kosztów była jak najmniejsza oraz osiągnięto globalne minimum. Dokładny przebieg tego procesu omówiony jest w dalszej części rozprawy. Taki algorytm optymalizacji wymaga ustalenia startowych wartości dla parametrów sieci, czyli inicjalizacji parametrów, a od tego, jak zostaną dobrane zależy skuteczność sieci neuronowej. Mogłoby się wydawać, że ustalenie parametrów jako jednakową liczbę np. zero jest prawidłowym rozwiązaniem, jednakże ze względu na symetrię paramentów i gradientów nie będzie można dokonać jakichkolwiek zmian podczas uczenia sieci neuronowej. Z kolei niektóre parametry wyjściowe mogą spowodować, że sieć nie znajdzie globalnego minimum i jej

skuteczność będzie zła. Jednym z rozwiązań jest wprowadzenie losowych parametrów lub odpowiedniej metody inicjalizacji [316]. Przykładem takiej metody jest metoda Xavier, która stosuje rozkład jednostajny do inicjalizowania parametrów i była powszechnie używana [317], jednakże metoda ta zakłada, że zastosowana funkcja aktywacji ma przebieg liniowy, co nie jest prawdą np. w stosunku do funkcji ReLU [318]. Alternatywą jest zastosowanie metody Kaiming (metoda ta funkcjonuje również pod nazwą He, obie nazwy pochodzą od imienia i nazwiska autora). Skuteczność tej metody inicjalizacji zwłaszcza przy jednoczesnym zastosowaniu funkcji aktywacji ReLU potwierdzili to m. in Li i wsp. [319] i Kumar i wsp. [316].

Funkcja kosztów

Rosenblatt w swojej pracy napisał : „O systemie można powiedzieć, że się uczy, jeżeli jego skuteczność jest mierzalna i ulega progresji wraz z doświadczeniem” [196]. Bardzo ważnym aspektem podczas projektowania sieci neuronowych jest odpowiedni dobór funkcji kosztów. Funkcja kosztów (ang. loss function) jest wartością mierzalną skuteczności sieci neuronowej. Wybór tej funkcji jest ściśle związany z danymi wyjściowymi, jakie chcemy uzyskać [320]. Najbardziej podstawowymi są dwie przedstawione na poniższej Rycina 15 [207].

$$L\left(Y, \hat{f}(X)\right) = \begin{cases} \left(Y - \hat{f}(X)\right)^2 & \text{błąd kwadratowy} \\ \left[Y - \hat{f}(X)\right] & \text{błąd absolutny} \end{cases}$$

Rycina 15 Wzór błędu kwadratowego i absolutnego

Aktualnie funkcja średniokwadratowa i absolutna funkcja średniokwadratowa, która była bardzo popularna w latach 1980 i 1990 często prowadziły do niezadawalających rezultatów i z czasem zostały one zastąpione przez funkcję cross-entropy loss. W sytuacji potrzeby użycia funkcji błędu średniokwadratowego zalecane jest jej użycie w problemach regresji [321]. Cross-entropy loss jest odwrotną logarytmiczną funkcją wiarygodności, która wykonuje warunkowe oszacowanie maksymalnego prawdopodobieństwa. W ten sposób klasyfikator osiąga najlepszy wynik dla $y = \hat{y}$ i prawdopodobieństwa 1 lub najgorszy dla $y \neq \hat{y}$ i prawdopodobieństwa 0 .Dla

klasyfikacji binarnej (są tylko dwie klasy) i regresji logistycznej, czyli w przybliżeniu pojedynczego perceptronu wzór dla tej funkcji wygląda jak na Rycina 16:

$$\log P\,(y|x) = y \log P\,(\hat{y} = 1) + (1 - y) \log P\,(\hat{y} = 0)$$

Rycina 16 Funkcja cross entropy dla klasyfikacji binarnej i regresji logistycznej, y - docelowa wartość wyjściowa, ŷ - przewidywana wartość wyjścioway - docelowa wartość wyjściowa [322]

W środowisku PyTorch [323] będący jednym z dwóch wiodących środowisk do prototypowania rozwiązań z dziedziny sztucznej inteligencji zaimplementowana jest wersja weighted cross entropy loss z zaimplementowaną funkcją softmax dla problemu klasyfikacji przy większej puli klas, wzór ten dla pojedynczej próbki przedstawiony jest na Rycina 17.

$$l_n = -\sum_{c=1}^{C} w_c \log \frac{\exp(x_{n,c})}{\sum_{i=1}^{C} \exp(x_{n,i})} y_{n,c}$$

Rycina 17 Wzór cross entropy loss dla pojedynczej próbki zaimplementowany w PyTorch

Natomiast średnią wartości tej funkcji dla całego zbioru przedstawia Rycina 18

$$l(x, y) = \frac{\sum_{n=1}^{N} l_n}{N}$$

Rycina 18 Wzór średniej wartości cross entropy loss dla całego zbioru zaimplementowany w PyTorch

Należy zwrócić uwagę, że wpływ na wynik tej funkcji ma zmienna $w_{\square}$ (waga), która, jeżeli jedna z klas jest nadreprezentowana, może zmniejszyć znaczenie wyniku funkcji kosztów. Równoważy to problem omówiony wcześniej dotyczący braku równowagi w liczbie próbek w poszczególnych klasach. Kolejną rzeczą, na którą należy zwrócić uwagę jest fakt, że w liczniku i mianowniku wzoru znajduje się funkcja wykładnicza. Jest to zaimplementowana funkcja softmax, której wzór przedstawia Rycina 19.

$$\text{softmax}(x)i = \frac{\exp(x_i)}{\sum j = 0^{K-1} \exp(x_j)}$$

Rycina 19 Wzór funkcji softmax

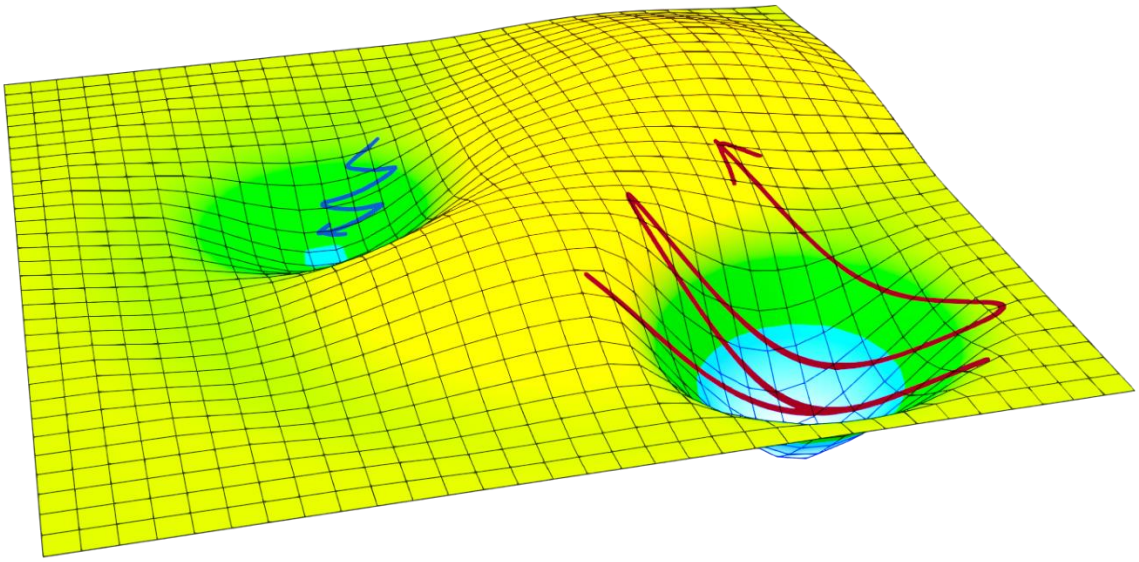
Nieliniowość funkcji softmax jest tu używana do normalizacji wyniku wyjściowego do rozkładu prawdopodobieństwa [324].

Algorytm propagacji wstecznej

Żeby zrozumieć jak uczą się sieci neuronowe trzeba zacząć od propagacji wstecznej [325]. Algorytm ten pozwala na obliczenie wstecznego przebiegu przez sieć neuronową wartości kosztu. Mylnie, często ta metoda uważana jest za sam proces uczenia. W rzeczywistości służy ona tylko do obliczenia gradientu, natomiast zaktualizowanie parametrów sieci neuronowej zajmuje się oddzielna funkcja nazywana optymalizatorem [320]. Zwiększając złożoność sieci zaczynamy stosować kompozycję funkcji to znaczy wartość wyjściową jednej z funkcji składowej włączamy jako wartość wejściowa innej [326]. Propagacja wsteczna składa się z dwóch faz. Propagacji wejść, gdy sieć przyjmuje wartość x i zwraca wartość y. Przechodząc przez poszczególne warstwy sieci zostaje pozostawiona pewna informacja na temat zależności poszczególnych jej elementów i obliczona funkcja kosztu [320]. W drugim etapie, czyli tzw. propagacji błędu przy pomocy reguły łańcuchowej, czy reguły opartej o twierdzenie o pochodnej funkcji złożonej obliczamy pochodne błędu [202]. Tak więc obliczany gradient to wektor funkcji kosztu względem składowych wektora parametrów sieci, czyli. miara i kierunek tego, jak szybko zmienia się wartość funkcji w każdym punkcie sieci [327]. Wzór uzyskany w ten sposób obrazuje Rycina 20

$$\frac{\partial f}{\partial h^{(l-1)}} = \frac{\partial h^{(l)}}{\partial h^{(l-1)}} \frac{\partial f}{\partial h^{(l)}} = \frac{\partial \sigma\left(W^{(l)} h^{(l-1)} + b^{(l)}\right)}{\partial h^{(l)}} \frac{\partial f}{\partial h^{(l)}}$$

Rycina 20 Wzór propagacji wstecznej. f wynik ostateczny dla sieci, $h^{(l-1)}$ wane wyjsciowe w warstwie $(l-1)$, $h^{(l)}$ - dane, wyjściowe w warstwie l , $W^{(l)}$ and $b^{(l)}$ parametry w warstwie l , σ - funkcja atywacji.



Rycina 21 Wizualizacja modelu gradientu. Skala kolorów od żółtego najwyższego do białego najniższego. Linia czerwona przedstawia duży współczynnik uczenia. Linia niebieska mały współczynnik uczenia

Algorytmy optymalizacji

Jak wspomniano wcześniej proces uczenia sieci neuronowej oparty jest nie tylko na propagacji wstecznej, ale wymaga również odpowiedniego algorytmu optymalizacji, czyli algorytmu, który zaktualizuje parametry sieci na podstawie obliczonego wcześniej gradientu w efekcie pokieruje procesem uczenia [320]. Na ogólnym modelu gradientu przedstawiającym płaszczyznę (Rycina 21) łatwiej zobrazować złożoność propagacji wstecznej i algorytmu optymalizacji. Znajdując się w najwyższym punkcie tej płaszczyzny w procesie uczenia chcielibyśmy znaleźć punkt najniższy. Opisując proces kolokwialnie obliczony gradient funkcji kosztu określa nam nachylenie powierzchni natomiast algorytm optymalizacji „poszukuje odpowiedniej drogi" robiąc to stopniowo. Najbardziej podstawowym wariantem algorytmu optymalizacji jest metoda gradientu wstecznego prostego. (ang. batch gradient) [328]. Polega ona na znalezieniu na podstawie całego zbioru treningowego i zgodnie ze wzorem (Rycina 22) takich parametrów sieci, by wynik funkcji kosztu był jak najmniejszy.

$$\theta = \theta - \alpha \cdot \nabla J(\theta)$$

Rycina 22 Wzór algorytmu optymalizacji metodą gradientu wstecznego prostego $\nabla J(\theta)$ - gradient funkcji kosztów, ∇ - grecki symbol nabla, symbol gradientu , θ - parametry, które bedą podlegały optymalizacji, α - współczynnik uczenia, J - funkcja kosztów [328]

Symbol α oznacza w tym wzorze współczynnik uczenia. Jest to wartość, która wpływa na stopień w jakim parametry są modyfikowane przez algorytm optymalizacji z użyciem gradientu funkcji kosztów [328]. Należy zwrócić uwagę na fakt, iż gradient jest wektorem, a więc ma kierunek, a wartość gradientu nie mówi o tym, jak zmienić parametry, ale w jakim stopniu, ponieważ jest wynikiem funkcji kosztów mówiącym o tym jak bardzo odbiega wynik szacowany od prawdziwego. Współczynnik uczenia moduluje sposób w jaki algorytm pod wpływem gradientu modyfikuje parametry. Celem pierwotnym treningu jest osiągnięcie globalnego minimum, czyli takiego stanu, w którym funkcje kosztów będą najniższe, czyli sieć będzie najoptymalniej dostosowana do danych, które jej przedstawiono. Jeżeli współczynnik uczenia będzie za duży to algorytm optymalizacji ominie/przeskoczy to minimum (Rycina 21) a w odwrotnej sytuacji trening będzie bardzo czasochłonny. Możliwe jest też sytuacja, w której algorytm optymalizacji utknie w tzw. lokalnym minimum [329] (Rycina 21), czyli miejscu, gdzie wartości funkcji są najmniejsze lokalnie, ale nie najmniejsze w kontekście wszystkich wartości funkcji(globalne minimum [329]). Innym nowszym algorytmem jest stochastyczny gradient wsteczny (ang. stochastic gradient descent, SGD) [330]. Główna różnica pomiędzy prostymi polega na zakresie danych, na podstawie których parametry są zoptymalizowane. Dla gradientu prostego jest to cały zbiór treningowy, a dla stochastycznego wybrana próbka lub w modyfikacji tego algorytmu wybrane próbki (ang. mini-batch). Widać tę zależność we wzorze (Rycina 23).

$$\theta = \theta - \alpha \cdot \nabla J\left(\theta; x^{(i)}; y^{(i)}\right)$$

Rycina 23 Wzór stochastycznego gradientu wstecznego[328]

We wzorze stochastycznego gradientu wyrażenie $\nabla J\left(\theta; x^{(i)}; y^{(i)}\right)$ reprezentuje gradient funkcji kosztów J w odniesieniu do parametrów θ , które oszacowywane są na podstawie próbki $\left(x^{(i)}, y^{(i)}\right)$. Zaletą tego algorytmu jest prostota oraz prędkość i oszczędność w kontekście mocy obliczeniowej. SGD nie jest jedynym algorytmem

optymalizacji parametrów sieci neuronowej. Powstało wiele różnych ich odmian, które miały ten proces usprawnić pod różnym kątem. Jednym z nich jest metoda Adam [331]. Metoda ta została wprowadzona przez Kingma i wsp. Adam jest modyfikacją stochastycznego algorytmu obejmującą następujące funkcje i parametry: adaptacyjne tempo uczenia, momentum, średnie bieżące gradientu i niezerowy współczynniki uczenia. Adaptacyjne tempo uczenia zaimplementowane do modułu Adam dostosowuje tempo uczenia na podstawie przeszłych gradientów i średnich bieżących tych gradientów. Różni się to od metody zastosowanej w SGD, w której tempo jest stałe. Dodatkowo współczynniki uczenia nigdy nie są zerowe. W efekcie Adam ma mniejsze ryzyko utknięcia w tzw. lokalnym minimum. Zaimplementowane momentum [332] stabilizuje proces uczenia, a funkcja średnich bieżących pozwala na odniesienie się do gradientów z poprzednich iteracji. Adam uważany jest za bardziej zaawansowany algorytm od SGD, aczkolwiek w 2017 roku Loshchilova i wsp. wprowadzili dodatkową modyfikację, czyli rozkład wagowy (ang. weight decay). W procesie tym dodawany jest współczynnik, który ma zapobiec nadmiernemu dopasowaniu się modelu do danych treningowych i chociaż wydaje się na pierwszy rzut oka niepraktycznym zabiegiem to w rzeczywistości zapobiega to niekorzystnemu zjawisku jakim jest przetrenowanie (ang. overfitting) i poprawia zdolność do generalizacji. AdamW w zależności od zadania wykazuje najlepszą skuteczność [333]. Należy pamiętać, że celem optymalizacji nie jest osiągnięcie najwyższego stopnia zbieżności ze zbiorem danych treningowych, ani nawet ewaluacyjnych, ale ostatecznym i najbardziej pożądanym jest najwyższy poziom generalizacji, czyli zgodnie z opisaną wcześniej teorią funkcji aproksymacji najlepszej umiejętności dopasowania danych wejściowych do danych wyjściowych zwłaszcza tych jeszcze niewidzianych.

Manipulator wielkości współczynnika uczenia (ang. learning rate scheduler)

Wspominano wcześniej, że istotnym hiperparametrem w uczeniu sieci neuronowej jest współczynnik uczenia. Jeżeli jest on za duży może dojść do przeoczenia globalnego minimum, jeżeli za mały to w wyniku działania optymalizator funkcji kosztów utknie w lokalnym minimum nigdy nie osiągając tego globalnego. Najprostszym sposobem na rozwiązanie tego problemu jest stopniowe zmniejszanie współczynnika o stałą wartość lub procent wraz z przebiegiem treningu (ang. learning

rate decay) [202] . Smith i wsp. zastosowali odmienną metode 1cycle learning rate policy. Polega ona na tym, że współczynnik uczenia jest w pewnej wartości wyjściowej, następnie zwiększany jest on do ustalonej wartości maksymalnej i zmniejszany poniżej wartości początkowej [334].

Problemy przy szkoleniu sieci neuronowej

Nadmierne dopasowanie (ang. overfitting).

Jednym z częstszych problemów sieci neuronowych jest tzw. nadmierne dopasowanie. Fenomen ten polega na nadmiernym dostosowaniu sieci neuronowej do treningowych danych bez osiągnięcia odpowiedniej generalizacji i w efekcie brak skuteczności przy ekspozycji na dane niebędące treningowymi [335]. Metody przeciwdziałania temu zjawisku można podzielić na trzy klasy. Jedne z nich dotyczą odpowiedniego przygotowania danych treningowych, jak augmentacja obrazu, czy odpowiednia ich reprezentatywność, jak opisano to rozleglej we wcześniejszej części tej rozprawy. Druga klasa metod dotyczy samego modelu sieci neuronowej, oraz jego architektury i jest to np. zaimplementowany w algorytmie optymalizacji AdamW proces regularyzacji, jakim jest rozkład wag [336] lub zastosowanie warstw dropout. [337]. Drugi jest to proces podobny do procesu rozpadu wag i podobnie sprzeczny z celem treningu, a polega on na losowym wyłączeniu połączeń pomiędzy poszczególnymi neuronami. Trzecia z kolei polega na odpowiednim przeprowadzeniu procesu treningowego i obejmuje m.in techniki takie jak „early stop”, gdzie proces treningowy zatrzymywany jest przed pełną konwergencją z danymi treningowymi [335].

Niedookreślenie (ang. Underspecification)

W 2020 roku naukowcy z Google [338] zwrócili uwagę, że niektóre sieci bardzo źle oszacowują realne dane mimo bardzo dobrych wyników w typowych testach. Zjawisko to często dotyczy obrazów medycznych, czy nawet danych dotyczących genomiki. Nazwali ten problem niedookreśleniem. Problem jest o tyle złożony, że jego przyczyna może leżeć nie tylko w złych danych treningowych, o czym autorzy wspominają, ale także w bardzo drobnych zmianach wprowadzonych do sieci na przykład podczas ponownego jej przeszkolenia w ten sam sposób jak poprzednio, ale zmieniając np. ziarno generatora liczb losowych. Sieć neuronowa, której dotyczy ten

problem będzie bardzo dobrze sprawdzała się w testach, ale odmiennie i często źle przy realnych danych.

Eksplozja gradientu

Problem dotyczy sieci neuronowych, których szkolenie związane jest z gradientem. Wraz z dużymi wartościami funkcji kosztów akumulowanych w trakcie propagacji wstecznej dochodzi do znaczących zmian w parametrach sieci, a sam proces przestaje być stabilny i efektywny [339].

Zanikanie gradientu

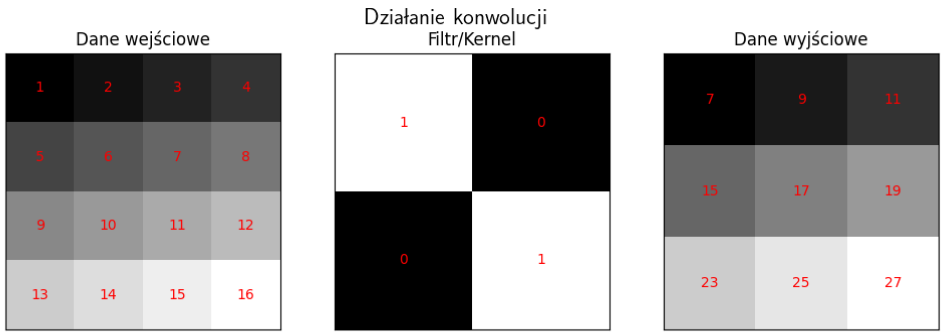
Odwrotnie do problemu eksplodującego gradientu mamy zanikanie gradientu, podobnie problem dotyczy sieci neuronowych szkolonych w oparciu o propagację wsteczną i optymalizatory związane z gradientem. W trakcie treningu parametry sieci neuronowej ustalane są na podstawie aktualnego stanu i proporcjonalnie do pochodnej funkcji kosztów. Jeżeli gradient będzie dążył do zera to w pewnym momencie dalszy trening może być niemożliwy [339, 340].

Charakterystyka wybranych architektur sieci neuronowych

Jest bardzo dużo modeli sieci neuronowych Zaczynając od najprostszej, czyli jednokierunkowa sieć neuronowa (ang. feed-forward network) przechodząc przez sieci konwolucyjne, rozszerzające budowę tej pierwszej o wprowadzenie warstw konwolucyjnych i pooling, dzięki czemu znalazły zastosowanie w dziedzinie nazywanej widzeniem komputerowym (ang. computer vision). W analizowaniu danych sekwencyjnych i robotyce znalazły zastosowanie sieci typu recurrent neural networks, które mają swojego rodzaju „pamięć”. Powstały też bardziej złożone typu transformers, w których zaimplementowano tzw. warstwę uwagi (ang. attention layer) i multi-head attention layer, co znacząco wpłynęło na sposób w jaki radzą sobie z celem, do którego zostały zaprojektowane, czyli przetwarzanie naturalnego języka (ang. natural language processing).

Konwolucyjne sieci neuronowe

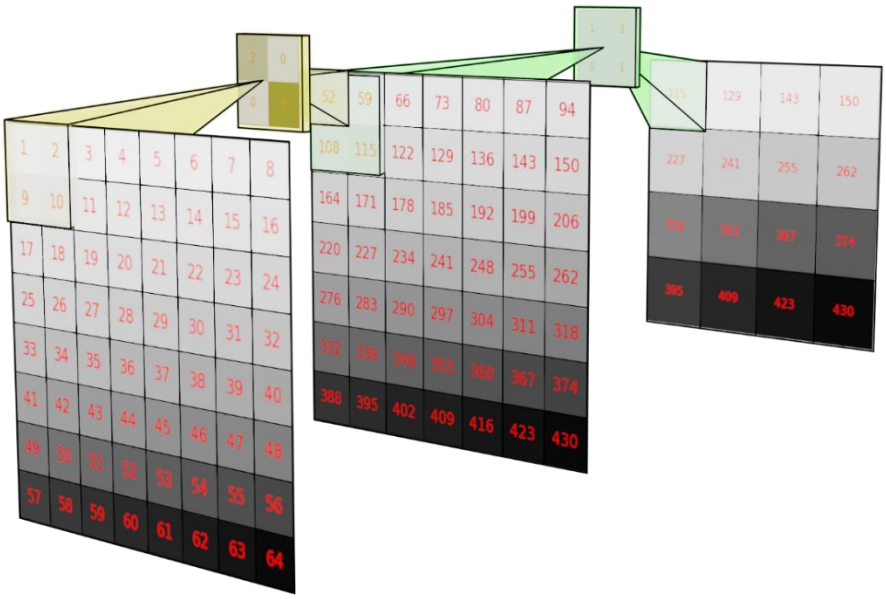
Konwolucyjne sieci neuronowe to rodzaj sztucznych sieci neuronowych, które główne zastosowanie znalazły w analizie obrazów i video pozwalając na efektywne wykonywanie zadań typu klasyfikacja czy segmentacja obrazów. Stanowią często jeden z elementów bardziej złożonych sieci neuronowych. Przykładem jest np. AlphaGo, będący algorytmem, który w 2016 pokonał człowieka, będącego mistrzem świata w grę Go. W tym przypadku warstwa konwolucyjna analizuje stan gry [222]. Patrząc historycznie, można powiedzieć, że pierwszą taką siecią neuronową był LeNet, której zadaniem było rozpoznawanie pisma odręcznego [199]. Za kolejny kamień milowy w rozwoju tej architektury można uznać sieć AlexNet, która wniosła ogromną innowację w dziedzinie rozpoznania obrazów [220] Podstawą działania tych sieci neuronowych jest proces konwolucji. W prosty sposób można to opisać jako przemieszczanie się filtra po danych wejściowych. Filtr z punktu widzenia matematycznego jest macierzą, tak samo jak odpowiadający fragment danych wejściowych. Oba te elementy ulegają przemnożeniu i zsumowaniu, dając w efekcie odpowiedni wynik. Efektem całego procesu jest powstanie mapy cech (ang. feature map). W przypadku analizy obrazu może być to kształt, a w przypadku np. wspomnianego AlphaGo stan gry (Rycina 24). Warto zwrócić uwagę, że w wielu implementacjach sieci konwolucyjnych nie stosuje się typowego mnożenia, a metodę korelacji wzajemnej (ang. cross-correlation) [341].



Rycina 24 Wzór i wizualizacja konwolucji

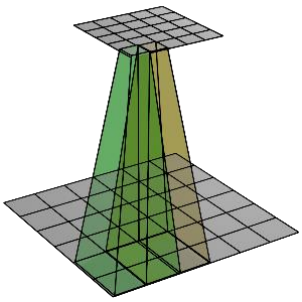
Struktura sieci konwolucyjnej była zainspirowana korą wzrokową, w której złożony zespół komórek odbiera bodźce wzrokowe i poddaje je analizie [342]. Podstawowa architektura sieci konwolucyjnej opiera się na warstwie konwolucyjnej z następującą po

niej warstwą łączącą (ang. pooling layer), funkcją aktywacji i następnie jednokierunkową siecią neuronową. Zależność między warstwą konwolucyjną, filtrami i warstwą łączącą przedstawiono na Rycina 25



Rycina 25 Wizualizacja działania sieci konwolucyjnej i zależności pomiędzy warstwą konwolucyjną, warstwą pooling, ich filtrami, oraz danymi wyjściowymi.

Filtr może przemieszczać się o zadany krok (ang. stride) np. pomijając pozycje. Na rycinach Rycina 26 i **Błąd! Nie można odnaleźć źródła odwołania.** przedstawiono ten proces.



Rycina 26 Wizualizacja działania filtra z krokiem równym jeden

Poza wyżej opisanym krokiem sieć konwolucyjną można dostosować z użyciem innych hiperparametrów jak:

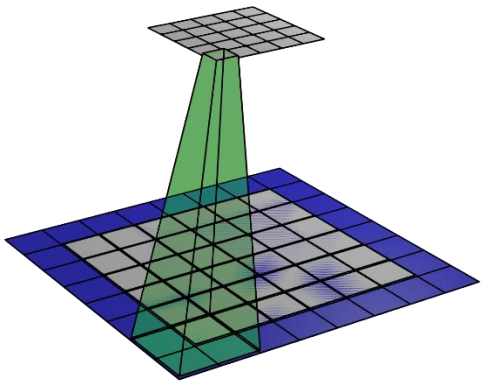
1. wypełnienie (ang. padding)
2. liczba zastosowanych filtrów (kanałów, ang. channels)
3. dylatacja (ang. dilation)
4. głębokość (ang. depth)

W efekcie działania filtra macierz wyjściowa jest mniejsza od macierzy wejściowej zgodnie ze wzorem Rycina 27

$$wymiar\_macierzy\_wyjściowej = \left\lfloor \frac{wymiar\_macierzy\_wejścowej + 2 * p - wymiar\_filtru}{wielkość\_kroku} \right\rfloor + 1$$

Rycina 27 Wzór przedstawiający wpływ wymiaru macierzy wejściowej, wymiaru filtra oraz wielkości kroku na wymiary macierzy wyjściowej

Jako przeciwdziałanie temu efektowi stosuję się metodę wypełniania, która polega na uzupełnieniu jednakowymi wartościami skrajnych wartości w macierzy (Rycina 28).



Rycina 28 Działanie wypełnienia (ang. padding)

Kolejnym hiperparametrem jest liczba zastosowanych filtrów. Każdy dodatkowy filtr powoduje powstanie dodatkowego kanału/warstwy macierzy, a wynik może być dostosowany do innego wzoru czy cechy w efekcie sieć może lepiej dostosować się do zadanych danych. Dylatacja to metoda polegająca na powiększeniu filtra przez umieszczenie w nim ciągów zer przed przeprowadzeniem konwolucji. Celem dylatacji jest zwiększenie pola recepcyjnego, w taki sposób by wychwycić bardziej globalne/uogólnione cechy. Ostatnim i jednocześnie bardzo istotnym elementem jest głębokość sieci konwolucyjnej tzn. jak wiele warstw konwolucyjnych jest ze sobą połączonych. W ten sposób powstaje pewna hierarchiczność w budowaniu mapy cech. Wspomniana wcześniej warstwa łącząca ma za zadanie skompresować dane uzyskane w

poprzedniej. Mechanizm działania jest podobny jak w warstwie konwolucyjnej przez przemieszczane się filtru po danych. Różnica polega na zastosowanej funkcji tzn. może być to średnia, minimalna lub największa wartość z danego obszaru. Powszechne jest stosowanie funkcji wybierającej wartość największą (ang. Max Pooling Layer). Kolejnym elementem jest funkcja aktywacji, która została opisana powyżej.. Ostatecznie w modelu sieci konwolucyjnej, której funkcja miałaby być klasyfikatorem dane trafiają do warstwy zbudowanej z jednokierunkowej sieci neuronowej, której ostatnie węzły odpowiadają zadanym klasom/kategoriom.

Sieci neuronowe typu transformers

Sieci neuronowe typu transformers osiągnęły ogromną wydajność nie tylko w przetwarzaniu naturalnego języka, do czego zostały zaprojektowane, ale również w widzeniu komputerowym i rozpoznawaniu mowy. Historia transformers rozpoczyna się w 2017 roku dzięki naukowcom z Google.[343] Przed ich pojawieniem się do przetwarzania tekstu używano głównie sieci opartych o architekturę rekurencyjnych sieci neuronowych, czy long-short-term memory (LSTM) [343]. Aczkolwiek w kontekście wyników, transformers znacząco je przewyższyły zarówno pod względem jakości, jak i kosztów uczenia [344]. Rekurencyjne sieci neuronowe są wrażliwe na problem zanikającego i eksplodującego gradientu [345]. Problem ten próbowano rozwiązać przy pomocy architektury LSTM [346], która wykorzystuje bloki pamięci. Taki blok zawiera komórki pamięci, które mają możliwość zapamiętywania tymczasowych stanów sieci. Przepływ informacji kontrolowany jest przez jednostki bramkujące (ang. gated units) [346]. Obie te architektury charakteryzuje tzw. zależność długoterminowa. W ogólnym zarysie na przykładzie analizy tekstu jeżeli mamy bardzo długi dokument, a istotne jest połączenie między pierwszym i ostatnim wyrazem to należy zakodować wszystkie słowa pomiędzy, nawet jeżeli są one nieistotne [347]. W efekcie nie ma możliwości zastosowania paralelizacji, czyli wykonywania kilku procesów na raz, a czas obliczenia wzrasta w raz z długością sekwencji [347]. Architektura typu transformers bazuje na zasadzie enkoder-dekoder [344]. Enkoder całą wprowadzoną sekwencję np. zdanie w języku angielskim koduje uzyskując liczbową reprezentację znaczenia sekwencji, a następnie w kolejnym etapie dekoder zamienia tę sekwencję na oczekiwany wynik np. zdanie w języku polskim [343]. Taka architektura

niestety wiąże się z wystąpieniem problemu wąskiego gardła tzn. najpierw zdanie musi być zakodowane, żeby je dekodować [344]. W transformers jednak rozwiązano ten problem przez zastosowanie mechanizmu nazwanego atencją/uwagą (ang. attention) [344]. Problem tzw. „wąskiego gardła" będący wynikiem użycia pojedynczego ukrytego stanu, co dotyczy typowej architektury enkoder-dekoder zastąpiono sytuacją, w której enkoder udostępnia dekoderowi stany ukryte dla każdego kroku [344]. Taki stan wiąże się z dużą ilością danych przechodzących pomiędzy enkoderem, a dekoderem, dlatego potrzebny był mechanizm priorytetujący – jest nim właśnie atencja wprowadzona do sieci neuronowych typu transformers [344]. Twórcy tej architektury opisują ten mechanizm jako mapowanie zapytania (ang. query) i zbioru par klucz-wartość (ang. key-value pair) do danych wyjściowych, gdzie zapytanie, klucz, wartość i dane wyjściowe są wektorami. Dane wyjściowe są wynikiem sumy ważonej, gdzie wagi sumy obliczane są przez funkcję kompatybilności zapytania do korespondującego klucza [343]. Autorzy Transformers funkcję uwagi nazywają dokładniej uwagą skalowanego iloczynu skalarnego (ang. Scaled Dot-Product Attention), ponieważ dane wejściowe składające się z zapytania i kluczy o wymiarach dk, a wartości o wymiarach dv. Iloczyn skalarny obliczany jest na podstawie zapytań i kluczy, a następnie dzielony przez √dk. Następnie przez zastosowanie funkcji softmax uzyskiwane są wagi dla wartości [343]. Dodatkowo zapytanie, klucz i wartość w transformers dzielone są na człony i analizowane jednoczasowo w wieloczłonowej warstwie uwagi. (ang. multi-head attention layer) [343]. Funkcja samouwagi (ang. self-attention) pozwala na przeanalizowanie elementów całej sekwencji między sobą i priorytetyzację znaczenia, a model sieci neuronowej w ten sposób „skupia uwagę" na istotnych elementach [347]. Podobnie ludzkie oko odbiera ogromną ilość danych w ciągu sekundy, ale nie analizuje ich w całości. William James (psycholog) mechanizm biologicznej uwagi rozdziela na dwa komponenty dobrowolny (ang. voluntary) i mimowolny (ang. involuntary), które wpływają na ostateczny odbiór bodźców. Mimowolny oparty jest na atrakcyjności i zauważalności bodźca, a dobrowolny na celowym skupieniu uwagi na bodźcu np. konkretnym obiekcie, na którym chcemy spojrzeć. Zdolność do skupienia uwagi np. na nieoczekiwanym zagrożeniu, albo na czynności pozwoliła ludziom na rozwój i przetrwanie [348]. Zarówno warstwa enkodera, jak i dekodera składa się z kilku wieloczłonowych mechanizmów uwagi połączonych warstwą jednokierunkową (ang. feed-forward network). Danymi wejściowymi dla pierwotnego zastosowania

transformers jest tekst. Tekst w procesie tokenizacji zamieniany jest na tokeny. Najprostszy sposób na uzyskanie tokenów z sekwencji wyrazów jest podział względem znaków przestankowych i odstępów [344]. W przeciwieństwie do sieci rekurencyjnych, które dane wejściowe podają analizie w sposób sekwencyjny (token po tokenie) w transformers zastosowano warstwę kodująca pozycję (ang. positional encoding layer) [343]. Tokeny zamieniane są na wektory kodujące znaczenie, a dokładnej zależność pomiędzy innymi tokenami w warstwie osadzenia (ang. embedding layer) [347]. Efekt tego procesu można w dużym przybliżeniu zobrazować jako wzór: król – mężczyzna + kobieta = królowa [349]. Od czasu opracowania architektury transformers powstało dużo jej modyfikacji. Możemy wyróżnić te składające się, jak pierwowzór z enkodera i dekodera. Przykładem jest Bart, a tego typu sieci wciąż głównie wykorzystywane są np. w tłumaczeniu tekstu. Innym przykładem są te składające się tylko z warstwy dekodera, którą można wykorzystać do automatycznego uzupełniania/generowania treści, jak to jest w modelu GPT. Modyfikacje transformers składające się tylko z enkodera dobrze sprawdzają się w zadaniach opartych na klasyfikacji [344]. Powyżej omówiono pierwotne zastosowanie architektury transformers, aczkolwiek model ten sprawdza się również w innych zastosowaniach np. siec neuronowa ViTs rozpoznaje obrazy [350]. Proces tokenizacji zastąpiono podziałem obrazu na fragmenty (ang. patches), a sama architektura opracowana została przez „Google Research, Brain Team” [350].

Cel pracy

Celem tej rozprawy jest ocena czy architektura sieci neuronowej oparta o kombinację sieci neuronowej konwolucyjną i typu transformers (ang. convolutional-transformers neural networks), która została przetrenowana z użyciem syntetycznych danych, jest porównywalnym lub lepszym od standardowych metod narzędziem do analizy metylacji DNA.

Materiały i metodyka

Ogólny opis narzędzia

Przedmiotem badania tej rozprawy jest sieć neuronowa oparta o sieć konwolucyjną i sieć typu transformers, którą przetrenowano z użyciem danych syntetycznych. Celem działania tego narzędzia jest wskazanie w badanych próbkach odmiennie metylowanych sekwencji CpG o możliwie największym znaczeniu biologicznym. Metoda ta jest częścią opracowanej przez autora biblioteki CTMeth służącej do analizy metylacji. Biblioteka CTMeth napisana jest w języku programowania Python [351], a w jej skład wchodzą dodatkowe narzędzia takie jak moduł do analizy głównych składowych, klastrowania hierarchicznego i autorski algorytm do analizy interakcji między genami powiązanymi z sekwencjami CpG (CpG-gene-gene-CpG algorithm). Opracowana biblioteka oraz pełne wyniki uzyskane przy jej pomocy znajdują się pod adresem https://www.ctmeth.com.

Ogólny opis procedury przetwarzania danych

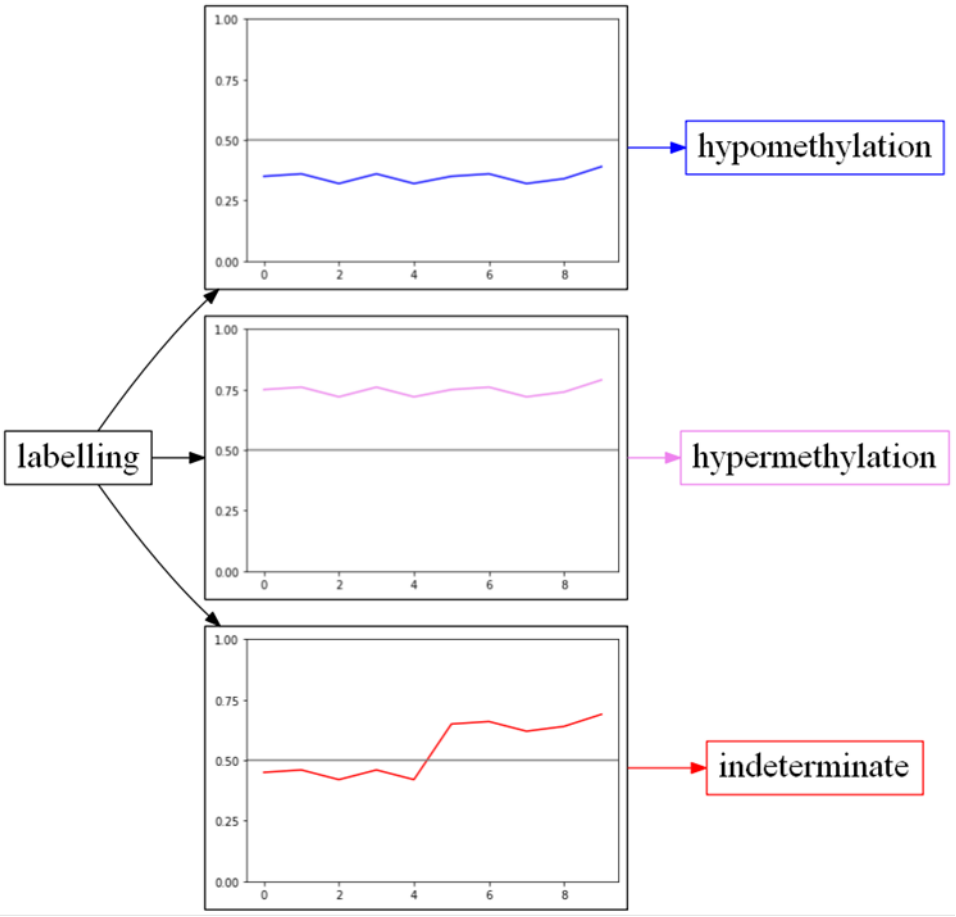
Opracowane narzędzie, które jest tematem tej rozprawy wymaga danych pierwotnie przetworzonych z surowych danych z pliku idat na wartości β w pliku comma-separated values files (CSV). Parametry analizy ustawiane są dla ułatwienia i reprodukcyjności badań w oddzielnym pliku YAML [352]. Jest to typowy format wykorzystywany w plikach konfiguracyjnych i zgodny z human-readable data-serialization language. Danymi wejściowymi dla algorytmu jest tabela danych zawierające wartości β. Wiersze reprezentują poszczególne sekwencje CpG, a kolumny próbki. W wspomnianym pliku YAML umieszczane są informacje o podziale próbek na

grupę kontrolną i badaną. Skrypt z użyciem opisanej sieci neuronowej klasyfikuje do jednej z trzech skwantyfikowanych etykiet (0, 1, 2) każdą sekwencję CpG oddzielnie dla grupy kontrolnej i badanej. Etykiety odpowiadają trzem stanom metylacji (hipermetylowane, hipometylowane, stan nieokreślony/częściowo metylowany) (Rycina 29). Etykieta nieokreślony nadawana jest sekwencjom CpG, których ocena pod względem statusu metylacji w grupie nie może być jasno zdefiniowana. W kolejnym etapie następuje odfiltrowanie odmiennie sklasyfikowanych sekwencji pomiędzy grupą kontrolną i badaną zgodnie z jednym z dwóch wariantów – CTMeth-hh i CTMeth-hhi. Pierwszy wariant wskazuje CpG o odmiennych etykietach w kontekście hipermetylacja i hipometylacja z pominięciem tych, które zostały ocenione jako nieokreślone (CTMeth-hh). Drugi wariant z kolei jako wynik wskazuje również te sekwencje, w których dla danej grupy nadano etykietę „nieokreślony" np. hipermetylowana grupa kontrolna i nieokreślona grupa badawcza (CTMeth-hhi). Dane wyjściowe (wyniki) zawierają wybrane sekwencje CpG, wartości β poszczególnych próbek oraz stopień ufności sieci, czyli parametr oceniający jak bardzo dane pasują do danej kategorii wg. opracowanego algorytmu(Rycina 30). Dane uzyskane w ten sposób można w prosty sposób poddać dalszej analizie z wykorzystaniem innych metod, jak te, które stanowią dodatkowe moduły biblioteki CTMeth.

Architektura zastosowanej sieci neuronowej

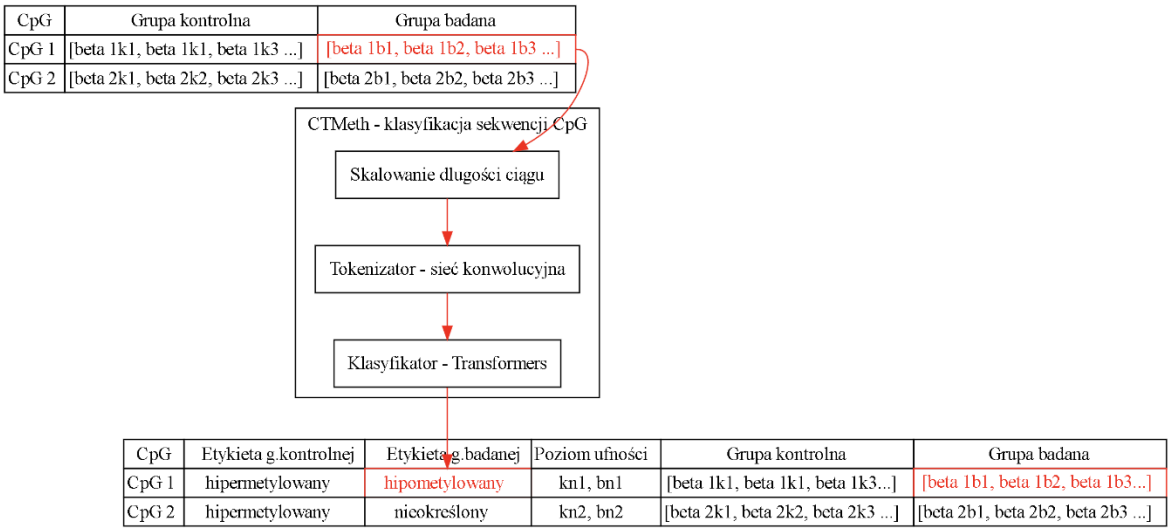
Architektura użytej sieci neuronowej oparta jest o sieć konwolucyjną oraz sieć typu transformers. Całość opracowana została przy użyciu PyTorch, będącym jednym z dwóch najczęściej wykorzystywanych środowisk w prototypowaniu sieci neuronowych. Sieć transformers zastosowana w opracowanym narzędziu pełni funkcję klasyfikatora i wymaga zamiany danych wejściowych na tokeny, czyli podzieleniu danych wejściowych na podstawowe jednostki. Za tę wymaganą zamianę sekwencji wartości β dla danego CpG na odpowiednie tokeny w przypadku zastosowanej sieci neuronowej odpowiada sieć konwolucyjna. Sieć konwolucyjna składa się z 1-wymiarowej warstwy konwolucyjnej, warstwy łączącej oraz funkcji ReLu, a dane przed wprowadzeniem do niej są ekstrapolowane, tak by składały się z ciągu 100 elementów (100 próbek). Klasyfikator, czyli sieć transformers została zmodyfikowana i składa się tylko z warstwy enkodera, a sam enkoder składa się z dwóch warstw enkodujących, dwóch

wieloczłonowych warstw uwagi oraz 64-wymiarowej sieci jednokierunkowej. Warstwa kodująca pozycję dodana jest przed enkoder i może ulegać optymalizacji w trakcie treningu. Jako prewencja problemu nadmiernego dopasowania zastosowano dodatkowo warstwę dropout



.

Rycina 29 Wizualizacja sposobu etykietowania sekwencji metylacji przez sieć neuronową



Rycina 30 Ogólny schemat działania narzędzia CTMeth

Trening sieci neuronowej

Trening sieci neuronowej przeprowadzono z użycie syntetycznych danych generowanych w czasie rzeczywistym w trakcie uczenia przez dodatkowy algorytm opisany w dalsze częsci rozprawy. Natomiast proces klasycznej ewaluacji po treningu przeprowadzono na podstawie ręcznie stworzonego słownika przedstawiającego dane w sposób zgeneralizowany (

Tabela 1). Należy zaznaczyć, że proces klasycznej ewaluacji w przypadku opracowanego narzędzia służy jedynie do oceny zdolności uniwersalnej aproksymacji sieci neuronowej i nie jest ostatecznym sprawdzeniem jej skuteczności i porównaniem do innych metod, co opisane w dalszej części rozprawy. Algorytm użyty w trakcie treningu i generujący syntetyczne ciągi wartości β użyte do jego przeprowadzenia tworzy je na podstawie wybranych zeskalowanych rozkładów danych (Tabela 2) odpowiednio dla jednej sekwencji CpG wraz ze zgodną skwantyfikowaną etykietą (0, 1, 2). Wygenerowany ciąg wartości zostaje następnie wzbogacony o zmienności o charakterze stochastycznym, co wprowadza do danych element naturalnej zmienności (szum), oraz maksymalnie do 10% wartości zamienianych jest na wartości bliskie skrajnym, które mogą nie spełniać zakresów zgodnych z przyjętymi założeniami dla danej etykiety, co z kolei odpowiadać ma tzw. outliers/spikes.

Tabela 1 Słownik zgeneralizowanych i ogólnych wartości sekwencji CpG z etykietami. Słownik użyto do etapu ewaluacji uczenia sieci neuronowej.

	G1	G2	G3	G4	Label
cg1	1	1	1	1	0
cg2	0.9	0.9	0.9	0.1	0
cg3	0.8	0.8	0.8	0.2	0
cg4	0.7	0.7	0.7	0.4	0
cg5	0.6	0.6	0.6	0.5	0
cg6	0.5	0.5	0.5	0.5	2
cg6	0.5	0.5	0.5	0.5	2
cg6	0.5	0.5	0.5	0.5	2
cg6	0.5	0.5	0.5	0.5	2
cg7	0.4	0.4	0.4	0.5	1
cg8	0.3	0.3	0.3	0.5	1
cg9	0.2	0.2	0.2	0.5	1
cg10	0.4	0.4	0.4	0.5	1

Tabela 2 Użyte rozkłady wartości w generatorze syntetycznych sekwencji β

Rozkład ciągły	Rozkład dyskretny
Rozkład Laplace'a	Rozkład Poissona
Rozkład normalny	Rozkład Beta-binomialny
Rozkład jednostajny	Rozkład Hyper-geom
Skośny rozkład normalny	
Rozkład n-modalny	
Rozkład Weibulla	

Przed przekazaniem ciągu do funkcji treningowej algorytm generujący sprawdza, czy wartości β spełniają następujące kryteria:

1. >80% spełnia kryteria dla zadanej etykiety (np. dla etykiety hipermetylowany warunek dla wartości β próbki to >0.5 lub odwrotnie <0.5 dla hipometylowany)
2. Żadna z wartości nie jest >1 lub <0.
3. Dla wartości nieokreślonych stosunek hipermetylowanych do hipometylowanych próbek jest zbliżony do 0.5

W przypadku niespełniania kryteriów, ciąg wartości β nie jest używany do treningu. Do inicjowania parametrów użyto metody Kaiming. Po każdej fazie uczenia następuje ewaluacja z wyżej opisanym słownikiem. Proces uczenia uległ zakończeniu, gdy

trafność (ang. accuracy) podczas treningu przekroczyła 90%, a podczas ewaluacji wynosiła 100%. Podczas ewaluacji nie następuje modyfikacja parametrów, jak wspomniano celem jest sprawdzenie zdolności do uniwersalnej aproksymacji. Podczas treningu zastosowano One Cycle Scheduler jako manipulator wielkości współczynnika uczenia rozpoczynając od współczynnika uczenia 0.0005 i maksymalny ustalając na 0.000005.

Ocena skuteczności metod w analizie metylacji

Metodę z użyciem sieci neuronowej porównano do dwóch standardowych metod używanych w analizie metylacji – liniowej regresji zaimplementowanej w powszechnie używanej bibliotece ChAMP[184] [raz różnicy w średniej metylacji (metoda delta) potwierdzonej testem statystycznym T-Studenta (wartość $p<0.05$). Metody porównano z użyciem następujących zbiorów danych zawierających sekwencje CpG podzielone na grupę kontrolną i badawczą:

1. Metylacja DNA limfocytów B oraz limfocytów T CD4+ (**B-Cell-CD4+)**
2. Metylacja DNA zdrowych limfocytów B i limfocytów B pacjentów chorujących na przewlekłą białaczkę limfatyczną (**B-Cell-CLL**)
3. Metylację DNA pacjentów z przewlekłą białaczką limfocytową z IGHV 100% i mniej (**CLL-100**)

Wspomniany parametr IGHV to fundamentalny czynnik różnicujący podstawowe biologiczno-kliniczne podtypy przewlekłej białaczki limfocytowej. Określa on obecność lub brak hipermutacji somatycznej rejonu zmiennego ciężkiego łańcucha immunoglobulin (IGHV, immunoglobulin heavy chain variable) [353]. Zbiory danych wybrane do testów pozyskano z publicznej bazy danych Gene Expression Omnibus z dostępów (ang. accession) - GSE110554 [354], GSE136724 [355]. Zbiór pierwszy pochodzi z dostępu GSE110554 [354], zbiór drugi jest połączeniem danych z GSE136724 [355] oraz GSE110554 stanowiących odpowiednio grupę kontrolną i badawczą. Zbiór trzeci uzyskano z dostępu GSE136724 [355]. Zbiory danych do analizy dobrano pod względem rosnącej złożoności, porównując jednocześnie trzy sytuacje badawcze:

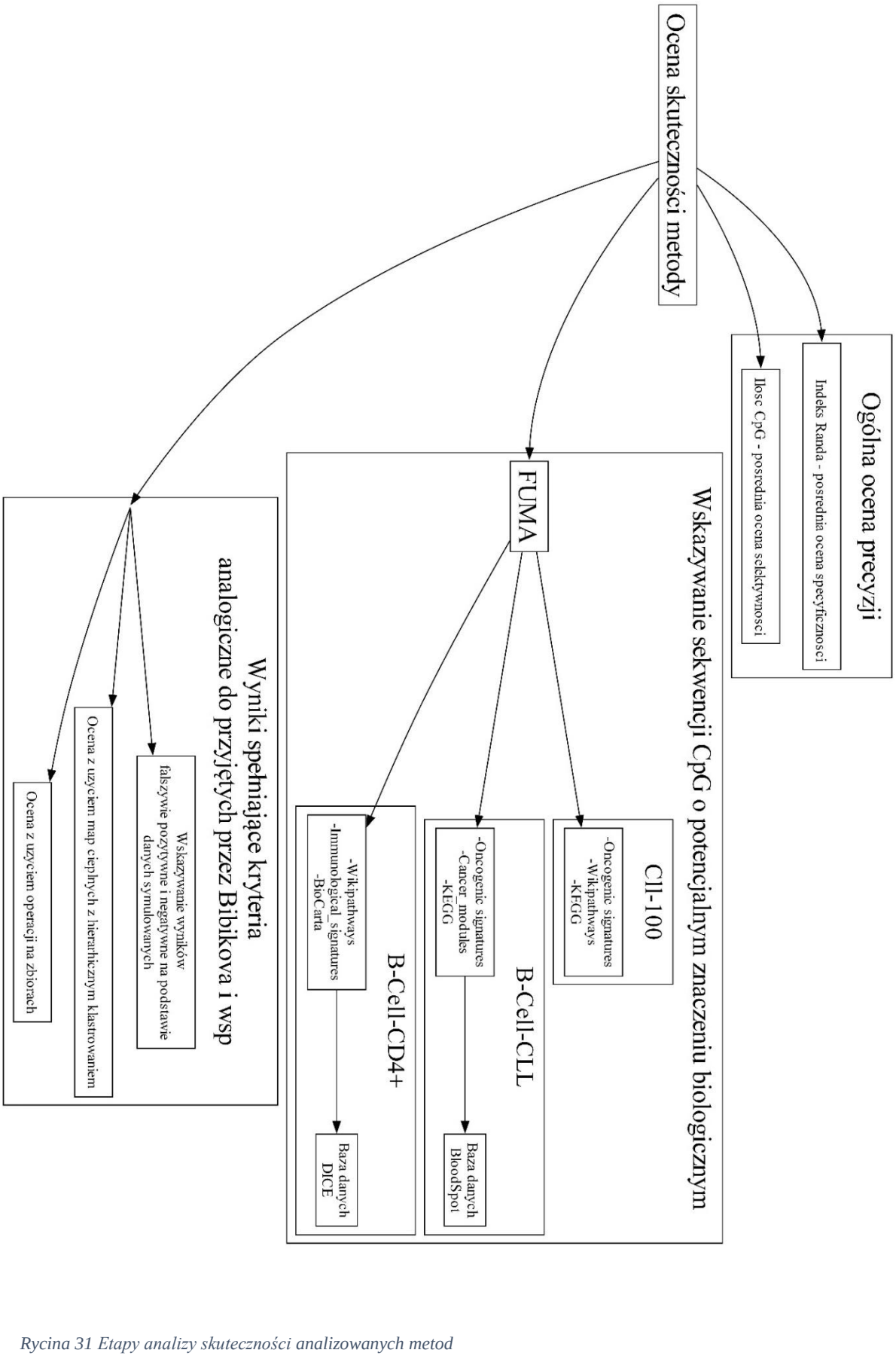
1. Symetryczny podział grup z przewidywanymi klarownymi różnicami w metylacj (B-Cell-CD4+).
2. Przewidywane klarowne różnice w metylacji przy asymetrycznym podziale na grupy – jedna z grup badanych znacząco przewyższa drugą pod względem ilości próbek(B-Cell-CLL).
3. Złożony heterogeniczny zbiór danych – niewielkie różnice w metylacji, obecność artefaktów(CLL-100).

Etapy analizy skuteczności metod i ich znaczenie

Analizę skuteczności metod i ich znaczenia przeprowadzono w kierunku:

1. Ilości wskazywanych odmiennie metylowanych sekwencji CpG – ocena selektywności
2. Zdolności do wskazywania optymalnej liczby sekwencji CpG pozwalających na utrzymanie różnicowania na grupę kontrolną i badaną – ocena specyficzności
3. Zdolności do wskazywania sekwencji CpG o potencjalnym znaczeniu biologicznym
4. Zdolności do uzyskania wyników spełniających kryteria analogiczne do przyjętych przez Bibikova i wsp.[159].
5. Oceny wskazywania przez metody wyników fałszywie pozytywnych i negatywnych na podstawie danych symulowanych

Etapy analizy przedstawiono na Rycina 31.

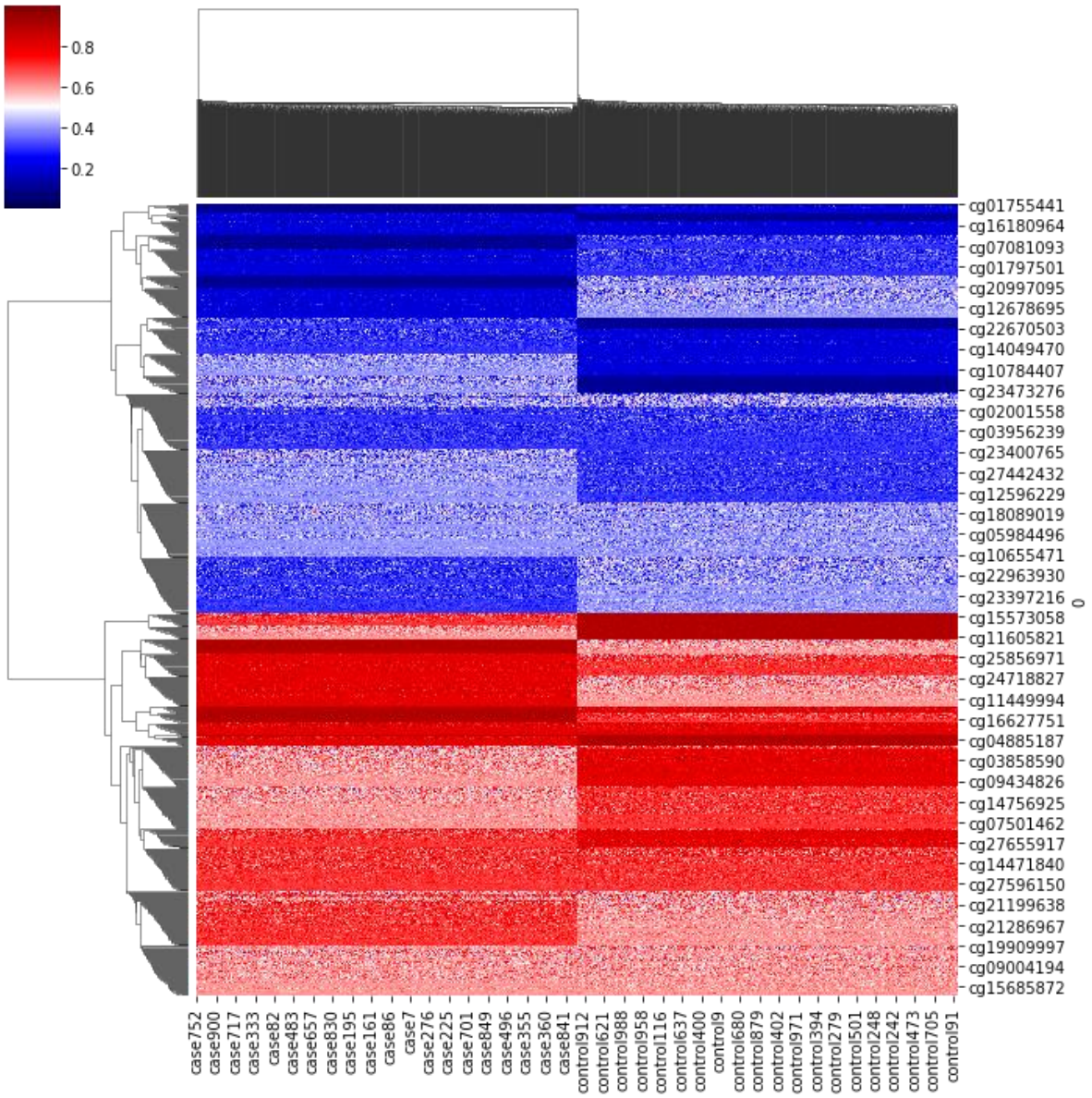


Rycina 31 Etapy analizy skuteczności analizowanych metod

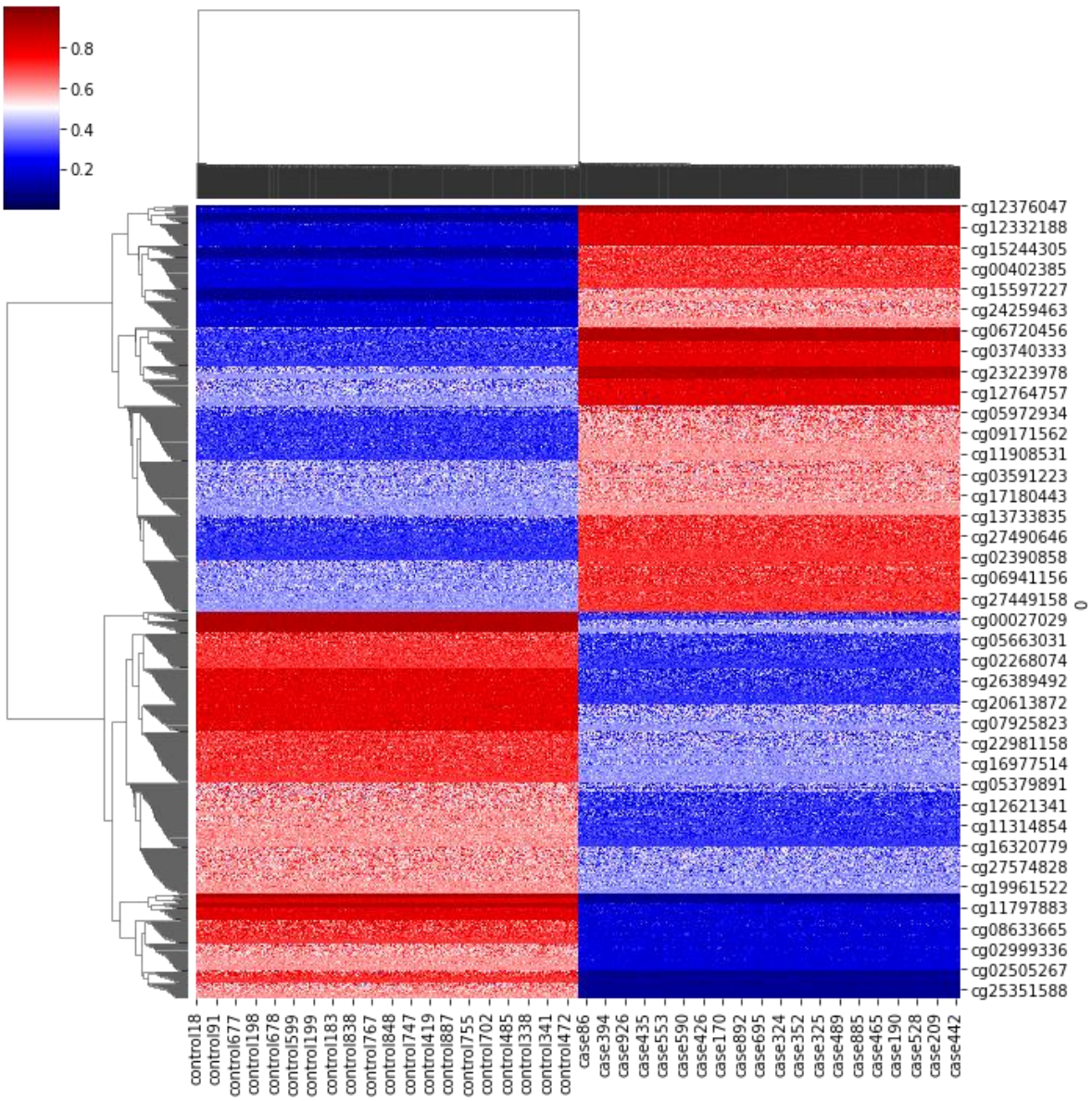
Metodę opartą o sieć neuronową oraz metody standardowe w pierwszej kolejności porównano pod względem wielkości puli sekwencji CpG wskazanych jako różnicujące grupę kontrolną i badawczą – pośrednia ocena selektywności. Kolejnym etapem porównania była ocena specyficzności, której dokonano analizując skuteczność klastrowania z użyciem wskaźnika Randa [356], który jest popularnym wskaźnikiem wydajności algorytmów klasteryzacji porównującym uzyskany podział z prawdziwym podziałem danych – w tym przypadku realnym podziałem na grupę kontrolną i badawczą. W przypadku tej rozprawy wszystkie wyniki uzyskane przez porównywane ze sobą metody zostały poddane grupowaniu hierarchicznemu z użyciem tego samego i powszechnie uznawanego algorytmu (algorytm grupowania metodą Warda), dlatego w efekcie ocenie podlegają sekwencje CpG pod względem jakości różnicowania na grupę kontrolną i badawczą, a w konsekwencji porównywana jest opracowana metoda do metod klasycznych pod względem zdolności do znajdowania cech (sekwencji CpG) dobrze różnicujących analizowane grupy próbek. W kolejnym etapie analizy skuteczności badanych metod uzyskane za ich pomocą wyniki przeanalizowano w kontekście potencjalnego znaczenia biologicznego. Znaczenie biologiczne jest pojęciem bardzo szerokim i na pewno wielowymiarowym, które może być trudne do precyzyjnego zdefiniowania. W zainteresowaniu tej rozprawy leży jednak konkretny aspekt tego znaczenia, mianowicie jak zmienność metylacji wpływa na ekspresję genów i czy wybrane metody wskazują te sekwencje CpG, które niosą ze sobą informację o konkretnych funkcjach biologicznych, różnicach w ekpresji genów, których można się spodziewać w wybranych próbkach. Do tego celu użyto platformy FUMA (ang. Functional Mapping and Annotation of Genome-Wide Association Studies), oraz dodatkowo z Database of Immune Cell Expression, Expression quantitative trait loci (eQTLs) and Epigenomics (DICE) [357] i BloodSpot [358]. Dana sekwencja CpG może być anotowana (powiązana) z określonym genem. Powiązanie do danego genu uzyskano z opracowanego przez firmę Illumina pliku manifest. Anotowanie genu nie jest jednoznacznym wykładnikiem funkcjonalnego powiązania genu z daną sekwencją CpG, aczkolwiek wskazuje potencjalne miejsce interakcji. W ten sposób uzyskano zestawy genów związanych z odmiennie metylowanymi sekwencjami CpG wskazywanymi przez porównywane metody. Utworzone listy genów wprowadzono do platformy FUMA uzyskując przy pomocy algorytmu GENE2FUNC listę biologicznych szlaków i funkcji, z którymi te zbiory genów się najlepiej pokrywają. Celem tego etapu

jest ocena znaczenia biologicznego (informacji biologicznej) wskazywanych przez porównywane metody sekwencji CpG poprzez to, jak anotowane do nich geny wpasowują się w odpowiednie szlaki genetyczne. Wyniki uzyskane ze zbioru sekwencji CpG B-Cell-CD4+ porównano na podstawie zestawów genów z kategorii Wikipathways i Immunological Signatures, oraz BioCarta. Wikipathways, jest to otwarta, międzynarodowa baza danych zawierająca informacje o tym, jakie geny wchodzą w skład poszczególnych szlaków. Natomiast kategoria Immunological Signatures gromadzi dane na temat sygnatur immunologicznych, czyli wzorców ekspresji genów, które charakteryzują specyficzne stany lub odpowiedzi układu odpornościowego. W tym przypadku uzyskane wyniki dodatkowo skorelowano z bazą danych DICE [357]), która zawiera różnice w ekspresji genów poszczególnych komórek układu odpornościowego. Rezultaty z analiz kolejnego zbioru sekwencji CpG, czyli B-Cell-CLL porównano w kontekście kategorii związanych z chorobami nowotworowymi tj. Oncogenic Signatures, Cancer Modules oraz dużego zbioru Kegg Pathways zawierającego liczne zbiory genów dotyczące istotnych szlaków biologicznych. Wyniki z tego porównania skorelowano z narzędziem BloodSpot [358]. Ze względu na największą złożoność zbioru CLL-100 do analizy wybrano zarówno Wikipathways i KEGG Pathways, czyli kategorie dotyczące szlaków biologicznych, jak również Oncogenic Signatures dotyczące chorób nowotworowych. Następny etap analizy skuteczności metod to porównanie z użyciem podstawowych operacji na zbiorach oraz z użyciem hierarchicznie sklastrowanych względem sekwencji CpG map cieplnych badając, które sekwencje CpG w wynikach poszczególnych metod pokrywają się ze sobą, a które występują tylko w wynikach danej metody i w jakim stopniu odpowiadają one zadanemu kryterium dla hiper- i hipometylacji analogicznemu do przyjętego przez Bibikova i wsp.[159]. Również istotnym pod względem badawczym jest to, jak metoda sprawdza się w kontekście wskazywania fałszywie dodatnich i fałszywie ujemnych wyników. Przeprowadzono taką analizę z użyciem danych symulowanych. Ocenę przeprowadzono pod względem różnicowania na hipermetylowane sekwencje i hipometylowane sekwencje w analogiczny sposób, jak opisano wyżej (rozdz. Interpretacja wartości β) a za punkt odcięcia wartości β pomiedzy wartością hipermetylacji i hipometylacji uznano 0.5. Dane symulowane uzyskano jako losowe wartości będące pochodnymi rozkładu normalnego i zeskalowanymi do przedziału od 0 do 1 ,czyli zgodnego z zakresem wartości β. Uzyskano w ten sposób

dwa zbiory danych, w których w jednym grupy kontrolna i badawcza nie różniły się pod względem stanu metylacji (grupa fałszywie dodatnia) oraz w drugim gdzie ta różnica występowała (grupa fałszywie ujemna). Oba zbiory danych testowych składały się z 1000 sekwencji CpG dla 1000 próbek kontrolnych i 1000 próbek badanych. Obie grupy fałszywie dodatnia i ujemna przedstawiono odpowiednio na Rycina 32 i Rycina 33.



Rycina 32 Grupowanie hierarchiczne symulowany danych. Zbiór fałszywie dodatni.



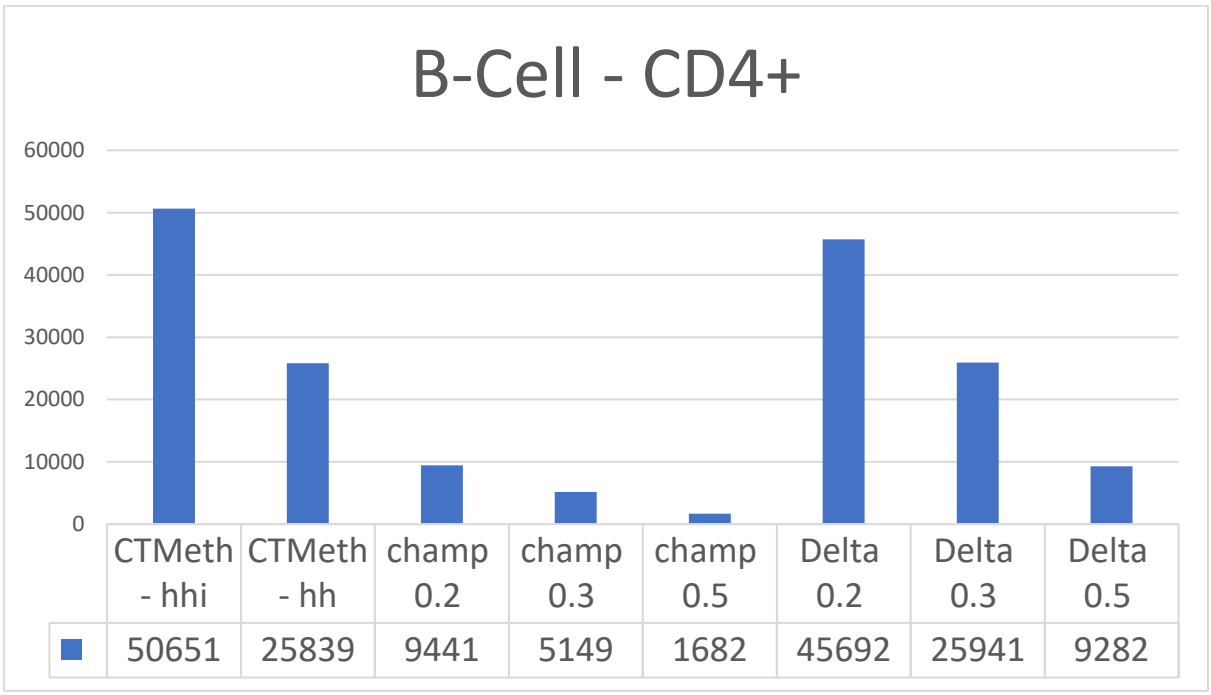
Rycina 33 Grupowanie hierarchiczne symulowany danych. Zbiór fałszywie ujemny.

Wyniki

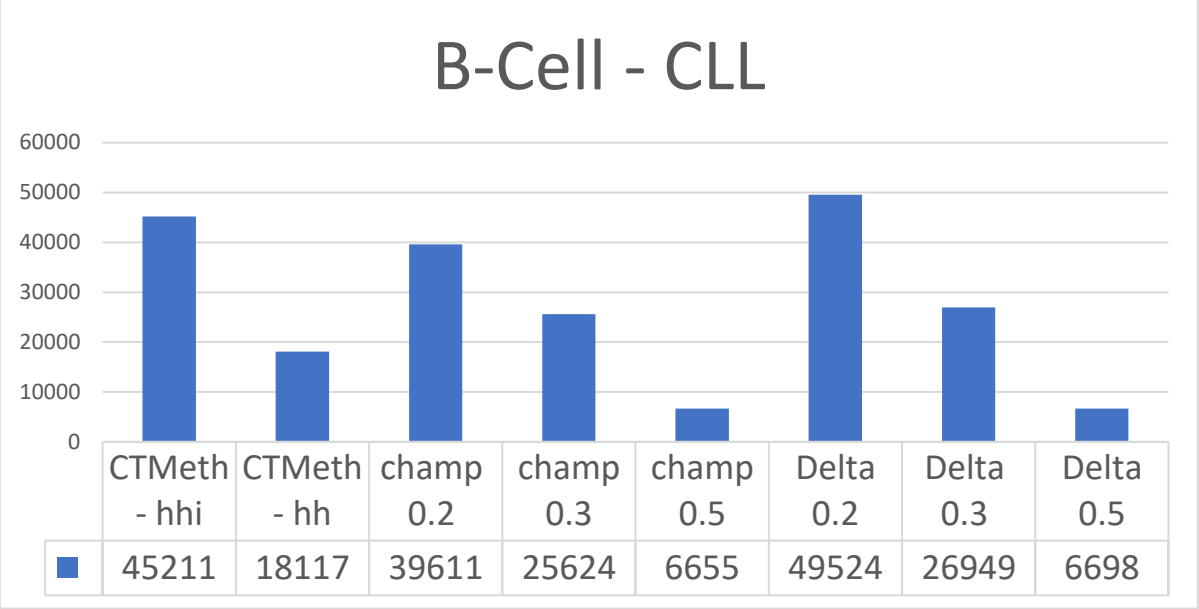
Liczba sekwencji CpG – ocena selektywności

Pełną analizę puli sekwencji CpG odmiennie metylowanych pomiędzy grupami kontrolnymi i badanymi dla wybranych zbiorów danych znalezionych przez poszczególne metodyprzedstawiono na Rycina 34, Rycina 35 i Rycina 36. Dystrybucja

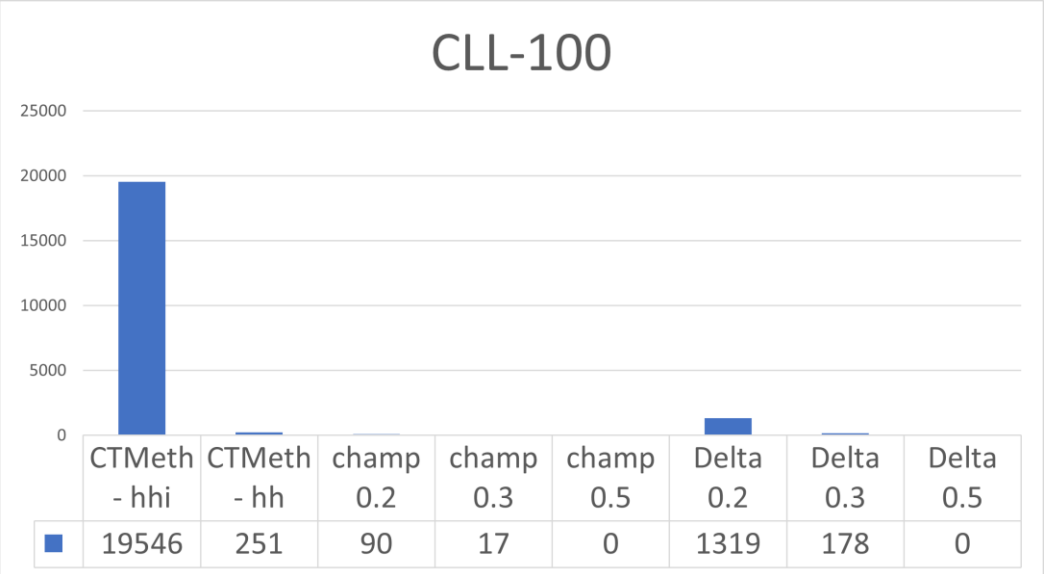
wyników dla grup B-Cell-CD4+ i B-Cell-CLL była porównywalna, pomiędzy zestawami analizowanych danych (tj. B-Cell-CD4+ a B-Cell-CLL), aczkolwiek różnicująca pomiędzy poszczególnymi metodami. W analizie zbioru CLL-100 dwie metody (ChAMP 0.5 i delta 0.5) nie wskazały żadnego CpG, a CTMeth-hhi wskazała ich znacznie większą liczbę w stosunku do innych metod. Pełna lista sekwencji CpG ze względu na swoją obszerność znajduje się pod adresem url: https://ctmeth.com.



Rycina 34 Liczba sekwencji CpG oznaczonych jako różnicujące grupę kontrolną i badaną przez wybrane metody w zbiorze B-Cell-CD4+



Rycina 35 Liczba sekwencji CpG oznaczonych jako różnicujące grupę kontrolną i badaną przez wybrane metody w zbiorze B-Cell-CLL



Rycina 36 Liczba sekwencji CpG oznaczonych jako różnicujące grupę kontrolną i badaną przez wybrane metody w zbiorze CLL-100

Wydajność klastrowania – ocena specyficzności

Dla zbioru B-Cell-CD4+ (Tabela 3) wszystkie metody wskazują sekwencje CpG pozwalające na uzyskanie grupowania hierarchicznego na podobnym poziomie ze współczynnikiem Randa 0,846154. Najmniejszą liczbę CpG wskazuje metoda ChAMP

0.5, a największą CTMeth–hhi. Inaczej prezentuje się analiza wyników pozyskanych ze zbioru B-Cell-CLL, gdzie najmniej sekwencji CpG wskazała metoda ChAMP 0.5, a najwięcej delta 0.2. Z kolei niższą niż inne metody wartość indeksu Randa uzyskała metoda CTMeth w wersji hh (Tabela 4). Jeszcze inaczej prezentują się wyniki dla najbardziej złożonego zbioru, czyli CLL-100 (Tabela 5). W tym przypadku największy współczynnik Randa osiągnęły metody CTMeth-hh i CTMeth-hhi. Metoda ChAMP 0.5 wraz z metodą delta 0.5 osiągnęły wynik 0, ze względu na brak sekwencji CpG wskazywanych jako odmiennie metylowane pomiędzy grupą kontrolną a badaną.

Tabela 3 Wydajność klastrowania dla B-Cell - CD4+

	indeks Randa	Liczba CpG
ChAMP 0.2 B-Cell vs CD4+	0.846154	9441
ChAMP 0.3 B-Cell vs CD4+	0.846154	5149
ChAMP 0.5 B-Cell vs CD4+	0.846154	1682
Delta 0.2 B-Cell vs CD4+	0.846154	45692
Delta 0.3 B-Cell vs CD4+	0.846154	25941
Delta 0.5 B-Cell vs CD4+	0.846154	9282
CTMeth B-Cell vs CD4+ - hh	0.846154	25839
CTMeth B-Cell vs CD4+ - hhi	0.846154	50651

Tabela 4 Wydajność klastrowania dla B-Cell-CLL

	Indeks Randa	Liczba CpG
ChAMP 0.2 B-cell - CLL	0.974359	39611
ChAMP 0.3 B-cell - CLL	0.974359	25624
ChAMP 0.5 B-cell - CLL	0.974359	6655
Delta 0.2 B-cell - CLL	0.974359	49524
Delta 0.3 B-cell - CLL	0.974359	26949
Delta 0.5 B-cell - CLL	0.974359	6698
CTMeth - hh	0.925075	18117
CTMeth - hhi	0.974359	45211

Tabela 5 Wydajność klastrowania dla CLL-100

	indeks Randa	Liczba CpG
ChAMP 0.2 CLL-100	0.580986	90
ChAMP 0.3 CLL-100	0.649452	17
ChAMP 0.5 CLL-100	0	0
Delta 0.2 CLL-100	0.634194	1319
Delta 0.3 CLL-100	0.619718	178
Delta 0.5 CLL-100	0	0
CTMeth CLL-100 - hh	0.665493	251
CTMeth CLL-100 – hhi	0.665493	19546

FUMA – Zdolność do wskazywania sekwencji CpG o potencjalnym znaczeniu biologicznym

B-Cell CD4+

Zbiór B-Cell-CD4+ przeanalizowano z algorytmem FUMA w kontekście następujących kategorii: Wikipathways, Immunological Signatures i BioCarta. Dla tego zbioru wśród metod, które wskazały sekwencje CpG, do których najwięcej anotowanych genów było powiązanych ze szlakami biologicznymi z Wikipathways znalazły się CTMeth-hhi i delta 0.2. Natomiast dominujące zestawienia genów Wikipathways to: PI3K-Akt Signaling Pathway, Focal Adhesion-PI3K-Akt-mTOR-signaling pathway, Nuclear Receptors Meta-Pathway, VEGFA-VEGFR2 Signaling Pathway, MAPK Signaling Pathway (Tabela 6). Ze względu na obszerność wymienionych zbiorów oraz ich biologiczne powiązanie, zwłaszcza w kontekście analizowanych komórek wybrano te geny, które pokrywają się pomiędzy wyżej wymienionymi szlakami. Do znalezienia elementów wspólnych nie użyto listy Nuclear Receptor Meta-Pathway, którą odrzucono, gdyż miała ona najmniej wspólnych genów z innymi ścieżkami W efekcie uzyskano listę 10 genów(Rycina 37), a następnie sprawdzono, czy są one obecne w wynikach anotowanych genów do znalezionych sekwencji CpG przez poszczególne metody i przeanalizowano pod względem ich różnicy w ekspresji pomiędzy limfocytami B i CD4+ (Tabela 7, Tabela 8, Tabela 9).

Dane o ekpresji pochodzą z bazy danych DICE [357]. Zwraca uwagę fakt, iż metoda CTMeth-hhi pokrywa się z większością tych 10 genów, oraz fakt, iż metody delta 0.2 i delta 0.3 pominęły sekwencję CpG anotowane do MAPK3, które wykazują dużą zmienność w ekspresji, zwłaszcza gdy porównujemy aktywowane limfocyty CD4+ do limfocytów B. Dalsza analiza Wikipathways pokazała, że cztery metody (CTMeth-hhi, CTMeth-hh, delta 0.2 i delta 0.3) wskazały sekwencje CpG powiązane z genami, które w ponad 70% pokrywają się ze szlakami genetycznymi typowymi dla limfocytów B i CD4+ (Tabela 10). W kategorii immunological signature z bazy danych FUMA najwięcej genów powiązanych było ze zbiorem GSE10325_CD4_TCELL_VS_BCELL_UP w wynikach metod CTMeth-hhi i delta 0.2 (Tabela 11). Pogłębiona analiza w kontekście kategorii immunological signatures, wykazał również, że metody CTMeth-hhi, CTMeth-hh i delta 0.2 oraz delta 0.3 identyfikują sekwencje CpG związane z genami, które wykazywały ponad 70% korelację z kolekcjami genów typowymi dla limfocytów B i CD4+. (Tabela 12). Zestawienie z repozytorium BioCarta wykazało, że najwięcej anotowanych genów pokrywało się z BIOCARTA_MAPK_PATHWAY i BIOCARTA_HIVNEF_PATHWAY w wynikach dla metod CTMETH-hhi, Delta 0.2 i CTMeth-hh (Tabela 13). Zbiór danych Biocarta_mapk_pathwayw, który odpowiada szlakowi MAPK, będący istotnym szlakiem pod względem funkcjonowania komórek, a jednocześnie spójny z FUMA Wikipathways poddano dodatkowej analizie konotując wskazane pośrednio przez metody geny z danymi z bazy danych DICE, z której pobrano informacje na temat różnic w ekspresji genów pomiędzy limfocytami B i CD4+ (Tabela 14). Zwraca uwagę fakt, iż metoda CTMeth-hhi wskazuje więcej genów i w większym stopniu korelujących z różnicami w ekspresji.

B-Cell-CLL

Wyniki uzyskane ze zbioru B-Cell-CLL przeanalizowano w kontekście kategorii Oncogenic Signatures, Kegg Pathways oraz Cancer Modules. Dla kategorii Kegg Pathways najczęstszym zbiorem i zgodnym z największą ilość genów anotowanych do wyników badanych metod był KEGG_PATHWAYS_IN_CANCER (Tabela 15, Tabela 16). Z genów, które korelowały z tym zestawem wybrano dla każdego porównywanego narzędzia do analizy metylacji te, które są unikalne tj. nie występują w wynikach innych

metod (Rycina 38) i porównano je z bazą danych BloodSpot [358] sprawdzając, które geny wg. tej bazy danych charakteryzują się odmienną ekspresją pomiędzy pacjentami z przewlekłą białaczką limfocytową i osobami zdrowymi (Tabela 17). Metoda CTMeth-hhi wskazała najwięcej wyników pokrywających się z bazą danych BloodSpot[358] tzn. wskazała te sekwencje CpG jako odmiennie metylowane, do których były anotowane geny o odmiennej ekpresji wg. BloodSpot [358], a nie wskazała tych gdzie ta różnica w ekpresji nie występuje i jednocześnie geny te korelowały z listą KEGG_PATHWAYS_IN_CANCER. Analiza w kontekście kategorii Oncogenic Signatures wykazała, że geny anotowane do wskazywanych sekwencji CpG przez badane narzędzia do analizy metylacji najbardziej odpowiadają zbiorom NFE2L2.V2, KRAS.600UP.V1_UP, KRAS.600.LUNG.BREAST_UP.V1_DN (Tabela 18). Na podstawie tych zbiorów utworzono listę wspólnych genów (NFE2L2.V2 z KRAS.600_UP.V1_UP oraz NFE2L2.V2 z KRAS.600.LUNG.BREAST_UP.V1_DN), które następnie sprawdzono pod względem różnic w ekpresji z bazą danych DICE oraz genami anotowanymi do wskazywanych przez porównywane metody sekwencji CpG. Największy wynik pod względem spójności osiągnęła tutaj metoda CTMeth-hhi tzn. wykazała ona te sekwencje CpG, jako odmiennie metylowane, do których anotowane geny różnią się ekspresją pomiędzy pacjentami chorymi na CLL a zdrowymi wg. bazy danych BloodSpot[358]. Pełne wyniki przedstawiono w Tabela 19. Wyniki uzyskane ze zbioru B-Cell-CLL sprawdzono również z kategorią FUMA Cancer Modules. Najczęstszym i z największą ilością wspólnych genów dla wyników analizowanych metod był GenSet Cancer_Module_88. Metoda CTMeth-hhi i delta 0.2 wskazują tę samą największą liczbę genów spójną z Cancer_Module_88 (Tabela 20). Analogicznie do poprzednich kroków porównania, unikalne anotowane geny z wyników uzyskanych za pomocą porównywanych metod do badania metylacji odpowiadające elementom zawartym na liście FUMA Cancer_Module_88, poddano weryfikacji z wykorzystaniem bazy danych BloodSpot[358], aby potwierdzić ich różnicę w ekpresji pomiędzy pacjentami zdrowymi i chorymi na CLL (Tabela 21). Najwięcej, 56 spójnych genów wykazała metoda CTMeth-hhi. Druga metoda to delta 0.2. Zwraca uwagę fakt, że obie metod CTMeth-hhi i delta 0.2 znalazły identyczną liczbę genów wspólną z listą Cancer_Module_88, ale to wyniki CTMeth-hhi wśród unikalnych genów, nie występujących w wynikach delta 0.2 wykazały się lepszą spójnością z bazą danych BloodSpot[358]. W kolejnym kroku wybrano z kategorii Cancer_Modules te listy, które

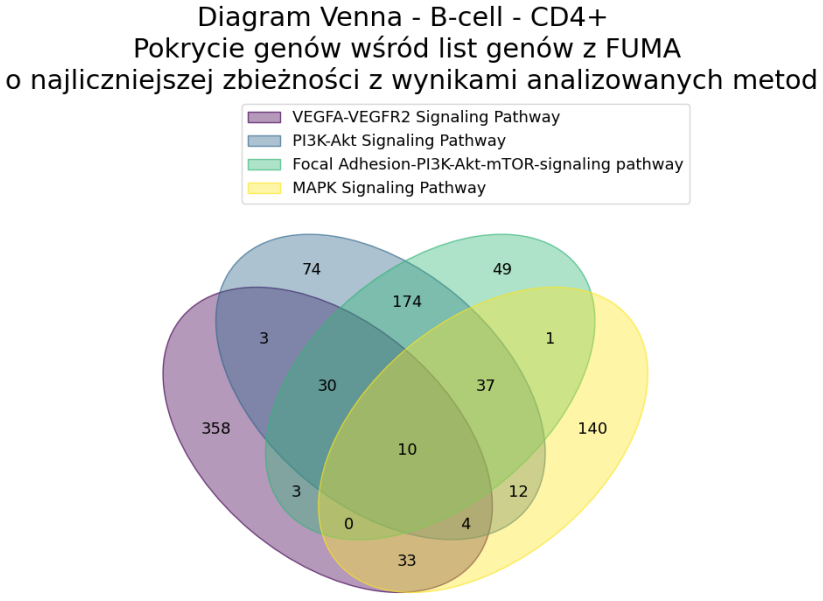
wg jej twórców [359] posiadały ponad 70% trafień (ang. hits) korelujących z przewlekłą białaczką limfocytową (Tabela 22)**.** Następnie sprawdzono, wyniki, których metod w analizie z użyciem FUMA zawierają te listy genów (Tabela 23). Tylko dwie listy genów korelują z wynikami uzyskanymi z użyciem analizowanych metod badawczych i obydwie znajdują się w wynikach każdej z nich, natomiast sumarycznie najwięcej genów spójnych znalazła metoda CTMeth-hhi (Tabela 24,Tabela 25). Raz jeszcze geny unikalne dla wyników poszczególnych metod poddano analizie, której na celu jest sprawdzenie spójności z bazą danych BloodSpot[358] (Tabela 26). Metoda CTmeth-hhi i delta 0.2 wskazały ich poprawnie najwięcej.

CLL-100

Dla danych opartych o wyniki ze zbioru CLL-100 przeprowadzono analizę zgodnie z Wikipathways, Kegg Pathways i Oncogenic Signatures. Ze względu na brak dostępu do publicznej bazy danych, analogicznej do DICE lub BloodSpot[358], która zawierałaby zweryfikowany spis genów i informacje o ich ekpresji w zależności od IGHV, w ramach procedury badawczej dla każdej kategorii FUMA i metody wyekstrahowano po cztery elementy charakteryzujące się najwyższą ilością spójnych genów. Następnie przeprowadzono analizę porównawczą w celu identyfikacji obecności tych list genów uzyskanych z wyników innych metod oraz oceny ilości anotowanych genów, które z nimi korelują. Rezultatem jest kompilacja listy, która inkorporuje cztery najwyższej jakości elementy z każdego zbioru oraz te elementy, które są wspólne dla danych uzyskanych z wyników innych analizowanych narzędzi. Listy dublujące się oznaczono kolorem czerwonym (Tabela 27,Tabela 28,Tabela 29). Wyniki uzyskane przy użyciu CTMeth-hhi, będące odpowiednikami czterech najlepszych list genów uzyskany z pozostałych metod, wykazują wyższą jakość pod względem ilości korelujących genów w każdej z trzech kategorii. Dodatkowo w kategorii Wikipathways wyniki uzyskane z użyciem metody CTMeth-hhi jako jedyne korelują z listą genów Wnt/beta-catenin in leukemia i najlepiej korelują z listą Wnt signaling pathway (Tabela 30). Zaburzenia w obrębie tej ścieżki biologicznej mają znaczenie dla różnicowania w kontekście IGHV[360–362]. Natomiast w wynikach uzyskanych z użyciem narzędzia CTMeth-hhi znalazły się unikalne listy np. KRAS.600_UP.V1_DN. Mutacje w obrębie KRAS mają związek z IGHV w przewlekłej białaczce limfocytowej[363].

Tabela 6 FUMA Wikipathways - B-Cell-CD4+. Wyniki posortowane względem największej liczby genów pokrywających się

Zbiór danych	Liczba pokrywających się genów	Procent z całego zbioru	Wartość p	Metoda
PI3K-Akt Signaling Pathway	219	63.66279	3.76E-33	CTMETH HHI
Focal Adhesion-PI3K-Akt-mTOR-signaling pathway	207	67.42671	1.83E-36	CTMETH HHI
PI3K-Akt Signaling Pathway	200	58.13953	7.09E-35	DELTA 0.2
Nuclear Receptors Meta-Pathway	194	60.625	1.18E-25	CTMETH HHI
Focal Adhesion-PI3K-Akt-mTOR-signaling pathway	183	59.60912	5.31E-34	DELTA 0.2
VEGFA-VEGFR2 Signaling Pathway	175	73.83966	1.07E-38	CTMETH HHI
MAPK Signaling Pathway	172	69.07631	1.19E-32	CTMETH HHI
VEGFA-VEGFR2 Signaling Pathway	170	71.72996	3.90E-47	DELTA 0.2
PI3K-Akt Signaling Pathway	169	49.12791	3.81E-30	CTMETH HH
Nuclear Receptors Meta-Pathway	159	49.6875	6.44E-19	DELTA 0.2



Rycina 37 FUMA Wikipathways - B-Cell-CD4+ - Diagram Venna - B-Cell-CD4+ Zidentyfikowane listy genów FUMA, które wykazały największą zbieżność genów z wynikami uzyskanymi przez poszczególne metody

Tabela 7 B-Cell-CD4+ analiza ekspresji 10 genów występujących w listach genów z FUMA o największej pod kątem genów zbieżności z wynikami porównywanych metod oraz ich obecność w wynikach poszczególnych tych metod

	CTMETH-hhi	CTMETH-hh	delta 0.2	delta 0.3
MAPK3	TAK	TAK	NIE	NIE
GRB2	TAK	TAK	TAK	TAK
MAP2K2	TAK	TAK	TAK	TAK
AKT1	TAK	NIE	NIE	NIE
ATF2	TAK	NIE	TAK	NIE
MAPK1	TAK	TAK	TAK	TAK
RAF1	TAK	TAK	TAK	TAK
HRAS	TAK	NIE	NIE	NIE
MAP2K1	TAK	TAK	TAK	TAK
ATF4	NIE	NIE	NIE	NIE

Tabela 8 B-Cell - CD4+ Naive Nieaktywowane - analiza ekspresji 10 genów występujących w listach genów z FUMA o największej pod kątem genów zbieżności z wynikami porównywanych metod

Gene	Biotype	B cell, naive Mean Expression (TPM)	T cell, CD4, naive Mean Expression (TPM)	Log 2 Fold Change	Adjusted p-Value
GRB2	Protein coding	245.21	143.9	0.92	0
MAPK1	Protein coding	73.52	66.25	0.36	4.80E-52
HRAS	Protein coding	11.54	9.96	0.35	0.00023
ATF4	Protein coding	343.67	326.53	0.27	7.80E-19
MAP2K2	Protein coding	27.65	26.02	0.21	0.0014
ATF2	Protein coding	61.06	61.83	0.19	1.60E-10
RAF1	Protein coding	111.4	121.65	0.06	0.00089
MAP2K1	Protein coding	53.05	68.7	-0.16	1.40E-12
MAPK3	Protein coding	29.88	49.78	-0.58	2.10E-30

Tabela 9 B-Cell - CD4+ Naive Aktywowane - analiza ekspresji 10 genów występujących w listach genów z FUMA o największej pod kątem genów zbieżności z wynikami porównywanych metod

Gene	Biotype	B cell, naive Mean Expression (TPM)	T cell, CD4, naive [activated]Mean Expression (TPM)	Log 2 Fold Change	Adjusted p-Value
MAPK3	Protein coding	29.88	7.62	1.97	0
GRB2	Protein coding	245.21	137.9	0.71	0
MAP2K2	Protein coding	27.65	17.18	0.59	6.10E-20
AKT1	Protein coding	15.44	11.75	0.3	0.000027
ATF2	Protein coding	61.06	49.33	0.25	3.30E-17
MAPK1	Protein coding	73.52	61.19	0.2	6.50E-17
RAF1	Protein coding	111.4	110.41	-0.05	0.0044
HRAS	Protein coding	11.54	15.26	-0.51	3.90E-08
MAP2K1	Protein coding	53.05	73.2	-0.52	6.00E-116
ATF4	Protein coding	343.67	556.13	-0.72	2.10E-123

Tabela 10 B-Cell-CD4+ - Wikipathways, Korelacja wyników uzyskanych z poszczególnych metod z listami genów z FUMA, które dotyczą limfocytów B i CD4+

Zbiór danych	n genów w zbiorze	Procent z całego zbioru	Wartość p	Metoda
B Cell Receptor Signaling Pathway	98	80.6122449	3.29E-22	CTMETH HHI
B Cell Receptor Signaling Pathway	98	80.6122449	4.33E-28	DELTA 0.2
B Cell Receptor Signaling Pathway	98	71.42857143	3.21E-25	CTMETH HH
B Cell Receptor Signaling Pathway	98	67.34693878	1.13E-24	DELTA 0.3
B Cell Receptor Signaling Pathway	98	50	6.43E-23	DELTA 0.5
B Cell Receptor Signaling Pathway	98	42.85714286	5.11E-16	CHAMP 0.2
B Cell Receptor Signaling Pathway	98	33.67346939	2.58E-14	CHAMP 0.3
B Cell Receptor Signaling Pathway	98	22.44897959	8.99E-13	CHAMP 0.5
T-Cell antigen Receptor (TCR) Signaling Pathway	90	80	2.29E-25	DELTA 0.2
T-Cell antigen Receptor (TCR) Signaling Pathway	90	78.88888889	4.67E-19	CTMETH HHI
T-Cell antigen Receptor (TCR) Signaling Pathway	90	67.77777778	2.25E-20	CTMETH HH
T-Cell antigen Receptor (TCR) Signaling Pathway	90	63.33333333	1.06E-19	DELTA 0.3
T-Cell antigen Receptor (TCR) Signaling Pathway	90	55.55555556	6.25E-26	DELTA 0.5
T-Cell antigen Receptor (TCR) Signaling Pathway	90	31.11111111	1.56E-07	CHAMP 0.2
T-Cell antigen Receptor (TCR) Signaling Pathway	90	22.22222222	5.17E-06	ChAMP 0.3
T-Cell antigen Receptor (TCR) Signaling Pathway	90	13.33333333	1.09E-04	CHAMP 0.5
T-Cell Receptor and Co-stimulatory Signaling	29	89.65517241	6.47E-10	CTMETH HHI
T-Cell Receptor and Co-stimulatory Signaling	29	89.65517241	6.07E-12	DELTA 0.2
T-Cell Receptor and Co-stimulatory Signaling	29	82.75862069	1.44E-11	CTMETH HH
T-Cell Receptor and Co-stimulatory Signaling	29	72.4137931	1.79E-09	DELTA 0.3
T-Cell Receptor and Co-stimulatory Signaling	29	58.62068966	6.65E-10	DELTA 0.5

Tabela 11 B-Cell-CD4+ - Listy genów FUMA immunological signatures korelujące najbardziej z wynikami uzyskanymi przy użyciu poszczególnych metod

Zbiór danych	Liczba genów znalezionych	Procent z całego zbioru	Wartość p-value	Metoda
GSE10325_CD4_TCELL_VS_BCELL_UP	167	84.77157	4.40E-51	CTMETHHI
GSE10325_CD4_TCELL_VS_BCELL_UP	166	84.26396	3.58E-63	DELTA 0.2
GSE10325_LUPUS_CD4_TCELL_VS_LUPUS_BCELL_UP	166	84.26396	2.65E-50	CTMETHHI
GSE11057_NAIVE_VS_CENT_MEMORY_CD4_TCELL_DN	164	82.82828	1.75E-60	DELTA 0.2
GSE10325_LUPUS_CD4_TCELL_VS_LUPUS_BCELL_UP	162	82.2335	5.22E-59	DELTA 0.2
GSE30083_SP2_VS_SP4_THYMOCYTE_DN	162	81	2.04E-57	DELTA 0.2
GSE5542_UNTREATED_VS_IFNA_TREATED_EPITHELIAL_CELLS_24H_UP	162	81	7.95E-45	CTMETHHI
GSE11057_NAIVE_VS_CENT_MEMORY_CD4_TCELL_DN	161	81.31313	7.95E-45	CTMETHHI
GSE5542_UNTREATED_VS_IFNA_TREATED_EPITHELIAL_CELLS_24H_UP	161	80.5	2.07E-56	DELTA 0.2
GSE30083_SP2_VS_SP4_THYMOCYTE_DN	160	80	4.73E-43	CTMETHHI

Tabela 12 B-Cell-CD4+ - Listy genów FUMA immunological signatures typowych dla badanych limfocytów B i limfocytów T CD4+ korelujące najbardziej z wynikami uzyskanymi przy użyciu poszczególnych metod

Zbiór danych	Liczba genów znalezionych	Procent z całego zbioru	Wartość p-value	Metoda
GSE10325_CD4_TCELL_VS_BCELL_DN	158	81.86528497	7.95E-45	CTMETHHHI
GSE10325_CD4_TCELL_VS_BCELL_DN	157	81.34715026	3.15E-56	DELTA 0.2
GSE10325_CD4_TCELL_VS_BCELL_DN	148	76.68393782	1.38E-59	CTMETHHH
GSE10325_CD4_TCELL_VS_BCELL_DN	144	74.61139896	3.95E-63	DELTA 0.3
GSE10325_CD4_TCELL_VS_BCELL_DN	119	61.65803109	4.65E-67	CHAMP 0.2
GSE10325_CD4_TCELL_VS_BCELL_DN	111	57.51295337	8.31E-61	DELTA 0.5
GSE10325_CD4_TCELL_VS_BCELL_DN	95	49.22279793	4.24E-60	CHAMP 0.3
GSE10325_CD4_TCELL_VS_BCELL_DN	64	33.16062176	9.78E-51	CHAMP 0.5
GSE10325_CD4_TCELL_VS_BCELL_UP	167	84.7715736	4.40E-51	CTMETHHHI
GSE10325_CD4_TCELL_VS_BCELL_UP	166	84.26395939	3.58E-63	DELTA 0.2
GSE10325_CD4_TCELL_VS_BCELL_UP	150	76.14213198	7.99E-68	DELTA 0.3
GSE10325_CD4_TCELL_VS_BCELL_UP	147	74.61928934	1.26E-56	CTMETHHH
GSE10325_CD4_TCELL_VS_BCELL_UP	105	53.29949239	4.11E-53	DELTA 0.5
GSE10325_CD4_TCELL_VS_BCELL_UP	57	28.93401015	1.21E-13	CHAMP 0.2
GSE10325_CD4_TCELL_VS_BCELL_UP	33	16.75126904	2.40E-07	CHAMP 0.3
GSE10325_CD4_TCELL_VS_BCELL_UP	11	5.583756345	0.01638866	CHAMP 0.5

Tabela 13 B-Cell-CD4+ - Listy genów FUMA BioCarta korelujące najbardziej z wynikami uzyskanymi przy użyciu poszczególnych metod

Zbiór danych	Liczba pokrywających się genów	Procent z całego zbioru	Wartość p	Metoda
BIOCARTA_MAPK_PATHWAY	61	75.30864	3.88E-18	DELTA 0.2
BIOCARTA_MAPK_PATHWAY	60	74.07407	8.61E-13	CTMETH HHI
BIOCARTA_MAPK_PATHWAY	51	62.96296	3.94E-14	CTMETH HH
BIOCARTA_MAPK_PATHWAY	43	53.08642	1.65E-10	DELTA 0.3
BIOCARTA_HIVNEF_PATHWAY	42	75	2.11E-09	CTMETH HHI
BIOCARTA_HIVNEF_PATHWAY	42	75	2.42E-12	DELTA 0.2
BIOCARTA_HIVNEF_PATHWAY	39	69.64286	4.76E-13	CTMETH HH
BIOCARTA_HIVNEF_PATHWAY	36	64.28571	7.60E-12	DELTA 0.3
BIOCARTA_KERATINOCYTE_PATHWAY	36	78.26087	5.20E-09	CTMETH HHI
BIOCARTA_TCR_PATHWAY	36	81.81818	9.31E-10	CTMETH HHI

Tabela 14 Liczba genów pokrywających się z BIOCARTA_MAPK_PATHWAY wskazanych przez analizowane metody i z rożną ekspresją dla CD4+ i limfocytów B wg DICE

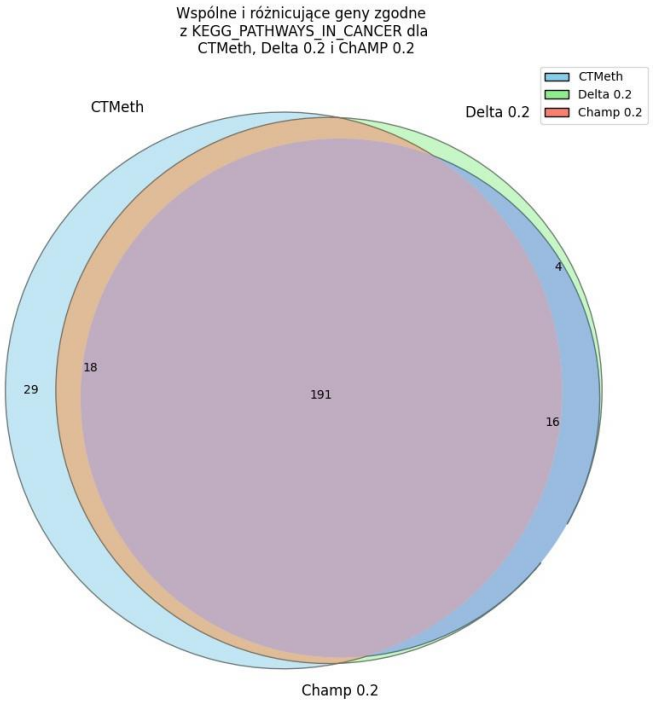
	CTMETH-hhi	delta 0.2	CTMETH-hh	delta 0.3
Liczba genów pokrywających się z BIOCARTA_MAPK _PATHWAY	60	61	51	43
Liczba genów pokrywających się z BIOCARTA_MAPK_PATHWAY i rożną ekspresją dla CD4+ i limfocytów B wg DICE	50	49	42	36
Procent	83.33	80.33	82.35	83.72

Tabela 15 B-Cell-CLL - Listy genów FUMA z kategorii KEGG korelujące najbardziej z wynikami uzyskanymi przy użyciu poszczególnych metod

GeneSet	Liczba pokrywających się genów	p-value	Metoda	Procent całości zbioru
KEGG_PATHWAYS_IN_CANCER	238	1.370770616039709e-32	CTMeth-hhi	74.375
KEGG_PATHWAYS_IN_CANCER	229	1.7828281100791575e-24	Delta 0.2	71.5625
KEGG_PATHWAYS_IN_CANCER	207	4.5571588580404746e-21	ChAMP 0.2	64.6875
KEGG_OLFACTORY _TRANSDUCTION	202	7.566533617742562e-05	Delta 0.2	53.2981
KEGG_NEUROACTIVE_LIGAND_ RECEPTOR_INTERACTION	196	6.648069494489237e-27	CTMeth-hhi	74.2424
KEGG_OLFACTORY _TRANSDUCTION	186	2.784369583761712e-05	ChAMP 0.2	49.0765
KEGG_MAPK _SIGNALING_PATHWAY	182	1.705812826782416e-21	CTMeth-hhi	71.09375
KEGG_NEUROACTIVE_LIGAND_ RECEPTOR_INTERACTION	182	5.089305104666866e-17	Delta 0.2	68.9393
KEGG_MAPK _SIGNALING_PATHWAY	181	9.818491694489088e-19	Delta 0.2	70.70313

Tabela 16 B-Cell-CLL - Lista genów KEGG_PATHWAYS_IN_CANCER z FUMA i stopień korelacji z wynikami uzyskanymi przy użyciu poszczególnych metod

GeneSet	Liczba pokrywających się genów	p-value	Metoda	Procent całości zbioru
KEGG_PATHWAYS_IN_CANCER	238	2.55E-30	CTMeth-hhi	74.375
KEGG_PATHWAYS_IN_CANCER	229	1.77E-22	delta 0.2	71.5625
KEGG_PATHWAYS_IN_CANCER	207	4.24E-19	ChAMP 0.2	64.6875
KEGG_PATHWAYS_IN_CANCER	180	1.14E-17	delta 0.3	56.25
KEGG_PATHWAYS_IN_CANCER	175	1.71E-17	ChAMP 0.3	54.6875
KEGG_PATHWAYS_IN_CANCER	150	1.88E-14	CTMETH-hh	46.875
KEGG_PATHWAYS_IN_CANCER	83	3.33E-10	Delta 0.5	25.9375
KEGG_PATHWAYS_IN_CANCER	82	3.16E-10	ChAMP 0.5	25.625



Rycina 38 B-cell – CLL – Wspólne i różnicujące geny zgodne z KEGG_PATHWAYS_IN_CANCER dla CTMeth-hhi, delta 0.2 oraz ChAMP 0.2. Metody delta 0.3 i 0.5, ChAMP 0.3 i 0.5 oraz CTMeth-hh nie zawierały unikalnych genów ze względów na ich mniej selektywny charakter

Tabela 17 B-Cell-CLL - Korelacja pomiędzy różnicą w ekspresji poszczególnych genów u pacjentów zdrowych i chorych na przewlekłą białaczkę limfocytową a występowaniem tych genów w wynikach uzyskanych z poszczególnych metod. Geny te występują w liście genów KEGG_PATHWAYS_IN_CANCER i są unikalne dla wyników poszczególnych metod. Informacja o ekpresji pochodzi z bazy danych BloodSpot

		CTMeth		Delta			ChAMP			BloodSpot
Gene	P-value	HHI	HH	0.2	0.3	0.5	0.2	0.3	0.5	<0.05
DVL2_log2	1.25E-36	Nie	Nie	Tak	Nie	Nie	Nie	Nie	Nie	TAK
APC2_log2	0.031427601	Tak	Tak	Nie	Nie	Nie	Nie	Nie	Nie	TAK
AXIN2_log2	0.069971576	Tak	Tak	Nie	Nie	Nie	Nie	Nie	Nie	NIE
CBLC_log2	0.00504372	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie	TAK
CCNA1_log2	4.39E-51	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie	TAK
CDK4_log2	3.25E-56	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie	TAK
COL4A4_log2	1.28E-25	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie	TAK
FADD_log2	1.10E-16	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie	TAK
FGF17_log2	0.654576687	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie	NIE
FGF19_log2	0.09183394	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie	NIE
FGF3_log2	0.005618233	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie	TAK
FGF4_log2	8.03E-07	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie	TAK
FGF6_log2	0.271683852	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie	NIE
FGF7_log2	1.43E-08	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie	TAK
FLT3LG_log2	2.02E-07	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie	TAK
FOS_log2	1.86E-24	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie	TAK
FZD9_log2	1.58E-07	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie	TAK
HSP90AA1_log2	0.010894607	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie	TAK
ITGA3_log2	5.58E-29	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie	TAK
KLK3_log2	5.37E-117	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie	TAK
MLH1_log2	1.99E-36	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie	TAK
MSH2_log2	8.47E-26	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie	TAK
PGF_log2	5.27E-17	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie	TAK
RALBP1_log2	3.11E-30	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie	TAK
WNT2_log2	0.084155856	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie	NIE
WNT6_log2	1.99E-05	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie	TAK
WNT9B_log2	0.452836542	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie	NIE
MSH6_log2	2.30E-32	Nie	Nie	Tak	Nie	Nie	Nie	Nie	Nie	TAK
RBX1_log2	2.88E-65	Nie	Nie	Tak	Nie	Nie	Nie	Nie	Nie	TAK
VEGFA_log2	6.04E-21	Nie	Nie	Tak	Nie	Nie	Nie	Nie	Nie	TAK
Ilość spójnych elementów z różnicą w ekspresji z bazy BloodSpot		20	6	10	6	6	6	6	6	31

Tabela 18 B-Cell-CLL - Listy genów FUMA Oncogenic signatures korelujące najbardziej z wynikami uzyskanymi przy użyciu poszczególnych metod

Kategoria	GeneSet	Liczba pokrywających się genów	adjP	Metoda	Procent całości zbioru
Oncogenic_signatures	NFE2L2.V2	250	5.49E-12	CTMeth-hhi	58.96226
Oncogenic_signatures	NFE2L2.V2	247	1.82E-09	delta 0.2	58.25472
Oncogenic_signatures	NFE2L2.V2	223	1.86E-08	ChAMP 0.2	52.59434
Oncogenic_signatures	KRAS.600_UP.V1_UP	193	6.90E-21	CTMeth-hhi	71.48148
Oncogenic_signatures	KRAS.600.LUNG. BREAST_UP.V1_DN	193	2.86E-20	CTMeth-hhi	70.69597
Oncogenic_signatures	KRAS.600_UP.V1_UP	190	5.89E-17	Delta 0.2	70.37037
Oncogenic_signatures	NFE2L2.V2	188	1.49E-07	Delta 0.2	44.33962
Oncogenic_signatures	TBK1.DF_UP	187	8.19E-14	Delta 0.2	66.54804
Oncogenic_signatures	KRAS.600_UP.V1_DN	186	1.98E-16	Delta 0.2	69.66292
Oncogenic_signatures	KRAS.600.LUNG. BREAST_UP.V1_DN	186	3.63E-15	Delta 0.2	68.13187

Tabela 19 B-Cell-CLL - Korelacja pomiędzy różnicą w ekspresji poszczególnych genów u pacjentów zdrowych i chorych na przewlekłą białaczkę limfocytową a występowaniem tych genów w wynikach uzyskanych z poszczególnych metod. Geny te pokrywają się pomiędzy listami genów NFE2L2.V2, KRAS.600_UP.V1_UP, KRAS.600.LUNG.BREAST_UP.V1_DN. Informacja o ekpresji pochodzi z bazy danych BloodSpot

		CTMeth		delta			ChAMP			BloodSpot
Gene	P-value	hhi	hh	0.2	0.3	0.5	0.2	0.3	0.5	P<0.05
AKR1B10_log2	0.000114	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	TAK
CDH12_log2	0.56866	Tak	Nie	Tak	Tak	Nie	Tak	Tak	Nie	NIE
DEFB1_log2	1.11E-92	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie	TAK
FLT1_log2	0.792103	Tak	Tak	Tak	Tak	Nie	Tak	Tak	Nie	NIE
HRK_log2	2.17E-26	Tak	Tak	Nie	Nie	Nie	Nie	Nie	Nie	TAK
IL18R1_log2	1.23E-46	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	TAK
PDE6B_log2	1.27E-24	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Tak	TAK
PLXND1_log2	2.32E-05	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Tak	TAK
PPFIA2_log2	0.583988	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Tak	NIE
SDS_log2	3.19E-07	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie	TAK
KIF5C_log2	0.477878	Tak	Tak	Tak	Nie	Nie	Tak	Nie	Nie	NIE
KIT_log2	1.06E-158	Tak	Nie	Tak	Tak	Nie	Tak	Tak	Nie	TAK
PEG3_log2	0.381535	Tak	Tak	Tak	Tak	Nie	Tak	Tak	Nie	NIE
SMPX_log2	0.551852	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	NIE
SPP1_log2	2.80E-14	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	TAK
SRGN_log2	1.93E-56	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	TAK
TERT_log2	7.88E-07	Tak	Tak	Tak	Tak	Nie	Tak	Tak	Nie	TAK
TLR4_log2	6.26E-63	Tak	Nie	Tak	Nie	Nie	Nie	Nie	Nie	TAK
TNFSF15_log2	0.006616	Nie	Nie	Tak	Tak	Nie	Tak	Tak	Nie	TAK
Ilość spójnych elementów z różnicą w ekspresji z bazy BloodSpot		9	6	7	7	7	6	7	7	20

Tabela 20 B-Cell-CLL - Listy genów FUMA Cancer_Modules korelujące najbardziej z wynikami uzyskanymi przy użyciu poszczególnych metod

Category	GeneSet	N_genes	N_overlap	adjP	SourceFile	Procent
Cancer_Modules	MODULE_88	802	564	4.10E-59	CTMeth-hhi	70.32418953
Cancer_modules	MODULE_88	802	564	4.26E-52	Delta 0.2	70.32418953
Cancer_modules	MODULE_55	800	557	1.17E-49	Delta 0.2	69.625
Cancer_modules	MODULE_55	800	555	2.88E-55	CTMeth-hhi	69.375
Cancer_modules	MODULE_88	802	518	3.49E-48	ChAMP 0.2	64.58852868
Cancer_modules	MODULE_55	800	508	2.29E-44	ChAMP 0.2	63.5
Cancer_modules	MODULE_117	697	502	1.97E-51	Delta 0.2	72.02295552
Cancer_modules	MODULE_117	697	484	2.07E-48	CTMeth-hhi	69.44045911
Cancer_modules	MODULE_117	697	455	2.29E-44	ChAMP 0.2	65.27977044
Cancer_modules	MODULE_88	802	435	7.15E-38	Delta 0.3	54.2394015

Tabela 21 B-Cell-CLL - Korelacja pomiędzy różnicą w ekspresji poszczególnych genów u pacjentów zdrowych i chorych na przewlekłą białaczkę limfocytową a występowaniem tych genów w wynikach uzyskanych z poszczególnych metod. Geny te pokrywają się z listą genów Cancer_Module_88 i są dobrane pod względem unikalności w wynikach poszczególnych metod

		CTMeth		delta			ChAMP			BloodSpot
Gene	P-value	hh	hhi	0.2	0.3	0.5	0.2	0.3	0.5	P<0.05
ZYX_log2	8.50E-59	Nie	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Tak
VEGFA_log2	6.04E-21	Nie	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Tak
ALDH1A3_log2	5.86E-22	Nie	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Tak
MEF2C_log2	5.12E-55	Nie	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Tak
COL1A2_log2	3.29E-24	Nie	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Tak
IL1R1_log2	2.08E-13	Nie	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Tak
GOT1_log2	1.97E-06	Nie	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Tak
APOA1_log2	0.896939531	Nie	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie
MDK_log2	0.269172612	Nie	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Tak
CLDN8_log2	0.631562988	Nie	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie
DAO_log2	0.600791104	Nie	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie
GREM1_log2	0.096744288	Nie	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Tak
PPL_log2	0.046914844	Nie	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Tak
SSTR2_log2	0.031286004	Nie	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Tak

TM4SF1_log2	8.12E-19	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
G0S2_log2	6.63E-17	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
LTF_log2	5.57E-108	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
ATF5_log2	5.54E-12	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
BDH1_log2	5.12E-40	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
ARL4A_log2	4.44E-06	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
SCNN1B_log2	4.13E-05	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
HSPA2_log2	3.86E-11	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
CD14_log2	3.80E-25	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
CCNF_log2	3.58E-106	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
AMBP_log2	3.54E-08	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
SCP2_log2	3.53E-42	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
SDS_log2	3.19E-07	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
FCGRT_log2	2.77E-06	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
C4BPA_log2	2.68E-10	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
CASP2_log2	2.53E-18	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
T_log2	2.28E-08	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
ENG_log2	2.23E-59	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
PRSS8_log2	2.11E-08	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
TMEM97_log2	2.04E-78	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
FOS_log2	1.86E-24	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
EPB41L3_log2	1.85E-17	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
GSTM5_log2	1.85E-10	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
ASGR2_log2	1.84E-21	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
CA4_log2	1.64E-89	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
SLC2A3_log2	1.53E-27	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
COL6A1_log2	1.45E-06	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
UCN_log2	1.37E-27	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
LTBR_log2	1.13E-51	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
DEFB1_log2	1.11E-92	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
SLC9A3R1_log2	1.02E-41	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
PAEP_log2	0.717902259	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie
ADRA2C_log2	0.689475379	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie
REN_log2	0.665060024	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie

NNAT_log2	0.559358818	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Nie
SEZ6L_log2	0.3004153	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
CYP2A7_log2	0.253302535	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
COL2A1_log2	0.233434411	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
GDF15_log2	0.16779146	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
FKBP1B_log2	0.161213165	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
JAK3_log2	0.117706746	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
LYPD3_log2	0.115042457	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
TBX1_log2	0.11103017	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
CDK5R1_log2	0.037636284	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
GPRC5B_log2	0.02298774	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
PTGIS_log2	0.009901286	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
CNKSR1_log2	0.007654411	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
KRT19_log2	0.006587717	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
CYP2E1_log2	0.006159969	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
SLC22A1_log2	0.004289841	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
SLC17A3_log2	0.00419252	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
PTPRN_log2	0.002896899	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
PHYH_log2	0.002622846	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
HCN3_log2	0.002146626	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
CRYAB_log2	0.002041929	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
RHOB_log2	0.000377542	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
SERPINE1_log2	0.000217832	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Nie	Tak
Ilość spójnych elementów z różnicą w ekspresji z bazy BloodSpot		7	56	15	7	7	7	7	7	71

Tabela 22 B-Cell-CLL Listy genów Cancer_modules, które zawierają >70% trafień (ang. hits) dla przewlekłej białaczki limfocytowej

Module	Hits dla Chronic Lymphocytic Leukemia
MODULE_156	85.7
MODULE_240	85.7
MODULE_254	73.6
MODULE_313	85.7
MODULE_439	81.2
MODULE_537	83.8
MODULE_547	79.3
MODULE_573	73.9

Tabela 23 B-Cell-CLL Listy genów Cancer_modules, które zawierają >70% trafień (ang. hits) dla przewlekłej białaczki limfocytowej i ich występowanie w wynikach FUMA dla poszczególnych metod

	delta 0.2	delta 0.3	delta 0.5	ChAMP 0.2	ChAMP 0.3	ChAMP 0.5	CTMeth-hhi	CTMeth-hh
MODULE_156	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie
MODULE_240	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie
MODULE_254	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Tak
MODULE_313	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie
MODULE_439	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie
MODULE_537	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Tak
MODULE_547	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie
MODULE_573	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie

Tabela 24 B-Cell-CLL Liczba anotowanych genów do sekwencji CpG wskazywanych przez poszczególne metody i korelujących z listą Cancer MODULE_254

Category	GeneSet	N_genes	N_overlap	adjP	SourceFile
Cancer_modules	MODULE_254	60	37	0.003474	CTMeth-hhi
Cancer_modules	MODULE_254	60	34	0.046775	Delta 0.2
Cancer_modules	MODULE_254	60	32	0.029036	ChAMP 0.2
Cancer_modules	MODULE_254	60	30	0.003403	ChAMP 0.3
Cancer_modules	MODULE_254	60	30	0.006407	Delta 0.3
Cancer_modules	MODULE_254	60	26	0.006439	CTMeth-hh
Cancer_modules	MODULE_254	60	18	0.000705	ChAMP 0.5
Cancer_modules	MODULE_254	60	18	0.000737	Delta 0.5

Tabela 25 B-Cell-CLL Liczba anotowanych genów do sekwencji CpG wskazywanych przez poszczególne metody i korelujących z listą Cancer MODULE_537

Category	GeneSet	N_genes	N_overlap	adjP	SourceFile
Cancer_modules	MODULE_537	17	12	0.027541	CTMeth-hhi
Cancer_modules	MODULE_537	17	12	0.038527	delta 0.2
Cancer_modules	MODULE_537	17	11	0.04738	ChAMP 0.2
Cancer_modules	MODULE_537	17	11	0.009011	ChAMP 0.3
Cancer_modules	MODULE_537	17	11	0.012387	Delta 0.3
Cancer_modules	MODULE_537	17	9	0.034544	CTMeth-hh
Cancer_modules	MODULE_537	17	7	0.008573	ChAMP 0.5
Cancer_modules	MODULE_537	17	7	0.00896	Delta 0.5

Tabela 26 B-Cell-CLL - Korelacja pomiędzy różnicą w ekspresji poszczególnych genów u pacjentów zdrowych i chorych na przewlekłą białaczkę limfocytową a występowaniem tych genów w wynikach uzyskanych z poszczególnych metod. Geny te pokrywają się z listą genów Cancer_module_537 i Cancer_module 254 i są dobrane pod względem unikalności w wynikach poszczególnych metod

Gene	P-value	CTMeth		delta			ChAMP			BloodSpot
		hhi	hh	0.2	0.3	0.5	0.2	0.3	0.5	<0.05
BCL2A1_log2	1.32E-23	Tak	Nie	Tak	Tak	Nie	Tak	Tak	Nie	Tak
BCL2L1_log2	2.57E-17	Tak	Nie	Tak	Nie	Nie	Nie	Nie	Nie	Tak
BIRC2_log2	0.584312082	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie
BIRC3_log2	3.62E-52	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Tak
BIRC5_log2	8.83E-209	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Tak
BNIP3_log2	0.987590696	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Nie
CD2_log2	9.55E-06	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Tak
CFLAR_log2	0.009695408	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Tak
IER3_log2	3.53E-41	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Tak
IL1A_log2	0.348205858	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Nie
MYBL2_log2	1.41E-31	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Tak	Tak
NFKB1_log2	0.015515298	Tak	Tak	Tak	Tak	Nie	Tak	Tak	Nie	Tak
SERPINB2_log2	1.14E-25	Tak	Nie	Tak	Tak	Nie	Tak	Tak	Nie	Tak
SON_log2	3.88E-22	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Nie	Tak
Ilość spójnych elementów z różnicą w ekspresji z bazy BloodSpot		8	5	8	7	4	7	7	4	14

Tabela 27 CLL-100 – Wikipathways - Trzy listy genów najlepiej skorelowane z wynikami poszczególnych metod oraz ich stopień powiązania z wynikami uzyskanymi z innych metod

Lp.	GeneSet	N_genes	N_overlap	p	Metoda	Procent
1	PI3K-Akt Signaling Pathway	344	155	5.87E-29	CTMeth-hhi	45.058
2	Focal Adhesion-PI3K-Akt-mTOR-signaling pathway	307	147	4.83E-31	CTMeth-hhi	47.883
3	Nuclear Receptors Meta-Pathway	320	133	3.80E-21	CTMeth-hhi	41.563
4	VEGFA-VEGFR2 Signaling Pathway	237	106	5.73E-20	CTMeth-hhi	44.726
5	MAPK Signaling Pathway	249	106	4.75E-18	CTMeth-hhi	42.570
8	Ras Signaling	185	85	3.49E-17	CTMeth-hhi	45.946
18	Chemokine signaling pathway	165	74	1.80E-14	CTMeth-hhi	44.848
27	Insulin Signaling	161	67	2.00E-11	CTMeth-hhi	41.615
108	PI3K-Akt Signaling Pathway	344	26	8.53E-08	delta 0.2	7.558
128	VEGFA-VEGFR2 Signaling Pathway	237	23	4.77E-09	delta 0.2	9.705
139	Focal Adhesion-PI3K-Akt-mTOR-signaling pathway	307	21	7.00E-06	delta 0.2	6.840
160	Chemokine signaling pathway	165	18	3.70E-08	delta 0.2	10.909
172	Ras Signaling	185	17	1.03E-06	delta 0.2	9.189
182	Nuclear Receptors Meta-Pathway	320	16	0.002514	delta 0.2	5.000
211	MAPK Signaling Pathway	249	14	0.001572	delta 0.2	5.622
223	Insulin Signaling	161	13	7.30E-05	delta 0.2	8.075
292	Transcription factor regulation in adipogenesis	22	9	0.013749	CTMeth-hhi	40.909
357	Ras Signaling	185	7	4.73E-06	delta 0.3	3.784
358	MAPK Signaling Pathway	249	7	3.23E-05	delta 0.3	2.811
388	VEGFA-VEGFR2 Signaling Pathway	237	6	0.000212	delta 0.3	2.532
389	Insulin Signaling	161	5	9.17E-06	ChAMP 0.2	3.106
390	VEGFA-VEGFR2 Signaling Pathway	237	5	5.85E-05	ChAMP 0.2	2.110
411	Chemokine signaling pathway	165	5	0.00032	delta 0.3	3.030
412	Chemokine signaling pathway	165	4	0.000199	ChAMP 0.2	2.424
476	Transcription factor regulation in adipogenesis	22	2	0.002767	delta 0.3	9.091

Tabela 28 CLL-100 – KEGG - Trzy listy genów najlepiej skorelowane z wynikami poszczególnych metod oraz ich stopień powiązania z wynikami uzyskanymi z innych metod

	GeneSet	N_genes	N_overlap	p	Metoda	Procent
1	KEGG_PATHWAYS_IN_CANCER	325	152	1.28E-30	CTMeth -hhi	46.76923
2	KEGG_NEUROACTIVE_LIGAND _RECEPTOR_INTERACTION	270	144	4.30E-37	CTMeth -hhi	53.33333
3	KEGG_MAPK_SIGNALING _PATHWAY	267	116	1.83E-20	CTMeth -hhi	43.44569
6	KEGG_CYTOKINE_CYTOKINE _RECEPTOR_INTERACTION	265	99	1.12E-12	CTMeth -hhi	37.35849
8	KEGG_CHEMOKINE _SIGNALING_PATHWAY	189	80	7.89E-14	CTMeth -hhi	42.32804
11	KEGG_ENDOCYTOSIS	181	70	3.70E-10	CTMeth -hhi	38.67403
35	KEGG_ADHERENS_JUNCTION	73	41	1.37E-12	CTMeth -hhi	56.16438
85	KEGG_PATHWAYS_IN_CANCER	325	21	1.66E-05	delta 0.2	6.461538
89	KEGG_CYTOKINE_CYTOKINE _RECEPTOR_INTERACTION	265	20	2.65E-06	delta 0.2	7.54717
95	KEGG_CHEMOKINE _SIGNALING_PATHWAY	189	19	5.85E-08	delta 0.2	10.05291
103	KEGG_ENDOCYTOSIS	181	16	3.53E-06	delta 0.2	8.839779
116	KEGG_MAPK _SIGNALING_PATHWAY	267	14	0.002979	delta 0.2	5.243446
143	KEGG_ADHERENS_JUNCTION	73	9	3.62E-05	delta 0.2	12.32877
157	KEGG_ENDOCYTOSIS	35	7	1.06E-05	delta 0.2	20
170	KEGG_MAPK _SIGNALING_PATHWAY	267	7	5.02E-05	delta 0.3	2.621723
183	KEGG_ENDOCYTOSIS	181	4	0.000283	ChAMP 0.2	2.209945
184	KEGG_CHEMOKINE _SIGNALING_PATHWAY	189	4	0.000333	ChAMP 0.2	2.116402
188	KEGG_ADHERENS_JUNCTION	73	3	0.000284	ChAMP 0.2	4.109589

Tabela 29 CLL-100 – Oncogenic signatures - Trzy listy genów najlepiej skorelowane z wynikami poszczególnych metod oraz ich stopień powiązania z wynikami uzyskanymi z innych metod

	GeneSet	N_genes	N_overlap	p	SourceFile	Procent
1	NFE2L2.V2	465	180	7.66E-24	CTMeth-hhi	38.70968
2	KRAS.600.LUNG. BREAST_UP.V1_UP	274	136	7.66E-31	CTMeth-hhi	49.63504
3	KRAS.600_UP.V1_UP	276	134	3.52E-29	CTMeth-hhi	48.55072
4	KRAS.600_UP.V1_DN	278	131	5.62E-27	CTMeth-hhi	47.1223
6	TBK1.DF_UP	287	114	1.37E-16	CTMeth-hhi	39.72125
17	STK33_SKM_UP	272	96	1.00E-10	CTMeth-hhi	35.29412
20	STK33_UP	284	94	6.86E-09	CTMeth-hhi	33.09859
37	RAF_UP.V1_UP	194	85	1.17E-15	CTMeth-hhi	43.81443
49	CAMP_UP.V1_DN	199	83	7.32E-14	CTMeth-hhi	41.70854
51	CYCLIN_D1_KE_.V1_UP	188	82	5.04E-15	CTMeth-hhi	43.61702
56	EGFR_UP.V1_UP	192	80	2.18E-13	CTMeth-hhi	41.66667
117	ESC_J1_UP_EARLY.V1_DN	179	64	6.99E-08	CTMeth-hhi	35.75419
154	RB_P130_DN.V1_UP	130	50	1.38E-07	CTMeth-hhi	38.46154
169	CORDENONSI_YAP _CONSERVED_SIGNATURE	57	26	3.55E-06	CTMeth-hhi	45.61404
174	TBK1.DF_UP	287	23	1.66E-07	delta 0.2	8.013937
177	CAMP_UP.V1_DN	199	20	2.65E-08	delta 0.2	10.05025
179	STK33_SKM_UP	272	19	1.41E-05	delta 0.2	6.985294
180	STK33_UP	284	19	2.57E-05	delta 0.2	6.690141
192	EGFR_UP.V1_UP	192	16	7.52E-06	delta 0.2	8.333333
193	RAF_UP.V1_UP	194	16	8.57E-06	delta 0.2	8.247423
203	KRAS.600_UP.V1_UP	276	13	0.00981	delta 0.2	4.710145
222	CYCLIN_D1_KE_.V1_UP	188	11	0.003555	delta 0.2	5.851064
240	RB_P130_DN.V1_UP	130	9	0.00267	delta 0.2	6.923077
253	RB_P130_DN.V1_UP	130	6	5.51E-05	CTMeth-hh	4.615385
254	CORDENONSI_YAP _CONSERVED_SIGNATURE	57	6	0.001691	delta 0.2	10.52632
255	EGFR_UP.V1_UP	192	5	2.14E-05	ChAMP 0.2	2.604167
258	RAF_UP.V1_UP	194	4	0.000367	ChAMP 0.2	2.061856
259	ESC_J1_UP_EARLY.V1_DN	179	2	0.000533	Delta 0.4	1.117318
260	CYCLIN_D1_KE_.V1_UP	188	2	0.000587	Delta 0.4	1.06383
261	EGFR_UP.V1_UP	192	2	0.000612	Delta 0.4	1.041667

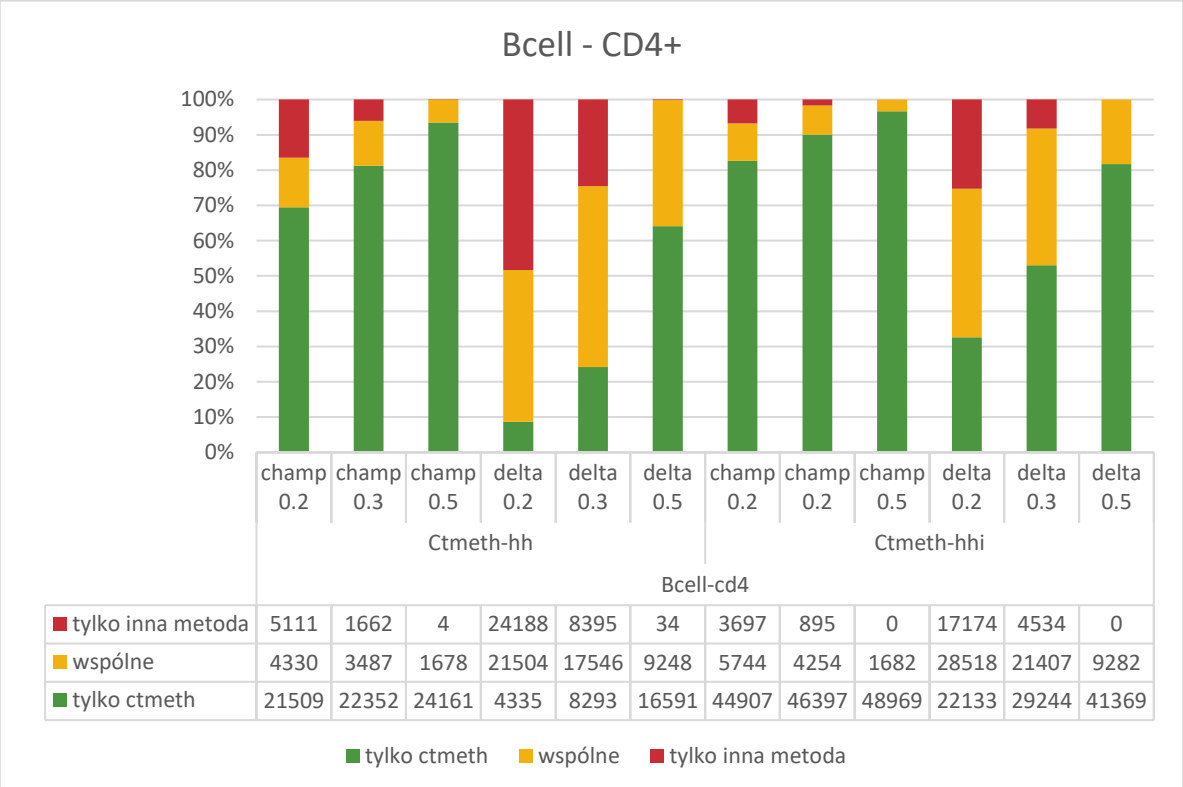
Tabela 30 CLL-100 – Wikipathways – Wnt Signaling

GeneSet	Category	N_genes	N_overlap	p	SourceFile	Procent
Wnt Signaling	Wikipathways	115	55	1.62E-12	CTMeth-hhi	47.826
Wnt/beta-catenin Signaling Pathway in Leukemia	Wikipathways	26	14	6.77E-05	CTMeth-hhi	53.846
Wnt Signaling	Wikipathways	115	8	4.39E-03	delta 0.2	6.957

Porównanie z użyciem podstawowych operacji na zbiorach

B-Cell-CD4+

Dla zbioru B-Cell-CD4+ metody z użyciem sieci neuronowej wskazują więcej sekwencji CpG, które są pominięte przez standardowe techniki. Wyjątkiem jest użycie metody delta 0.2 i 0.3 w porównaniu z narzędziem CTMeth-hh. Dokładniejsza analiza stosunków różnic w wynikach sieci neuronowej i standardowych metod przedstawiono na Rycina 39.

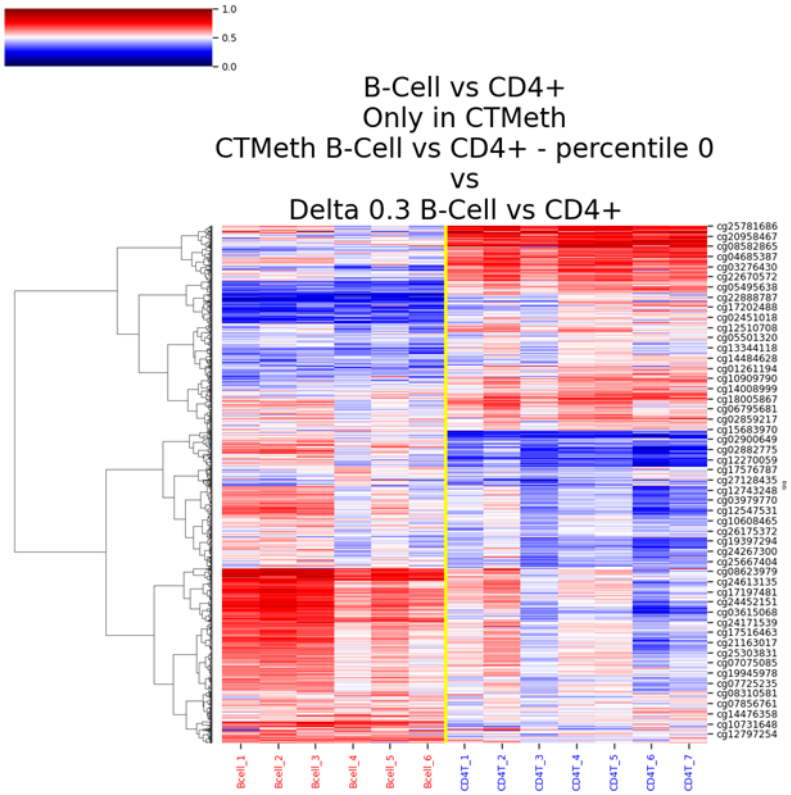


Rycina 39 Stosunek różnic i podobieństw w sekwencjach znalezionych przez poszczególne metody dla zbioru B-Cell-CD4+

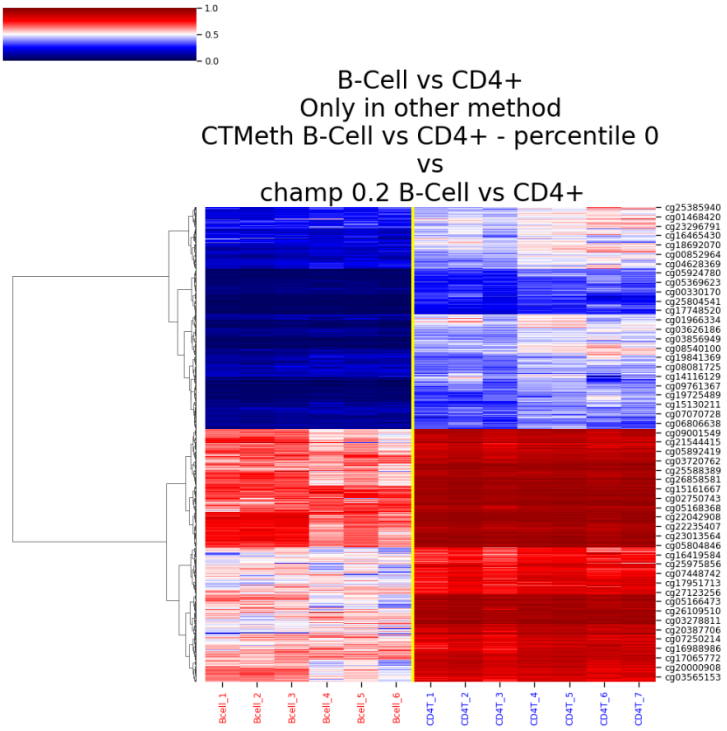
Pomimo zastosowania mniejszej różnicy dla modeli regresji i metody delta znacząca liczba sekwencji CpG, które różnią się metylacją analogicznie do modelu przyjętego przez Bibikova i wsp. [159] jest przez te metody pomijana przy jednoczesnym

wskazaniu sekwencji, które nie spełniają tego kryterium – niezależnie od porównania do CTMeth-hh, czy CTMeth-hhi co przedstawiono na hierarchicznie sklastrowanych względem sekwencji CpG mapach cieplnych (Rycina 40,Rycina 41,Rycina 42,Rycina 43,Rycina 44,

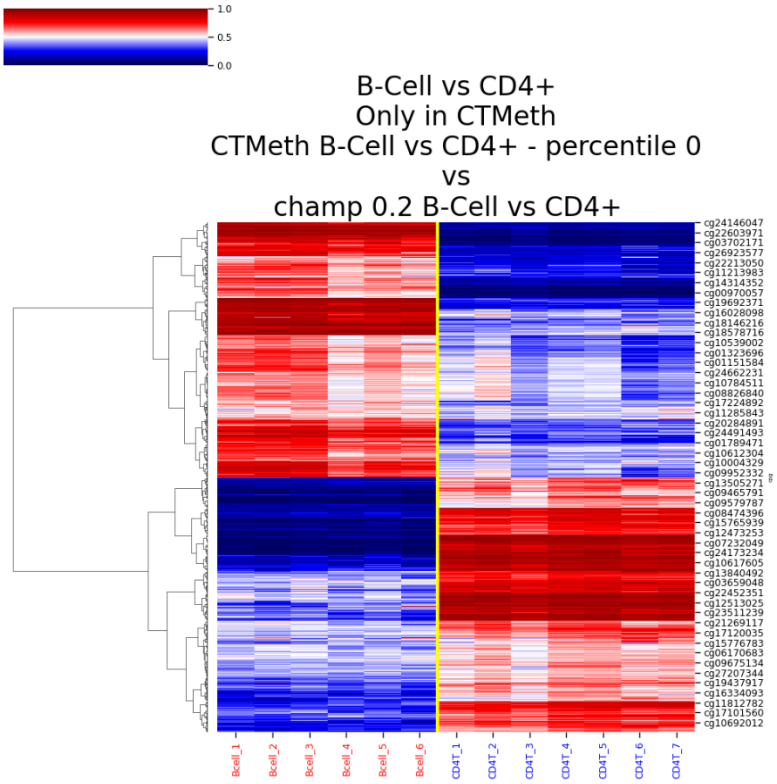
Rycina 45,Rycina 46,Rycina 47,Rycina 48,Rycina 49). Użyto właśnie tej formy prezentacji wyników, ze względu na najbardziej przejrzystą formę przedstawienia zależności przy tej ilości danych.



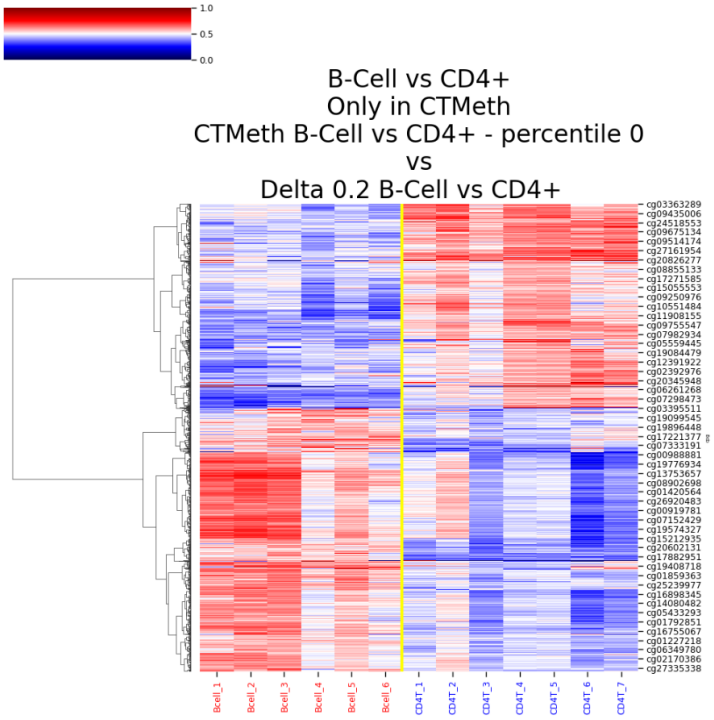
Rycina 40 Zbiór B-Cell-CD4+ - porównanie metody CTMeth -hh z Delta 0.3 - sekwencje znalezione tylko przez metodę CTMeth-hhi



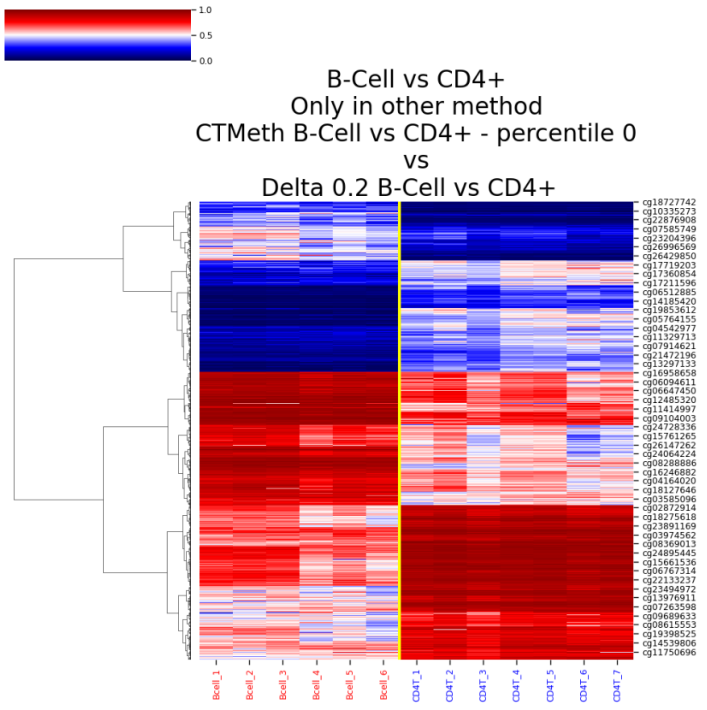
Rycina 41 Zbiór B-Cell-CD4+ - porównanie metody CTMeth -hh z ChAMP 0.2 - sekwencje znalezione tylko przez metodę ChAMP 0.2



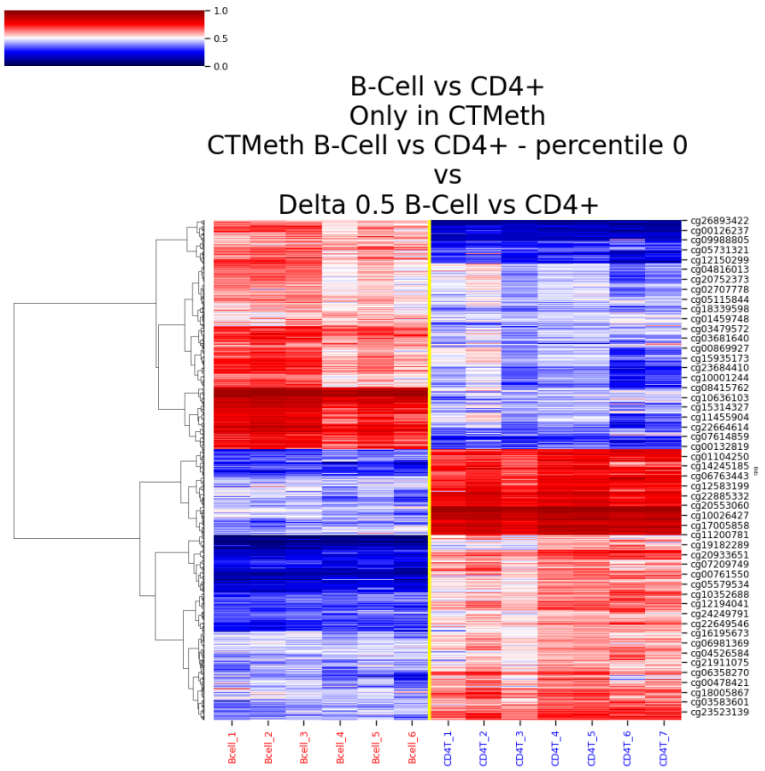
Rycina 42 Zbiór B-Cell-CD4+- porównanie metody CTMeth -hh z ChAMP 0.2 - sekwencje znalezione tylko przez metodę CTMeth-hh



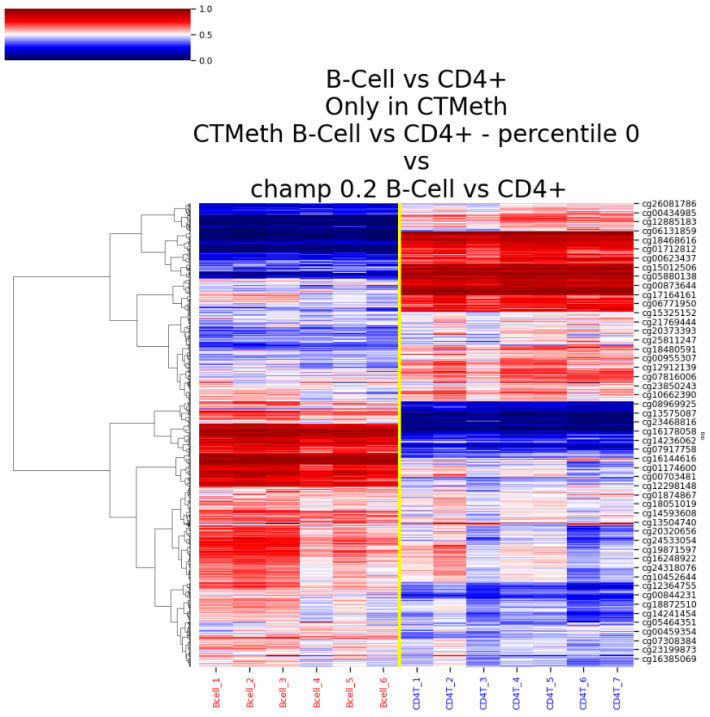
Rycina 43 Zbiór B-Cell-CD4+ - porównanie metody CTMeth -hh z Delta 0.2 - sekwencje znalezione tylko przez metodę CTMeth-hh



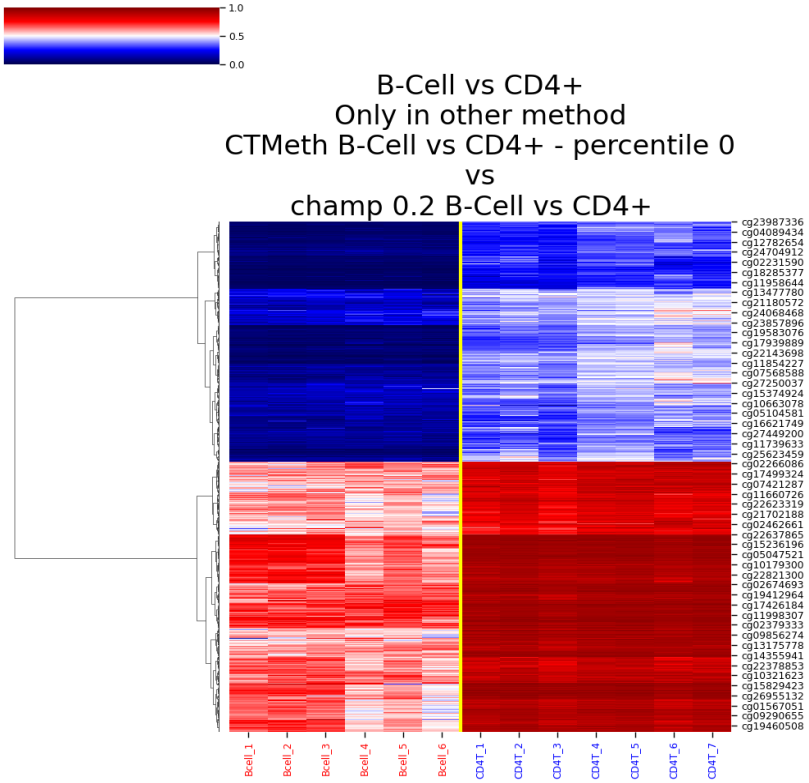
Rycina 44 Zbiór B-Cell-CD4+ - porównanie metody CTMeth -hh z Delta 0.2 - sekwencje znalezione tylko przez metodę Delta 0.2



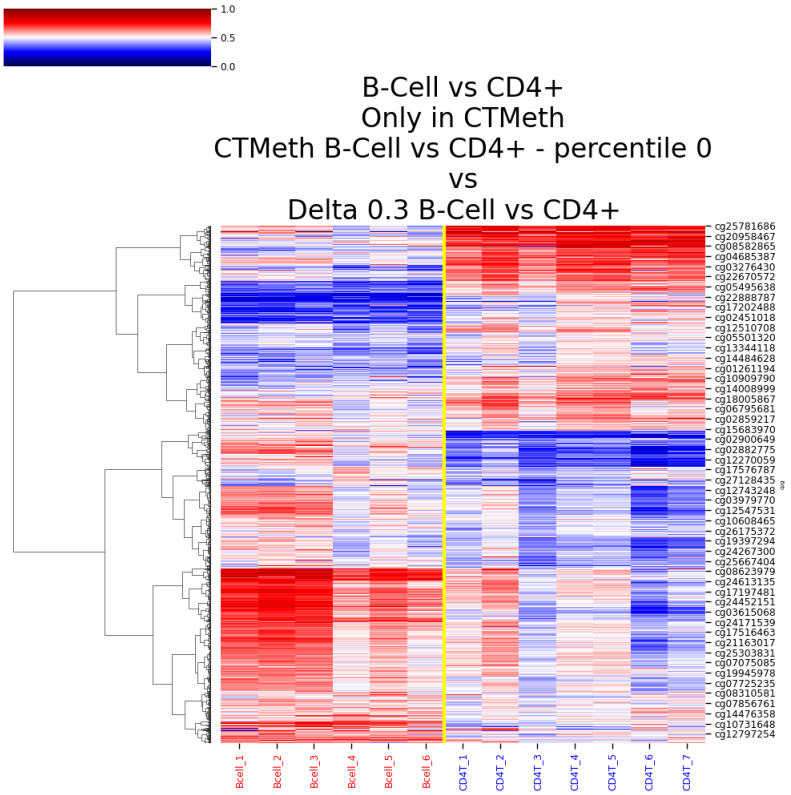
Rycina 45 Zbiór B-Cell-CD4+- porównanie metody CTMeth -hh z Delta 0.5 - sekwencje znalezione tylko przez metodę Ctmeth-hh



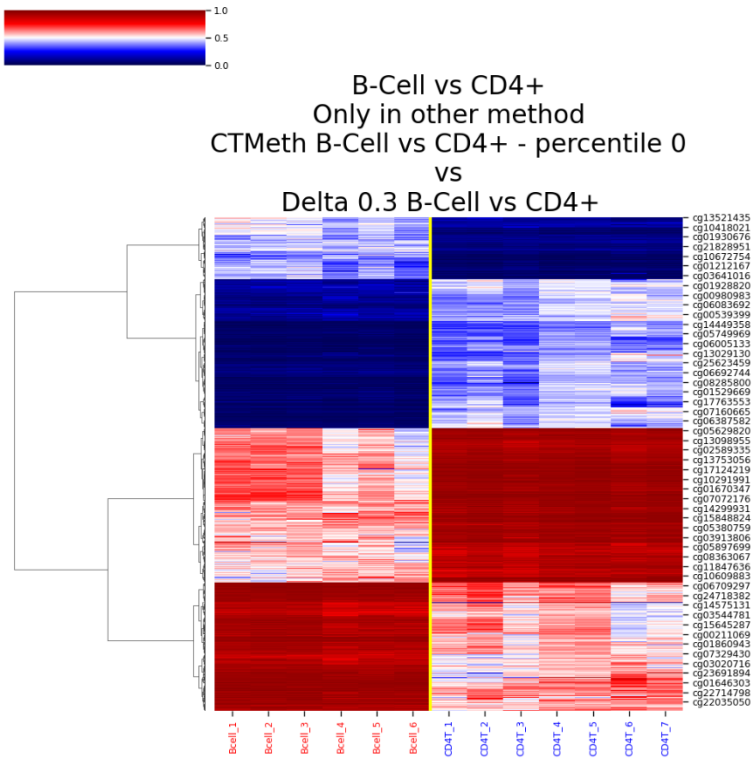
Rycina 46 Zbiór B-Cell-CD4+- porównanie metody CTMeth -hhi z ChAMP 0.2 - sekwencje znalezione tylko przez metodę CTMeth - hhi



Rycina 47 Zbiór B-Cell-CD4+- porównanie metody CTMeth -hhi z ChAMP 0.2 - sekwencje znalezione tylko przez metodę ChAMP 0.2



Rycina 48 Zbiór B-Cell-CD4+ - porównanie metody CTMeth -hhi z Delta 0.3 - sekwencje znalezione tylko przez metodę CTMeth - hhi

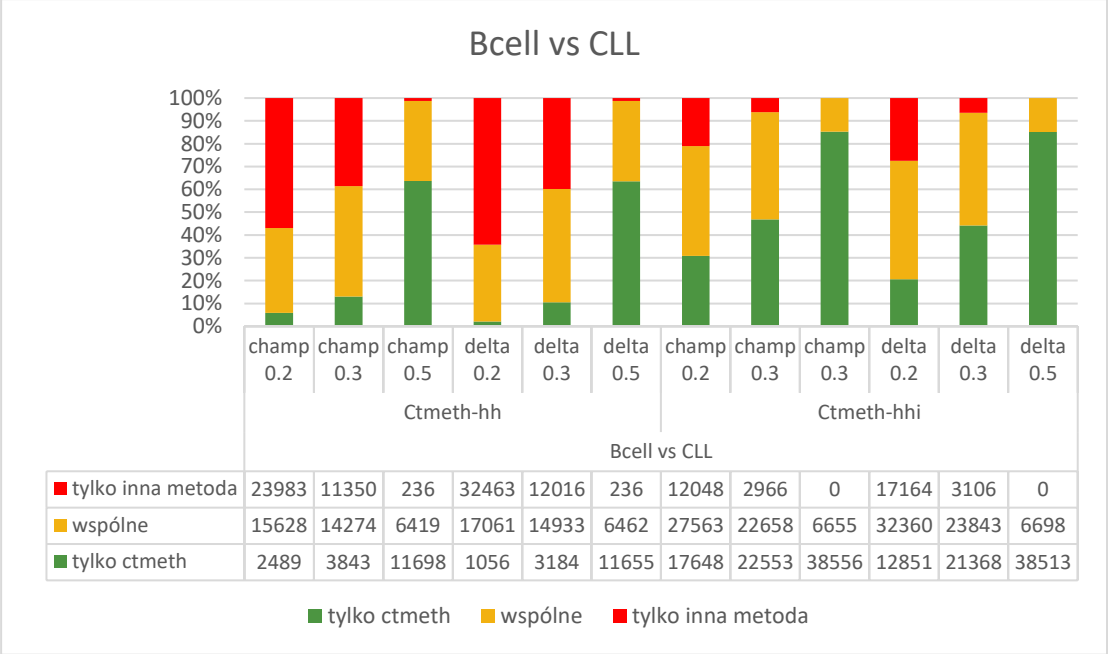


Rycina 49 Zbiór B-Cell-CD4+ - porównanie metody CTMeth -hhi z Delta 0.3 - sekwencje znalezione tylko przez metodę Delta 0.3

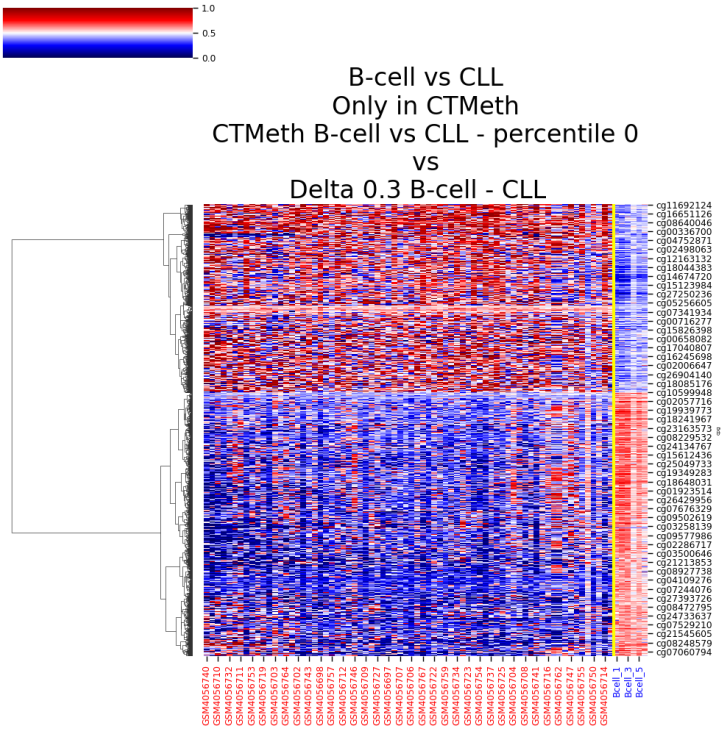
B-Cell – CLL

Rozkład wspólnych i unikalnych sekwencji CpG dla metod CTMeth i delta jest podobny pomiędzy wynikami uzyskanymi ze zbiorów B-Cell-CLL i B-Cell-CD4+(Rycina 50). W przypadku porównania metod CTMeth do ChAMP to standardowa metoda do analizy metylacji wskazuje więcej unikalnych rekordów podczas analizy danych B-Cell-CLL. W analizie map cieplnych z użyciem klastrowania hierarchicznego w kontekście oznaczania sekwencji analogicznie do założeń Bibikova i wsp. [159] w unikalnych wynikach uzyskanych z metody CTMeth-hh widać dobrze odgraniczone klastry z dobrze widocznymi różnicami w metylacji pomiędzy grupą kontrolną a badaną (Rycina 51). W przypadku analizy map cieplnych uzyskanych dla unikalnych wyników powstałych z użycia narzędzia CTMeth-hhi obraz jest mniej jednoznaczny, aczkolwiek widoczna jest odmienność w metylacji pomiędzy badanymi grupami (Rycina 52). Sekwencje CpG wskazane przez metody standardowe, a nie występujące w wynikach metod uzyskanych z zastosowaniem sieci neuronowych przedstawiono w formie map cieplnych z hierarchicznym grupowaniem na rycinach (Rycina 53,Rycina 54). W tym przypadku widać dobrze odgraniczone klastry,

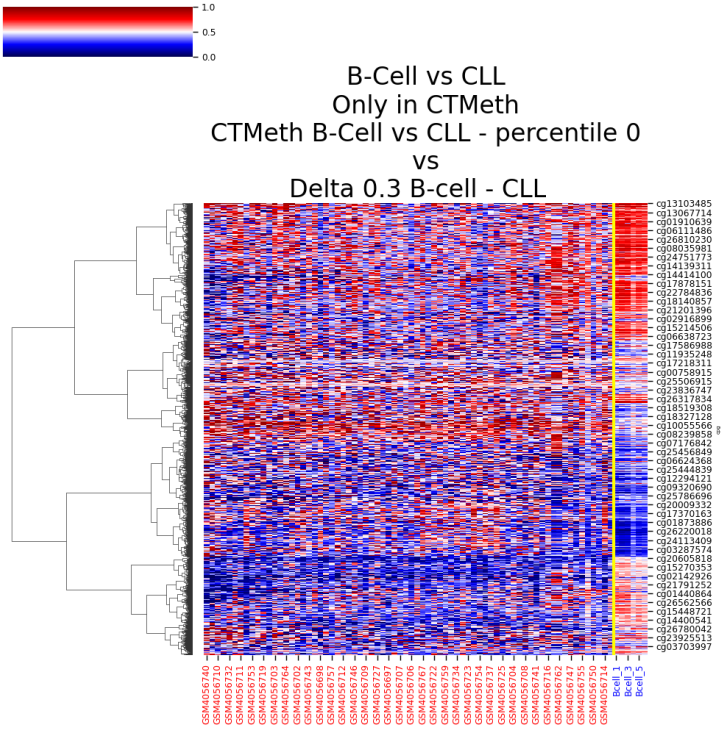
aczkolwiek wyniki nie spełniają kryteriów analogicznych do Bibikova i wsp. [159]. Przykład wyników wspólnych dla metod opartych o sieci neuronowe i klasyczne przedstawiono na Rycina 55.



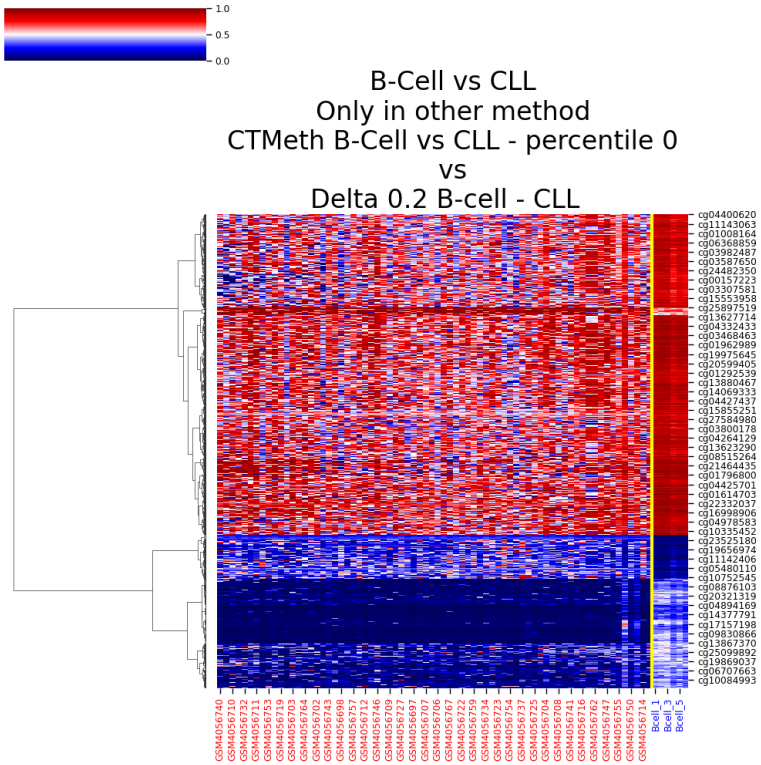
Rycina 50 Stosunek różnic i podobieństw w sekwencjach znalezionych przez poszczególne metody dla zbioru B-Cell-CLL



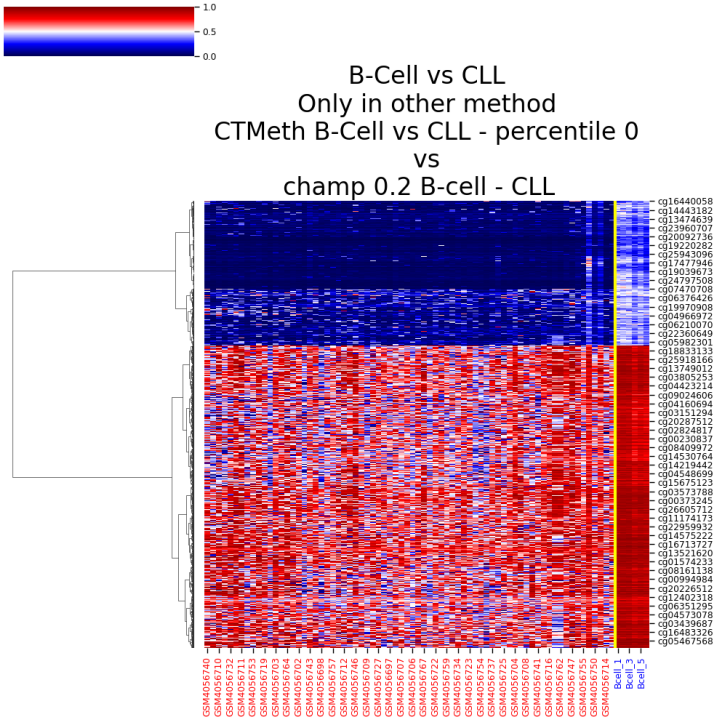
Rycina 51 Zbiór B-Cell-CLL - porównanie metody CTMeth -hh z Delta 0.3 - sekwencje znalezione tylko przez metodę CTMeth-hh



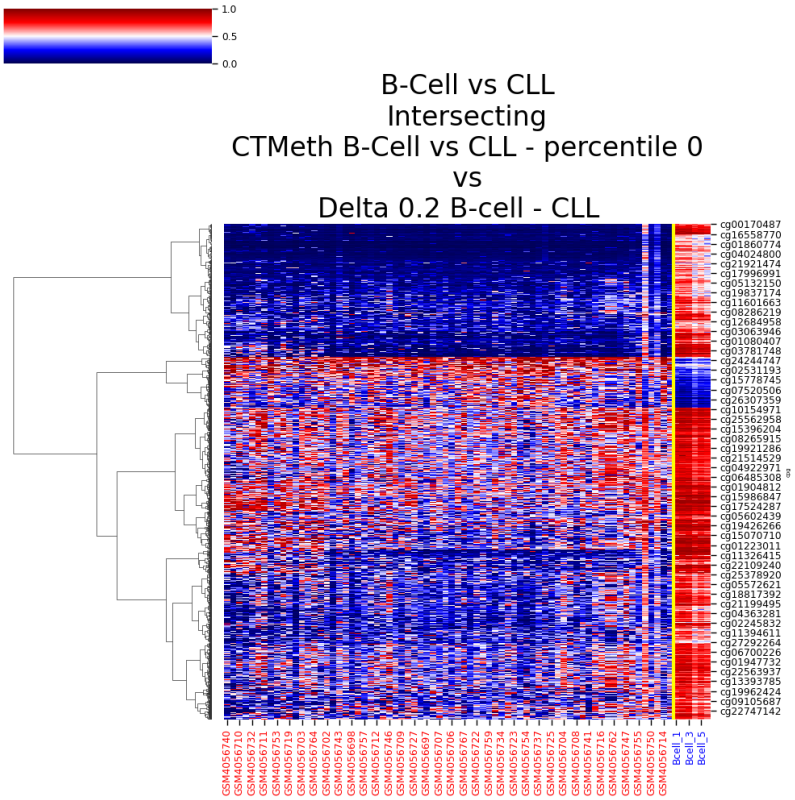
Rycina 52 Zbiór B-Cell-CLL - porównanie metody CTMeth -hhi z Delta 0.3 - sekwencje znalezione tylko przez metodę CTMeth-hhi



Rycina 53 Zbiór B-Cell-CLL - porównanie metody CTMeth -hhi z Delta 0.3 - sekwencje znalezione tylko przez metodę delta 0.3



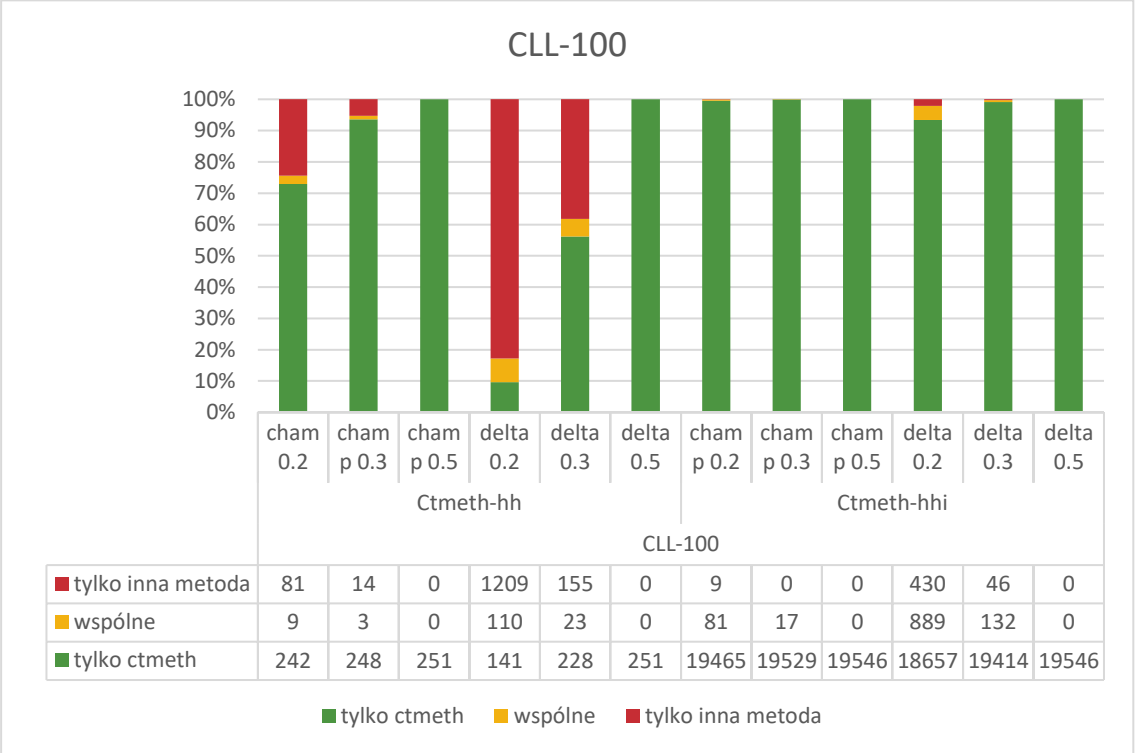
Rycina 54 Zbiór B-Cell-CLL - porównanie metody CTMeth -hhi z ChAMP 0.2 - sekwencje znalezione tylko przez metodę ChAMP 0.2



Rycina 55 Zbiór B-Cell-CLL - porównanie metody CTMeth -hhi z Delta 0.2 - sekwencje wspólne dla obu metod

CLL-100

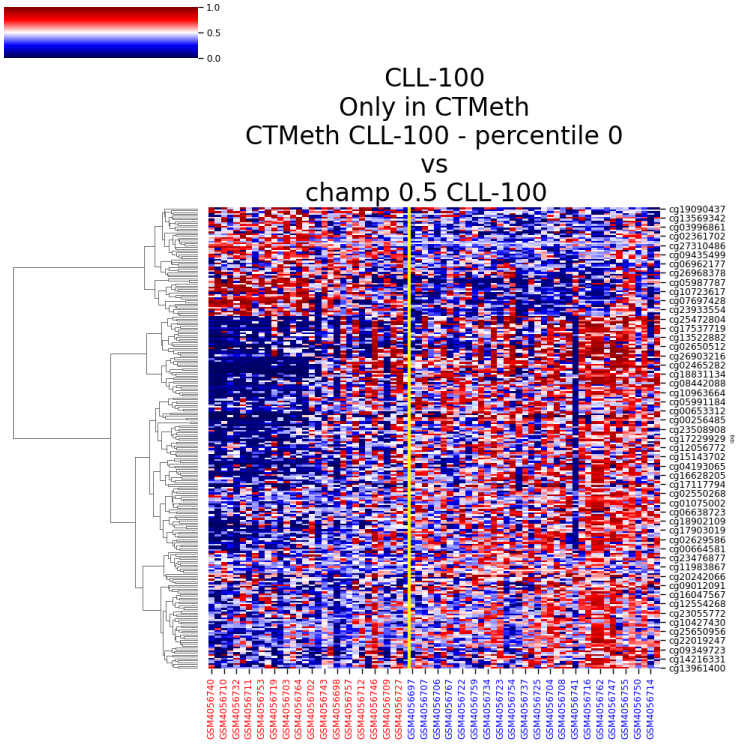
Analiza bardziej skomplikowanego zapisu metylacji, jakim jest różnica w metylacji pomiędzy pacjentami chorymi na przewlekłą białaczkę limfocytową pod względem IGHV obrazuje inne zachowanie poszczególnych metod. Dwie z zastosowanych metod nie wykazały żadnych różnic w metylacji pomiędzy grupami kontrolną i badaną (ChAMP 0.5, Delta 0.5) (Rycina 56).



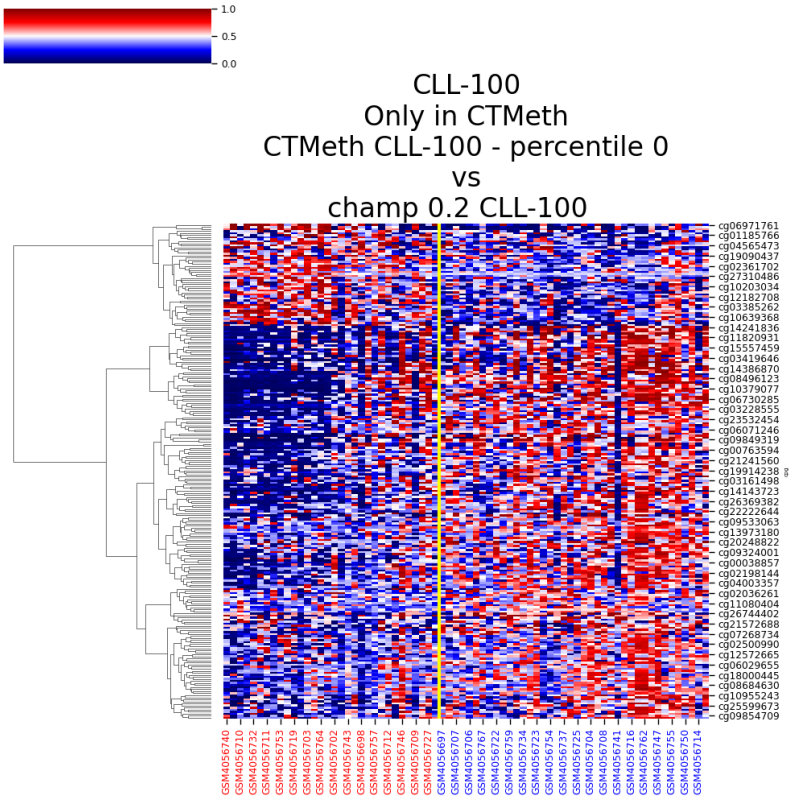
Rycina 56 Stosunek różnic i podobieństw w sekwencjach znalezionych przez poszczególne metody dla zbioru CLL-100

Pominięte przez te metody sekwencje CpG w porównaniu do metody CTMeth-hh przedstawiono na Rycina 57. Metoda CTMeth-hh wskazuje sekwencje odpowiadające poszukiwanemu kryterium, które są bardziej jednolite w obrębie grupy (Rycina 58), a metody standardowe wskazały sekwencję, w których w jednej z grup widoczna jest podgrupa (Rycina 59 ,Rycina 60). Wersja sieci neuronowej CTMeth-hhi w wynikach również wskazuje sekwencje, które świadczą o istnieniu podgrupy (Rycina 61). Wykres pominiętych przez metodę delta 0.2 CpG, a wskazanych przez metodę CTMeth -hhi pokazuje Rycina 62, a odwrotnie pominiętych przez CTMeth-hhi, a wskazanych przez delta 0.2 Rycina 63. W tych dwóch przypadkach wykonano dodatkową analizę sporządzając wykresy dla pojedynczych sekwencji CpG. Dla lepszej wizualizacji przykładowe wyniki przedstawiono na wykresach, gdzie wartości w obrębie grupy posegregowano od najmniejszej do największej wartości, jak na Rycina 64. Bardziej szczegółowa analiza do poziomu poszczególnych CpG pokazuje, że delta 0.2 jest metodą wrażliwą na problem outliers (Rycina 65), natomiast CTMeth-hhi mimo

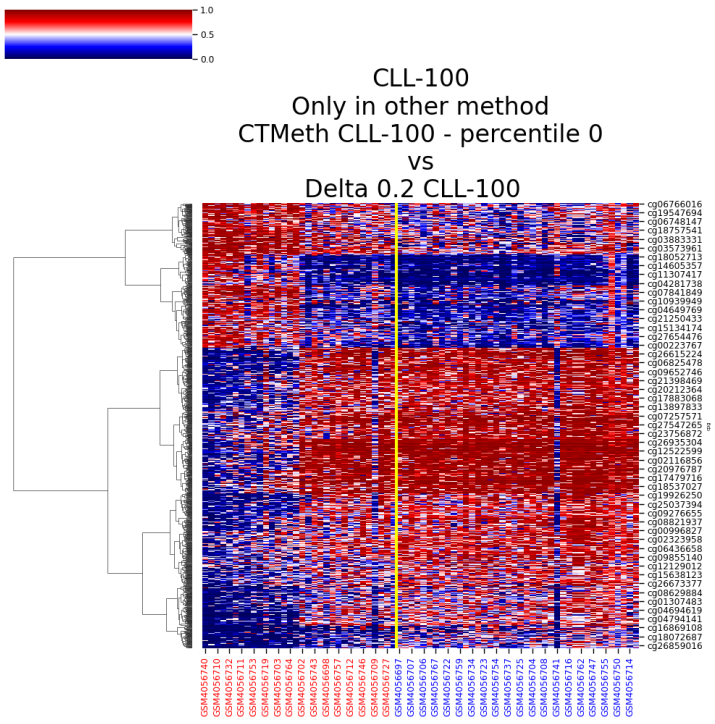
niejednoznacznego obrazu rezultatów na Rycina 62 wskazuje sekwencje CpG o ogólnych różnicach w metylacji pomiędzy grupami kontrolną i badaną (Rycina 66).



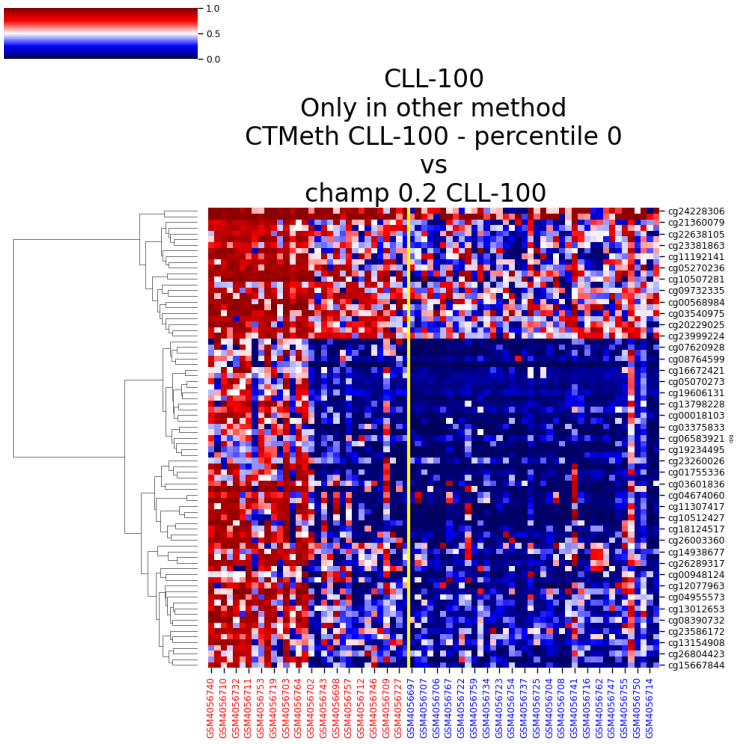
Rycina 57 Zbiór CLL-100 porównanie metody CTMeth -hh z ChAMP 0.5 - sekwencje znalezione tylko przez CTMeth - hh



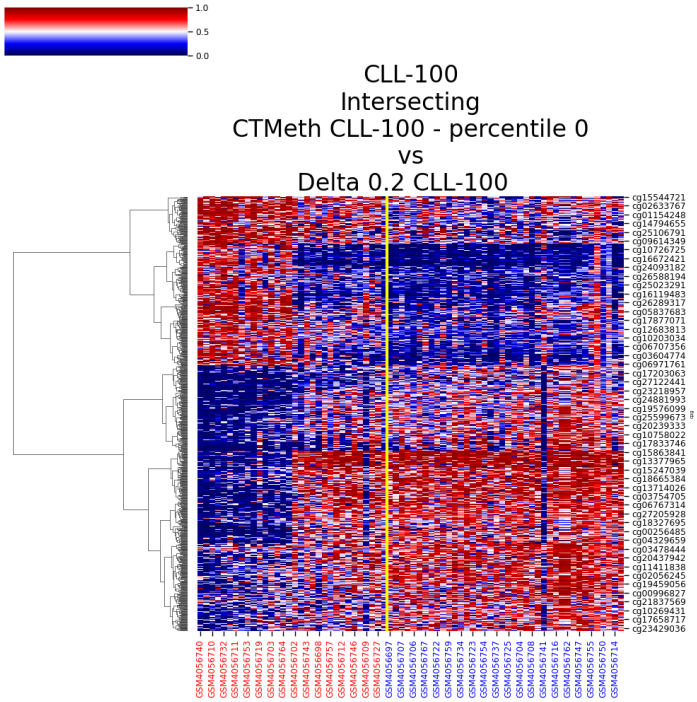
Rycina 58 Zbiór CLL-100 porównanie metody CTMeth -hh z ChAMP 0.2 - sekwencje znalezione tylko przez CTMeth - hh



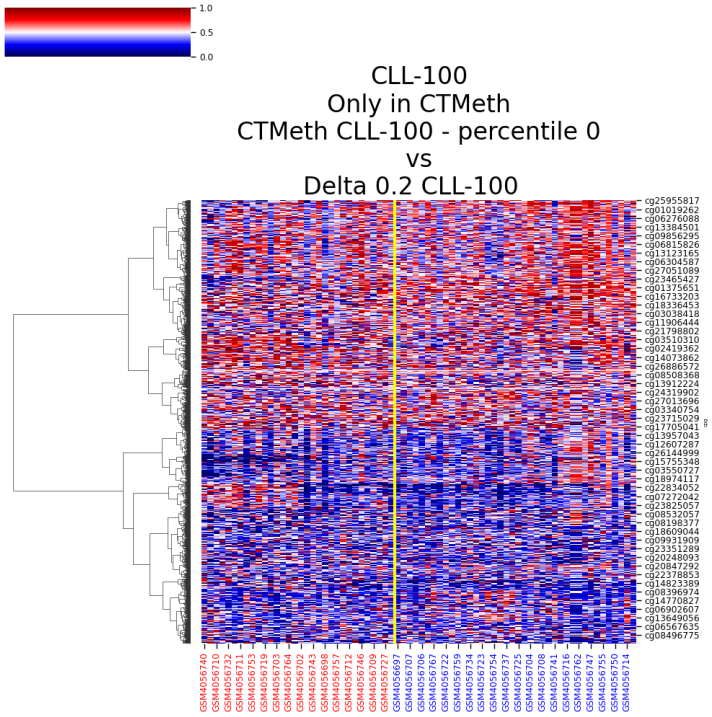
Rycina 59 Zbiór CLL-100 porównanie metody CTMeth -hh z Delta 0.2 - sekwencje znalezione tylko przez Delta 0.2



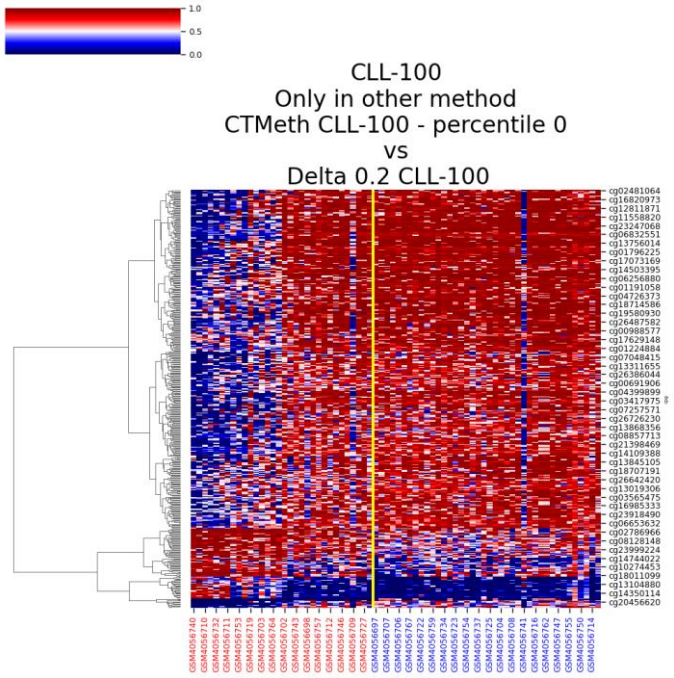
Rycina 60 Zbiór CLL-100 porównanie metody CTMeth -hh z ChAMP 0.2 - sekwencje znalezione tylko przez ChAMP 0.2



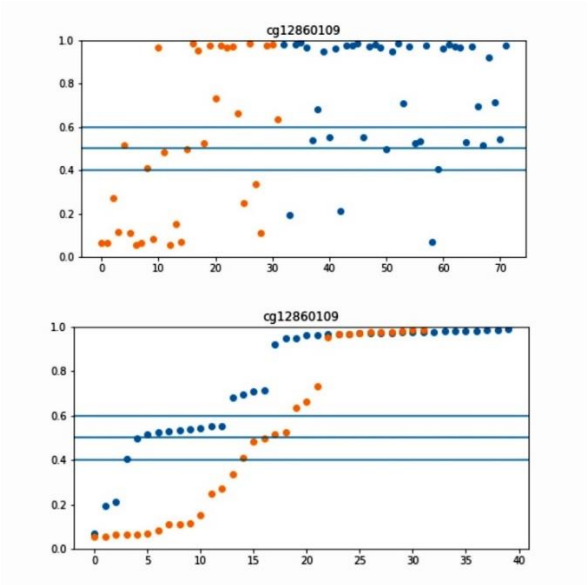
Rycina 61 Zbiór CLL-100 porównanie metody CTMeth -hhi z delta 0.2 - sekwencje znalezione wspólnie przez obie metody



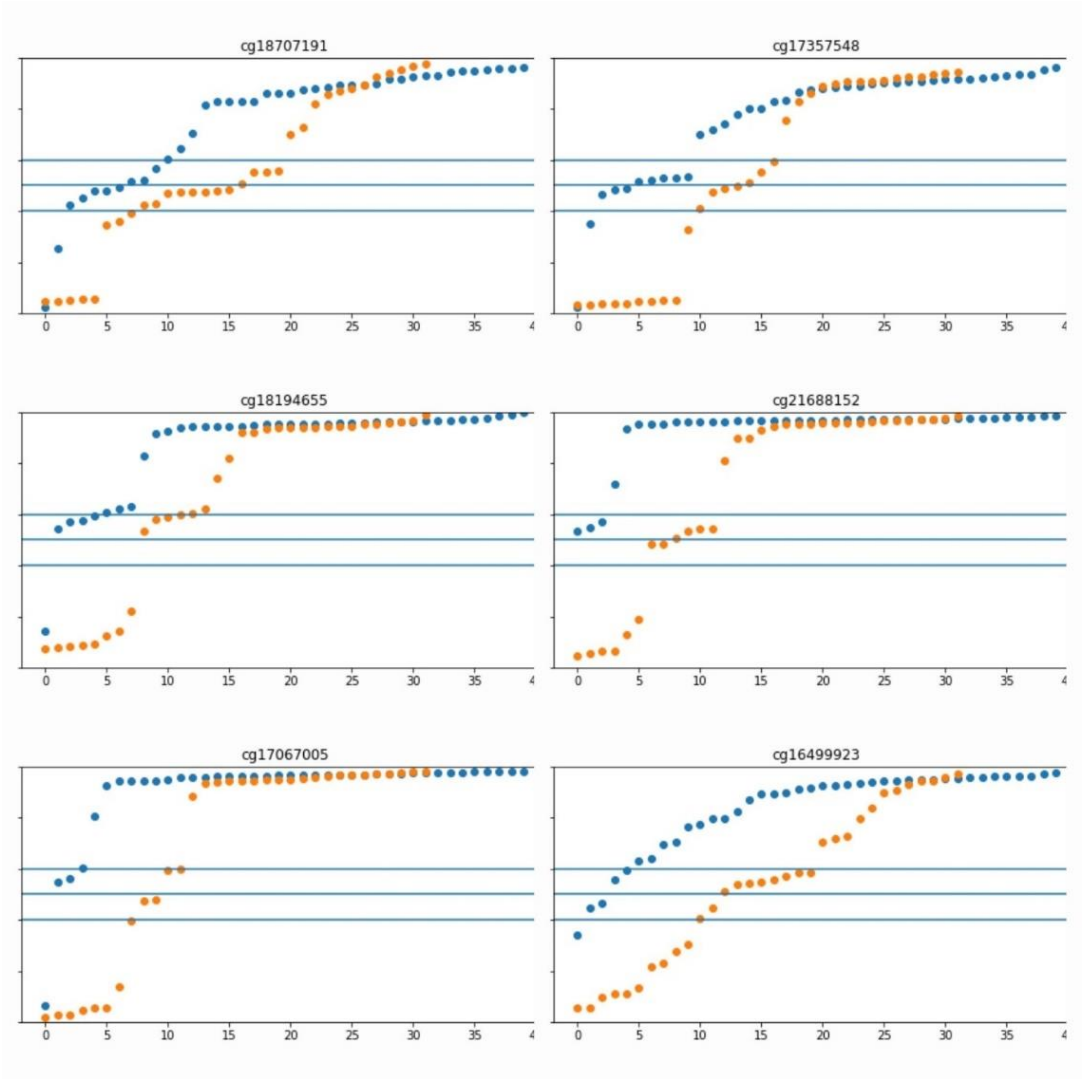
Rycina 62 Zbiór CLL-100 porównanie metody CTMeth -hhi z delta 0.2 - sekwencje znalezione przez CTMeth – hhi



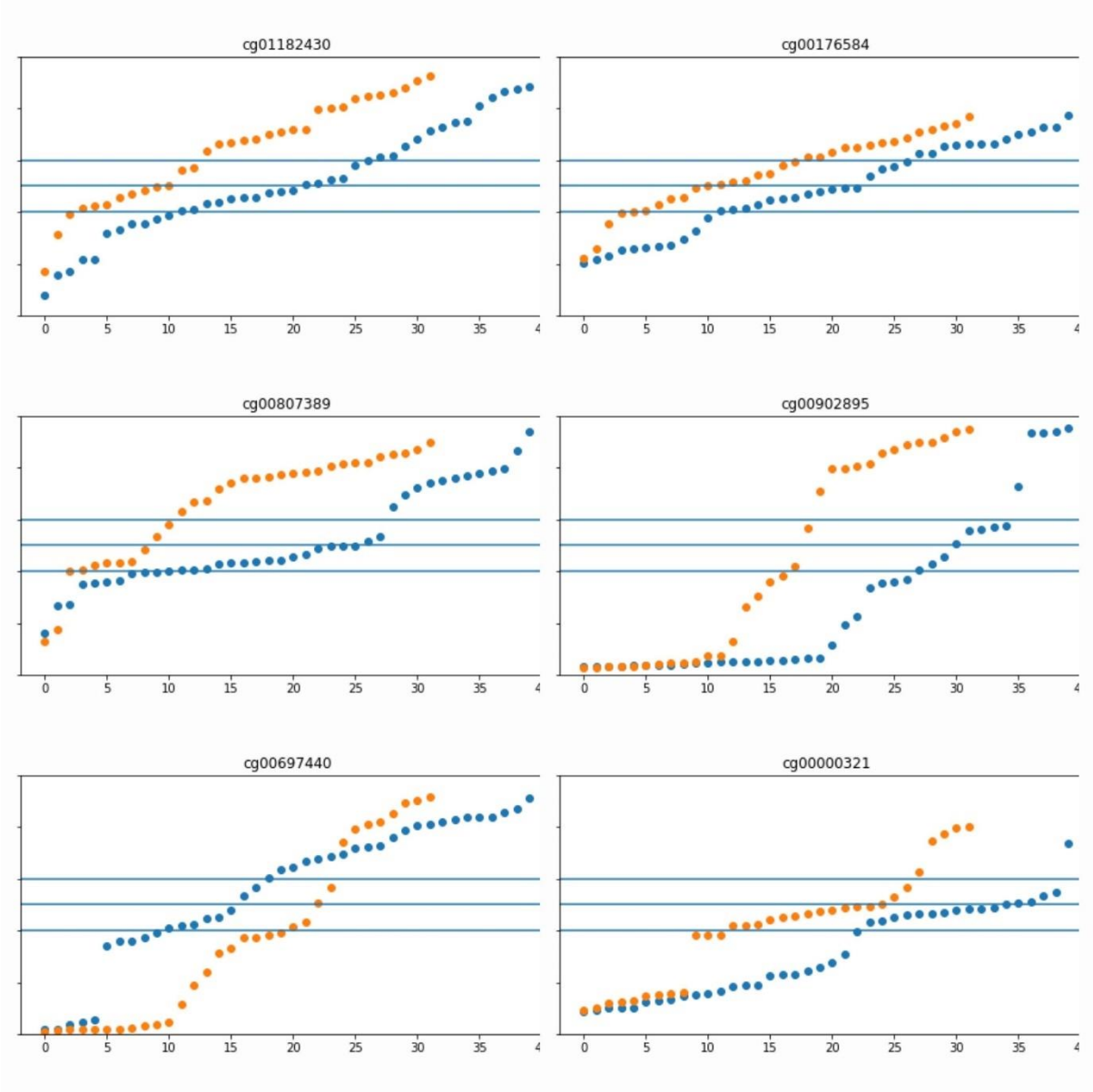
Rycina 63 Zbiór CLL-100 porównanie metody CTMeth -hhi z delta 0.2 sekwencje znalezione przez delta 0.2



Rycina 64 Dla poprawy czytelności wartości β w obrębie grup posegregowane od najmniejszej do największej wartości. Kolorem pomarańczowym oznaczono - grupę kontrolną, a niebieskim badaną – prezentacja techniki



Rycina 65 Przykładowe sekwencje CpG ze zbioru CLL-100 wskazywane przez metodę delta 0.2



Rycina 66 Przykładowe sekwencje CpG ze zbioru CLL-100 wskazywane przez metodę CTMETH-hhi

Analiza z użyciem symulowanych danych

Analiza z użyciem symulowanych danych wykazała, że metody CTMeth-hh i CTMeth-hhi jako jedyne nie dają wyników zarówno fałszywie negatywnych i fałszywie dodatnich. Jako kryterium przyjęto punkt odcięcia analogiczny do Bibikova i wsp. [159]. Pełne dane przedstawiono w Tabela 31.

Tabela 31 Analiza z użyciem symulowanych danych

Metoda	Fałszywie negatywne	Fałszywie dodatnie
ChAMP 0.5	758	0
ChAMP 0.3	279	35
ChAMP 0.2	65	167
Delta 0.5	783	0
Delta 0.3	294	32
Delta 0.2	70	167
CTMeth -hhi	0	0
CTMeth-hh	0	0

Wnioski

Do stworzenia prototypu tej sieci neuronowej wykorzystano środowisko PyTorch [364] i język programowania Python [351]. Język programowania Python jest opracowanym w 1991 roku językiem wysokiego poziomu, który dzięki swojej czytelnej składni stał się liderem pod względem popularności, a co za tym idzie jest to język o bardzo szerokim spektrum zastosowania począwszy od analizy danych, przez publikację stron internetowych, budowę aplikacji, czy rozwój technologii związanych ze sztuczną inteligencją. PyTorch z kolei to jedno z dwóch dominujących środowisk do rozwoju sieci neuronowych i szeroko pojętej sztucznej inteligencji. Stanowi ono zespół gotowych i sprawdzonych narzędzi, które pozwalają na budowę bardziej złożonych lub odpowiednio dostosowanych rozwiązań. Użyta architektura sieci neuronowej jest połączeniem dwóch typów architektur, którymi są sieć konwolucyjna [303] i typu transformers [343]. Sieć typu transformers to nowoczesne rozwiązanie, które przez zastosowanie mechanizmu uwagi charakteryzuje się wysoką efektywnością, wydajnością oraz zdolnością generalizacji. Dodatkowo jest to architektura elastyczna, którą można dostosować do wielu zadań. Jednakże, ze względu na fakt, że została ona stworzona z myślą o zadaniach związanych z przetwarzaniem naturalnego języka jako dane wejściowe wymaga ona odpowiedniego przetworzenia danych do tzw. tokenów. Dane na temat metylacji nie są danymi, które można w łatwy sposób poddać tokenizacji w podobny sposób, jak to jest wykonywane w przypadku przetwarzania naturalnego języka, gdzie tokeny tworzone są przez rozdzielenie ciągu wyrazów za pomocą przecinków, odstępów czy kropek. Z tego powodu użyto sieci konwolucyjnej, która ma zdolność do znajdowania różnicujących elementów przez tworzenie mapy cech. Zastosowano warstwę jednowymiarową, mając na uwadze, iż ciąg danych o metylacji nie wymaga zastosowania typowej dwuwymiarowej warstwy konwolucyjnej oraz

redukując w ten sposób koszt związany z mocą obliczeniową. Sama architektura transformers w przypadku CTMeth została pozbawiona warstwy dekodera, ponieważ celem jest klasyfikacja danych na temat metylacji, a nie ich przekształcenie w inne dane. Ma to również znaczenie w kontekście optymalizacji wykorzystania zasobów obliczeniowych. Celem prewencji nadmiernego dopasowania do treningowego zbioru danych do architektury sieci neuronowej włączono warstwy droput oraz wykorzystano optymalizator AdamW [365]. Sieć neuronowa użyta w opracowanej bibliotece CTMeth umożliwia użytkownikowi korzystanie z dwóch rodzajów wyników - jednego koncentrującego się na różnicach skrajnych jakimi jest różnicowanie pod względem hiper- i hipometylacji (CTMeth-hh), oraz drugiego, który włącza w to stany pośrednie i nieokreślone (niedominującej) w obrębie grupy metylacji. Dzięki temu użytkownicy mają większą swobodę w wyborze wyników, które najlepiej odpowiadają ich potrzebom. Zastosowane punkty odcięcia przy tworzeniu treningowego zbioru danych dla wartości hiper- i hipometylowanych odzwierciedlają te stosowane przez wielu innych badaczy m. in Bibikova i wsp. [159], którzy wartość β dla sekwencji metylowanych określają jako >0.75 i <0.2 dla niemetylowanych, jednocześnie wyróżniając grupę sekwencji CpG oscylujących wokół wartości 0.5 nazwaną częściowo metylowaną. Trening sieci neuronowej oparto o syntetyczny zbiór treningowy ograniczony regułami tak by wpasować się w wyżej wymienione punkty odcięcia oraz z wykorzystaniem opisanego wcześniej słownika zgeneralizowanych pojęć wykorzystując zdolność sieci neuronowych do funkcjonowania jako uniwersalny aproksymator. Jako funkcję aktywacji wybrano funkcję ReLu. Jest to powszechnie używana funkcja aktywacji mimo nie spełniania przez nią wszystkich założeń dla funkcji wg. Datta [306] i faktu, że powstało wiele jej udoskonalonych wersji. Udoskonalone wersje testowane były dla innych modeli, dlatego postanowiono o pozostaniu przy powszechnym i sprawdzonym narzędziu. Wybór metody Kaiming do inicjalizacji parametrów podyktowany był użyciem funkcji ReLu i oparty na badaniach m. in Li i wsp. [319] i Kumar i wsp. [316]. Zastosowanie wyżej wspominanych danych syntetycznych pozwala na równomierne rozłożenie elementów w poszczególnych klasach, jak również rozwiązuje problem niedoboru dobrze sklasyfikowanych danych rzeczywistych oraz pozwala na dostosowanie funkcjonowania sieci w zależności od potrzeb badawczych. Warto również zwrócić uwagę, że dane syntetyczne rozwiązują

też problem prywatności, gdyż nie pochodzą od rzeczywistych pacjentów, a więc nie są danymi wrażliwymi.

.

Analiza Wyników

Jak wspomniano w poprzednim rozdziale tej rozprawy i przedstawiono na Rycina 31 porównywane metody przeanalizowano w kontekście:

1. Ilość wskazywanych odmiennie metylowanych sekwencji CpG – ocena selektywności
2. Zdolności do wskazywania optymalnej liczby sekwencji CpG pozwalających na utrzymanie różnicowania na grupę kontrolną i badaną – ocena specyficzności
3. Zdolności do wskazywania sekwencji CpG o potencjalnym znaczeniu biologicznym
4. Zdolności do uzyskania wyników spełniających kryteria analogiczne do przyjętych przez Bibikova i wsp.[159]
5. Oceny wskazywania przez metody wyników fałszywie pozytywnych i negatywnych na podstawie danych symulowanych

W każdym teście metoda CTMeth-hhi była porównywalna lub skuteczniejsza od metod klasycznych.

Zdolność do wskazywania optymalnej liczby sekwencji CpG pozwalających na utrzymanie różnicowania na grupę kontrolną i badaną – ocena specyficzności i selektywności

Metody CTMeth-hh i CTMeth-hhi cechują się porównywalną selektywnością odpowiednio do delta 0.3 i delta 0.2 przy analizie zbiorów danych (B-Cell-CD4+, B-Cell-CLL), gdzie przewiduje się obecność znaczących, wyraźnie widocznych zmian (Rycina 34,Rycina 35). W tej grupie danych najbardziej selektywne są delta 0.5 i ChAMP 0.5, aczkolwiek przy analizie bardziej złożonego zbioru danych (CLL-100) gdzie przewiduje się obecność nieznacznych lub trudno wykrywalnych zmian nie wskazują żadnych odmiennie metylowanych sekwencji CpG. Dodatkowo w przypadku

CLL-100 metoda CTMeth-hh wykazuje wyraźnie większą selektywność w stosunku do delta 0.3, a mniejszą niż delta 0.2, a metoda CTMeth-hhi jest najmniej selektywna i wskazuje najwięcej odmiennie metylowanych sekwencji CpG (Rycina 36). Kolejny etap badania, czyli użycie analizy wydajności klastrowania ma za zadanie określić, czy wybrane przez poszczególne metody sekwencje CpG pozwalają na prawidłowe różnicowanie pomiędzy grupą kontrolną a badaną – ocena specyficzności. Najlepszy wynik, czyli identyczny z prawdziwym podział na grupy kontrolne i badane to w wybranej metodzie Randa wartość 1. Dla zbioru B-Cell-CD4+, w którym różnice między grupą kontrolną a badaną powinny być najbardziej wyraźne i istotne wszystkie metody osiągnęły wynik 0,8461, przy czym metoda ChAMP 0.5 wykazała najmniejszą liczba sekwencji CpG, a metoda CTMeth-hhi największą liczbę (Tabela 3). Inaczej prezentuje się analiza wyników pozyskanych ze zbioru B-Cell-CLL, gdzie najmniej sekwencji CpG wskazała metoda ChAMP 0.5, a najwięcej delta 0.2. Z kolei niższą niż inne metody wartość indeksu Randa uzyskała metoda CTMeth w wersji hh (Tabela 4). Metoda CTMeth-hhi uzyskała taki sam wynik indeksu Randa, jak pozostałe metody. Kolejny porównywany zbiór sekwencji CpG, czyli CLL-100 (Tabela 5) to najbardziej złożony zbiór z wyżej wymienionych. Analizowane tutaj są wartości metylacji u pacjentów z białaczką limfocytową podzielonych pod względem wartości IGHV. W tym przypadku metody ChAMP 0.5 i delta 0.5 nie wskazały żadnych różnic pomiędzy grupą kontrolną i badaną dla swojego punktu odcięcia, jednocześnie otrzymując wynik 0 dla metody Randa – jest to wynik niekorzystny. Z kolei najlepszy w tym porównaniu wynik osiągnęły metoda CTMeth-hh i CTMeth-hhi. Rezultaty dla analizy specyficzności i selektywności dla wszystkich trzech zbiorach wskazują, że metody oparte na opracowanej sieci neuronowej dają porównywalne wyniki dla mniej złożonych zbiorów w porównaniu z analogicznymi metodami standardowymi, a metoda CTMeth-hhi przewyższa je pod względem specyficzności przy analizie zbiorów bardziej skomplikowanych. Jednocześnie nasuwa się obserwacja, że dobry wynik uzyskany z analizy metodą Randa przy mniejszej puli wskazanych sekwencji CpG nie jest jednoznaczny z porównywalną skutecznością danej metody w innych przypadkach, jak to miało miejsce w przypadku ChAMP 0.5. Metoda ta w przypadku zbioru danych o niewielkich spodziewanych różnicach w metylacji okazała się zbyt selektywna i nie wykazała żadnych sekwencji CpG.

Zdolności do wskazywania sekwencji CpG o potencjalnym znaczeniu biologicznym

Zdolności do wskazywania sekwencji CpG o potencjalnym znaczeniu biologicznym jest najistotniejszym elementem tej analizy. Metoda CTMeth-hhi w ocenie zdolności do wskazywania sekwencji CpG o potencjalnym znaczeniu biologicznym okazała się najskuteczniejszą z metod. Dla zbioru B-Cell-CD4+ wśród metod, które wskazały sekwencje CpG powiązane z największą pulą funkcyjnych zbiorów Wikipathways znalazły się głównie CTMeth-hhi i delta 0.2. Wśród 10 genów pokrywających się pomiędzy dominującymi w wynikach zbiorami genów Wikipathways, czyli PI3K-Akt Signaling Pathway, Focal Adhesion-PI3K-Akt-mTOR-signaling pathway, VEGFA-VEGFR2 Signaling Pathway i MAPK Signaling Pathway CTMeth-hhi osiągnęła najlepszy wynik wskazując ich 9 (Tabela 7). Wybrane 10 genów występuje we wszystkich wymienionych ścieżkach (Rycina 37), a szlaki te są istotne dla różnicowania, aktywności i proliferacji komórek, od których zależy prawidłowe funkcjonowanie limfocytów B i CD4+ [366–370]. Znaczenie tych wyników potwierdzono z oddzieloną bazą danych DICE [371], która zawiera informacje o różnicach w ekspresji poszczególnych genów w analizowanych komórkach. Zwraca również uwagę fakt, że metoda CTMeth-hhi nie ustępowała innym metodom, a zwłaszcza delta 0.2 we wskazywaniu największej liczby genów z list typowych dla badanych limfocytów B, oraz T CD4+(np. T-Cell antigen Receptor (TCR) Signaling Pathway) w kategoriach Wikipathways, czy Immunological Signatures (Tabela 10,Tabela 11,Tabela 12), a dalsza analiza z użyciem list genów BioCarta i porównaniem ich z bazą danych ekspresji DICE [357] wykazała, że wskazywane przez nią geny są istotniejsze pod względem ich znaczenia biologicznego w kontekście różnicy w ekspresji genów pomiędzy limfocytami B a CD4+ (Tabela 14). Podczas analizy B-cell - CLL najlepiej wypadła metoda CTMeth-hhi, która wskazuje najwięcej sekwencji CpG, do których anotowane geny mają największą pod względem ilości spójność z istotnymi listami genów FUMA w analizowanych kategoria Kegg Pathways, Oncogenic signatures i Cancer_Modules. Dodatkowo należy zwrócić uwagę, że geny anotowane do sekwencji CpG wskazywanych przez metodę CTMeth-hhi i będące unikalnymi (nie występujące w wynikach innych metod), które pokrywają się z tymi listami FUMA, w dodatkowym potwierdzeniu z bazą danych BloodSpot [358] zawierającą dane na temat

różnic w ekspresji genów u osób zdrowych i chorych na CLL, wykazują z tym repozytorium największą spójność (Tabela 17, Tabela 19, Tabela 21, Tabela 25). W analizie najbardziej złożonego zbioru, jakim jest CLL-100 metoda CTMeth-hhi wskazała znacząco więcej genów anotowanych do sekwencji CpG. Jak już wspomniano w tej rozprawie ze względu na brak dostępu do publicznej bazy danych, takiej jak BloodSpot, która zawierałaby zweryfikowany spis genów i informacje o ich ekpresji w zależności od IGHV w ramach procedury badawczej z każdej kategorii i metody wyekstrahowano po cztery elementy charakteryzujące się najwyższą ilością spójnych genów. Następnie skorelowano je pomiędzy metodami zauważając, że wyniki uzyskane przy użyciu CTMeth-hhi, będące odpowiednikami czterech najbardziej spójnych list genów uzyskanych z analizy przy użyciu pozostałych metod, wykazują wyższą jakość pod względem ilości korelujących genów w każdej z trzech kategorii. Każdy element występujący wśród tych 4 najlepszych z pozostałych metod występuje w wynikach CTMeth-hhi. Natomiast w wynikach uzyskanych z użyciem narzędzia CTMeth-hhi znalazły się unikalne listy np. KRAS.600_UP.V1_DN. Mutacje w obrębie KRAS mają związek z IGHV w przewlekłej białaczce limfocytowej[363]. Na uwagę również zasługuje fakt, że metoda ta, wskazała sekwencje CpG, do których anotowane jest 14 genów ze szlaku Wnt/beta-catenin Signaling Pathway in Leukemia (Tabela 30) i 55 genów z Wnt signaling Pathway. Zaburzenia w obrębie tej ścieżki biologicznej mają znaczenie dla różnicowania w kontekście IGHV[360–362].

Zdolność do uzyskania wyników zgodnych z przyjętymi kryteriami dla metylacji

Celem tego etapu analizy było sprawdzenie, czy dana metoda wskazuje sekwencje CpG będące zgodnymi z przyjętymi kryteriami analogicznymi do tych zdefiniowanych przez Bibikova i wsp. [159], którzy przyjęli wartości $\beta < 0.2$ jako sekwencje CpG niemetylowane, a > 0.75 jako metylowane, a wartości oscylujące wokół 0.5 jako częściowo metylowane. W pierwszej kolejności w porównaniu z użyciem zbiorów B-Cell-CD4+ oraz B-Cell-CLL metod CTMeth-hh i CTMeth-hhi ze standardowymi metodami przy użyciu operacji na zbiorach (sekwencje CpG wspólne i unikalne dla metody) zwraca uwagę duża rozbieżność we wskazywanych CpG (Rycina 39, Rycina 50, Rycina 56). Dokładniejsza analiza z zastosowaniem map cieplnych z hierarchicznym grupowaniem wykazała, że sekwencje CpG znajdowane przez metody

CTMeth lepiej wskazują wyniki odpowiadające wyżej opisanym kryteriom zdefiniowanym przez Bibikova i wsp. [159]. Analiza map cieplnych z hierarchicznym grupowaniem jest metodą obserwacyjną i eksploracyjną, która w prosty i intuicyjny sposób pozwala na interpretację zależności i wzorców obecnych w wielkoskalowych wynikach takich jak analiza macierzy metylacji. Przy zastosowaniu skali barwnej, gdzie wartości mniejsze niż 0.5 oznaczone są kolorem niebieskim a wartości większe czerwonym można zaobserwować, że klasyczne metody wskazują sekwencje CpG nie będące odmiennie metylowanymi sekwencjami CpG w kontekście kryteriów przyjętych przez Bibikova i wsp. [159]. Analiza bardziej skomplikowanego zapisu metylacji, jakim jest różnica w metylacji pomiędzy pacjentami chorymi na przewlekłą białaczką limfocytową pogrupowanym względem IGHV obrazuje inne zachowanie poszczególnych metod (zbiór CLL-100). Porównując różnice pomiędzy wynikami poszczególnych metod z użyciem hierarchicznego grupowania na mapach cieplnych można zaobserwować, że metoda CTMeth-hh jest metodą bardziej specyficzną, a wyniki uzyskane przy jej użyciu są bardzo spójne z kryteriami przyjętymi przez Bibikova i wsp.[159]. Dodatkowo metoda ta jest bardziej selektywna niż metoda delta 0.2 i na tyle czuła, by nie odrzucić sekwencji CpG z bardziej złożonego zbioru danych, jak to ma miejsce przy zastosowaniu metody delta 0.5, czy ChAMP 0.5. Inaczej sytuacja wygląda, jeżeli chodzi o analizę z użyciem metody CTMeth-hhi. Metoda ta oprócz wskazywania różnic w kontekście analogicznym do Bibikova i wsp. [159] wybiera również te, w których w danej grupie uczestniczącej w badaniu dominujący stan metylacji próbek nie jest możliwy do określenia lub oznaczany jako częściowo metylowany. CTMeth-hhi wskazuje najwięcej sekwencji CpG, także te występujące w wynikach alternatywnych metod opisywanych w tej rozprawie. Największa pula niespójnych z CTMeth-hhi sekwencji dotyczy tych uzyskanych za pomocą delta 0.2. Analiza z użyciem grupowania hierarchicznego na mapach cieplnych nie jest w tym przypadku jednoznaczna, jak we wcześniejszych porównaniach, a poszczególne klastry nie posiadają zdefiniowanej granic w wynikach wskazywanych tylko przez CTMeth-hhi. Niemniej jednak analizując poszczególne sekwencje CpG wyraźniej widać, że metoda ta wskazuje te, w których jest relatywna różnica między grupą badaną, a kontrolną. Ze względu na to, że analiza ta jest oparta na badaniu bardziej obserwacyjnym, a użycie klasycznych metod statystycznych nie jest tu możliwe, gdyż na nich oparte są metody, do których porównywana jest opracowana technika analizy

metylacji z użyciem sieci neuronowych w kolejnym etapie, jako pogłębienie zdolności do wskazywania sekwencji CpG różnicujących grupy kontrolną i badaną pod względem metylacji definiowanej przez Bibikova i wsp. [159] użyto danych syntetycznych wygenerowanych przez inny niż użyty do treningu algorytm celem oceny fałszywie dodatnich i ujemnych wyników. Analiza z użyciem symulowanych danych jest rozwinięciem sprawdzenia, czy dana metoda wskazuje wyniki odpowiadające kryteriom dla metylacji analogicznym do zdefiniowanych przez Bibikova i wsp. [159]. Analiza z użyciem symulowanych danych wskazuje, że metody z użyciem CTMeth znacząco przewyższają klasyczne metody swoją swoistością i czułością, ponieważ nie dają żadnych fałszywie dodatnich i fałszywie ujemnych wyników, czego nie można powiedzieć o klasycznych metodach (Tabela 31).

Wnioski - podsumowanie

Z opracowanych metod, metoda CTMeth-hhi wykazała wysoką wydajność w doborze cech (w tym przypadku sekwencji CpG) wymaganych do różnicowania pomiędzy grupami kontrolnymi i badanymi, co wykazano z użyciem indeksu Randa. Ilość wskazywanych CpG, a więc selektywność jest podobna do metody klasycznej delta 0.2 w przypadku analizy homogennych zbiorów danych, a dużo bardziej czuła w przypadku występowania heterogeniczności jak w przypadku danych CLL-100, a w tym przypadku wybrane przez metodę sekwencje pozwalają na lepsze różnicowanie grup, niż w przypadku wyników uzyskanych przez inne metody. Zbyt restrykcyjnie użyte metody klasyczne przy zbiorach heterogennych mogą nie dawać żadnych wyników. Anotowane geny do sekwencji CpG wskazanych przez tę metodę wykazują większe znaczenie biologiczne co wykazano porównując uzyskane wyniki do baz danych FUMA[372], DICE[357] oraz BloodSpot[358]. Obie metody CTMeth wykazują się lepszą zdolnością do wskazywania odmiennie metylowanych sekwencji CpG na zasadzie analogicznej do . Bibikova i wsp. [159]. Zastosowanie danych syntetycznych do treningu pozwala na rozwiązanie problemu prywatności danych medycznych oraz ich niedoboru i braku zbalansowania. Dodatkowo pozwala to na dalszą modyfikację działania sieci neuronowej i dostosowanie wraz z rozwojem wiedzy na temat metylacji, czy potrzeb badawczych. Podsumowując opracowana metoda do analizy metylacji CTMeth-hhi w dogłębnej analizie stanowi skuteczną alternatywę dla standardowych

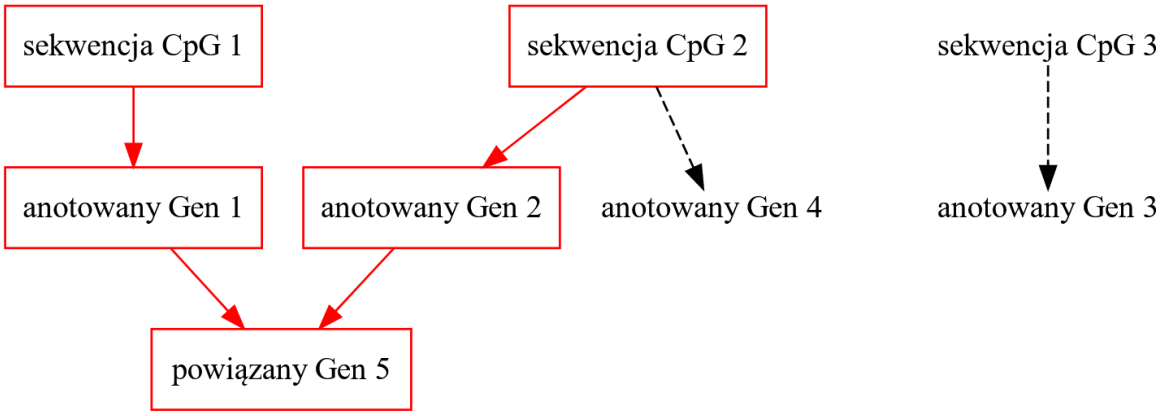
metod. Kod opisywanej biblioteki CTMeth oraz pełna lista sekwencji CpG znalezionych z użyciem metod CTMeth-hh i CTMeth-hhi znajduje się pod adresem url https://www.ctmeth.com. W ten sposób utworzono bazę danych sekwencji CpG o potencjalnym znaczeniu biologicznym dla pacjentów chorych na przewlekłą białaczkę limfocytową, jej podgrupy IGHV 100 oraz różnicującą limfocyty B od CD4+. Ścisły dowód matematyczny skuteczności sieci neuronowej przez jej znaczną złożoność i nieliniowość jest wysoce trudny. Sieci neuronowe często są unikalne pod kątem architektury, użytych danych treningowych, czy złożoności optymalizacji (jedna sieć może posiadać kilka minimów lokalnych). Jednocześnie brakuje ogólnej i zalecanej metodologii porównawczej tych metod w kontekście analizy metylacji. Wykorzystana w tej rozprawie metodyka może stanowić podwaliny do opracowania w przyszłości skutecznego sposobu klasyfikacji nowych metod do analizy metylacji, ponieważ zarówno bierze pod uwagę cechy typowo dotyczące statystyki, jak i biologiczną informację zakodowaną w metylacji, oraz spójność wyników z aktualną wiedzą w szerszym spektrum niż istotność statystyczna.

Dodatkowe moduły i dalszy plan rozwoju

Oprócz omówionej sieci neuronowej opracowana przez autora biblioteka CTMeth zawiera dodatkowe narzędzia takie jak funkcje do hierarchicznego klastrowania, moduł do analizy głównych składowych oraz algorytm do analizy interakcji i zależności pomiędzy powiązanymi z sekwencjami CpG genami (Moduł CpG-Gen-Gen-CpG).

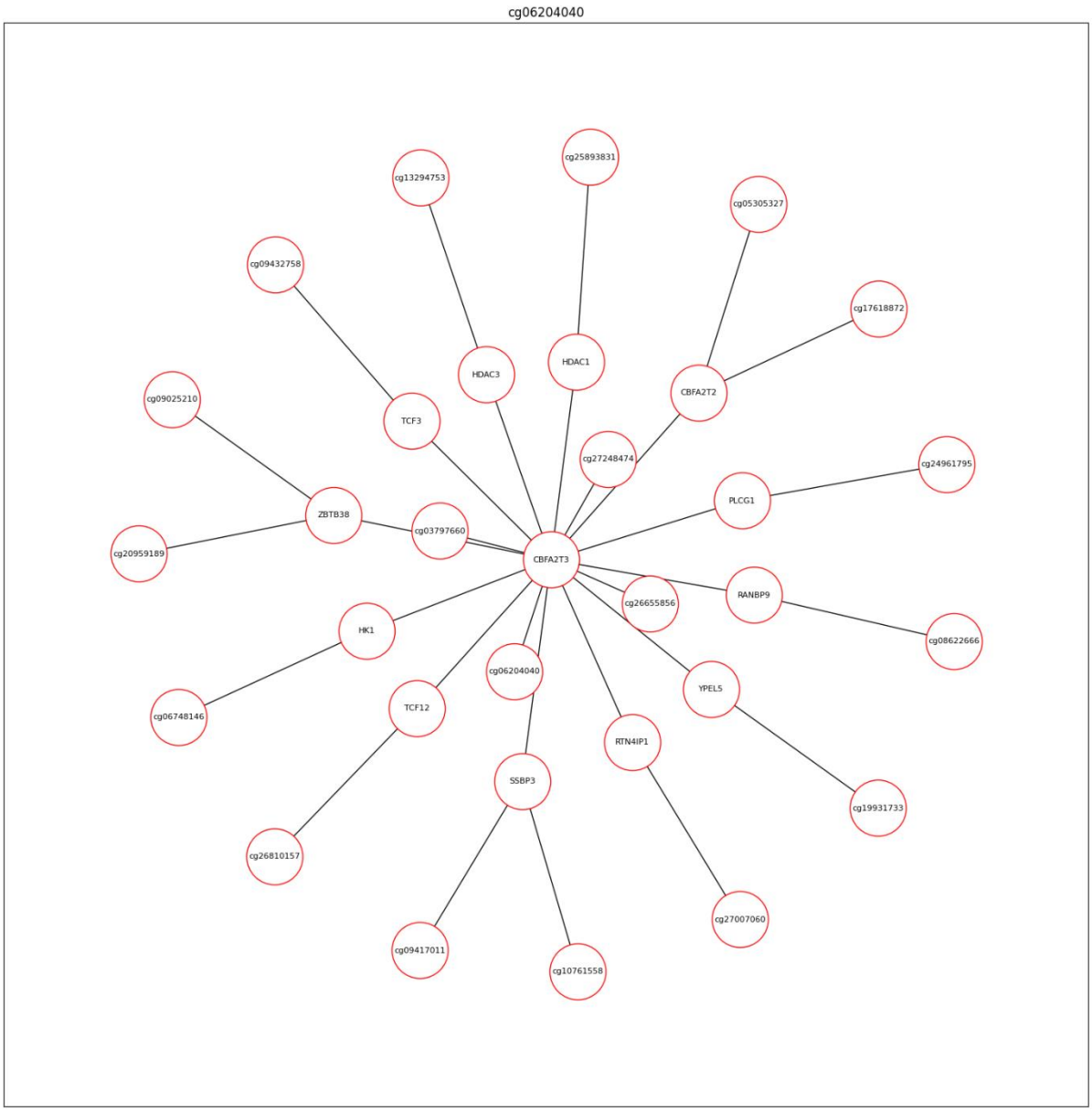
Moduł CpG-Gen-Gen-CpG

Moduł ten tworzy zbiór anotowanych do zadanych sekwencji CpG genów, a następnie przy użyciu danych z biblioteki Biogrid [373] wyszukuje potencjalne drogi interakcji.



Rycina 67 Zasada działania modułu CpG-Gene-Gene-CpG. Kolorem czerwonym oznaczono sekwencje przefiltrowane przez moduł CpG-Gen-Gen-CpG, oraz ścieżkę interakcji

Efektem działania algorytmu jest wskazanie sekwencji CpG oraz genów, przy których dochodzi do największej interakcji w ten sposób dodatkowo zawężając wyniki w poszukiwaniu potencjalnej patogenezy badanej jednostki chorobowej, czy interwencji.



Rycina 68 Przykładowy wynik działania modułu CpG-Gen-Gen-CpG

Dalszy rozwój siec neuronowej i biblioteki CTMeth

Planowane jest wprowadzenie szeregu licznych udogodnień w opracowanej platformie, m.in zwiększenie liczby formatów plików, które można poddać analizie, wprowadzenie metody do pozyskiwania wartości β z plików. „idat”, czy wprowadzenie łatwego w obsłudze systemu graficznego. Niewykluczone jest dodanie innych modeli sieci neuronowych.

Bibliografia

1. Al Aboud NM, Tupper C, Jialal I. Genetics, Epigenetic Mechanism. In: StatPearls. Treasure Island (FL): StatPearls Publishing; 2022.

2. Bourc'his D, Xu GL, Lin CS, Bollman B, Bestor TH. Dnmt3L and the establishment of maternal genomic imprints. Science. 2001;294:2536–9.

3. Pervjakova N, Kasela S, Morris AP, Kals M, Metspalu A, Lindgren CM, et al. Imprinted genes and imprinting control regions show predominant intermediate methylation in adult somatic tissues. Epigenomics. 2016;8:789–99.

4. Sharp AJ, Stathaki E, Migliavacca E, Brahmachary M, Montgomery SB, Dupre Y, et al. DNA methylation profiles of human active and inactive X chromosomes. Genome Res. 2011;21:1592–600.

5. Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature. 2010;466:253–7.

6. Mallik S, Seth S, Bhadra T, Zhao Z. A Linear Regression and Deep Learning Approach for Detecting Reliable Genetic Alterations in Cancer Using DNA Methylation and Gene Expression Data. Genes. 2020;11:931.

7. Taryma-Leśniak O, Sokolowska KE, Wojdacz TK. Current status of development of methylation biomarkers for in vitro diagnostic IVD applications. Clin Epigenet. 2020;12:100.

8. Shu C, Zhang X, Aouizerat BE, Xu K. Comparison of methylation capture sequencing and Infinium MethylationEPIC array in peripheral blood mononuclear cells. Epigenetics & Chromatin. 2020;13:51.

9. Lodish H, Berk A, Kaiser CA, Kaiser C, Krieger M, Scott MP, et al. Molecular cell biology. Macmillan; 2008.

10. Song X, Reif J. Nucleic Acid Databases and Molecular-Scale Computing. ACS Nano. 2019;13:6256–68.

11. Ceze L, Nivala J, Strauss K. Molecular digital data storage using DNA. Nat Rev Genet. 2019;20:456–66.

12. Wąsiewicz P, Malinowski A, Nowak R, Mulawka JJ, Borsuk P, Węgleński P, et al. DNA computing: implementation of data flow logical operations. Future Generation Computer Systems. 2001;17:361–78.

13. Lewis R. Human genetics: concepts and applications. Thirteenth edition. New York, NY: McGraw-Hill Education; 2021.

14. Bersaglieri C, Santoro R. Genome Organization in and around the Nucleolus. Cells. 2019;8:579.

15. Cohen I, Poręba E, Kamieniarz K, Schneider R. Histone Modifiers in Cancer: Friends or Foes? Genes & Cancer. 2011;2:631–47.

16. Suganuma T, Workman JL. Signals and Combinatorial Functions of Histone Modifications. Annual Review of Biochemistry. 2011;80:473–99.

17. Alexander RP, Fang G, Rozowsky J, Snyder M, Gerstein MB. Annotating non-coding regions of the genome. Nat Rev Genet. 2010;11:559–71.

18. Necsulea A, Kaessmann H. Evolutionary dynamics of coding and non-coding transcriptomes. Nat Rev Genet. 2014;15:734–48.

19. Burley SK. The TATA box binding protein. Current Opinion in Structural Biology. 1996;6:69–75.

20. Struhl K. Helix-turn-helix, zinc-finger, and leucine-.zipper motifs for eukaryotic transcriptional regulatory proteins. 1989.

21. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The Human Transcription Factors. Cell. 2018;172:650–65.

22. Chen WJ, Zhu T. Networks of transcription factors with roles in environmental stress response. Trends in Plant Science. 2004;9:591–6.

23. Lee TI, Young RA. Transcriptional Regulation and Its Misregulation in Disease. Cell. 2013;152:1237–51.

24. Singh H, Khan AA, Dinner AR. Gene regulatory networks in the immune system. Trends in Immunology. 2014;35:211–8.

25. Cramer P. Organization and regulation of gene transcription. Nature. 2019;573:45–54.

26. Spielmann M, Mundlos S. Looking beyond the genes: the role of non-coding variants in human disease. Human Molecular Genetics. 2016;25:R157–65.

27. Eisenberg E, Levanon EY. Human housekeeping genes, revisited. Trends in Genetics. 2013;29:569–74.

28. MacPhail TM. The viral gene: an undead metaphor recoding life. Science as Culture. 2004;13:325–45.

29. Zuckerkandl E, Cavalli G. Combinatorial epigenetics, “junk DNA”, and the evolution of complex organisms. Gene. 2007;390:232–42.

30. Kim Y-J, Lee J, Han K. Transposable Elements: No More “Junk DNA.” Genomics Inform. 2012;10:226–33.

31. Palazzo AF, Gregory TR. The Case for Junk DNA. PLOS Genetics. 2014;10:e1004351.

32. Zhao L-Y, Song J, Liu Y, Song C-X, Yi C. Mapping the epigenetic modifications of DNA and RNA. Protein Cell. 2020;11:792–808.

33. Emery AE. Pierre Louis Moreau de Maupertuis (1698-1759). J Med Genet. 1988;25:561–4.

34. Darwin C. On the Origin of Species, 1859. London: Routledge; 2003.

35. MENDEL G. Versuche uber Pflanzen-Hybriden. Verh Naturforsch Ver Brunn. 1866;4:3–47.

36. Sutton WS. THE CHROMOSOMES IN HEREDITY. The Biological Bulletin. 1903;4:231–50.

37. Griffith F. The Significance of Pneumococcal Types. The Journal of Hygiene. 1928;27:113.

38. McCarty M, Avery OT. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: II. Effect of desoxyribonuclease on the biological activity of the transforming substance. The Journal of experimental medicine. 1946;83:89.

39. Watson JD, Crick FH, others. A structure for deoxyribose nucleic acid. 1953.

40. Holliday R, Pugh JE. DNA Modification Mechanisms and Gene Activity During Development: Developmental clocks may depend on the enzymic modification of specific bases in repeated DNA sequences. Science. 1975;187:226–32.

41. Riggs AD. X inactivation, differentiation, and DNA methylation. CGR. 1975;14:9–25.

42. Bird AP. CpG-rich islands and the function of DNA methylation. Nature. 1986;321:209–13.

43. Chial H. DNA sequencing technologies key to the Human Genome Project. Nature Education. 2008;1.

44. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.

45. Moore LD, Le T, Fan G. DNA methylation and its basic function. Neuropsychopharmacology. 2013;38:23–38.

46. Pelizzola M, Ecker JR. The DNA methylome. FEBS Letters. 2011;585:1994–2000.

47. Rechache NS, Wang Y, Stevenson HS, Killian JK, Edelman DC, Merino M, et al. DNA Methylation Profiling Identifies Global Methylation Differences and Markers of Adrenocortical Tumors. J Clin Endocrinol Metab. 2012;97:E1004–13.

48. Kim JK, Samaranayake M, Pradhan S. Epigenetic mechanisms in mammals. Cell Mol Life Sci. 2009;66:596–612.

49. Vavouri T, Lehner B. Human genes with CpG island promoters have a distinct transcription-associated chromatin organization. Genome Biology. 2012;13:R110.

50. Razin A, Cedar H. DNA methylation and gene expression. Microbiological Reviews. 1991;55:451–8.

51. Doi A, Park I-H, Wen B, Murakami P, Aryee MJ, Irizarry R, et al. Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. Nat Genet. 2009;41:1350–3.

52. Deaton AM, Bird A. CpG islands and the regulation of transcription. Genes Dev. 2011;25:1010–22.

53. Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LT-Y, Kohlbacher O, et al. Charting a dynamic DNA methylation landscape of the human genome. Nature. 2013;500:477–81.

54. Farlik M, Sheffield NC, Nuzzo A, Datlinger P, Schönegger A, Klughammer J, et al. Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. Cell Rep. 2015;10:1386–97.

55. van Eijk KR, de Jong S, Boks MP, Langeveld T, Colas F, Veldink JH, et al. Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. BMC Genomics. 2012;13:636.

56. Lock LF, Takagi N, Martin GR. Methylation of the Hprt gene on the inactive X occurs after chromosome inactivation. Cell. 1987;48:39–46.

57. Sharp AJ, Stathaki E, Migliavacca E, Brahmachary M, Montgomery SB, Dupre Y, et al. DNA methylation profiles of human active and inactive X chromosomes. Genome Res. 2011;21:1592–600.

58. Li E, Beard C, Jaenisch R. Role for DNA methylation in genomic imprinting. Nature. 1993;366:362–5.

59. Pervjakova N, Kasela S, Morris AP, Kals M, Metspalu A, Lindgren CM, et al. Imprinted genes and imprinting control regions show predominant intermediate methylation in adult somatic tissues. Epigenomics. 2016;8:789–99.

60. Vilain A, Bernardino J, Gerbault-Seureau M, Vogt N, Niveleau A, Lefrançois D, et al. DNA methylation and chromosome instability in lymphoblastoid cell lines. Cytogenet Cell Genet. 2000;90:93–101.

61. Robertson KD. DNA methylation and human disease. Nat Rev Genet. 2005;6:597–610.

62. Subramaniam D, Thombre R, Dhar A, Anant S. DNA Methyltransferases: A Novel Target for Prevention and Therapy. Front Oncol. 2014;4:80.

63. Lyko F. The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. Nat Rev Genet. 2018;19:81–92.

64. Sharifi-Zarchi A, Gerovska D, Adachi K, Totonchi M, Pezeshk H, Taft RJ, et al. DNA methylation regulates discrimination of enhancers from promoters through a H3K4me1-H3K4me3 seesaw mechanism. BMC Genomics. 2017;18:964.

65. Yang X, Han H, De Carvalho DD, Lay FD, Jones PA, Liang G. Gene Body Methylation can alter Gene Expression and is a Therapeutic Target in Cancer. Cancer Cell. 2014;26:577–90.

66. Ge Y-Z, Pu M-T, Gowher H, Wu H-P, Ding J-P, Jeltsch A, et al. Chromatin Targeting of de Novo DNA Methyltransferases by the PWWP Domain*. Journal of Biological Chemistry. 2004;279:25447–54.

67. Dhayalan A, Rajavelu A, Rathert P, Tamas R, Jurkowska RZ, Ragozin S, et al. The Dnmt3a PWWP Domain Reads Histone 3 Lysine 36 Trimethylation and Guides DNA Methylation*. Journal of Biological Chemistry. 2010;285:26114–20.

68. Baubec T, Colombo DF, Wirbelauer C, Schmidt J, Burger L, Krebs AR, et al. Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. Nature. 2015;520:243–7.

69. Molenaar TM, van Leeuwen F. SETD2: from chromatin modifier to multipronged regulator of the genome and beyond. Cell Mol Life Sci. 2022;79:346.

70. Bourc'his D, Xu G-L, Lin C-S, Bollman B, Bestor TH. Dnmt3L and the Establishment of Maternal Genomic Imprints. Science. 2001;294:2536–9.

71. Ooi SKT, Qiu C, Bernstein E, Li K, Jia D, Yang Z, et al. DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. Nature. 2007;448:714–7.

72. Bestor T, Laudano A, Mattaliano R, Ingram V. Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells: The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. Journal of Molecular Biology. 1988;203:971–83.

73. Li E, Bestor TH, Jaenisch R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell. 1992;69:915–26.

74. Hermann A, Goyal R, Jeltsch A. The Dnmt1 DNA-(cytosine-C5)-methyltransferase Methylates DNA Processively with High Preference for Hemimethylated Target Sites *. Journal of Biological Chemistry. 2004;279:48350–9.

75. Bostick M, Kim JK, Estève P-O, Clark A, Pradhan S, Jacobsen SE. UHRF1 plays a role in maintaining DNA methylation in mammalian cells. Science. 2007;317:1760–4.

76. Sharif J, Koseki H. Recruitment of Dnmt1 roles of the SRA protein Np95 (Uhrf1) and other factors. Prog Mol Biol Transl Sci. 2011;101:289–310.

77. Vaughan RM, Rothbart SB, Dickson BM. The finger loop of the SRA domain in the E3 ligase UHRF1 is a regulator of ubiquitin targeting and is required for the maintenance of DNA methylation. J Biol Chem. 2019;294:15724–32.

78. Cai Y, Geutjes E-J, de Lint K, Roepman P, Bruurs L, Yu L-R, et al. The NuRD complex cooperates with DNMTs to maintain silencing of key colorectal tumor suppressor genes. Oncogene. 2014;33:2157–68.

79. Wu H, Zhang Y. Reversing DNA Methylation: Mechanisms, Genomics, and Biological Functions. Cell. 2014;156:45–68.

80. Lorsbach RB, Moore J, Mathew S, Raimondi SC, Mukatira ST, Downing JR. TET1, a member of a novel protein family, is fused to MLL in acute myeloid leukemia containing the t(10;11)(q22;q23). Leukemia. 2003;17:637–41.

81. Ito S, D'Alessio AC, Taranova OV, Hong K, Sowers LC, Zhang Y. Role of Tet proteins in 5mC to 5hmC conversion, ES cell self-renewal, and ICM specification. Nature. 2010;466:1129–33.

82. He Y-F, Li B-Z, Li Z, Liu P, Wang Y, Tang Q, et al. Tet-Mediated Formation of 5-Carboxylcytosine and Its Excision by TDG in Mammalian DNA. Science. 2011;333:1303–7.

83. Rasmussen KD, Helin K. Role of TET enzymes in DNA methylation, development, and cancer. Genes Dev. 2016;30:733–50.

84. Otani J, Kimura H, Sharif J, Endo TA, Mishima Y, Kawakami T, et al. Cell Cycle-Dependent Turnover of 5-Hydroxymethyl Cytosine in Mouse Embryonic Stem Cells. PLOS ONE. 2013;8:e82961.

85. Hashimoto H, Liu Y, Upadhyay AK, Chang Y, Howerton SB, Vertino PM, et al. Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. Nucleic Acids Research. 2012;40:4841–9.

86. He Y-F, Li B-Z, Li Z, Liu P, Wang Y, Tang Q, et al. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. Science. 2011;333:1303–7.

87. Maiti A, Drohat AC. Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. J Biol Chem. 2011;286:35334–8.

88. Koh KP, Yabuuchi A, Rao S, Huang Y, Cunniff K, Nardone J, et al. Tet1 and Tet2 regulate 5-hydroxymethylcytosine production and cell lineage specification in mouse embryonic stem cells. Cell Stem Cell. 2011;8:200–13.

89. Huang Y, Chavez L, Chang X, Wang X, Pastor WA, Kang J, et al. Distinct roles of the methylcytosine oxidases Tet1 and Tet2 in mouse embryonic stem cells. Proc Natl Acad Sci U S A. 2014;111:1361–6.

90. Costa Y, Ding J, Theunissen TW, Faiola F, Hore TA, Shliaha PV, et al. NANOG-dependent function of TET1 and TET2 in establishment of pluripotency. Nature. 2013;495:370–4.

91. Neri F, Incarnato D, Krepelova A, Rapelli S, Pagnani A, Zecchina R, et al. Genome-wide analysis identifies a functional association of Tet1 and Polycomb repressive complex 2 in mouse embryonic stem cells. Genome Biol. 2013;14:R91.

92. Yin Y, Morgunova E, Jolma A, Kaasinen E, Sahu B, Khund-Sayeed S, et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science. 2017;356:eaaj2239.

93. Liu Y, Olanrewaju YO, Zheng Y, Hashimoto H, Blumenthal RM, Zhang X, et al. Structural basis for Klf4 recognition of methylated DNA. Nucleic Acids Res. 2014;42:4859–67.

94. Boyes J, Bird A. DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein. Cell. 1991;64:1123–34.

95. Leighton G, Williams DC. The Methyl-CpG–Binding Domain 2 and 3 Proteins and Formation of the Nucleosome Remodeling and Deacetylase Complex. Journal of Molecular Biology. 2020;432:1624–39.

96. Finnegan DJ. Eukaryotic transposable elements and genome evolution. Trends in Genetics. 1989;5:103–7.

97. Eickbush TH. Transposing without ends: the non-LTR retrotransposable elements. New Biol. 1992;4:430–40.

98. Konkel MK, Walker JA, Batzer MA. LINEs and SINEs of primate evolution. Evolutionary Anthropology: Issues, News, and Reviews. 2010;19:236–49.

99. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8:973–82.

100. Yoder JA, Walsh CP, Bestor TH. Cytosine methylation and the ecology of intragenomic parasites. Trends Genet. 1997;13:335–40.

101. Barau J, Teissandier A, Zamudio N, Roy S, Nalesso V, Hérault Y, et al. The DNA methyltransferase DNMT3C protects male germ cells from transposon activity. Science. 2016;354:909–12.

102. Ichiyanagi K, Li Y, Watanabe T, Ichiyanagi T, Fukuda K, Kitayama J, et al. Locus- and domain-dependent control of DNA methylation at mouse B1 retrotransposons during male germ cell development. Genome Res. 2011;21:2058–66.

103. Sharma S, Kelly TK, Jones PA. Epigenetics in cancer. Carcinogenesis. 2010;31:27–36.

104. Mendis SR, Topham JT, Titmuss E, Williamson L, Pleasance ED, Culibrk L, et al. Comprehensive transcriptome analysis reveals link between epigenetic dysregulation, endogenous retrovirus expression and immunogenicity in metastatic colorectal carcinoma. JCO. 2019;37 15_suppl:3535–3535.

105. Crouse HV. The Controlling Element in Sex Chromosome Behavior in Sciara. Genetics. 1960;45:1429–43.

106. Solter D. Differential imprinting and expression of maternal and paternal genomes. Annu Rev Genet. 1988;22:127–46.

107. Hannula-Jouppi K, Muurinen M, Lipsanen-Nyman M, Reinius LE, Ezer S, Greco D, et al. Differentially methylated regions in maternal and paternal uniparental disomy for chromosome 7. Epigenetics. 2014;9:351–65.

108. Monk D, Mackay DJG, Eggermann T, Maher ER, Riccio A. Genomic imprinting disorders: lessons on how genome, epigenome and environment interact. Nat Rev Genet. 2019;20:235–48.

109. Peters J. The role of genomic imprinting in biology and disease: an expanding view. Nat Rev Genet. 2014;15:517–30.

110. Millership SJ, Van de Pette M, Withers DJ. Genomic imprinting and its effects on postnatal growth and adult metabolism. Cell Mol Life Sci. 2019;76:4009–21.

111. Ferguson-Smith AC. Genomic imprinting: the emergence of an epigenetic paradigm. Nat Rev Genet. 2011;12:565–75.

112. Sanli I, Feil R. Chromatin mechanisms in the developmental control of imprinted gene expression. The International Journal of Biochemistry & Cell Biology. 2015;67:139–47.

113. Rougeulle C, Glatt H, Lalande M. The Angelman syndrome candidate gene, UBE3A/E6-AP, is imprinted in brain. Nat Genet. 1997;17:14–5.

114. Lopez SJ, Laufer BI, Beitnere U, Berg EL, Silverman JL, O'Geen H, et al. Imprinting effects of UBE3A loss on synaptic gene networks and Wnt signaling pathways. Hum Mol Genet. 2019;28:3842–52.

115. Fridman C, Koiffmann CP. Genomic imprinting: genetic mechanisms and phenotypic consequences in Prader-Willi and Angelman syndromes. Genet Mol Biol. 2000;23:715–24.

116. Joshi RS, Garg P, Zaitlen N, Lappalainen T, Watson CT, Azam N, et al. DNA Methylation Profiling of Uniparental Disomy Subjects Provides a Map of Parental Epigenetic Bias in the Human Genome. Am J Hum Genet. 2016;99:555–66.

117. Mackay DJG, Eggermann T, Buiting K, Garin I, Netchine I, Linglart A, et al. Multilocus methylation defects in imprinting disorders. Biomol Concepts. 2015;6:47–57.

118. Sanchez-Delgado M, Riccio A, Eggermann T, Maher ER, Lapunzina P, Mackay D, et al. Causes and Consequences of Multi-Locus Imprinting Disturbances in Humans. Trends Genet. 2016;32:444–55.

119. Sparago A, Verma A, Patricelli MG, Pignata L, Russo S, Calzari L, et al. The phenotypic variations of multi-locus imprinting disturbances associated with maternal-effect variants of NLRP5 range from overt imprinting disorder to apparently healthy phenotype. Clin Epigenetics. 2019;11:190.

120. Hara-Isono K, Matsubara K, Hamada R, Shimada S, Yamaguchi T, Wakui K, et al. A patient with Silver-Russell syndrome with multilocus imprinting disturbance, and Schimke immuno-osseous dysplasia unmasked by uniparental isodisomy of chromosome 2. J Hum Genet. 2021;66:1121–6.

121. Narusawa H, Sasaki S, Hara-Isono K, Matsubara K, Fukami M, Nagasaki K, et al. A boy with overgrowth caused by multi-locus imprinting disturbance including hypomethylation of MEST:alt-TSS-DMR. Eur J Med Genet. 2022;65:104502.

122. Seisenberger S, Peat JR, Hore TA, Santos F, Dean W, Reik W. Reprogramming DNA methylation in the mammalian life cycle: building and breaking epigenetic barriers. Philos Trans R Soc Lond B Biol Sci. 2013;368:20110330.

123. Parry A, Rulands S, Reik W. Active turnover of DNA methylation during cell fate decisions. Nat Rev Genet. 2021;22:59–66.

124. Cedar H, Sabag O, Reizel Y. The role of DNA methylation in genome-wide gene regulation during development. Development. 2022;149:dev200118.

125. Ko YA, Mohtat D, Suzuki M, Park ASD, Izquierdo MC, Han SY, et al. Cytosine methylation changes in enhancer regions of core pro-fibrotic genes characterize kidney fibrosis development. Genome Biology. 2013;14.

126. Pidsley R, Viana J, Hannon E, Spiers H, Troakes C, Al-Saraj S, et al. Methylomic profiling of human brain tissue supports a neurodevelopmental origin for schizophrenia. Genome Biol. 2014;15:483.

127. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, et al. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biol. 2011;12:R10.

128. Zhang T, Choi J, Dilshat R, Einarsdóttir BÓ, Kovacs MA, Xu M, et al. Cell-type-specific meQTLs extend melanoma GWAS annotation beyond eQTLs and inform melanocyte gene-regulatory mechanisms. Am J Hum Genet. 2021;108:1631–46.

129. Xue A, Wu Y, Zhu Z, Zhang F, Kemper KE, Zheng Z, et al. Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. Nat Commun. 2018;9:2941.

130. Fogarty MP, Cannon ME, Vadlamudi S, Gaulton KJ, Mohlke KL. Identification of a Regulatory Variant That Binds FOXA1 and FOXA2 at the CDC123/CAMK1D Type 2 Diabetes GWAS Locus. PLOS Genetics. 2014;10:e1004633.

131. Sanchez-Mut JV, Heyn H, Silva BA, Dixsaut L, Garcia-Esparcia P, Vidal E, et al. PM20D1 is a quantitative trait locus associated with Alzheimer's disease. Nat Med. 2018;24:598–603.

132. Pihlstrøm L, Berge V, Rengmark A, Toft M. Parkinson's disease correlates with promoter methylation in the α-synuclein gene. Mov Disord. 2015;30:577–80.

133. Cerrato F, Sparago A, Ariani F, Brugnoletti F, Calzari L, Coppedè F, et al. DNA Methylation in the Diagnosis of Monogenic Diseases. Genes (Basel). 2020;11:355.

134. Christensen BC, Houseman EA, Marsit CJ, Zheng S, Wrensch MR, Wiemels JL, et al. Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. PLoS Genet. 2009;5:e1000602.

135. Li S, Wong EM, Bui M, Nguyen TL, Joo JHE, Stone J, et al. Causal effect of smoking on DNA methylation in peripheral blood: A twin and family study. Clinical Epigenetics. 2018;10:18.

136. Meng W, Zhu Z, Jiang X, Too CL, Uebe S, Jagodic M, et al. DNA methylation mediates genotype and smoking interaction in the development of anti-citrullinated peptide antibody-positive rheumatoid arthritis. Arthritis Res Ther. 2017;19:71.

137. Sharma S, Kelly TK, Jones PA. Epigenetics in cancer. Carcinogenesis. 2010;31:27–36.

138. Hur K, Cejas P, Feliu J, Moreno-Rubio J, Burgos E, Boland CR, et al. Hypomethylation of long interspersed nuclear element-1 (LINE-1) leads to activation of proto-oncogenes in human colorectal cancer metastasis. Gut. 2014;63:635–46.

139. Zelic R, Fiano V, Grasso C, Zugna D, Pettersson A, Gillio-Tos A, et al. Global DNA hypomethylation in prostate cancer development and progression: a systematic review. Prostate Cancer Prostatic Dis. 2015;18:1–12.

140. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012;13:484–92.

141. Xie W, Schultz MD, Lister R, Hou Z, Rajagopal N, Ray P, et al. Epigenomic Analysis of Multilineage Differentiation of Human Embryonic Stem Cells. Cell. 2013;153:1134–48.

142. Baylin SB, Jones PA. Epigenetic Determinants of Cancer. Cold Spring Harb Perspect Biol. 2016;8:a019505.

143. Bennett RL, Licht JD. Targeting Epigenetics in Cancer. Annual Review of Pharmacology and Toxicology. 2018;58:187–207.

144. Hibi K, Goto T, Mizukami H, Kitamura Y-H, Sakuraba K, Sakata M, et al. Demethylation of the CDH3 gene is frequently detected in advanced colorectal cancer. Anticancer Res. 2009;29:2215–7.

145. Thomas R, Trapani D, Goodyer-Sait L, Tomkova M, Fernandez-Rozadilla C, Sahnane N, et al. The polymorphic variant rs1800734 influences methylation acquisition and allele-specific TFAP4 binding in the MLH1 promoter leading to differential mRNA expression. Sci Rep. 2019;9:13463.

146. Vinciguerra M, Agodi A, Barchitta M, Quattrocchi A, Maugeri A. DAPK1 Promoter Methylation and Cervical Cancer Risk: A Systematic Review and a Meta-Analysis. PLoS ONE. 2015;10:e0135078.

147. Jayaprakash C, Varghese VK, Bellampalli R, Radhakrishnan R, Ray S, Kabekkodu SP, et al. Hypermethylation of Death-Associated Protein Kinase (DAPK1) and its association with oral carcinogenesis - An experimental and meta-analysis study. Archives of Oral Biology. 2017;80:117–29.

148. Wang X-B, Cui N-H, Liu X-N, Ma J-F, Zhu Q-H, Guo S-R, et al. Identification of DAPK1 Promoter Hypermethylation as a Biomarker for Intra-Epithelial Lesion and Cervical Cancer: A Meta-Analysis of Published Studies, TCGA, and GEO Datasets. Front Genet. 2018;9:258.

149. Zhang W, Xu J. DNA methyltransferases and their roles in tumorigenesis. Biomarker Research. 2017;5:1.

150. Kong X, Chen J, Xie W, Brown SM, Cai Y, Wu K, et al. Defining UHRF1 Domains that Support Maintenance of Human Colon Cancer DNA Methylation and Oncogenic Properties. Cancer Cell. 2019;35:633-648.e7.

151. Jelinic P, Shaw P. Loss of imprinting and cancer. The Journal of Pathology. 2007;211:261–8.

152. Rasmussen KD, Helin K. Role of TET enzymes in DNA methylation, development, and cancer. Genes Dev. 2016;30:733–50.

153. Ambatipudi S, Horvath S, Perrier F, Cuenin C, Hernandez-Vargas H, Le Calvez-Kelm F, et al. DNA methylome analysis identifies accelerated epigenetic ageing associated with postmenopausal breast cancer susceptibility. Eur J Cancer. 2017;75:299–307.

154. Vafadar A, Mokaram P, Erfani M, Yousefi Z, Farhadi A, Elham Shirazi T, et al. The effect of decitabine on the expression and methylation of the PPP1CA, BTG2, and PTEN in association with changes in miR-125b, miR-17, and miR-181b in NALM6 cell line. Journal of Cellular Biochemistry. 2019;120:13156–67.

155. Spencer DH, Russler-Germain DA, Ketkar S, Helton NM, Lamprecht TL, Fulton RS, et al. CpG Island Hypermethylation Mediated by DNMT3A Is a Consequence of AML Progression. Cell. 2017;168:801-816.e13.

156. Wei S, Tao J, Xu J, Chen X, Wang Z, Zhang N, et al. Ten Years of EWAS. Advanced Science. 2021;8:2100727.

157. Stirzaker C, Taberlay PC, Statham AL, Clark SJ. Mining cancer methylomes: prospects and challenges. Trends Genet. 2014;30:75–84.

158. Kernaleguen M, Daviaud C, Shen Y, Bonnet E, Renault V, Deleuze J-F, et al. Whole-Genome Bisulfite Sequencing for the Analysis of Genome-Wide DNA Methylation and Hydroxymethylation Patterns at Single-Nucleotide Resolution. Methods Mol Biol. 2018;1767:311–49.

159. Bibikova M, Le J, Barnes B, Saedinia-Melnyk S, Zhou L, Shen R, et al. Genome-wide DNA methylation profiling using Infinium® assay. Epigenomics. 2009;1:177–200.

160. Sandoval J, Heyn H, Moran S, Serra-Musach J, Pujana MA, Bibikova M, et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenetics. 2011;6:692–702.

161. Moran S, Arribas C, Esteller M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. Epigenomics. 2016;8:389–99.

162. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. Genomics. 2011;98:288–95.

163. Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, et al. Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biology. 2015;16:22.

164. GenomeStudio Methylation Module v1.8 User Guide (11319130).

165. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30:1363–9.

166. Hop PJ, Zwamborn RAJ, Hannon EJ, Dekker AM, van Eijk KR, Walker EM, et al. Cross-reactive probes on Illumina DNA methylation arrays: a large study on ALS shows that a cautionary approach is warranted in interpreting epigenome-wide association studies. NAR Genomics and Bioinformatics. 2020;2:lqaa105.

167. Sun Z, Chai H, Wu Y, White WM, Donkena KV, Klein CJ, et al. Batch effect correction for genome-wide methylation data with Illumina Infinium platform. BMC Medical Genomics. 2011;4.

168. Jiao C, Zhang C, Dai R, Xia Y, Wang K, Giase G, et al. Positional effects revealed in Illumina methylation array and the impact on analysis. Epigenomics. 2018;10:643–59.

169. Pidsley R, Y Wong CC, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. BMC Genomics. 2013;14:293.

170. Campagna MP, Xavier A, Lechner-Scott J, Maltby V, Scott RJ, Butzkueven H, et al. Epigenome-wide association studies: current knowledge, strategies and recommendations. Clin Epigenetics. 2021;13:214.

171. Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010;11:587.

172. Saadati M, Benner A. Statistical challenges of high-dimensional methylation data. Stat Med. 2014;33:5347–57.

173. Campagna MP, Xavier A, Lechner-Scott J, Maltby V, Scott RJ, Butzkueven H, et al. Epigenome-wide association studies: current knowledge, strategies and recommendations. Clinical Epigenetics. 2021;13:214.

174. Guéant J-L, Chéry C, Oussalah A, Nadaf J, Coelho D, Josse T, et al. A PRDX1 mutant allele causes a MMACHC secondary epimutation in cblC patients. Nat Commun. 2018;9:67.

175. Schema for HAIB Methyl450 - CpG Methylation by Methyl 450K Bead Arrays from ENCODE/HAIB. https://genome.ucsc.edu/cgi-bin/hgTables?db=hg19&hgta_group=regulation&hgta_track=wgEncodeHaibMethyl450

&hgta_table=wgEncodeHaibMethyl450HaeSitesRep1&hgta_doSchema=describe+table+schema. Accessed 9 Jan 2023.

176. Mao X, Ye Q, Zhang G, Jiang J, Zhao H, Shao Y, et al. Identification of differentially methylated genes as diagnostic and prognostic biomarkers of breast cancer. World Journal of Surgical Oncology. 2021;19:29.

177. Oshima G, Poli EC, Bolt MJ, Chlenski A, Forde M, Jutzy JMS, et al. DNA Methylation Controls Metastasis-Suppressive 14q32-Encoded miRNAs. Cancer Research. 2019;79:650–62.

178. Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong CT, et al. Dynamic changes in the human methylome during differentiation. Genome Res. 2010;20:320–31.

179. Lowe R, Slodkowicz G, Goldman N, Rakyan VK. The human blood DNA methylome displays a highly distinctive profile compared with other somatic tissues. Epigenetics. 2015;10:274–81.

180. Chen X-G, Ma L, Xu J-X. Abnormal DNA methylation may contribute to the progression of osteosarcoma. Molecular Medicine Reports. 2018;17:193–9.

181. Batra RN, Lifshitz A, Vidakovic AT, Chin S-F, Sati-Batra A, Sammut S-J, et al. DNA methylation landscapes of 1538 breast cancers reveal a replication-linked clock, epigenomic instability and cis-regulation. Nat Commun. 2021;12:5406.

182. Sae-Lee C, Barrow TM, Colicino E, Choi SH, Rabanal-Ruiz Y, Green D, et al. Genomic targets and selective inhibition of DNA methyltransferase isoforms. Clinical Epigenetics. 2022;14:103.

183. Jaffe AE, Murakami P, Lee H, Leek JT, Fallin MD, Feinberg AP, et al. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. International Journal of Epidemiology. 2012;41:200–9.

184. Morris TJ, Butcher LM, Feber A, Teschendorff AE, Chakravarthy AR, Wojdacz TK, et al. ChAMP: 450k Chip Analysis Methylation Pipeline. Bioinformatics. 2014;30:428–30.

185. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2021.

186. Tian Y, Morris TJ, Webster AP, Yang Z, Beck S, Feber A, et al. ChAMP: updated methylation analysis pipeline for Illumina BeadChips. Bioinformatics. 2017;33:3982–4.

187. Tian Y. ChAMP Package for DNA methylation analysis. 2022.

188. Johnson JL. Probability and statistics for computer science. Hoboken, NJ: Wiley Interscience; 2008.

189. Jafari M, Ansari-Pour N. Why, When and How to Adjust Your P Values? Cell J. 2019;20:604–7.

190. Mansell G, Gorrie-Stone TJ, Bao Y, Kumari M, Schalkwyk LS, Mill J, et al. Guidance for DNA methylation studies: statistical insights from the Illumina EPIC array. BMC Genomics. 2019;20:366.

191. Pak M, Kim S. A review of deep learning in image recognition. In: 2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT). 2017. p. 1–3.

192. Deldjoo Y, Elahi M, Cremonesi P, Garzotto F, Piazzolla P, Quadrana M. Content-Based Video Recommendation System Based on Stylistic Visual Features. J Data Semant. 2016;5:99–113.

193. Crawford M, Khoshgoftaar TM, Prusa JD, Richter AN, Al Najada H. Survey of review spam detection using machine learning techniques. Journal of Big Data. 2015;2:23.

194. TURING AM. Computing Machinery and Intelligence. Mind. 1950;LIX:433–60.

195. Samuel AL. Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development. 1959;3:210–29.

196. Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review. 1958;65:386–408.

197. Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences. 1982;79:2554–8.

198. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323:533–6.

199. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, et al. Backpropagation Applied to Handwritten Zip Code Recognition. Neural Computation. 1989;1:541–51.

200. Zhu X (Jerry). Semi-Supervised Learning Literature Survey. Technical Report. University of Wisconsin-Madison Department of Computer Sciences; 2005.

201. Russell SJ, Norvig P, Davis E. Artificial intelligence: a modern approach. 3rd ed. Upper Saddle River: Prentice Hall; 2010.

202. Mirtaheri SL, Shahbazian R. Machine Learning Theory to Applications. 1st edition. Boca Raton: CRC Press; 2022.

203. Rebala G, Ravi A, Churiwala S. An Introduction to Machine Learning. Cham: Springer International Publishing; 2019.

204. Zhu X, Goldberg AB. Introduction to Semi-Supervised Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning. 2009;3:1–130.

205. Sutton RS, Barto AG. Reinforcement learning: an introduction. Second edition. Cambridge, Massachusetts: The MIT Press; 2018.

206. Uc-Cetina V, Navarro-Guerrero N, Martin-Gonzalez A, Weber C, Wermter S. Survey on reinforcement learning for language processing. Artif Intell Rev. 2022. https://doi.org/10.1007/s10462-022-10205-5.

207. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media; 2013.

208. Jin X, Han J. K-Means Clustering. In: Sammut C, Webb GI, editors. Encyclopedia of Machine Learning. Boston, MA: Springer US; 2010. p. 563–4.

209. Shetty P, Singh S. Hierarchical Clustering: A Survey. International Journal of Applied Research. 2021;7:178–81.

210. Vijaya, Sharma S, Batra N. Comparative Study of Single Linkage, Complete Linkage, and Ward Method of Agglomerative Clustering. In: 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon). 2019. p. 568–73.

211. Nanga S, Bawah AT, Acquaye BA, Billa M-I, Baeta FD, Odai NA, et al. Review of Dimension Reduction Methods. Journal of Data Analysis and Information Processing. 2021;9:189–231.

212. Yan X, Su X. Linear regression analysis: theory and computing. Singapore ; Hackensack, NJ: World Scientific; 2009.

213. Rosopa PJ, Schaffer MM, Schroeder AN. Managing heteroscedasticity in general linear models. Psychol Methods. 2013;18:335–51.

214. Jiang H. Machine Learning Fundamentals: A Concise Introduction. 1st edition. Cambridge University Press; 2021.

215. Schmidhuber J. Deep Learning in Neural Networks: An Overview. Neural Networks. 2015;61:85–117.

216. Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. Neural Networks. 1989;2:359–66.

217. Heinecke A, Ho J, Hwang W-L. Refinement and Universal Approximation via Sparsely Connected ReLU Convolution Nets. IEEE Signal Processing Letters. 2020;27:1175–9.

218. Le N, Rathour VS, Yamazaki K, Luu K, Savvides M. Deep Reinforcement Learning in Computer Vision: A Comprehensive Survey. 2021.

219. Chen Y, Mancini M, Zhu X, Akata Z. Semi-Supervised and Unsupervised Deep Visual Learning: A Survey. 2022.

220. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. In: Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2012.

221. Oord A van den, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, et al. WaveNet: A Generative Model for Raw Audio. 2016.

222. Mastering the game of Go with deep neural networks and tree search | Nature.

223. Matsubara H, Iida H, Grimbergen R. Natural Developments in Game Research: From CHESS to SHOGI to Go. ICG. 1996;19:103–12.

224. Casado-García Á, Domínguez C, García-Domínguez M, Heras J, Inés A, Mata E, et al. CLoDSA: a tool for augmentation in classification, localization, detection, semantic segmentation and instance segmentation tasks. BMC Bioinformatics. 2019;20:323.

225. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. 2017.

226. Hosseini-Asl E, Ghazal M, Mahmoud A, Aslantas A, Shalaby AM, Casanova MF, et al. Alzheimer's disease diagnostics by a 3D deeply supervised adaptable convolutional network. Front Biosci (Landmark Ed). 2018;23:584–96.

227. Irmak E. Multi-Classification of Brain Tumor MRI Images Using Deep Convolutional Neural Network with Fully Optimized Framework. Iranian Journal of Science and Technology, Transactions of Electrical Engineering. 2021;45:1015.

228. Hosny KM, Kassem MA, Foaud MM. Skin Cancer Classification using Deep Learning and Transfer Learning. 2018 9th Cairo International Biomedical Engineering Conference (CIBEC). 2018;:90–3.

229. Benhammou Y, Achchab B, Herrera F, Tabik S. BreakHis based breast cancer automatic diagnosis using deep learning: Taxonomy, survey and insights. Neurocomputing. 2020;375:9–24.

230. Liu Z, Mao H, Wu C-Y, Feichtenhofer C, Darrell T, Xie S. A ConvNet for the 2020s. 2022.

231. Roth HR, Lee CT, Shin H-C, Seff A, Kim L, Yao J, et al. Anatomy-specific classification of medical images using deep convolutional nets. In: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI). 2015. p. 101–4.

232. Zhao W, Shen L, Han B, Yang Y, Cheng K, Toesca DAS, et al. Markerless Pancreatic Tumor Target Localization Enabled By Deep Learning. Int J Radiat Oncol Biol Phys. 2019;105:432–9.

233. Dascalu A, David EO. Skin cancer detection by deep learning and sound analysis algorithms: A prospective clinical study of an elementary dermoscope. EBioMedicine. 2019;43:107–13.

234. Adegun A, Viriri S. Deep learning techniques for skin lesion analysis and melanoma cancer detection: a survey of state-of-the-art. Artif Intell Rev. 2021;54:811–41.

235. Chouhan V, Singh SK, Khamparia A, Gupta D, Tiwari P, Moreira C, et al. A Novel Transfer Learning Based Approach for Pneumonia Detection in Chest X-ray Images. Applied Sciences. 2020;10:559.

236. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. 2017. p. 4700–8.

237. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2016. p. 770–8.

238. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. 2016. p. 2818–26.

239. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going Deeper With Convolutions. 2015. p. 1–9.

240. Gurgel MSC, Bedone AJ, de Angelo Andrade LAL, Panetta K. Microinvasive Carcinoma of the Uterine Cervix: Histological Findings on Cone Specimens Related to Residual Neoplasia on Hysterectomy. Gynecologic Oncology. 1997;65:437–40.

241. Rivera C, Venegas B. Histological and molecular aspects of oral squamous cell carcinoma (Review). Oncology Letters. 2014;8:7–11.

242. Cireşan DC, Giusti A, Gambardella LM, Schmidhuber J. Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks. In: Mori K, Sakuma I, Sato Y, Barillot C, Navab N, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013. Berlin, Heidelberg: Springer; 2013. p. 411–8.

243. Józefowicz M, Wielisław Papierz. Histologiczne kryteria diagnostyczne oponiaków atypowych (GII WHO) i ich związek z potencjałem proliferacyjnym. 2012.

244. Lei H, Liu S, Elazab A, Gong X, Lei B. Attention-Guided Multi-Branch Convolutional Neural Network for Mitosis Detection From Histopathological Images. IEEE Journal of Biomedical and Health Informatics. 2021;25:358–70.

245. Sirinukunwattana K, Ahmed Raza SE, Yee-Wah Tsang null, Snead DRJ, Cree IA, Rajpoot NM. Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images. IEEE Trans Med Imaging. 2016;35:1196–206.

246. Celik Y, Talo M, Yildirim O, Karabatak M, Acharya UR. Automated invasive ductal carcinoma detection based using deep transfer learning with whole-slide images. Pattern Recognition Letters. 2020;133:232–9.

247. Thaha MM, Kumar KPM, Murugan BS, Dhanasekeran S, Vijayakarthick P, Selvi AS. Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images. J Med Syst. 2019;43:294.

248. Pereira S, Pinto A, Alves V, Silva CA. Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images. IEEE Transactions on Medical Imaging. 2016;35:1240–51.

249. Hu K, Gan Q, Zhang Y, Deng S, Xiao F, Huang W, et al. Brain Tumor Segmentation Using Multi-Cascaded Convolutional Neural Networks and Conditional Random Field. IEEE Access. 2019;7:92615–29.

250. Fang F, Fan S, Zhang X, Zhang MQ. Predicting methylation status of CpG islands in the human brain. Bioinformatics. 2006;22:2204–9.

251. Tian Q, Zou J, Tang J, Fang Y, Yu Z, Fan S. MRCNN: a deep learning model for regression of genome-wide DNA methylation. BMC Genomics. 2019;20:192.

252. Kapourani C-A, Sanguinetti G. Melissa: Bayesian clustering and imputation of single-cell methylomes. Genome Biology. 2019;20:61.

253. Souza CPE de, Andronescu M, Masud T, Kabeer F, Biele J, Laks E, et al. Epiclomal: Probabilistic clustering of sparse single-cell DNA methylation data. PLOS Computational Biology. 2020;16:e1008270.

254. Fan S, Li C, Ai R, Wang M, Firestein GS, Wang W. Computationally expanding infinium HumanMethylation450 BeadChip array data to reveal distinct DNA methylation patterns of rheumatoid arthritis. Bioinformatics. 2016;32:1773–8.

255. Bardowell SA, Parker J, Fan C, Crandell J, Perou CM, Swift-Scanlan T. Differential methylation relative to breast cancer subtype and matched normal tissue reveals distinct patterns. Breast Cancer Res Treat. 2013;142:365–80.

256. Houseman EA, Christensen BC, Yeh R-F, Marsit CJ, Karagas MR, Wrensch M, et al. Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. BMC Bioinformatics. 2008;9:365.

257. Amor R del, Colomer A, Monteagudo C, Naranjo V. A deep embedded refined clustering approach for breast cancer distinction based on DNA methylation. Neural Comput & Applic. 2022;34:10243–55.

258. Si Z, Yu H, Ma Z. Learning Deep Features for DNA Methylation Data Analysis. IEEE Access. 2016;4:2732–7.

259. Bahado-Singh RO, Vishweswaraiah S, Er A, Aydas B, Turkoglu O, Taskin BD, et al. Artificial intelligence and the detection of pediatric concussion using epigenomic analysis. Brain Research. 2020;1726.

260. Bahado-Singh RO, Vishweswaraiah S, Aydas B, Yilmaz A, Saiyed NM, Mishra NK, et al. Precision cardiovascular medicine: artificial intelligence and epigenetics for the pathogenesis and prediction of coarctation in neonates. J Matern Fetal Neonatal Med. 2022;35:457–64.

261. Albaradei S, Napolitano F, Thafar MA, Gojobori T, Essack M, Gao X. MetaCancer: A deep learning-based pan-cancer metastasis prediction model developed using multi-omics data. Comput Struct Biotechnol J. 2021;19:4404–11.

262. Xia C, Xiao Y, Wu J, Zhao X, Li H. A Convolutional Neural Network Based Ensemble Method for Cancer Prediction Using DNA Methylation Data. In: Proceedings of the 2019 11th International Conference on Machine Learning and Computing. New York, NY, USA: Association for Computing Machinery; 2019. p. 191–6.

263. Zhang M, Pan C, Liu H, Zhang Q, Li H. An Attention-Based Deep Learning Method for Schizophrenia Patients Classification Using DNA Methylation Data. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2020. p. 172–5.

264. Levy JJ, Titus AJ, Petersen CL, Chen Y, Salas LA, Christensen BC. MethylNet: an automated and modular deep learning approach for DNA methylation analysis. BMC Bioinformatics. 2020;21:108.

265. Titus AJ, Bobak CA, Christensen BC. A New Dimension of Breast Cancer Epigenetics - Applications of Variational Autoencoders with DNA Methylation: In: Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies. Funchal, Madeira, Portugal: SCITEPRESS - Science and Technology Publications; 2018. p. 140–5.

266. Wang Z, Wang Y. Extracting a biologically latent space of lung cancer epigenetics with variational autoencoders. BMC Bioinformatics. 2019;20:568.

267. Hao J, Kosaraju SC, Tsaku NZ, Song DH, Kang M. PAGE-Net: Interpretable and Integrative Deep Learning for Survival Analysis Using Histopathological Images and Genomic Data. Pac Symp Biocomput. 2020;25:355–66.

268. Lemsara A, Ouadfel S, Fröhlich H. PathME: pathway based multi-modal sparse autoencoders for clustering of patient-level multi-omics data. BMC Bioinformatics. 2020;21:146.

269. Müller F, Scherer M, Assenov Y, Lutsik P, Walter J, Lengauer T, et al. RnBeads 2.0: comprehensive analysis of DNA methylation data. Genome Biology. 2019;20:55.

270. Park Y, Figueroa ME, Rozek LS, Sartor MA. MethylSig: a whole genome DNA methylation analysis pipeline. Bioinformatics. 2014;30:2414–22.

271. Zhang B, Zhou Y, Lin N, Lowdon RF, Hong C, Nagarajan RP, et al. Functional DNA methylation differences between tissues, cell types, and across individuals discovered using the M&M algorithm. Genome Res. 2013;23:1522–40.

272. Su J, Yan H, Wei Y, Liu H, Liu H, Wang F, et al. CpG_MPs: identification of CpG methylation patterns of genomic regions from high-throughput bisulfite sequencing data. Nucleic Acids Research. 2013;41:e4.

273. Zhang Y, Liu H, Lv J, Xiao X, Zhu J, Liu X, et al. QDMR: a quantitative method for identification of differentially methylated regions by entropy. Nucleic Acids Research. 2011;39:e58.

274. Shen L, Zhu J, Robert Li S-Y, Fan X. Detect differentially methylated regions using non-homogeneous hidden Markov model for methylation array data. Bioinformatics. 2017;33:3701–8.

275. Reel PS, Reel S, Pearson E, Trucco E, Jefferson E. Using machine learning approaches for multi-omics data analysis: A review. Biotechnology Advances. 2021;49:107739.

276. Noé F, De Fabritiis G, Clementi C. Machine learning for protein folding and dynamics. Current Opinion in Structural Biology. 2020;60:77–84.

277. Baskin II, Winkler D, Tetko IV. A renaissance of neural networks in drug discovery. Expert Opinion on Drug Discovery. 2016;11:785–95.

278. Wang Z, Majewicz Fey A. Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. Int J CARS. 2018;13:1959–70.

279. Kaur T, Gandhi TK. Deep convolutional neural networks with transfer learning for automated brain image classification. Machine Vision and Applications. 2020;31:20.

280. Huang F, Zhang J, Zhou C, Wang Y, Huang J, Zhu L. A deep learning algorithm using a fully connected sparse autoencoder neural network for landslide susceptibility prediction. Landslides. 2020;17:217–29.

281. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. J Digit Imaging. 2013;26:1045–57.

282. Aksac A, Demetrick DJ, Ozyer T, Alhajj R. BreCaHAD: a dataset for breast cancer histopathological annotation and diagnosis. BMC Research Notes. 2019;12:82.

283. Krizhevsky A, Hinton G, others. Learning multiple layers of features from tiny images. 2009.

284. Althnian A, AlSaeed D, Al-Baity H, Samha A, Dris AB, Alzakari N, et al. Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain. Applied Sciences. 2021;11:796.

285. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. Journal of Big Data. 2019;6:27.

286. Shorten C, Khoshgoftaar TM. A survey on Image Data Augmentation for Deep Learning. Journal of Big Data. 2019;6:60.

287. Wang F, Wang H, Wang H, Li G, Situ G. Learning from simulation: An end-to-end deep-learning approach for computational ghost imaging. Opt Express, OE. 2019;27:25560–72.

288. Wood E, Baltrušaitis T, Hewitt C, Dziadzio S, Cashman TJ, Shotton J. Fake It Till You Make It: Face Analysis in the Wild Using Synthetic Data Alone. 2021. p. 3681–91.

289. Falconi LG, Perez M, Aguila WG, Conci A. Transfer Learning and Fine Tuning in Breast Mammogram Abnormalities Classification on CBIS-DDSM Database. Adv sci technol eng syst j. 2020;5:154–65.

290. Wang K-C, Fu Y, Li K, Khisti A, Zemel R, Makhzani A. Variational Model Inversion Attacks. 2022.

291. Ye D, Zhu T, Zhou S, Liu B, Zhou W. Label-only Model Inversion Attack: The Attack that Requires the Least Information. 2022.

292. Shokri R, Stronati M, Song C, Shmatikov V. Membership Inference Attacks against Machine Learning Models. 2017.

293. Carlini N, Chien S, Nasr M, Song S, Terzis A, Tramer F. Membership Inference Attacks From First Principles. 2022.

294. Yaacoub J-PA, Noura HN, Salman O, Chehab A. A Survey on Ethical Hacking: Issues and Challenges. 2021.

295. Costan V, Devadas S. Intel SGX Explained. 2016.

296. Xu R, Baracaldo N, Joshi J. Privacy-Preserving Machine Learning: Methods, Challenges and Directions. 2021.

297. Ji Z, Lipton ZC, Elkan C. Differential Privacy and Machine Learning: a Survey and Review. 2014.

298. Acar A, Aksu H, Uluagac AS, Conti M. A Survey on Homomorphic Encryption Schemes: Theory and Implementation. 2017.

299. Slijepčević D, Henzl M, Klausner LD, Dam T, Kieseberg P, Zeppelzauer M. k-Anonymity in Practice: How Generalisation and Suppression Affect Machine Learning Classifiers. Computers & Security. 2021;111:102488.

300. Azizi Z, Zheng C, Mosquera L, Pilote L, Emam KE. Can synthetic data be a proxy for real clinical trial data? A validation study. BMJ Open. 2021;11:e043497.

301. Chen RJ, Lu MY, Chen TY, Williamson DFK, Mahmood F. Synthetic data in machine learning for medicine and healthcare. Nat Biomed Eng. 2021;5:493–7.

302. Kuo NI-H, Garcia F, Sönnerborg A, Zazzi M, Böhm M, Kaiser R, et al. Generating Synthetic Clinical Data that Capture Class Imbalanced Distributions with Generative Adversarial Networks: Example using Antiretroviral Therapy for HIV. 2023.

303. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 1998;86:2278–324.

304. Schmidhuber J. Deep Learning in Neural Networks: An Overview. Neural Networks. 2015;61:85–117.

305. Dubey SR, Singh SK, Chaudhuri BB. Activation Functions in Deep Learning: A Comprehensive Survey and Benchmark. 2022.

306. Datta L. A Survey on Activation Functions and their relation with Xavier and He Normal Initialization. 2020.

307. Karlik B, Olgac AV. Performance Analysis of Various Activation Functions in Generalized MLP Architectures of Neural Networks.

308. Zou S, Chen W, Chen H. Image Classification Model Based on Deep Learning in Internet of Things. Wireless Communications and Mobile Computing. 2020;2020:e6677907.

309. Castaneda G, Morris P, Khoshgoftaar TM. Evaluation of maxout activations in deep learning across several big data domains. Journal of Big Data. 2019;6:72.

310. Glorot X, Bordes A, Bengio Y. Deep Sparse Rectifier Neural Networks. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings; 2011. p. 315–23.

311. Lu L, Shin Y, Su Y, Karniadakis GE. Dying ReLU and Initialization: Theory and Numerical Examples. CiCP. 2020;28:1671–706.

312. Klambauer G, Unterthiner T, Mayr A, Hochreiter S. Self-Normalizing Neural Networks. 2017.

313. Wang X, Ren H, Wang A. Smish: A Novel Activation Function for Deep Learning Methods. Electronics. 2022;11:540.

314. Goodfellow IJ, Warde-Farley D, Mirza M, Courville A, Bengio Y. Maxout Networks. 2013.

315. Bolya D, Zhou C, Xiao F, Lee YJ. YOLACT: Real-Time Instance Segmentation. 2019. p. 9157–66.

316. Kumar SK. On weight initialization in deep neural networks. 2017.

317. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks.

318. He K, Zhang X, Ren S, Sun J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. IEEE International Conference on Computer Vision (ICCV 2015). 2015;1502.

319. Li H, Krček M, Perin G. A Comparison of Weight Initializers in Deep Learning-Based Side-Channel Analysis. In: Zhou J, Conti M, Ahmed CM, Au MH, Batina L, Li Z, et al., editors. Applied Cryptography and Network Security Workshops. Cham: Springer International Publishing; 2020. p. 126–43.

320. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, Massachusetts: The MIT Press; 2016.

321. Raschka S, Mirjalili V. Python machine learning: machine learning and deep learning with python, scikit-learn, and tensorflow 2. Third edition. Birmingham: Packt Publishing, Limited; 2019.

322. Jiang J, Cui B, Zhang C. Distributed Machine Learning and Gradient Optimization. Singapore: Springer Singapore; 2022.

323. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, et al. Automatic differentiation in PyTorch. 2017.

324. Li L, Doroslovački M, Loew MH. Approximating the Gradient of Cross-Entropy Loss Function. IEEE Access. 2020;8:111626–35.

325. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323:533–6.

326. Unpingco J. Python for Probability, Statistics, and Machine Learning. Cham: Springer International Publishing; 2022.

327. Michelucci U. Applied Deep Learning: A Case-Based Approach to Understanding Deep Neural Networks. Berkeley, CA: Apress; 2018.

328. Ruder S. An overview of gradient descent optimization algorithms. 2017.

329. Kawaguchi K. Deep Learning without Poor Local Minima. In: Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2016.

330. Keskar NS, Mudigere D, Nocedal J, Smelyanskiy M, Tang PTP. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. 2017.

331. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. 2017.

332. Sutskever I, Martens J, Dahl G, Hinton G. On the importance of initialization and momentum in deep learning. In: Proceedings of the 30th International Conference on Machine Learning. PMLR; 2013. p. 1139–47.

333. Llugsi R, Yacoubi SE, Fontaine A, Lupera P. Comparison between Adam, AdaMax and Adam W optimizers to implement a Weather Forecast based on Neural Networks for the Andean city of Quito. In: 2021 IEEE Fifth Ecuador Technical Chapters Meeting (ETCM). 2021. p. 1–6.

334. Smith LN, Topin N. Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates. 2018.

335. Caruana R, Lawrence S, Giles C. Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping. In: Leen T, Dietterich T, Tresp V, editors. Advances in Neural Information Processing Systems. MIT Press; 2000.

336. Zhang G, Wang C, Xu B, Grosse R. Three Mechanisms of Weight Decay Regularization. 2018.

337. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research. 2014;15:1929–58.

338. D'Amour A, Heller K, Moldovan D, Adlam B, Alipanahi B, Beutel A, et al. Underspecification Presents Challenges for Credibility in Modern Machine Learning. 2020.

339. Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks. In: Proceedings of the 30th International Conference on Machine Learning. PMLR; 2013. p. 1310–8.

340. Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks. 1994;5:157–66.

341. PyTorch documentation — PyTorch 1.13 documentation. https://pytorch.org/docs/1.13/. Accessed 9 Mar 2023.

342. Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol Cybernetics. 1980;36:193–202.

343. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. 2017.

344. Tunstall L, von Werra L, Wolf T. Natural Language Processing with Transformers. O'Reilly; 2022.

345. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on

Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings; 2010. p. 249–56.

346. Gao C, Yan J, Zhou S, Varshney PK, Liu H. Long short-term memory-based deep recurrent neural networks for target tracking. Information Sciences. 2019;502:279–96.

347. Yildirim S, Asgari-Chenaghlu M. Mastering Transformers: Build state-of-the-art models from scratch with advanced natural language processing techniques. Packt Publishing; 2021.

348. Johnston WA, Dark VJ. Selective Attention. Annual Review of Psychology. 1986;37:43–75.

349. Vylomova E, Rimell L, Cohn T, Baldwin T. Take and Took, Gaggle and Goose, Book and Read: Evaluating the Utility of Vector Differences for Lexical Relation Learning. 2016.

350. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. 2021.

351. Foundation PS. Python Language Reference. 2021.

352. YAML Ain't Markup Language. 2021.

353. Białopiorowicz E, Juszczyński P. Molekularna patogeneza przewlekłej białaczki limfocytowej. Hematologia. 2017;7:273–86.

354. Salas LA, Koestler DC, Butler RA, Hansen HM, Wiencke JK, Kelsey KT, et al. An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. Genome Biol. 2018;19:64.

355. Yosifov DY, Bloehdorn J, Döhner H, Lichter P, Stilgenbauer S, Mertens D. DNA methylation of chronic lymphocytic leukemia with differential response to chemotherapy. Sci Data. 2020;7:133.

356. Rand WM. Objective Criteria for the Evaluation of Clustering Methods. Journal of the American Statistical Association. 1971;66:846–50.

357. Schmiedel BJ, Singh D, Madrigal A, Valdovino-Gonzalez AG, White BM, Zapardiel-Gonzalo J, et al. Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. Cell. 2018;175:1701-1715.e16.

358. Bagger FO, Sasivarevic D, Sohi SH, Laursen LG, Pundhir S, Sønderby CK, et al. BloodSpot: a database of gene expression profiles and transcriptional programs for healthy and malignant haematopoiesis. Nucleic Acids Research. 2016;44:D917–24.

359. Segal E, Friedman N, Koller D, Regev A. A module map showing conditional activity of expression modules in cancer. Nat Genet. 2004;36:1090–8.

360. Janovska P, Poppova L, Plevova K, Plesingerova H, Behal M, Kaucka M, et al. Autocrine Signaling by Wnt-5a Deregulates Chemotaxis of Leukemic Cells and Predicts Clinical Outcome in Chronic Lymphocytic Leukemia. Clinical Cancer Research. 2016;22:459–69.

361. Bagacean C, Iuga CA, Bordron A, Tempescul A, Pralea I-E, Bernard D, et al. Identification of altered cell signaling pathways using proteomic profiling in stable and progressive chronic lymphocytic leukemia. Journal of Leukocyte Biology. 2022;111:313–25.

362. Stevenson FK, Forconi F, Kipps TJ. Exploring the pathways to chronic lymphocytic leukemia. Blood. 2021;138:827–35.

363. Vendramini E, Bomben R, Pozzo F, Benedetti D, Bittolo T, Rossi FM, et al. KRAS, NRAS, and BRAF mutations are highly enriched in trisomy 12 chronic lymphocytic leukemia and are associated with shorter treatment-free survival. Leukemia. 2019;33:2111–5.

364. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. 2019.

365. Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. 2019.

366. Fruman D, Limon J. Akt and mTOR in B Cell Activation and Differentiation. Frontiers in Immunology. 2012;3.

367. Handi J, Patterson S, Levings M. The Role of the PI3K Signaling Pathway in CD4+ T Cell Differentiation and Function. Frontiers in Immunology. 2012;3.

368. Han JM, Patterson SJ, Levings MK. The Role of the PI3K Signaling Pathway in CD4+ T Cell Differentiation and Function. Front Immunol. 2012;3:245.

369. Abdelrasoul H, Werner M, Setz C, Okkenhaug K, Jumaa H. PI3K induces B-cell development and regulates B cell identity. Scientific Reports. 2018;8.

370. Abdelrasoul H, Werner M, Setz CS, Okkenhaug K, Jumaa H. PI3K induces B-cell development and regulates B cell identity. Sci Rep. 2018;8:1327.

371. Aristizabal MJ, Anreiter I, Halldorsdottir T, Odgers CL, McDade TW, Goldenberg A, et al. Biological embedding of experience: A primer on epigenetics. Proc Natl Acad Sci USA. 2020;117:23261–9.

372. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun. 2017;8:1826.

373. Oughtred R, Rust J, Chang C, Breitkreutz B-J, Stark C, Willems A, et al. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci. 2021;30:187–200.

Spis użytych skrótów i tłumaczeń:

Język polski	Oryginalna pisownia	Użyty skrót
algorytm k-średnich	k-means	-
analiza głównych składowych	principal component analysis	PCA
analiza skupień, grupowanie	data clustering	-
atak metodą odwrócenia modelu	model inversion attack	-
atak na modele uczenia maszynowego oparte na wnioskowaniu o członkostwie	membership inference attacks	-
augmentacja obrazów	image augmentation	-
autoenkoder	autoencoder	AE
badania asocjacyjne całego epigenomu	epigenome-wide association study	EWAS
białkiem wiążącym TATA	TATA-binding protein	-
częściowo metylowane	hemi-methylated	-
dioksygenazy TET	ten-eleven translocation dioxygenases	TET
dobrowolny	voluntary	-
domena SRA	SET- and RING-finger associated domain	-
domeny ATRX-DNMT3D-DNMT3L	-	ADD
domeny Pro-Trp-Trp-Pro	-	PWWP
dostęp	accession	-
działanie szlaku naprawy przez wycinanie zasad	base excision repair pathway	BER
efekt pozycyjny	positional effect	-
efekt serii	batch effect	-
etykiety	labels	-
funkcja kosztów	loss function	-
funkcja samouwagi	self-attention	-
funkcja wartości	value function	-
geny referencyjne	house keeping genes	-
głębokie nauczanie	deep learning	-
glikozylaza tymina DNA	thymine DNA glycosylase	TDG

grupa metylowa	-	-CH3
helisa-zwrot-helisa	helix-turn-helix	-
hierarchiczny model mieszany	hierarchical mixture model	-
hipermutacja somatyczna rejonu zmiennego ciężkiego łańcucha immunoglobulin	immunoglobulin heavy chain variable	IGHV
histon 3 trimetylowany na lizynie 4	tri-methylation at the 4th lysine residue of the histone H3 protein	H3K4me3
jądro konwolucji	kernel	-
jednostki bramkujące	gated units	-
klasteryzacja hierarchiczna	hierarchical clustering	-
kofaktor UHRF1	ubiquitin-like, containing PHD and RING finger domains 1	UHRF1
komórki prapłciowe	primoridal germ cells	PGCs
krok	stride	-
krzyżowa hybrydyzacja	cross-hybridization	-
kwas dezoksyrybonukleinowy	deoxyribonucleic acid	DNA
kwas rybonukleinowy	ribonucleic acid	RNA
lasów losowych	random forest	-
ligaza ubikwitynowa E3	E3 ubiquitin ligase	-
manipulator wielkości współczynnika uczenia	learning rate scheduler	-
mapa cech	feature map	-
maszyna wektorów nośnych	support vector machine	SVM
matrycowe RNA	messanger RNA	mRNA
metoda dostrajania	fine tuning	-
metoda gradientu wstecznego postego	batch gradient	-
metylotransferazy DNA	DNA methyltransferases	DNMTs
mimowolny	involuntary	-
model regresji zaimplemtowany w bibliotekę ChAMP zastosowną różnicą 0.2	-	ChAMP 0.2
model regresji zaimplemtowany w bibliotekę ChAMP zastosowną różnicą 0.3	-	ChAMP 0.3

model regresji zaimplemtowany w bibliotekę ChAMP zastosowną różnicą 0.5	-	ChAMP 0.5
model sieci neuronowej CTMeth - hypermethylation - hypomethylation	-	CTMeth -hh
model sieci neuronowej CTMeth - hypermethylation - hypomethylation - indeterminate	-	CTMeth - hhi
niedookreślenie	underspecification	-
nukleosomowy czynnik remodelujący	nucleosome remodeling deacetylase	NuRD
odmiennie metylowane regiony komórek macierzystych	germline differentially methylated regions	gDMRs
odmiennie metylowane sondy	differentially methylated probes	DMPs
odmiennie metylowany region	differentially methylated region	DMR
palec cynkowy	zinc finger	-
para klucz-wartość	key-value pair	-
parametry	weight/ parameter	-
patch	fragment	-
polityka w uczeniu przez wzmacnianie	policy	-
poziom ufności	confidence level	-
Projekt Poznania Ludzkiego Genomu	Human Genome Project	-
prywatyzacja różnicująca	differential privacy	-
przeprogramowanie epigenetyczne	epigenetic reprogramming	-
przetrenowanie	overfitting	-
przetwarzanie języka naturalnego	natural language processing	-
rak z niestabilnością mikrosatelitarną	microsatellite unstable cancer	-
redukcja wymiaru	dimensionality reduction	-
regiony CpG otwartego morza	open sea CpG regions	-
retrowirusy endogenne	endogenous retroviruses	ERVs
rozkład wagowy	weight decay	-
rozkład według wartości szczególnych	singular value decomposition	SVD
różnica w średniej metylacji potwierdzonej testem statystycznym T-Studenta >0.2	-	delta 0.2

różnica w średniej metylacji potwierdzonej testem statystycznym T-Studenta >0.3	-	deta 0.3
różnica w średniej metylacji potwierdzonej testem statystycznym T-Studenta >0.5	-	delta 0.5
samo organizujące się mapy	self-organizing map	-
sekwencja wiążaca TATA	TATA box	-
sekwencjonowanie całego genomu z użyciem wodorosiarczynu	whole genome bisulphite sequencing	WGBS
sekwencjonowanie z użyciem wodorosiarczynu	bisulfite sequencing	bis-seq
sekwncja CpG	-	CpG
sieć neuronowa w pełni połączona	feed forward neural network	-
stochastyczna metoda porządkowania sąsiadów w oparciu o rozkład t	t-distributed stochastic neighbour embedding	t-SNE
stochastyczny gradient wsteczny	stochastic gradient descent	SGD
stochastyczny gradient wsteczny dla wybranych próbek	mini-batch stochastic gradient descent	-
stopniowe zmniejszanie współczynnika uczenia	learning rate decay	-
sygnał nagrody	reward signal	-
szacowanie poziomów metylacji	computational estimation of methylation levels	-
szelfy CpG	CpG shores	-
sztuczne sieci neuronowe	artificial neural networks	-
szyfrowanie homomorficzne	homomorphic encryption	-
trafność	accuracy	-
treningowy zbiór danych	training dataset	-
uczenie częściowo nadzorowane	semi-supervised learning	-
uczenie nadzorowane	supervised learning	-
uczenie nienadzorowane	unsupervised learning	-
uczenie przez wzmacnianie	reinforcement learning	-
uczenie z użyciem transferu wiedzy	transfer learning	-
uwaga skalowanego iloczynu skalarnego	Scaled Dot-Product Attention	-
uwaga/atencja	attention	-

walidacyjny zbiór danych	validation dataset	-
wariacyjne autoenkodery	variational autoencoder	VAE
warstwa łącząca	pooling layer	-
warstwa osadzenia	embedding layer	-
warstwa uwagi	attention layer	-
warstwę kodująca pozycję	positional encoding layer	-
wartości odstające	outliers	-
widzenie komputerowe	computer vision	-
wieloczłonowa warstwa uwagi	multi-head attention layer	-
wielokaskadowej konwolucyjnej sieci neuronowej	multi-cascaded convolutional neural network	-
wodorosiarczynu sodu	sodium bisulphite	-
wskaźnik fałszywego wykrywania	false discovery rate	FDR
współautorzy	-	wsp.
wybrzeża CpG	Cpg shelves	-
wyspy CpG	CpG islands	-
wzbogacanie/rozszerzanie	expanding	-
wzmacniaczne transkrypcji	enhancers	-
zaburzenia metylacji w kilku locus genomu	multi-locus imprinting disorder	MLID
zamek leucynowy	leucine zipper	-
zapytanie	query	-
	Database of Immune Cell Expression, Expression quantitative trait loci (eQTLs) and Epigenomics	DICE
	Functional Mapping and Annotation of Genome-Wide Association Studies	FUMA
	gaussin error linear unit	GELU
	long interspersed elements	LINEs
	long short-term memory	LSTM
	long terminal repeat	LTR

	methylation binding domains, methyl-CpG-binding domain	MBD
	methylated DNA immunoprecipitation followed by sequencing	MeDIP-seq
	methylation quantitive trait loci	meQTLs
	methylation-sensitive restriction enzyme sequencing	MRE-seq
	polycomb repressive complex 2	PRC2
	scaled exponential linear unit	SeLU
	shapley Additive ExPlanation	SHAP
	short interspersed elements	SINEs

Spis rycin

Spis tabel

Tabela 12 B-Cell-CD4+ - Listy genów FUMA immunological signatures typowych dla badanych limfocytów B i limfocytów T CD4+ korelujące najbardziej z wynikami uzyskanymi przy użyciu poszczególnych metod

Tabela 13 B-Cell-CD4+ - Listy genów FUMA BioCarta korelujące najbardziej z wynikami uzyskanymi przy użyciu poszczególnych metod

Tabela 14 Liczba genów pokrywających się z BIOCARTA_MAPK_PATHWAY wskazanych przez analizowane metody i z rożną ekspresją dla CD4+ i limfocytów B wg DICE

Tabela 15 B-Cell-CLL - Listy genów FUMA z kategorii KEGG korelujące najbardziej z wynikami uzyskanymi przy użyciu poszczególnych metod

Tabela 16 B-Cell-CLL - Lista genów KEGG_PATHWAYS_IN_CANCER z FUMA i stopień korelacji z wynikami uzyskanymi przy użyciu poszczególnych metod

Tabela 17 B-Cell-CLL - Korelacja pomiędzy różnicą w ekspresji poszczególnych genów u pacjentów zdrowych i chorych na przewlekłą białaczkę limfocytową a występowaniem tych genów w wynikach uzyskanych z poszczególnych metod. Geny te występują w liście genów KEGG_PATHWAYS_IN_CANCER i są unikalne dla wyników poszczególnych metod. Informacja o ekpresji pochodzi z bazy danych BloodSpot

Tabela 18 B-Cell-CLL - Listy genów FUMA Oncogenic signatures korelujące najbardziej z wynikami uzyskanymi przy użyciu poszczególnych metod

Tabela 19 B-Cell-CLL - Korelacja pomiędzy różnicą w ekspresji poszczególnych genów u pacjentów zdrowych i chorych na przewlekłą białaczkę limfocytową a występowaniem tych genów w wynikach uzyskanych z poszczególnych metod. Geny te pokrywają się pomiędzy listami genów NFE2L2.V2, KRAS.600_UP.V1_UP, KRAS.600.LUNG.BREAST_UP.V1_DN. Informacja o ekpresji pochodzi z bazy danych BloodSpot

ANALYSIS OF CG METHYLATION SEQUENCES WITH MACHINE LEARNING AND NEURAL NETWORKS

Doctoral thesis in the field of medical sciences and health sciences

Field of study: Medical Sciences

lek. Tomasz Falgowski

Introduction

Human genome complexity relays not only on specific composition of billions of base pairs, but also on its alteration through chemical reactions, which can be read and interpreted by enzymes and different molecular factors. These chemical modifications depend on epigenetic mechanisms. DNA methylation of cytosine is one of the best-described factors in this complex system, defined as an epigenetics. It involves the covalent attachment of methyl groups (-CH3) at the 5th position of the pyrimidine ring of cytosine in DNA-by-DNA methyltransferases (DNMTs). This process is important for proper development and functioning and plays a significant role: in genomic imprinting, X chromosome inactivation, transcription regulation, and in the pathogenesis of many diseases, including cancerous ones, where it serves as a potential biomarker for detection and classification. One of the most popular platforms for methylation analysis is Illumina Infinium BeadChips Beta value is the primary value used to measure the level of methylation. Standard and widely used methods for comparing differences in methylation between control and test groups include linear regression models implemented in the ChAMP library and delta methylation confirmed by a Student's t-test (e.g., T-Student). However, beta value does not meet some standard statistical assumptions, such as normal distribution or homoscedasticity, making it difficult to analyze with their use and associated with a certain degree of error risk. Therefore, new methods for its analysis are being sought. In recent years, there has been

significant progress in the field of artificial intelligence and neural networks. Neural networks can perform the function of a universal approximator, meaning they can approximate any function with possibly significant accuracy, which is why their use is increasingly common in various fields. This development raises hopes for their application in DNA methylation analysis as well.

Aim of a study:

The purpose of this thesis is to evaluate whether a neural network architecture based on a combination of convolutional neural networks and transformer-type networks (convolutional-transformers neural networks), which has been trained using synthetic data, is a comparable or better tool for analyzing DNA methylation than standard methods.

Methodology

The subject of this study is a neural network based on a convolutional neural network and transformer-type networks, which has been trained using synthetic data. The purpose of the tool's operation is to identify CpG sequences in the studied samples that are differently methylated between the control group and the test group and have the highest possible biological significance. The input data for the algorithm are beta values, and the developed tool analyzes sets of data, assigning each CpG sequence for the control group and the tested group a label: "hypomethylated", "hypermethylated" or "indetermined/partially methylated", based on its chain of beta values in the samples. The algorithm operates in two versions: CTMeth-hh and CTMeth-hhi. The CTMeth-hh version focuses on identifying sets of data where values in one group are hypomethylated, while in the other group they are hypermethylated or vice versa. The CTMeth-hhi version selects sets where one of the control or test groups contains extreme values, and the other either indetermined/partially-methylated values or extreme values in the opposite range. The neural network was trained using synthetic data generated to reflect the likely distribution of beta value data, including extreme outlier values and small, random variations. Evaluations were conducted using a generalized dictionary to fully utilize the potential of neural networks as a universal approximation tool. The neural network method was compared to two standard methods

used in methylation analysis: linear regression implemented in the widely used ChAMP library and the delta methylation difference method confirmed by a Student's t-test ($p<0.05$). To conduct a detailed analysis, the study relied on the use of three diverse datasets containing information about DNA methylation in B lymphocytes and CD4+ T cells (B-cell-CD4+), healthy B lymphocytes and B lymphocytes from patients with chronic lymphocytic leukemia (B-cell-CLL), as well as DNA methylation in patients with chronic lymphocytic leukemia with IGHV 100% and less (CLL-100). The data was selected to represent three analytical problems: a symmetric division into control and test groups with homogeneous differences (B-cell-CD4+), an asymmetric division into groups (B-cell-CLL), and a complex, heterogeneous set of data (CLL-100). The datasets were obtained from public sources on the Gene Expression Omnibus website. (accession GSE110554, GSE136724). The research methodology was designed in five stages of verification: evaluating the selectivity of the method based on the number of CpG sequences identified as differently methylated, evaluating specificity based on the ability to identify the optimal number of CpG sequences that allow for differentiation between the control group and the test group, the ability to identify CpG sequences with potential biological significance, the ability to obtain results meeting criteria similar to those accepted by Bibikova et al., and evaluating the indication of false positive and negative results based on simulated data.

In this way, the effectiveness of each tool was thoroughly assessed, and the results were compared between different datasets.

Wyniki

The CTMeth-hhi method was comparable or better than standard methods on all stages of the assessment. In cases of analyzing more homogeneous data, the selectivity of CTMeth-hhi was comparable to the standard delta 0.2 method, while in the most complex dataset (CLL-100), CTMeth-hhi identified a larger number of differently methylated CpG sequences and exhibited higher efficiency in selecting the required ones for differentiating between control and test groups, as demonstrated using the Rand index. Based on the analysis of annotated genes to CpG sequences, it was found that the results obtained using the CTMeth-hhi method showed greater biological significance, which was assessed using FUMA, DICE, and BloodSpot databases. Both CTMeth

methods exhibited better ability to identify differently methylated CpG sequences in a manner similar to Bibikova et al., both in the test using real data and simulated data.

Conclusions

The CTMeth-hhi method can be an alternative and effective approach to DNA methylation analysis. Using synthetic data for training allows solving the problem of medical data privacy, as well as their scarcity and imbalance. Additionally, it enables further modification of the neural network's operation and adaptation in line with the development of knowledge about methylation or research needs.

Keywords

epigenetics, DNA methylatoin, neural networks, chronic lymphocytic leukemia

ANALIZA METYLACJI SEKWENCJI CpG W OPARCIU O UCZENIE MASZYNOWE I SIECI NEURONOWE

Rozprawa doktorska w dziedzinie nauk medycznych i nauk o zdrowiu

Dyscyplina nauki medyczne

lek. Tomasz Falgowski

Wprowadzenie

Złożoność ludzkiego genomu, polega nie tylko na określonej kompozycji miliardów par zasad, ale także chemicznej modyfikacji, która może być odczytywana i interpretowana przez enzymy i inne czynniki molekularne. Te chemiczne modyfikacje zależne są od epigenetycznych mechanizmów, z których jednym z najlepiej opisanych jest metylacja kwasu deoksyrybonukleinowego (DNA) polegająca na kowalencyjnym przyłączeniu grup metylowych (-CH3) w pozycji 5 pierścienia pirymidynowego cytozyny w DNA przez metylotransferazy DNA (DNMTs). Jest to ważny proces dla prawidłowego rozwoju i funkcjonowania organizmu, krtóry odgrywa znaczącą rolę m.in. w genomowym imprintingu, regulacji transkrypcji DNA, oraz w patogenezie wielu chorób w tym, dla których stanowi potencjalny biomarker detekcji i klasyfikacji.Poszechną metodą do analizy metylacji jest platforma Illumina Infinium BeadChips, a wartość beta jest podstawową wartością używaną do pomiaru stopnia metylacji. Standardowymi metodami do porównania różnic w metylacji pomiędzy grupą kontrolną i badaną są modele liniowej regresji zaimplementowanej w bibliotece ChAMP oraz różnice w średniej metylacji potwierdzonej testem statystycznym. Wartość beta, jednakże nie spełnia niektórych założeń standardowych testów statystycznych, jak rozkład normalny, czy homoskedastyczność dlatego analiza z ich użyciem jest obarczona ryzykiem błędu. W ostatnim czasie obserwujemy znaczący rozwój w dziedzinie sztucznej inteligencji i sieci neuronowych. Sieci neuronowe mogą pełnić funkcję uniwersalnego

aproksymatora, co oznacza, że mogę aproksymować dowolną funkcję z możliwie istotną dokładnością, dlatego współcześnie widoczne jest ich coraz częstsze zastosowanie w różnorakich dziedzinach. Fakt ten budzi nadzieję na ich wykorzystanie również w analizie metylacji DNA.

Cel pracy

Celem tej rozprawy jest ocena czy architektura sieci neuronowej oparta o kombinację sieci neuronowej konwolucyjną i typu transformers (ang. convolutional-transformers neural networks), która została przetrenowana z użyciem syntetycznych danych, jest porównywalnym lub lepszym od standardowych metod narzędziem do analizy metylacji DNA.

Metodologia

Przedmiotem badania tej rozprawy jest sieć neuronowa oparta o sieć konwolucyjną i sieć typu transformers, którą przetrenowano z użyciem danych syntetycznych. Celem działania tego narzędzia jest wskazanie w badanych próbkach sekwencji CpG , które są odmiennie metylowane pomiędzy grupą kontrolną a badaną i o możliwie największym znaczeniu biologicznym. Danymi wejściowymi dla algorytmu są wartości β, a opracowane narzędzie analizuje zestawy danych, przypisując każdej sekwencji CpG dla grupy kontrolnej i badanej etykietę: "hipometylowany", "hipermetylowany" lub "nieokreślony/częściowo-metylowany", na podstawie jej ciągu wartości B w próbkach. Algorytm działa w dwóch wersjach CTMeth-hh oraz CTMeth-hhi. Wersja CTMeth-hh skupia się na wyodrębnieniu zestawów danych, gdzie wartości w jednej grupie są hipometylowane, a w drugiej hipermetylowane, lub odwrotnie. Wersja CTMeth-hhi wybiera zestawy, w których jedna z grup kontrolna lub badana zawiera wartości skrajne, a druga albo wartości nieokreślone/częściowo-metylowane, albo skrajne w przeciwnym zakresie. Sieć neuronową przetrenowano z użyciem danych syntetycznych wygenerowanych tak by odzwierciedlały prawdopodobny rozkład danych wartości **B** włączając w to wartosci skrajne typu outliers i niewielkie, losowe zmienności, a ewaluacje przeprowadzono z ogólnym zgeneralizowanym słownikiem, aby w pełni wykorzystać potencjał sieci neuronowych jako narzędzia do uniwersalnej aproksymacji. Metodę z użyciem sieci neuronowej porównano do dwóch standardowych metod

używanych w analizie metylacji – liniowej regresji zaimplementowanej w powszechnie używanej bibliotece ChAMP oraz różnicy w średniej metylacji (metoda delta) potwierdzonej testem statystycznym T-Studenta (wartość $p<0.05$). W celu przeprowadzenia szczegółowej analizy, badanie oparto na wykorzystaniu trzech zróżnicowanych zbiorów danych zawierających odpowiednio dane na temat metylacji DNA limfocytów B oraz limfocytów T CD4+ (B-cell-CD4+), metylacji DNA zdrowych limfocytów B i limfocytów B pacjentów chorujących na przewlekłą białaczkę limfatyczną (B-cell-CLL) oraz metylacji DNA pacjentów z przewlekłą białaczką limfatyczną z IGHV 100% i mniej (CLL-100). Dane dobrane tak, aby reprezentować trzy problemy analityczne: symetryczny podział na grupy kontrolną i badawczą o homogennych różnicach(B-cell-CD4+), asymetryczny podział na grupy(B-cell-CLL), złożonch heterogeniczny zbiór danych(CLL-100). Bazy danych pozyskano z publicznych zbiorów Gene Expression Omnibus z dostępów (ang. accession) - GSE110554, GSE136724. Metodologia badawcza została zaprojektowana w pięciu etapach weryfikacji: ocena selektywności metody na podstawie ilości wskazywanych odmiennie metylowanych sekwencji CpG, ocena specyficzności na podstawie zdolności do wskazywania optymalnej liczby sekwencji CpG pozwalających na utrzymanie różnicowania na grupę kontrolną i badaną, zdolności do wskazywania sekwencji CpG o potencjalnym znaczeniu biologicznym, zdolności do uzyskania wyników spełniających kryteria analogiczne do przyjętych przez Bibikova i wsp., oceny wskazywania przez metody wyników fałszywie pozytywnych i negatywnych na podstawie danych symulowanych. W ten sposób szczegółowo oceniono skuteczność każdego narzędzia oraz porównano wyniki między różnymi zbiorami danych.

Wyniki

Metoda CTMeth-hhi na wszystkich etapach była porównywalna lub lepsza niż metody standardowe.W przypadkach analizy danych bardziej homogennych selektywność CTMeth-hhi była porównywalna do metody standardowej delta 0.2 , natomiast w przypadku najbardziej złożonego zbioru danych CLL-100, CTMeth-hhi zidentyfikowała większą ilość odmiennie metylowany sekwencji CpG, a jednocześnie wykazała się wyższą wydajność w ich doborze wymaganych do różnicowania pomiędzy grupami kontrolnymi i badanymi, co wykazano z użyciem indeksu Randa. Na

podstawie analizy anotowany genów do sekwencji CpG stwierdzono, że wyniki uzyskane przy użyciu metody CTMeth-hhi wykazują większe znaczenie biologiczne. Ocenę przeprowadzono z użyciem baz danych FUMA. DICE, oraz BloodSpot. Obie metody CTMeth wykazują się lepszą zdolnością do wskazywania odmiennie metylowanych sekwencji CpG na zasadzie analogicznej do . Bibikova i wsp. zarówno w teście z uzyciem danych rzeczywistych , jak i symulowanych.

Wnioski

Metoda CTMeth-hhi może być alternatywną i skuteczną metodą do analizy metylacji DNA. Zastosowanie danych syntetycznych do treningu pozwala na rozwiązanie problemu prywatności danych medycznych oraz ich niedoboru i braku zbalansowania. Dodatkowo pozwala to na dalszą modyfikację działania sieci neuronowej i dostosowanie wraz z rozwojem wiedzy na temat metylacji, czy potrzeb badawczych.

Słowa kluczowe

epigenetyka, metylacja DNA, sieci neuronowe, przewlkeła białaczka limfocytowa