**Wykorzystanie metody głębokiego uczenia w analizie sekwencji genomu zwierząt hodowlanych**

Głębokie uczenie jest dziedziną uczenia maszynowego, metody która w ostatnich latach szybko zyskuje na znaczeniu w wielu dziedzinach nauki. Pierwotnie, algorytmy głębokiego uczenia były wykorzystywane głównie w rozpoznawaniu obrazów, lecz obecnie są one coraz częściej używane również w innych dziedzinach, w tym w genomice. Wg opinii edytorów czasopisma Nature Genetics ze stycznia 2019 r. algorytmy głębokiego uczenia „mają zrewolucjonizować analizę genomu". Ich zastosowania sięgają od analizy ekspresji genów, poprzez modulację regulacji ekspresji genów, aż do proteomiki. Jednak w genomice zwierząt hodowlanych zastosowanie metody uczenia głębokiego pozostają bardzo rzadkie. Również sekwencjonowanie nowej generacji pozwalające na poznanie indywidualnej zmienności sekwencji DNA na poziomie całych genomów jest stosunkowo nowym obszarem nauki. Sekwencje DNA całych genomów mają bardzo wysoką zawartość informacyjną, mogą więc być używane do wykrycia powiązań pomiędzy zmiennością DNA, a zmiennością fenotypów, czy też w badaniach filogenetycznych. Jednakże, należy pamiętać, że z uwagi na wysoko przepustowy charakter pozyskiwania tych danych, są one obarczone wyższym prawdopodobieństwem błędu, niż ma to miejsce w przypadku analiz zmienności DNA prowadzonych na małą skalę. Celem zgłaszanego projektu jest wprowadzenie wykorzystania algorytmów uczenia głębokiego do genomiki zwierząt hodowlanych, z naciskiem na analizę sekwencji DNA całych genomów bydła.

W szczególności, projekt zmierza do wykorzystanie algorytmów uczenia głębokiego w czterech różnych aspektach analizy sekwencji DNA całych genomów. **Pierwszy**, najmniejszy zbiór danych obejmuje *klasyfikację* polimorfizmów pojedynczego nukleotydu (SNP) uzyskanych na podstawie sekwencji DNA całych genomów czterech buhajów uzyskanych za pomocą dwóch technologii (sekwencjonowanie nowej generacji i mikromacierzy oligonukleotydowej) *na poprawnie i błędnie zidentyfikowane*. Błędnie zidentyfikowane polimorfizmy są reprezentowane przez te SNP, które mają genotypy niezgodne między obiema technologiami genotypowania. Algorytm klasyfikacyjny zidentyfikuje cechy sekwencji DNA o największym wpływie na błędną detekcję SNP. **Drugi** zestaw danych obejmuje *klasyfikację* krów *na odporne lub podatne na zapalenie wymienia* w oparciu o mutacje punktowe (SNP) i strukturalne (warianty liczby kopii, CNV), zidentyfikowane na podstawie sekwencji DNA całych genomów 32 krów rasy Polska Holsztyńsko-Fryzyjska. Algorytm klasyfikacji zidentyfikuje SNP i CNV o największym wpływie na odporność na zapalenie wymienia. **Trzeci** zestaw danych zostanie wykorzystany do wielopoziomowego problemu klasyfikacji, w którym polimorfizmy SNP i InDel zidentyfikowane w sekwencji DNA całych genomów, zostaną wykorzystane do *klasyfikacji osobników do poszczególnych ras*. W tym celu wykorzystane zostaną dane z projektu 1 000 Bull Genomes (run7) obejmujące 3 103 osobników reprezentujących różne rasy bydła. Algorytm klasyfikacji zidentyfikuje polimorfizmy charakterystyczne dla poszczególnych ras. **Czwarty** zestaw danych wykorzystuje osobniki i ich genotypy z pierwszego etapu, lecz tym razem w kontekście stworzenia modelu *predykcji*, opartego na rekurencyjnych sieciach neuronowych – pozwalającego na *imputację* pełnego zestawu genotypów SNP dla całego genomu na podstawie podzbioru genotypów SNP uzyskanych z mikromacierzy oligonukleotydowej.

Podsumowując, klasyfikator statusu SNP wybierze cechy sekwencji DNA o największym wpływie na nieprawidłowe wykrywanie polimorfizmu i może być w przyszłości użyty do oszacowania prawdopodobieństwa poprawności genotypów SNP zidentyfikowanych w analizach sekwencji DNA całych genomów. Predyktory opracowane dla klinicznego zapalenia wymienia zidentyfikują SNP i CNV związane z ryzykiem zapalenia wymienia i pozwolą na ocenę tego ryzyka dla każdej krowy z dostępnymi genotypami SNP lub CNV, nawet przed jej wejściem do systemu produkcyjnego. Klasyfikator rasy wybiera SNP, które są najbardziej charakterystyczne dla każdej rasy i w konsekwencji wskazuje na regiony genomowe przyczyniające się do specyficznych rasowo cech.

W ramach projektu pragniemy również pozyskać nowe, niezależne od wykorzystywanych w części metodycznej, dane, które posłużą do walidacji utworzonych algorytmów klasyfikacyjnych (etapy 1-3) oraz algorytmu imputacji genotypów SNP (etap 4). Uzyskanie dobrych wyników walidacji, manifestujących się niskim poziomem błędnych przyporządkowań do klas, na niezależnym zbiorze danych znacznie podnosi wiarygodność uzyskanych wyników oraz przyszły zakres ich stosowania.