Streszczenie

Dzięki szybkiemu rozwojowi uczenia maszynowego i biologii molekularnej, dziedziny te zaczęły się mocno przeplatać. Eksperymenty biologiczne mają kluczowe znaczenie dla zrozumienia podłoża wielu złożonych chorób – wymagają jednak dużo czasu i pieniędzy. Doktorat ten pokazuje zastosowania różnych metod uczenia maszynowego, w szczególności głębokich sieci neuronowych, do przewidywania eksperymentów biologicznych. Jednym z najtańszych i najpowszechniejszych eksperymentów biologicznych, na którym ten doktorat jest oparty, jest sekwencjonowanie DNA. Wraz z rozwojem technologii staje się ono coraz bardziej przystępne cenowo, do tego stopnia, że w Internecie dostępne są setki tysięcy zsekwencjonowanych genomów ludzkich. Rozprawa zaczyna się się od analizy sekwencji – i pokazania jej złożonego związku z różnymi chorobami. Następnie zaprezentowano algorytm wykrywania wariantów – aby zapewnić badaczom możliwość uzyskania interesujących wariantów genetycznych. Algorytm wykorzystuje sieć neuronową i jest oparty na 8 najnowocześniejszych narzędzi do wykrywania wariantów. W dalszej części rozprawy zaprezentowano model oparty na DNABERT, który ma za zadanie przewidywanie przestrzennej konformacji chromatyny – czyli pętli obserwowanych w doświadczeniach trójwymiarowej genomiki ChIA-PET. Model został następnie rozszerzony, omijając ograniczenia algorytmu DNABERT. Zaproponowano model HiCDiffusion – algorytm, który łączy nowoczesne podejście architektury enkodera-dekodera wraz z transformerem (w celu wzmocnienia uczenia kontekstowego) i poprawia wyniki poprzez zastosowanie dyfuzji. Dzięki temu wyniki in silico reprezentujące strukturę przestrzenną chromatyny są nieodróżnialne od oryginalnych danych doświadczalnych Hi-C. Praca doktorska kończy się wykazaniem zależności pomiędzy konformacją przestrzenną chromatyny w jądrze komórkowym a ekspresją genów. Dotychczasowo opublikowane modele komputerowe stosowane do przewidywania genów bardzo często uwzględniają jedynie lokalną sekwencję DNA (np. 20 kbp wokół początku genu) i nie uwzględniają segmentów sekwencji położonych znacznie dalej wzdłuż łańcucha DNA. Z badań biologicznych wiadomo jednak, że elementy regulatorowe są są położone blisko siebie w przestrzeni trójwymiarowej, oddziałując z kompleksami białkowymi realizującymi proces trakskrypcji. Nasze badania obejmują zatem całe spektrum analiz genomicznych: od sekwencji DNA i identyfikacji wariantów strukturalnych, poprzez wykorzystanie spersonalizowanych genomów w celu uzyskania

konformacji przestrzennej, aż do przewidywania ekspresji genów. Ekspresja ta jest istotna, ponieważ nadprodukcja lub brak danego białka często leży u podstaw chorób genetycznych.

Słowa kluczowe: Sztuczna Inteligencja (SI), Uczenie Maszynowe, Głębokie Sieci Neuronowe, Duże Modele Językowe, Genomika 3D, Chromatyna, Ekspresja Genów, Warianty Strukturalne