Predykcja lokalizacji subkomórkowej

Michał Burdukiewicz, Piotr Sobczyk

1. Przygotowanie danych
   - Sekwencje pochodzenia eukariotycznego
   - Sekwencje pochodzenia prokariotycznego
2. Przyjęty model
   - Wyszukiwanie regionów n, h i c
   - (
3. Zaproponowany algorytm

Do stworzenia pierwszej wersji programu SignalP użyto bazy Swiss-Prot 29 (czerwiec 1994) zawierającej 38303 białek.

Zapytanie

created:[1950 TO 1995] AND reviewed:yes

Powyższe zapytanie znajduje 38440 białek w bazie UniProt.
Do uczenia programu SignalP wykorzystano tylko sekwencje z jednym, znanym miejscem cięcia, dlatego ze zbioru znalezionych rekordów usunięto te, w których informacja o peptydzie sygnałowym zawierała symbole: '<1', '?' oraz alternatywne miejsca cięcia.

Baza Swiss-Prot 29 po oczyszczeniu zawierała 2282 białek eukariotycznych zawierających sekwencje sygnałowe sekrecji, których obecność potwierdzono eksperymentalnie.

Poniższe zapytanie dalszym oczyszczeniu pozwala otrzymać 2382 białek:

Zapytanie - eukarionty

select: (keyword:signal) AND reviewed:yes AND created:[1950 TO 1995] AND taxonomy:'Eukaryota [2759]' AND annotation:(type:signal confidence:experimental)

Baza Swiss-Prot 29 po oczyszczeniu zawierała 579 białek prokariotycznych zawierających sekwencje sygnałowe sekrecji, których obecność potwierdzono eksperymentalnie. Ze zbioru uczącego wykluczono sekwencje odcinane przez peptydazę sygnałową II. Poniższe zapytanie dalszym oczyszczeniu pozwala otrzymać 603 białka:

Zapytanie - prokarionty

select: (keyword:signal) AND reviewed:yes AND created:[1950 TO 1995] AND annotation:(type:signal confidence:experimental) NOT keyword:'Lipoprotein [KW-0449]' AND (taxonomy:'Bacteria [2]' OR taxonomy:'Archaea [2157]')

NIE usunięto sekwencji redundatnych, ponieważ nie było jasne w jaki sposób dokonano tego podczas przygotowania zbioru uczącego dla programu SignalP.

- n-region: 1-5 aminokwasów, preferowane aminokwasy z dodatnim ładunkiem.
- h-region: ***0*** - ***20*** aminokwasów, preferowane aminokwasy hydrofobowe.
- c-region: 3 - 7 polarnych aminokwasów bez ładunku.
- miejsce cięcia: na pozycjach -3 i -1 w stosunku do miejsca cięcia małe, obojętne aminokwasy.

Wyznaczenie c-regionu:

- z pozycji -1 przesunąć się na pozycję -3,
- przesuwać się do N-końca do momentu napotkania bloku co najmniej dwóch hydrofobowych aminokwasów.

Wyznaczenie h-regionu:

- przenieść się 6 aminokwasów w stronę N-końca białka od 5'-końca c-regionu, *jeśli pozycja osiągnięta w ten sposób jest mniejsza niż 1, to przenieść się do pierwszego aminokwasu i przejść do kroku 3.*
- przesuwać się do N-końca do momentu napotkania aminokwasu z ładunkiem lub bloku co najmniej trzech niehydrofobowych aminokwasów *lub do momentu napotkania pierwszego aminokwasu.*
- jeśli N-koniec h-regionu nie jest hydrofobowym aminokwasem, przesuwać się w stronę 3'-końca do momentu napotkania hydrofobowego aminokwasu **lub początku c-regionu**.

Krok 1

przenieść się 6 aminokwasów w stronę N-końca białka od 5'-końca c-regionu, *jeśli pozycja osiągnięta w ten sposób jest mniejsza niż 1, to przenieść się do pierwszego aminokwasu i przejść do kroku 3*

Niekiedy początek c-regionu wyznaczany jest w okolicy pierwszego aminokwasu, wtedy założenie o minimalnej długości 6 aminokwasów h-regionu nie jest spełnione.

Krok 2

przesuwać się do N-końca do momentu napotkania aminokwasu z ładunkiem lub bloku co najmniej trzech niehydrofobowych aminokwasów *lub do momentu napotkania pierwszego aminokwasu*.

h-region może być tak krótki, że nie ma szansy na spełnienie innego warunku kończącego niż napotkanie pierwszego aminokwasu.

Krok 3

jeśli N-koniec h-regionu nie jest hydrofobowym aminokwasem, przesuwać się w stronę 3'-końca do momentu napotkania hydrofobowego aminokwasu **lub początku c-regionu**.

Bez dodatkowego warunku początek h-regionu może się znajdować za początkiem c-regionu.

Zgodność)

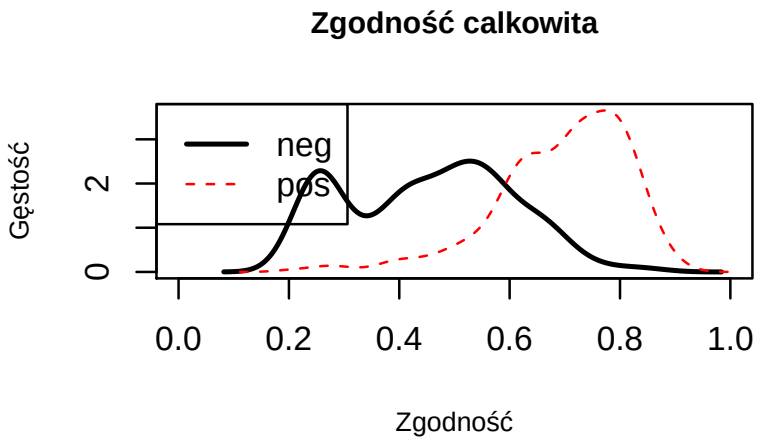
$$z = \sum_{i=1}^{l} I(k_i = l_i) \tag{1}$$

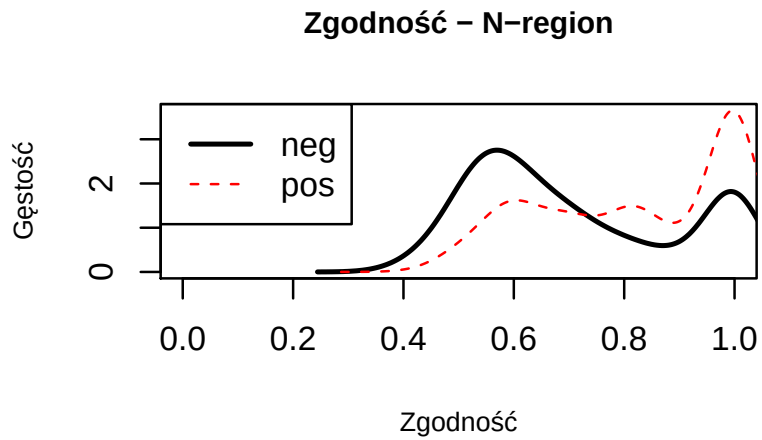
,

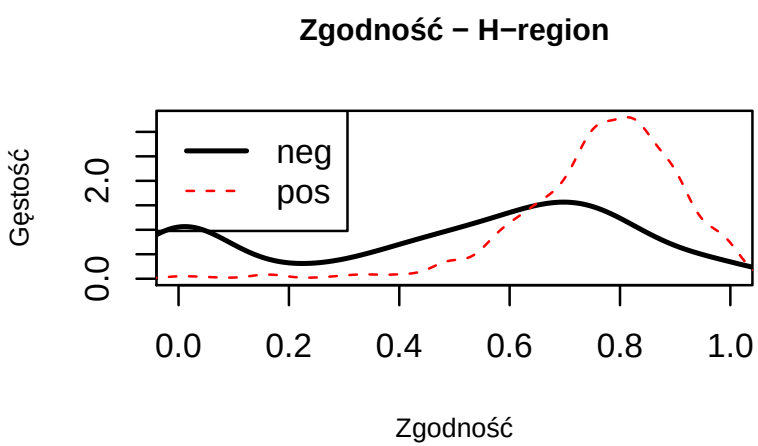
gdzie: k

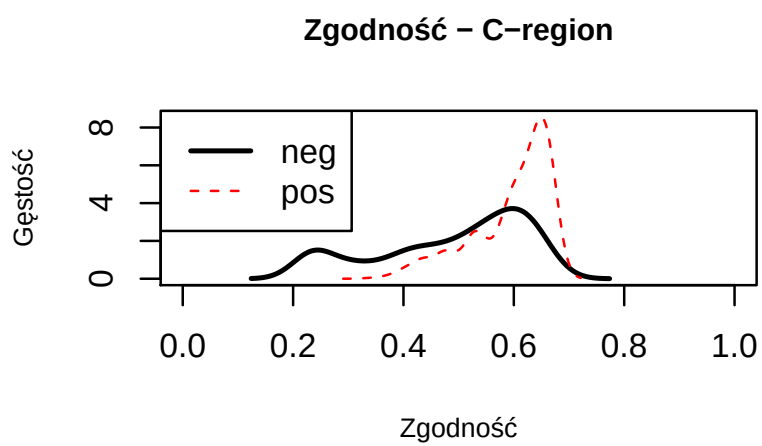
- k - predykcja przynależności regionalnej z modelu teoretycznego,
- l - predykcja przynależności regionalnej z ukrytego modelu markowa.

Poprawić te rysunki

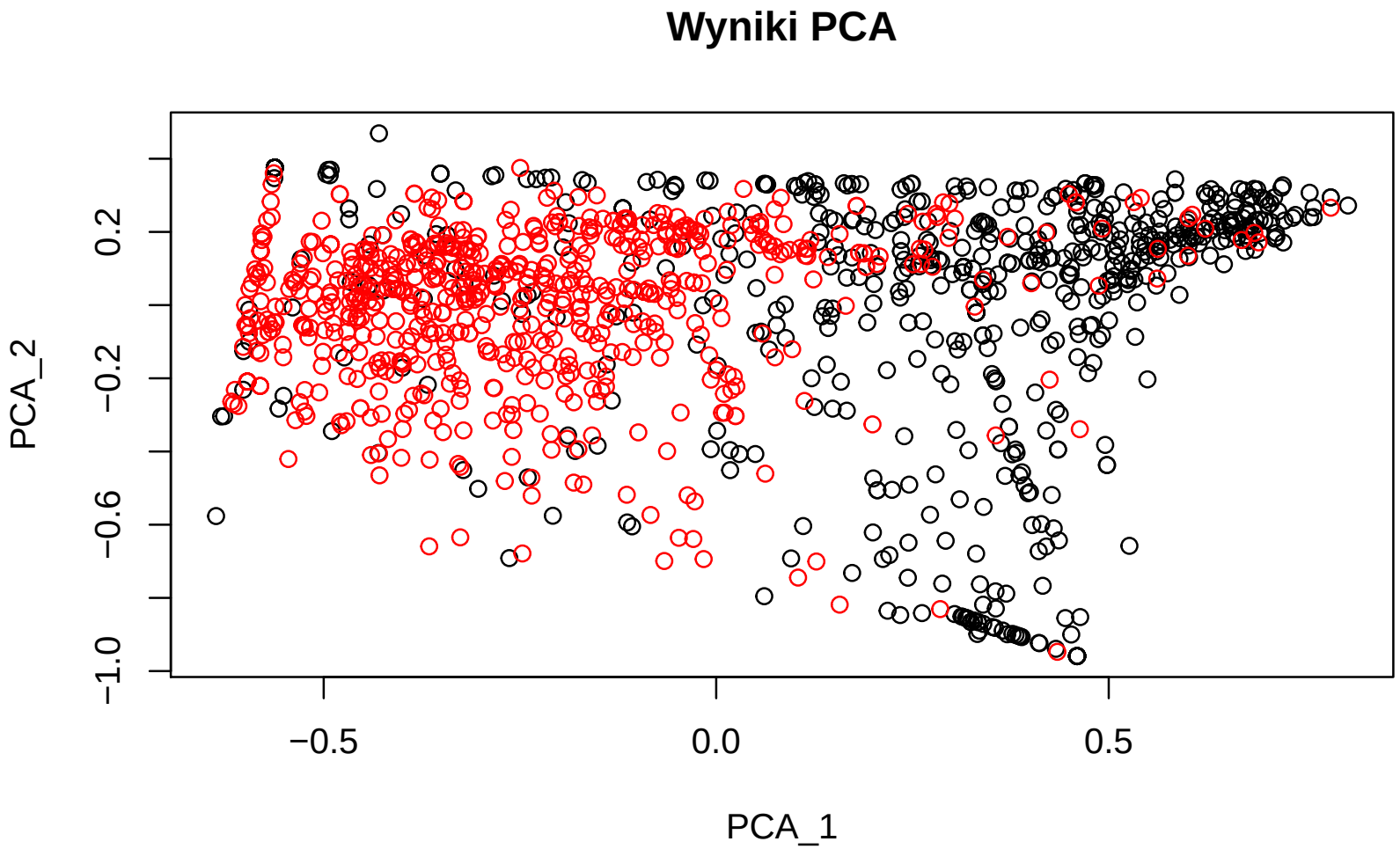




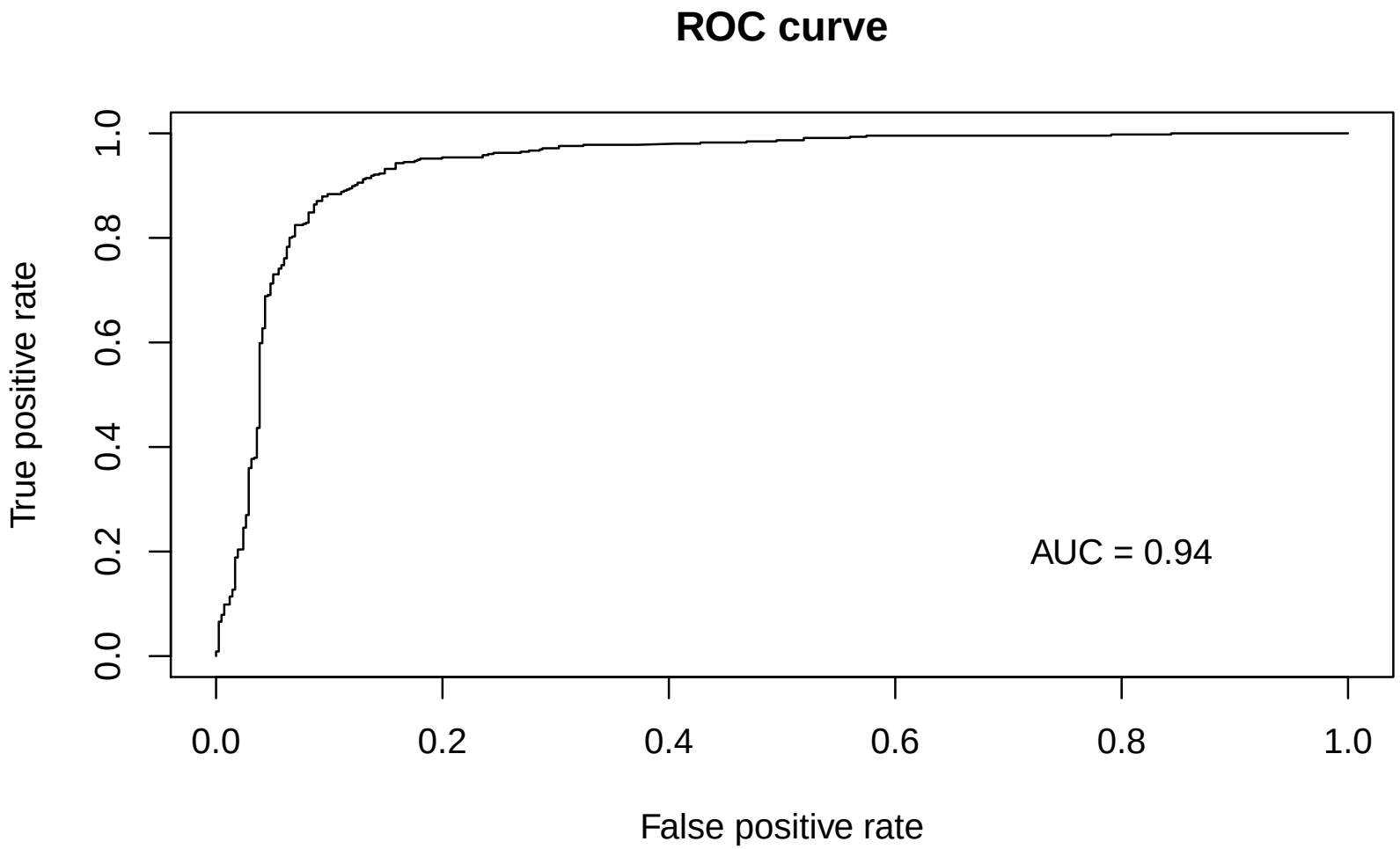




	Comp.1	Comp.2	Comp.3	Comp.4
Standard deviation	0.39	0.29	0.17	0.07
Proportion of Variance	0.57	0.31	0.10	0.02
Cumulative Proportion	0.57	0.88	0.98	1.00


Wyniki PCA
PCA_2
PCA_1
−1.0
−0.6
−0.2
0.2
−0.5
0.0
0.5

- na podstawie zbioru uczącego policzono częstości występowania aminokwasów w różnych regionach peptydu sygnałowego,
- dla każdego białka w zbiorze walidacyjnych dla każdego potencjalnego miejsca cięcia policzono zgodności,
- na podstawie uzyskanych wyników nauczono SVM.


ROC curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.94