1. Motywacja

Aby białka pełniły właściwie swoje funkcje, muszę znajdować się w odpowiednim regionie komórki. „Adres” miejsca przeznaczenia białka znajduje się w obrębie jego sekwencji, w postaci tak zwanego sygnału kierującego. Do takich sygnałów należą peptydy sygnałowe, które są odpowiedzialne za kierowanie białek do systemu błon wewnątrz komórki i w konsekwencji za ich wydzielanie na zewnątrz. Białka wyposażone w peptydy sygnałowe stanowią bardzo duży udział wśród wszystkich białek w komórce i odgrywają istotną rolę w metabolizmie, tworzeniu struktur tkankowych, odpowiedzi odpornościowej i regulacji funkcji organizmu. Ponadto, przejście wielu białek przez system wewnątrzbłonowy jest niezbędne do przyjmowania przez nie właściwej struktury i różnorodnych modyfikacji, jak glikozylacja (dodawania reszt cukrowych) i fosforylacja. Peptydy sygnałowe występują na początku białka i mają długość zwykle 20-30 reszt aminokwasowych. Powszechny model typowych peptydów sygnałowych zakłada, że składają się z trzech części o różnym składzie aminokwasowym z czego szczególnie istotny jest region z dużą ilością aminokwasów hydrofobowych. Na końcu znajduje się miejsce odcinania peptydu od dojrzałego białka po wykonaniu przez niego funkcji.

Dotychczasowe programy rozpoznające peptydy sygnałowe były głównie oparte na sekwencjach najczęściej badanych gatunków jak drożdże, roślinny wyższe i ssaki. Jednakże nasze analizy dużej liczby sekwencji reprezentujących wiele różnorodnych grup organizmów gatunków wskazują na dużą zmienność peptydów sygnałowych. W chwili obecnej nie ma uniwersalnego modelu peptydu sygnałowego. Popularne programy przewidujące peptydy sygnałowe są z reguły oparte na metodach uczenia maszynowego, które uniemożliwiają badaczowi zapoznanie się z cechami peptydów służące do ich rozpoznawania. Ponadto rozwiązania tego rodzaju nie są w stanie oddać pełnego zróżnicowania peptydów sygnałowych i źle rozpoznają nietypowe przypadki. Nie dostarczają one również odpowiedzi na pytania dotyczące budowy i zmiennością międzygatunkowej peptydów sygnałowych, co utrudnia stosowanie ich do projektowania nowych peptydów.

2. Cel pracy

Dlatego celem naszego projektu jest stworzenie modelu statystycznego opisującego peptydy sygnałowe, który będzie uwzględniał wiedzę o ich organizacji, składzie aminokwasowym oraz zmienności między różnymi grupami taksonomicznymi i typami peptydów. Składający się z reguł decyzyjnych model będzie mógł być również stosowany do analizy i wykrywania peptydów sygnałowych oraz do i projektowania ich sztucznych odpowiedników.

3. Metoda badawcza

Model peptydu sygnałowego proponowany w tym projekcie będzie oparty na tzw. ukrytych modelach semi-Markowa (HSMMs). W najprostszym ujęciu ukryty model Markowa obliczy na podstawie składu aminokwasowego badanej sekwencji do jakiego regionu peptydu sygnałowego ona należy. Dużą zaletą tej metody jest możliwość uwzględniania wiedzy biologicznej. Informacje o układzie regionów, ich długości oraz właściwości fizykochemiczne tworzących je aminokwasów dodane do modelu wnoszą istotny wkład do tworzonych reguł decyzyjnych. Stworzony w ten sposób model jest bardzo elastyczny i skutecznie opisuje również nietypowe peptydy sygnałowe, które posiadają różne domeny funkcjonalne albo pochodzą z białek specyficznych grup taksonomicznych takich jak np. pasożyty. Dzięki tym właściwościom model oparty na HSMMs nie musi być ciągle douczany z powodu ciągłego wzrostu liczby sekwencji z peptydami sygnałowymi w bazach danych. Nowa metoda kodowania aminokwasów oparta na podobieństwach ich cech fizykochemicznych umożliwia wydobycie z sekwencji peptydu sygnałowego najbardziej istotnych informacji. Dzięki temu możliwe jest tworzenie modeli peptydu sygnałowego dla nielicznych i specyficznych grup białek.

Stworzony model zostanie sprawdzony czy skutecznie wykrywa peptydy sygnałowe. Aby dokonać rzetelnego porównania modeli planujemy określić obszary stosowalności badanych algorytmów i stworzyć meta-predyktora peptydów sygnałowych wykorzystujący wiedzę z wielu dostępnych programów.

3. Wstępne wyniki

Pierwszy ze stworzonych modeli nazwany roboczo signalHsmm wykazał najwyższą wartość współczynnika AUC=0,98 w porównaniu z innymi programami (ang. Area Under the Curve, oznacza w najprostszym ujęciu wyznacznik jakości przewidywania w skali od 0 do 1, gdzie 0 to całkowicie niewłaściwa predykcja, 0,5 losowa, a 1 idealna). Ponadto model ten efektywnie wykrywał peptydy sygnałowe nawet po trenowaniu na małych zbiorach sekwencji. Przykładowo, wersja programu wyuczona na skrajnie małym zbiorze danych z 1987 roku (336 sekwencji) zidentyfikowała tak samo skutecznie peptydy jak wersja wyuczonej na o wiele większym zbiorze danych (2311 sekwencji).

Skuteczność stworzonego modelu sprawdzono dla peptydów sygnałowych z różnych słabo poznanych i reprezentowanych w bazach danych grup taksonomicznych. Przykładowo, dla peptydów pasożyta zarodźca malarii i spokrewnionych rodzajów, AUC wynosi 0.92, istotnie lepiej niż w przypadku popularnych programów (0.84).

4. Wpływ rezultatów

Ostateczny model będzie zaimplementowanych w formie programu komputerowego i stanie się użyteczny zarówno dla biologów eksperymentalnych, jak i bioinformatyków chcących analizować peptydy sygnałowe. Dodatkowo, opracowany model może zostać użyty do przyśpieszenia procesu projektowania sztucznych peptydów sygnałowych w kosztownych i czasochłonnych badaniach laboratoryjnych. Nasz model peptydu sygnałowego będzie również opublikowany jako narzędzie do wykrywania peptydów sygnałowych.

Białka wyposażone w peptydy sygnałowe pełnią istotne funkcje w komórce, a zaburzenie ich głównej roli związanej z importem białek do siateczki wewnątrzplazmatycznej jest przyczyną wielu chorób. Proponowany algorytm wykrywający skutecznie takie peptydy sygnałowe pozwoli znacznie zmniejszyć koszty i przyśpieszyć opracowywanie odpowiednich leków zorientowanych na te peptydy. Przykładem mogą być białka sekrecyjne pasożytów takich jak zarodziec malarii, które nie są dostatecznie rozpoznawane przez istniejące oprogramowanie.