Piotr Bałut

Piotr Dubiel

Metody bioinformatyki: Projekt 7

Automatyczne odtwarzanie sekwencji DNA kodującej białko za pomocą ukrytych Modeli Markowa

Opis problemu

Sekwencja DNA w postaci ciągu kodonów w sposób jednoznaczny definiuje sekwencję aminokwasów. Przekształcenie odwrotne nie ma już tej własności i jest zależne od gatunku organizmu. Odtworzenie sekwencji DNA za pomocą obliczonych prawdopodobieństw wystąpienia danego kodonu jest najproestszym rozwiązaniem. Przykładowo lizyna jest kodowana przez dwa kodony: AAA i AAG. Załóżmy jednak, że dla danego organizmu prawdopodobieństwo wystąpienia kodonu AAG jest większe. W związku z tym lizyna zostaje zdekodowana jako AAG.

Kodon	**Aminokwas**	**Kodon**	**Aminokwas**	**Kodon**	**Aminokwas**	**Kodon**	**Aminokwas**
UUU	fenyloalanina	**UCU**	seryna	**UAU**	tyrozyna	**UGU**	cysteina
UUC	fenyloalanina	**UCC**	seryna	**UAC**	tyrozyna	**UGC**	cysteina
UUA	leucyna	**UCA**	seryna	**UAA**	**Ochre** (*Stop*)	**UGA**	**Opal** (*Stop*)
UUG	leucyna	**UCG**	seryna	**UAG**	**Amber** (*Stop*)	**UGG**	tryptofan
CUU	leucyna	**CCU**	prolina	**CAU**	histydyna	**CGU**	arginina
CUC	leucyna	**CCC**	prolina	**CAC**	histydyna	**CGC**	arginina
CUA	leucyna	**CCA**	prolina	**CAA**	glutamina	**CGA**	arginina
CUG	leucyna	**CCG**	prolina	**CAG**	glutamina	**CGG**	arginina
AUU	izoleucyna	**ACU**	treonina	**AAU**	asparagina	**AGU**	seryna
AUC	izoleucyna	**ACC**	treonina	**AAC**	asparagina	**AGC**	seryna
AUA	izoleucyna	**ACA**	treonina	**AAA**	lizyna	**AGA**	arginina
AUG	metionina (*Start*)	**ACG**	treonina	**AAG**	lizyna	**AGG**	arginina
GUU	walina	**GCU**	alanina	**GAU**	asparaginian	**GGU**	glicyna
GUC	walina	**GCC**	alanina	**GAC**	asparaginian	**GGC**	glicyna
GUA	walina	**GCA**	alanina	**GAA**	glutaminian	**GGA**	glicyna
GUG	walina	**GCG**	alanina	**GAG**	glutaminian	**GGG**	glicyna

Jest to model Markowa zerowego rzędu, czyli taki, w którym kolejne stany procesu nie zależą od poprzednich stanów. Wprowadzenie modelu Markowa pierwszego rzędu powinno poprawić działanie algorytmu przez wzięcie pod uwagę zależności między kodonami. Jego działanie powinno być szczególnie widoczne przy kodonach o zbliżonych prawdopodobieństwach. Dla modelu Markowa zerowego rzędu zawsze bardziej prawdopodobny kodon będzie wybierany, dla modelu pierwszego rzędu mamy szansę na skorygowanie tego założenia o prawdopodobieństwo

wystąpienia po wcześniejszym kodonie w sekwencji.

Sposób realizacji problemu

Na potrzeby projektu przyjęliśmy model Markowa pierwszego rzędu. W tym modelu stanami modelu są kodony, zaś obserwacjami aminokwasy. Przyjęliśmy następujące parametry modelu Markowa:

- stany początkowe są równoprawdopodbne,
- prawdopodobieństwa przejść między stanami są unikalne dla każdego organizmu i będą otrzymywane na podstawie sekwencji DNA lub RNA tego organizmu użytej do nauki modelu
- prawdopodobieństwo emisji obserwacji otrzymywane jest na podstawie z góry ustalonej, stałej tabeli kodonów, która jednoznacznie definiuje białko otrzymywane z kodonu

Dekodowanie sekwencji aminokwasów zostanie wykonane za pomocą algorytmu Viterbiego. Określa on iteracyjnie dla każdej obserwacji z sekwencji stan procesu przez znalezienie najbardziej prawdopodobnej ścieżki od poprzedniego stanu.

Projekt wykonany zostanie w języku Python i będzie składał się z dwóch aplikacjikonsolowych. Pierwsza z nich będzie stanowić program uczący, który na podstawie zadanej sekwencji DNA dla określonego organizmu wygeneruje parametry modelu Markowa. Druga aplikacja będzie służyła do przewidywania sekwencji DNA oraz RNA na podstawie zadanej sekwencji aminokwasów, należącej do białka pochodzącego z określonego organizmu, na podstawie odnalezionych podczas nauki parametrów modelu Markowa.

W trybie nauki program będzie przyjmował nazwę pliku zawierającego zapisaną w formacie *plain* sekwencję DNA lub RNA organizmu, oraz nazwę pliku wyjściowego, w którym zachowane zostaną zapisane w formacie JSON otrzymane parametry modelu Markowa. Na podstawie podanej do programu sekwencji będzie on automatycznie określał, czy jest to sekwencja DNA czy RNA. Wygenerowane parametry modelu będą później służyły do przewidywania sekwencji takiego samego typu.

W trybie poszukiwania sekwencji program będzie przyjmował dwa pliki. Pierwszy z nich zawierać będzie parametry znalezione wcześniej paramtery modelu Markowa, drugi zaś zapisaną w formacie *plain* sekwencję aminokwasów dla której program będzie poszukiwał sekwencji DNA. Trzecim parametrem będzie plik wyjściowy do którego program zapisze znaleziona sekwencję.