**MACIEJ KAŁA**
**MATEUSZ PRZYBOROWSKI**
**SEBASTIAN GASPARIS**
**WACŁAW ORCZYK**
**ANNA NADOLSKA-ORCZYK**
Zakład Genomiki Funkcjonalnej, IHAR — PIB Radzików
m.kala@ihar.edu.pl

# Klasyfikacja podjednostek gluteninowych z wykorzystaniem metod uczenia maszynowego*

**Classification of glutenin subunits using machine learning methods**

Gluteniny to białka zapasowe występujące w bielmie ziarniaków pszenicy. Są one odpowiedzialne za rozciągliwość i elastyczność ciasta. Informacja o ich składzie jest wykorzystywana w procesie hodowlanym. Najprostszą i dającą najlepsze efekty metodą rozpoznawania podjednostek gluteninowych jest rozdział ekstraktów białkowych na żelu poliakrylamidowym w buforze zawierającym laurylosiarczan sodu (SDS-PAGE). Uzyskany obraz jest stosunkowo łatwy do analizy, jednakże duża ilość prób bardzo wydłuża czas opisywania poszczególnych ścieżek i tym samym zwiększa prawdopodobieństwo popełnienia błędu. Do zwiększenia precyzji i szybkości klasyfikacji podjednostek gluteninowych został napisany skrypt w języku Python 3. Zastosowano w nim algorytmy uczenia maszynowego i zbiór zdjęć z opisanymi wcześniej ścieżkami. Skrypt dzieli się na dwa moduły: w pierwszym następuje rozpoznanie ścieżek ze zdjęć, a w drugim budowanie modelu predykcyjnego. Zdjęcia po załadowaniu są wstępnie ujednolicane przez zastosowanie progowania adaptacyjnego a następnie usuwane są artefakty. Przygotowane w ten sposób obrazy są gotowe do wyodrębnienia ścieżek. W większości przypadków gluteniny wysokocząsteczkowe formują ścieżki równoległe co umożliwia zastosowanie prostego skryptu z wykorzystaniem wartości średnich ze wszystkich kolumn a następnie sum kumulatywnych dzięki czemu ścieżki mogą być od siebie odseparowane. W przypadku zniekształconych żeli zastosowano model DBSCAN (Density-based spatial clustering of applications with noise) do rozpoznania poszczególnych prążków, a następnie algorytm centroidów ( k-means) do znalezienia

* Prace zostały wykonane w ramach programu wieloletniego „Tworzenie naukowych podstaw postępu biologicznego i ochrona roślinnych zasobów genowych źródłem innowacji wsparcia zrównoważonego rolnictwa oraz bezpieczeństwa żywnościowego kraju" koordynowanego przez IHAR-PIB a finansowanego przez MRiRW.

centrów wyznaczonych uprzednio klastrów i aproksymacje wielomianową do wyznaczenia środków ścieżek. Uzyskany zbiór opisano ręcznie i nadano etykiety, a następnie znormalizowano do wartości w przedziale 0–1. W celu skrócenia czasu obliczeń i zapobiegnięciu przeuczenia zredukowano wymiary ścieżek przez zastosowanie analizy głównych składowych (PCA) i podzielono na część treningową i testową. Do testowania klasyfikacji wybrano cztery modele: lasy losowe (LL), maszynę wektorów nośnych (SVM), regresję logistyczną (RL) i perceptron wielowarstwowy (MLPC). Dopasowanie parametrów poszczególnych modeli zautomatyzowano funkcją GridSearchCV. Do pomiaru jakości predykcji zastosowano dwa mierniki: 1) średnia harmoniczna z precyzji i czułości (F1 score), 2) dokładność — prawdopodobieństwo prawidłowej klasyfikacji. Z wyjątkiem regresji logistycznej pozostałe modele klasyfikowały ścieżki ze zbioru testowego z wysoką dokładnością przy czym SVM i MLPC osiągnęły średni wynik 95%. Powyższy skrypt umożliwia wybór najlepszego modelu do klasyfikacji glutenin HMW w krótkim czasie i z dużą dokładnością. Tym samym może być cennym narzędziem w pracy hodowlanej.