

# n-gramy w analizie sekwencji biologicznych

Michał Burdukiewicz<sup>1</sup>, Piotr Sobczyk<sup>2</sup>, Paweł Mackiewicz<sup>1</sup>

<sup>1</sup>Zakład Genomiki, Uniwersytet Wrocławski

<sup>2</sup>Katedra Matematyki, Politechnika Wrocławska

# Outline

- 1 n-gramy (k-mery)
  - n-gramy (k-mery)
  - Informacja o pozycji
  - Nieciągłe n-gramy
  - Wybór informatywnych n-gramów - QuiPT
- 2 signalHSMM
  - Peptydy sygnałowe
  - Ukryte modele semi-Markowskie
  - Porównanie z innymi programami
  - Dodanie informacja o k-merach
- 3 Amyloidy
  - Zbiór danych
  - Wybór n-gramów
  - Predykcja w oparciu o n-gramy
  - Podsumowanie

## Definicja

n-gramy (k-mery, k-tuple) to wektory o długości  $n$  zawierające znaki z sekwencji wejściowych.

Pierwotnie analiza n-gramów rozwijana była na potrzeby analizy języka naturalnego, ale ma również zastosowania w genomice (Fang et al., 2011), transkryptomice (Wang et al., 2014) i proteomice (Guo et al., 2014).

	P1	P2	P3	P4	P5	P6
S1	C	T	T	A	G	C
S2	C	A	G	A	C	G
S3	G	T	G	A	T	T

Przykładowe sekwencje. S - sekwencje, P - pozycja nukleotydu.

	A	C	G	T
S1	1	2	1	2
S2	2	2	2	0
S3	1	0	2	3

### Zliczenia 1-gramów.

	P1	P2	P3	P4	P5	P6
S1	C	T	T	A	G	C
S2	C	A	G	A	C	G
S3	G	T	G	A	T	T

Przykładowe sekwencje. S - sekwencje, P - pozycja nukleotydu.

	AA	CA	GA	TA	AC	CC	GC	TC
S1	0	0	0	1	0	0	1	0
S2	0	1	1	0	1	0	0	0
S3	0	0	1	0	0	0	0	0

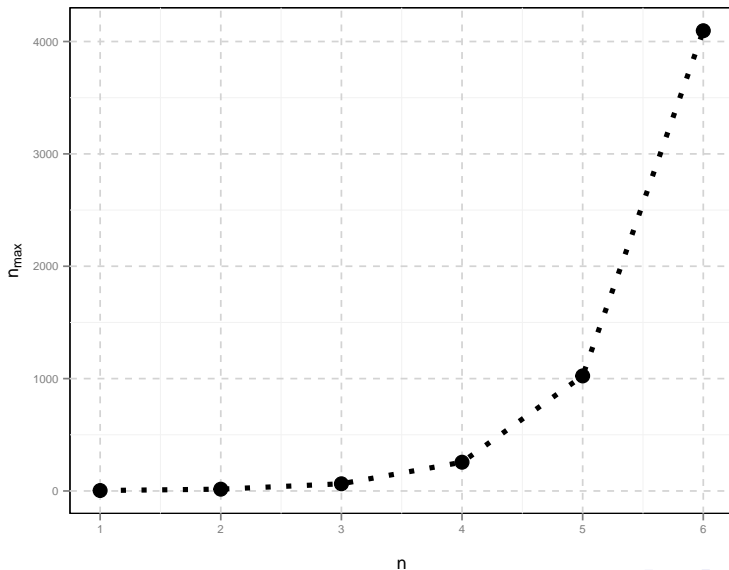
Zliczenia 2-gramów (fragment tabeli).

$$n_{\max} = u^n$$

$n_{\max}$ : liczba wszystkich możliwych n-gramów

$u$ : liczba liter w alfabecie.

$n$ : długość n-gramu



n-gramy mogą być przypisaną informację o pozycjach na których występują.



	P1	P2	P3	P4	P5	P6
S1	C	T	T	A	G	C
S2	C	A	G	A	C	G
S3	G	T	G	A	T	T

Przykładowe sekwencje. S - sekwencje, P - pozycja nukleotydu.

	1_A.A	2_A.A	3_A.A	4_A.A	5_A.A	1_C.A	2_C.A	3_C.A
S1	0	0	0	0	0	0	0	0
S2	0	0	0	0	0	1	0	0
S3	0	0	0	0	0	0	0	0

Zliczenia 2-gramów z informacją o pozycji (fragment tabeli).

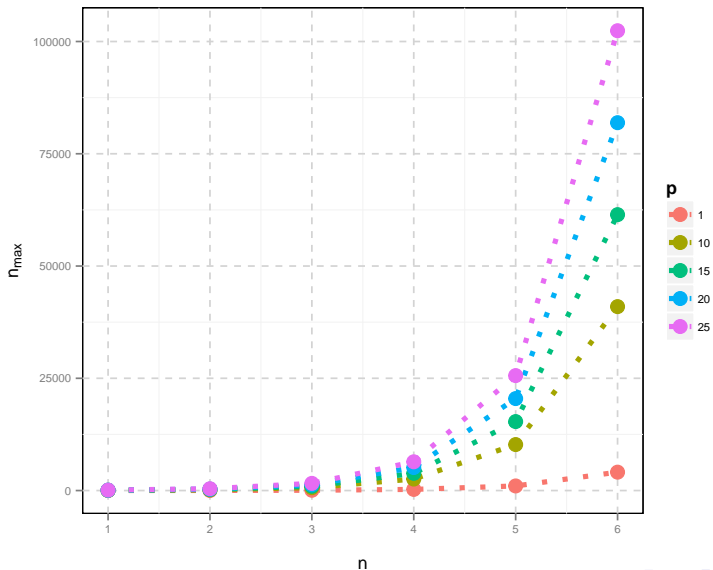
$$n_{\max} = p \times u^n$$

$n_{\max}$ : liczba wszystkich możliwych n-gramów

$p$ : liczba możliwych pozycji.

$u$ : liczba liter w alfabecie.

$n$ : długość n-gramu



n-gramy mogą być nieciągłe - pomiędzy elementami n-gramu mogą występować przerwy.

	P1	P2	P3	P4	P5	P6
S1	C	T	T	A	G	C
S2	C	A	G	A	C	G
S3	G	T	G	A	T	T

Przykładowe sekwencje. S - sekwencje, P - pozycja nukleotydu.

	A_A	C_A	G_A	T_A	A_C	C_C	G_C	T_C
S1	0	0	0	1	1	0	0	0
S2	1	0	0	0	0	0	1	0
S3	0	0	0	1	0	0	0	0

Zliczenia 2-gramów z przerwą 1 (fragment tabeli).

- Przed budową właściwego modelu, np. lasu losowego lub svm, należy dokonać wstępnej redukcji liczby zmiennych

- Przed budową właściwego modelu, np. lasu losowego lub svm, należy dokonać wstępnej redukcji liczby zmiennych
- Wielowymiarowa przestrzeń atrybutów filtrowana za pomocą testu statystycznego

- Przed budową właściwego modelu, np. lasu losowego lub svm, należy dokonać wstępnej redukcji liczby zmiennych
- Wielowymiarowa przestrzeń atrybutów filtrowana za pomocą testu statystycznego
- W tym celu stosujemy QuiPT (**Q**uick **P**ermutation **T**est)



W teście permutacyjnym oznaczenia klas są losowo mieszane na potrzeby obliczania statystyki testowej.

$$\text{p-value} = \frac{N_{T_P > T_R}}{N},$$

gdzie  $N_{T_P > T_R}$  to liczba losowań, kiedy  $T_P$  (permutowana statystyka testowa) miała wartość krytyczniejszą niż  $T_R$  (statystyka testowa dla niespermutowanych danych).

W teście permutacyjnym oznaczenia klas są losowo mieszane na potrzeby obliczania statystyki testowej.

$$\text{p-value} = \frac{N_{T_P > T_R}}{N},$$

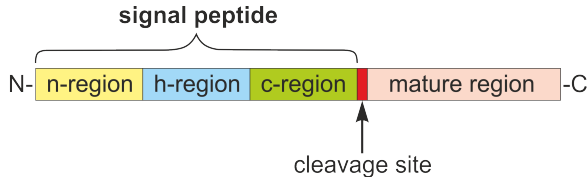
gdzie  $N_{T_P > T_R}$  to liczba losowań, kiedy  $T_P$  (permutowana statystyka testowa) miała wartość krytyczniejszą niż  $T_R$  (statystyka testowa dla niespermutowanych danych).

## QuiPT

Powyższy test permutacyjny jest równoważny dwustronnemu dokładnemu testowi Fishera. Ponieważ nasze wektory związane z n-gramami są bardzo rzadkie (wiele zer), to uzyskujemy bardzo istotne przyspieszenie i poprawę jakości obliczanych p-wartości.

# Outline

- 1 n-gramy (k-mery)
  - n-gramy (k-mery)
  - Informacja o pozycji
  - Nieciągłe n-gramy
  - Wybór informatywnych n-gramów - QuiPT
- 2 signalHSMM
  - Peptydy sygnałowe
  - Ukryte modele semi-Markowskie
  - Porównanie z innymi programami
  - Dodanie informacja o k-merach
- 3 Amyloidy
  - Zbiór danych
  - Wybór n-gramów
  - Predykcja w oparciu o n-gramy
  - Podsumowanie



- n-region: głównie zasadowe aminokwasy (Nielsen and Krogh, 1998),
- h-region: silnie hydrofobowe reszty aminokwasy (Nielsen and Krogh, 1998),
- c-region: kilka polarnych aminokwasów bez ładunku (Jain et al., 1994).

Istnieje szereg programów przewidujących występowanie peptydu sygnałowego:

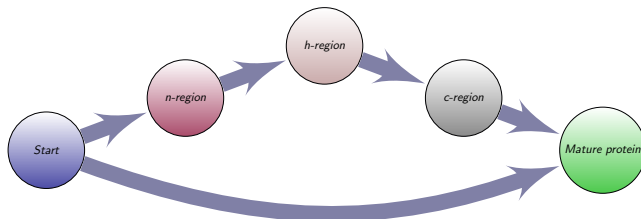
- signalP 4.1 (sieci neuronowe) (Petersen et al., 2011),
- PrediSi (Position Weight Matrix) (Hiller et al., 2004),
- Signal-3L (k-najbliższych sąsiadów) (Shen and Chou, 2007),
- Phobius (ukryte modele Markowskie) (Käll et al., 2004).

## Założenia modelu:

- obserwowany rozkład aminokasów jest wynikiem przebywania w określonym regionie (stanie),
- długość regionu (czas trwania stanu) jest modelowana poprzez rozkład prawdopodobieństwa (inny niż rozkład geometryczny jak w ukrytych modelach Markowskich).

1. Pozyskanie eukariotycznych białek z bazy UniProtKB 2014\_07 (po ocyszczeniu z nietypowych lub niedokładnie opisanych rekordów zbior danych liczy 3816 białek z peptydem sygnałowym i 9795 białek bez peptydu sygnałowego),
2. określenie granic n-, h- i c-regionów przez algorytm heurystyczny,
3. redukcja wymiarowości problemu poprzez zagregowanie aminokwasów na podstawie ich właściwości fizykochemicznych do kilku grup,
4. obliczenie częstości występowania grup aminokwasowych w danych regionie oraz długości regionów,
5. uczenie dwóch HSMM dla białek z peptydem sygnałowym i bez peptydu sygnałowego.

Podczas testowania, każde białko jest dopasowane do dwóch HSMM, które modelują odpowiednio białka bez peptydu sygnałowego i z peptydem sygnałowym. Prawdopodobieństwo obu dopasowań stanowią wynik działania programu.



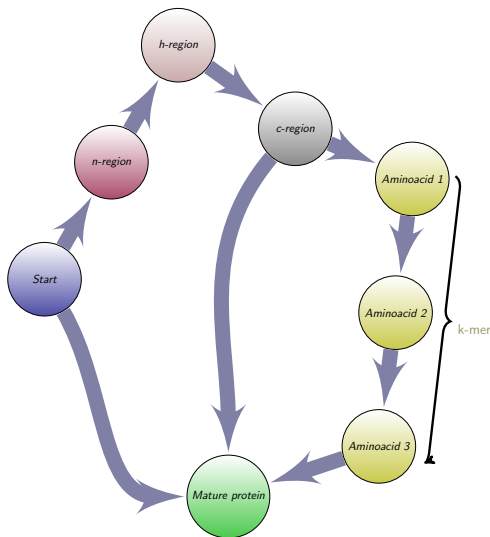


Zbiór danych do analizy porównawczej: 140 eukariotycznych białek z peptydem sygnałowym i 280 losowo wybranych eukariotycznych białek bez peptydu sygnałowego dodanych po 2010 do bazy UniProt.

signal.hsmm1987: wytrenowany na zbiorze 496 eukariotycznych białek z peptydem sygnałowym dodanych do bazy przed 1987.

signal.hsmm2010: wytrenowany na zbiorze 3676 eukariotycznych białek z peptydem sygnałowym dodanych do bazy przed 2010.

	AUC	H-measure	Gini index
Phobius	0.9643	0.8844	0.9286
PrediSi	0.9411	0.8238	0.8821
signalP 4.1 (no tm)	0.9679	0.8909	0.9357
signalP 4.1 (tm)	0.9750	0.9261	0.9500
signalhsmm2010	0.9893	0.8963	0.9786
signalhsmm1987	0.9889	0.8994	0.9778



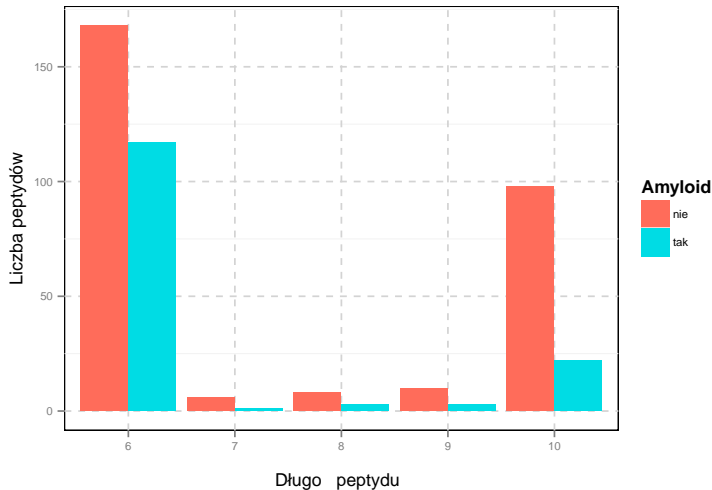
Aby polepszyć precyzję predykcji miejsca cięcia planujemy wykorzystać k-mery.

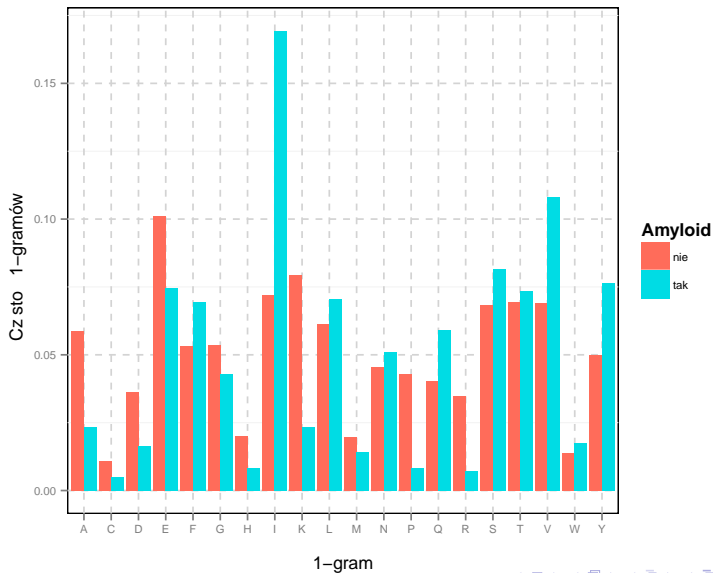
- Utworzone zostaną spozycjonowane 4-mery z dystansami
- **biogram** wybierze istotne k-mery
- Zostaną one dodane do modelu jak na schemacie obok

# Outline

- 1 n-gramy (k-mery)
  - n-gramy (k-mery)
  - Informacja o pozycji
  - Nieciągłe n-gramy
  - Wybór informatywnych n-gramów - QuiPT
- 2 signalHSMM
  - Peptydy sygnałowe
  - Ukryte modele semi-Markowskie
  - Porównanie z innymi programami
  - Dodanie informacja o k-merach
- 3 Amyloidy
  - Zbiór danych
  - Wybór n-gramów
  - Predykcja w oparciu o n-gramy
  - Podsumowanie

Zbiór danych: 146 amyloidów i 290 nieamyloidów (Gasior and Kotulska, 2014).





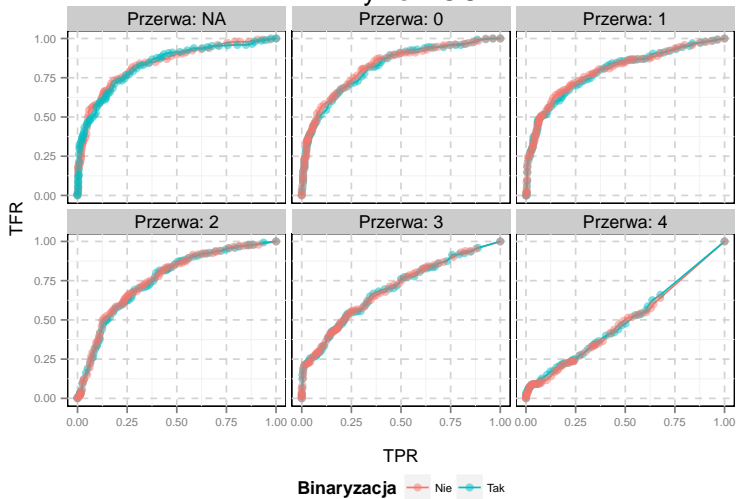
- Duże różnice w częstościach aminokwasów sugerują użycie n-gramów.
- Nierówne długości peptydów wykluczają n-gramy z informacją o pozycji.

Czy można zbinaryzować zliczenia 1- i 2-gramów bez utraty zbyt dużej ilości informacji?

W 5-krotnej walidacji krzyżowej porównano 6 klasyfikatorów (lasy losowe) dla 1-gramów i 2-gramów (rozważane przerwy: 0 - 4) dla zliczeń zwykłych i zbinaryzowanych.



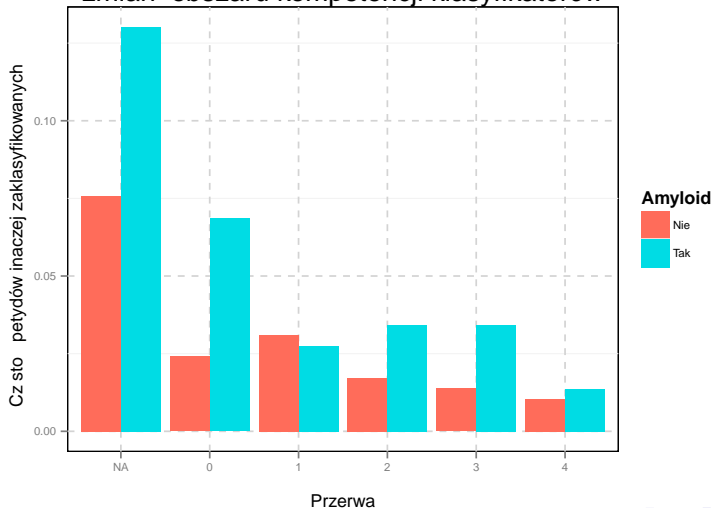
## Krzywe ROC

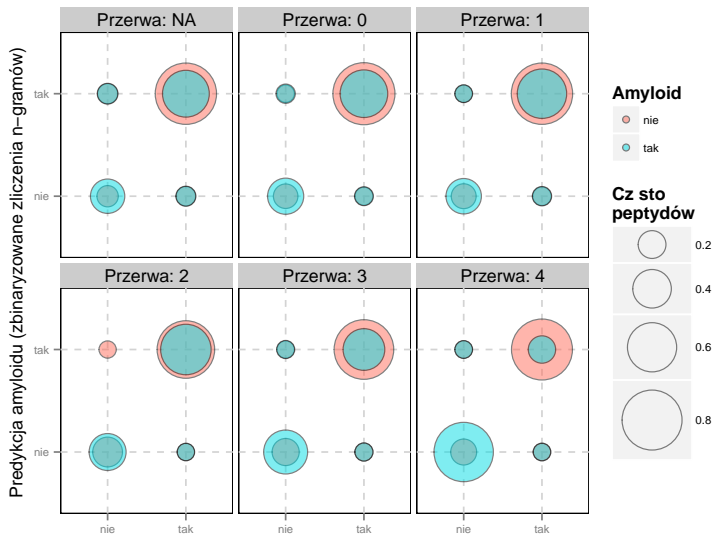


n	Przerwa	Binaryzacja	AUC	TNR	TPR
1	NA	nie	0.8351	0.8828	0.6094
1	NA	tak	0.8309	0.8621	0.6172
2	0	nie	0.8182	0.8862	0.6030
2	0	tak	0.8162	0.8690	0.6168
2	1	nie	0.7894	0.8276	0.6168
2	1	tak	0.7863	0.8310	0.6303
2	2	nie	0.7617	0.7966	0.5685
2	2	tak	0.7629	0.8000	0.5892
2	3	nie	0.7054	0.7966	0.4717
2	3	tak	0.6979	0.7897	0.4513
2	4	nie	0.4904	0.8345	0.1851
2	4	tak	0.4824	0.8241	0.1989

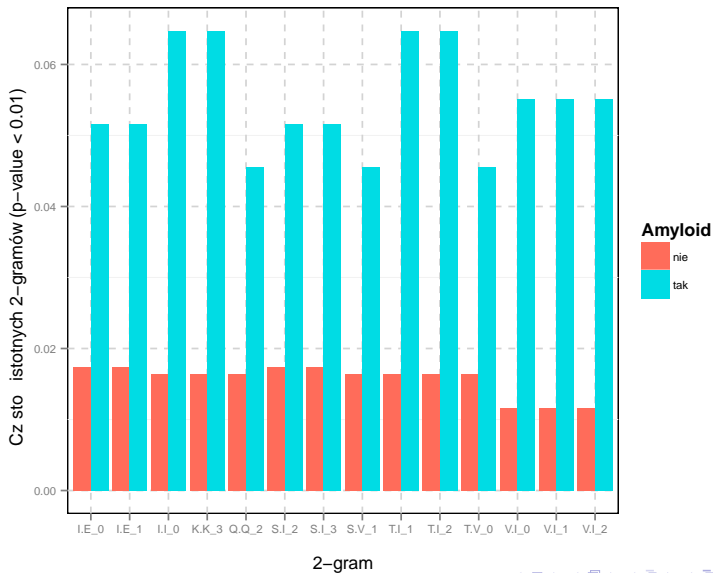
Ocena predykcji klasyfikatorów.

## Wpływ binaryzacji danych na zmian obszaru kompetencji klasyfikatorów





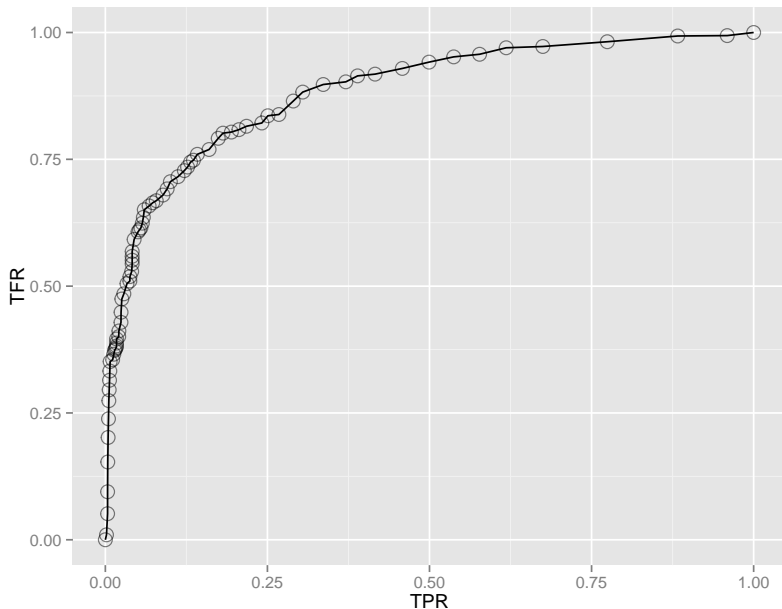
Predykcja amyloidu (zliczenia n-gramów)



- 1 W oparciu o wyniki poprzedniej analizy zliczono następujące n-gramy: 1-gramy, 2-gramy (przerwy od 0 do 3) i 3-gramy (przerwy od 0 do 1).
- 2 Wszystkie n-gramy (oprócz 1-gramów) zbinaryzowano.
- 3 Metodą QuiPT wybrano **214** z **25600** potencjalnych n-gramów (**0.0083**).
- 4 Wszystkie 1-gramy i wybrane 2 oraz 3-gramy użyto podczas 5-krotnej walidacji krzyżowej.

Iteracja	AUC	TPR	TFR
1	0.92	0.98	0.59
2	0.88	0.95	0.67
3	0.86	0.93	0.66
4	0.84	0.90	0.66
5	0.89	0.97	0.55

Predykcja ostatecznego klasyfikatora. Średnie AUC: **0.8798**. Średnie TNR: **0.9448**. Średnie TPR: **0.6230**.





- biogram: <http://github.com/michbur/biogram>
- QuiPT: <http://github.com/michbur/QuiPT>
- signalHSMM: <http://github.com/michbur/signalHSMM>

- Fang, Y.-C., Lai, P.-T., Dai, H.-J., and Hsu, W.-L. (2011). MeinfoText 2.0: gene methylation and cancer relation extraction from biomedical literature. *BMC Bioinformatics*, 12(1):471.
- Gasior, P. and Kotulska, M. (2014). Fish amyloid - a new method for finding amyloidogenic segments in proteins based on site specific co-occurrence of aminoacids. *BMC Bioinformatics*, 15(1):54.
- Guo, S.-H., Deng, E.-Z., Xu, L.-Q., Ding, H., Lin, H., Chen, W., and Chou, K.-C. (2014). inuc-pseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics*, 30(11):1522–1529.
- Hiller, K., Grote, A., Scheer, M., Münch, R., and Jahn, D. (2004). PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Research*, 32(suppl 2):W375–W379.
- Jain, R. G., Rusch, S. L., and Kendall, D. A. (1994). Signal peptide cleavage regions. functional limits on length and topological implications. *The Journal of Biological Chemistry*, 269(23):16305–16310.
- Käll, L., Krogh, A., and Sonnhammer, E. L. L. (2004). A combined transmembrane topology and signal peptide prediction method. *Journal of Molecular Biology*, 338(5):1027–1036.
- Nielsen, H. and Krogh, A. (1998). Prediction of signal peptides and signal anchors by a hidden markov model. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, 6:122–130.
- Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*, 8(10):785–786.
- Shen, H.-B. and Chou, K.-C. (2007). Signal-3L: A 3-layer approach for predicting signal peptides. *Biochemical and Biophysical Research Communications*, 363(2):297–303.
- Wang, Y., Liu, L., Chen, L., Chen, T., and Sun, F. (2014). Comparison of metatranscriptomic samples based on k-tuple frequencies. *PLoS ONE*, 9(1):e84348.