Przewidywanie właściwości sekwencji biologicznych w oparciu o analizę n-gramów

Mgr Michał Burdukiewicz

Przewidywanie właściwości białek i kwasów nukleinowych za pomocą metod *in silico* pozwala ograniczyć liczbę kosztownych i czasochłonnych eksperymentów. W tym celu stosuje się modele probabilistyczne, które łączą pierwszorzędową strukturę białka, czasem wzbogaconą o dodatkowe informacje, z daną cechą sekwencji (np. występowaniem miejsc modyfikacji potranslacyjnej, peptydów sygnałowych lub miejsc metylacji DNA). W mojej pracy opisuję opracowane przeze mnie metody analizy sekwencji, które zastosowałem w trzech problemach biologicznych. Moja metodologia opiera się głównie na analizie n-gramów (k-merów), czyli ciągów aminokwasów lub nukleotydów o długości n. W przypadku analiz białkowych stosuję również uproszczone alfabety, gdzie aminokwasy są grupowane ze względu na swoje podobieństwo.

Białka eukariotyczne zawierają informacje na temat swojej ostatecznej lokalizacji już w swojej sekwencji, w krótkich odcinkach nazywanych sygnałami sortującymi. Jednym z nich są peptydy sygnałowe odpowiedzialne za sekrecję białek. Najpopularniejsze programy do przewidywania peptydów sygnałowych nie jest w stanie skutecznie rozpoznawać peptydów o odmiennym składzie aminokwasowym. Takie peptydy występują np. w białkach wewnątrzkomórkowych pasożytów wywołujących malarię. Opracowane przeze mnie narzędzie, signalHsmm, identyfikuje typowe peptydy sygnałowe tak samo dobrze jak inne programy, a jednocześnie dużo lepiej radzi sobie z nietypowymi przypadkami.

Amyloidy to białka związane z wieloma schorzeniami klinicznymi, w tym chorobą Alzheimera i Creutzfeldta-Jakoba. Pomimo swojej różnorodności, wszystkie białka amyloidalne mogą ulegać agregacji inicjowanej przez krótkie fragmenty ich sekwencji. Aby znaleźć motywy decydujące o zdolności do agregacji, wytrenowałem predyktory amyloidogenności, wykorzystując n-gramy i

lasy losowe. Ponieważ amyloidogenność może nie zależeć od dokładnej sekwencji aminokwasów, ale od ich bardziej ogólnych właściwości, zidentyfikowałem uproszczony alfabet zapewniający najlepsze wyniki w walidacji krzyżowej. Predyktor oparty na tym alfabecie, zwany AmyloGramem, uzyskał najwyższa jakość predykcji w porównaniu z innymi programami (AUC: 0,90, MCC: 0,63).

Ważnym składnikiem ziemskiego ekosystemu są metanogeny, archea wytwarzająca metan. Mimo swojego dużego znaczenia, zarówno pod względem ekologicznym jak i gospodarczym, metanogeny są słabo poznane z powodu swoich bardzo różnorodnych warunków hodowli. Dlatego stworzyłem MethanoGram, który może być użyty do znacznego skrócenia czasu i obniżenia kosztów poszukiwania optymalnych warunków hodowli metanogenów poprzez przewidywanie ich na podstawie sekwencji 16S rRNA.

Opracowane przeze mnie programy są dostępne jako pakiety R i web serwery.