Głównym zadaniem projektu jest opracowanie metodologii służacej do predykcji własności białek. W tym celu zastosujemy metody regularyzacyjne, które są zdolne do modelowania rzadkich danych i kontrolują liczbę fałszywych odkryć (FDR), tj. SLOPE i adaptacyjny SLOPE. Białka i peptydy pełnią swoje funkcje, ponieważ mogą przyjmować określony przestrzenny kształt. Niestety, analiza eksperymentalna ich struktury jest kosztowna i czasochłonna. W celu jej przyspieszenia, często opieramy się na metodach obliczeniowych, które polegają na łatwo dostępnych sekwencjach aminokwasowych białek. Jednakże, uczenie modeli głębokich dla tego rodzaju danych wymaga ogromnej ilości obserwacji. W przypadku modeli peptydowych, gdzie ilość eksperymentalnie zdefiniowanych sekwencji jest mała, polegamy więc na klasycznych metodach wnioskowania statystycznego, jednak nawet najbardziej nowoczesne podejścia nie wykorzystują w pełni informacji zawartej w sekwencjach aminokwasowych z powodu mało wydajnej reprezentacji cech.

Użycie klasycznych metod wiąże się zatem z koniecznością przekonwertowania sekwencji biologicznych do formatu tabularycznego. W tym projekcie skupiamy się na kodowaniu k-merowym, gdzie sekwencja aminokwasowa jest reprezentowana jako kombinacja ciągłych lub nie ciągłych podrzędnych łańcuchów aminokwasów o długości k. Ponieważ standardowy alfabet aminokwasowy składa się aż z 20 różnych aminokwasów, liczba możliwych k-merów o długości k to aż 20^k, a nawet więcej, jeśli rozważamy k-mery z przerwami. Wskutek tego, przestrzeń k-merowa niesie za sobą problem dużej wymiarowości danych, a także zawiera w sobie wiele nieinformatywnych zmiennych. Dodatkowo, macierze k-merowe są niezwykle rzadkie, więc operacje na nich wymagają dedykowanych algorytmów. Problem ten jest jeszcze bardziej wyrazisty z powodu liczby zmiennych znacznie większej od liczby obserwacji ($p >> n$). Efektywny rozmiar próby jest nawet mniejszy biorąc pod uwagę rzadkość macierzy eksperymentalnej oraz niezbalansowanie lub skośność zmiennej odpowiedzi.W rezultacie, optymalne narzędzie dla danych bazujących na k-merach powinno dokonywać wstępnej selekcji zmiennych. Ponieważ taka reprezentacja pociąga za sobą binarną lub zliczeniową charakterystykę zmiennych niezależnych, wiele ze standardowych metod statystycznych wyboru zmiennych nie ma tutaj zastosowania.

Taki problem może być efektywnie zaadresowany do metod regularyzacji, które są kombinacją wyboru zmiennych i dopasowania modelu dzięki uwzględnieniu "kary" dla dużych współczynników regresji. Jedną z takich metod jest SLOPE, który ostatnio pojawił się jako solidne narzędzie regularyzacyjne i dzięki rankingowej karze (zgodnej z rozmiarem współczynnika) kontroluje liczbę fałszywych odkryć na niskim poziomie przewyższając tym samym inne narzędzia..

Aktualnie, powszechnym rozwiązaniem jest zastosowanie wstępnej metody filtrującej, a potem użycie klasycznego modelu takiego jak lasy losowe czy model LASSO używający prostej regularyzacji. Na chwilę obecną nie ma jeszcze żadnych badań w dziedzinie metod filtrujących. Nasze wstępne obliczenia pokazały, że wybór metody filtrującej ma znaczny wpływ na ostateczne wyniki klasyfikacji białek (stwierdzenie czy białko wykazuje daną cechę, czy nie). Badania zaczniemy porządkując wiedzę na temat istniejących metod filtrujących. W tej części pracy ziamplementujemy znalezione metody i przeprowadzimy symulację pokazującą ich działanie i wpływ na wyniki klasyfikacji tj. przyporządkowanie danego białka do odpowiedniej klasy (pozytywnej - wykazującej badaną cechę, albo negatywnej - nie wykazującej). Następnie zajmiemy się usprawnianiem metody SLOPE poprzez przyspieszenie jej za pomocą użycia algorytmu spadku po gradiencie, a także opracowanie dedykowanego solvera dla rzadkich danych. Ponadto opracujemy metodę przesiewową opartą o Hesjan dedykowaną dla metody SLOPE. Skupimy się także na logistycznej wersji adapcyjnego SLOPE, która pozwoli na założenie z góry wiedzy o pewnych k-merach (na przykład na podstawie wyników eksperymentalnych).

Wszystkie nowo opracowane metody zostaną zaimplementowane w środowisku R, a ich działanie będzie porównane z innymi istniejącymi metodami pod względem wyników klasyfikacji i czasu ewaluacji.