

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, który 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 obciążona 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 mogą aproksymować dowolną funkcję z możliwie istotną dokładnością, dlatego współcześnie widoczne jest ich coraz częstsze zastosowanie w różnorodnych 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 β 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 β włączając w to wartości skrajne typu outliers i niewielkie, losowe zmienności, a ewaluację 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), złożonych 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 użyciem 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, przewłkła białaczka limfocytowa