n-gramy w analizie sekwencji biologicznych

Michał Burdukiewicz[1], Piotr Sobczyk[2], Paweł Mackiewicz[1]

[1]Zakład Genomiki, Uniwersytet Wrocławski

[2]Instytut Matematyki i Informatyki, Politechnika Wrocławska

Outline

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

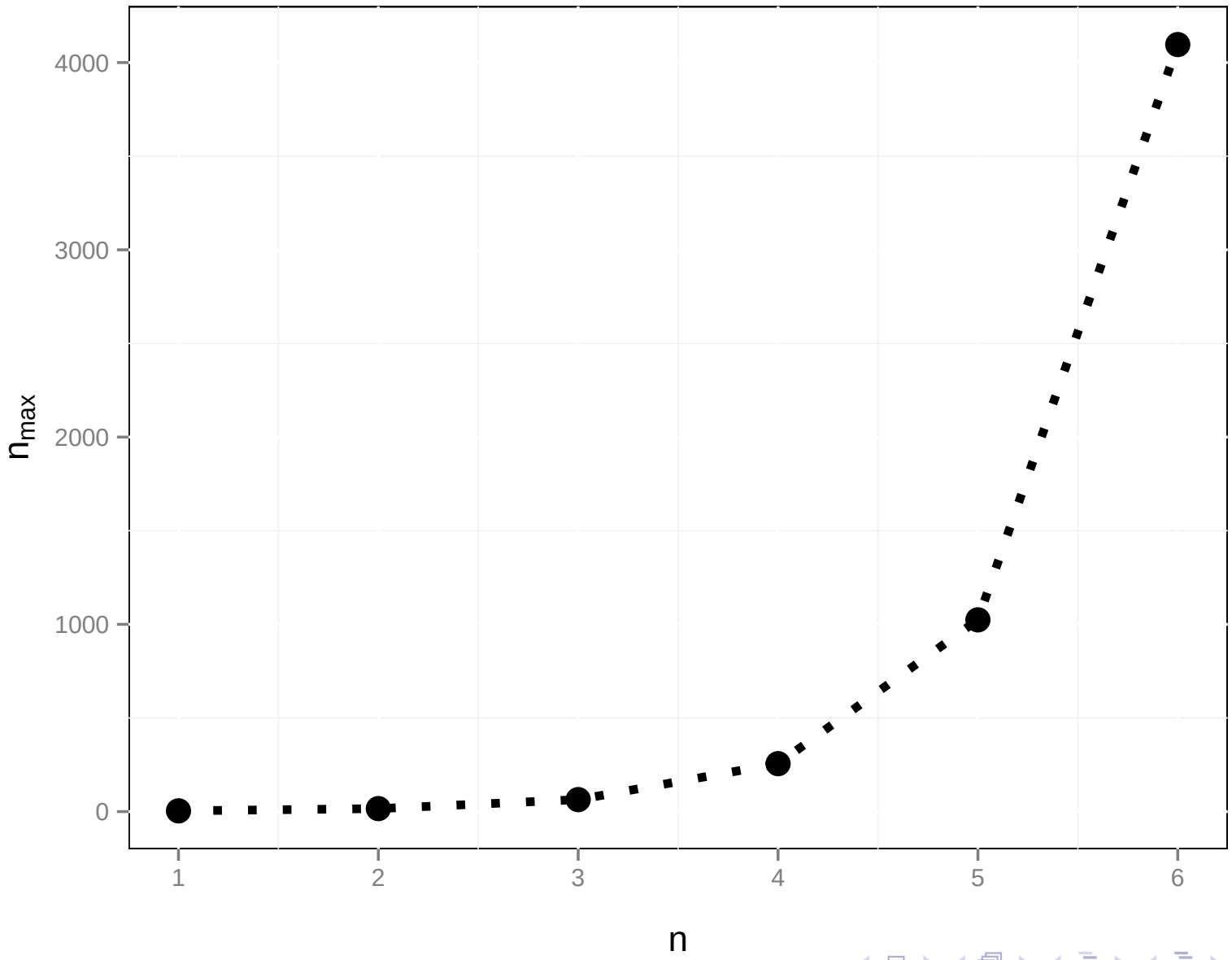
Przykładowe sekwencje. S - sekwencje, P - pozycja nukleotydu.

	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).

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

n_{max}: liczba wszystkich możliwych n-gramów
u: liczba liter w alfabecie.
n: długość n-gramu


4000
3000
2000
1000
0
n_{max}
1
2
3
4
5
6
n

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

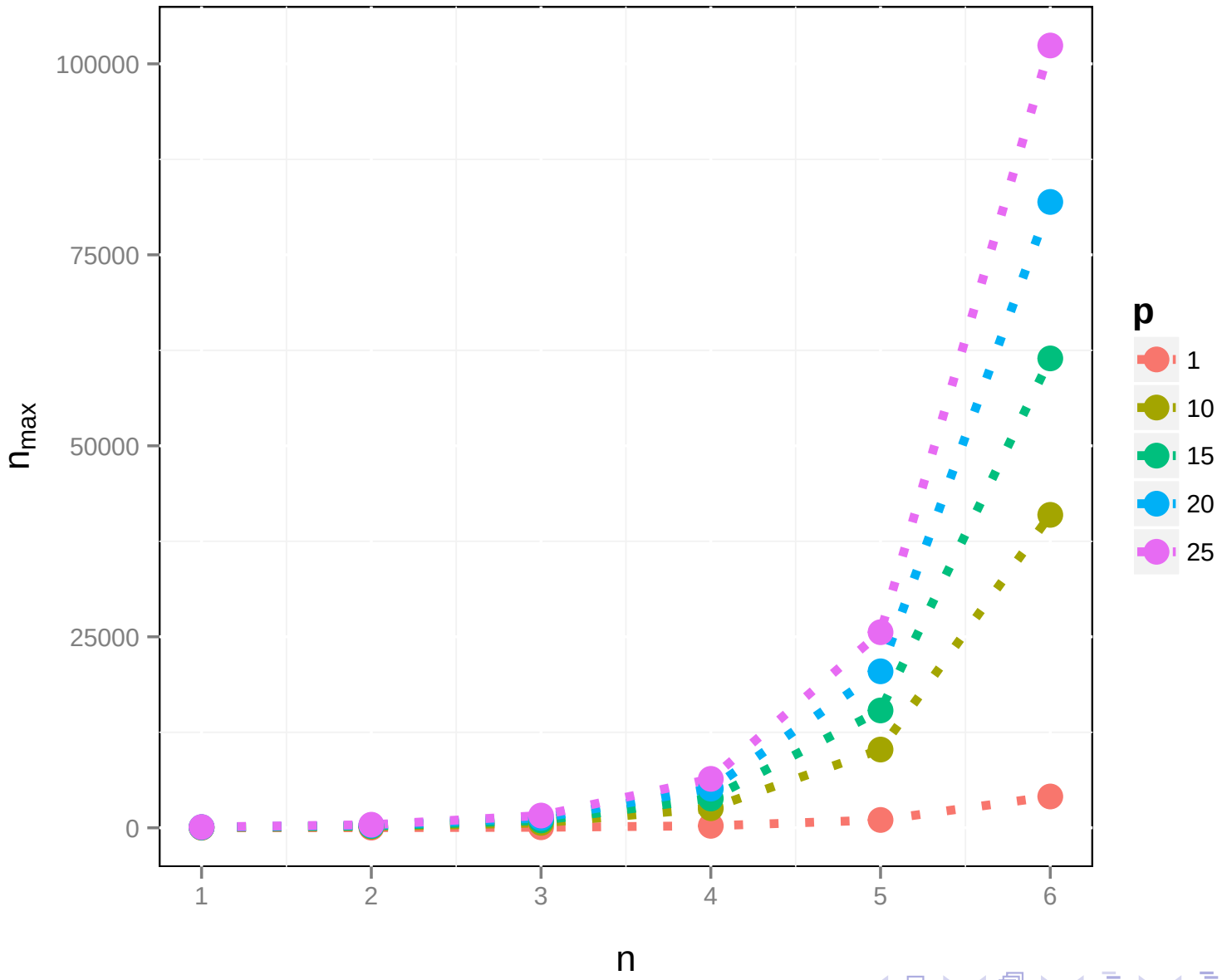
Przykładowe sekwencje. S - sekwencje, P - pozycja nukleotydu.

	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).

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

n_{max}: liczba wszystkich możliwych n-gramów
p: liczba możliwych pozycji.
u: liczba liter w alfabecie.
n: długość n-gramu


100000
75000
50000
25000
0
n_{max}
1
2
3
4
5
6
n
p
1
10
15
20
25

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).

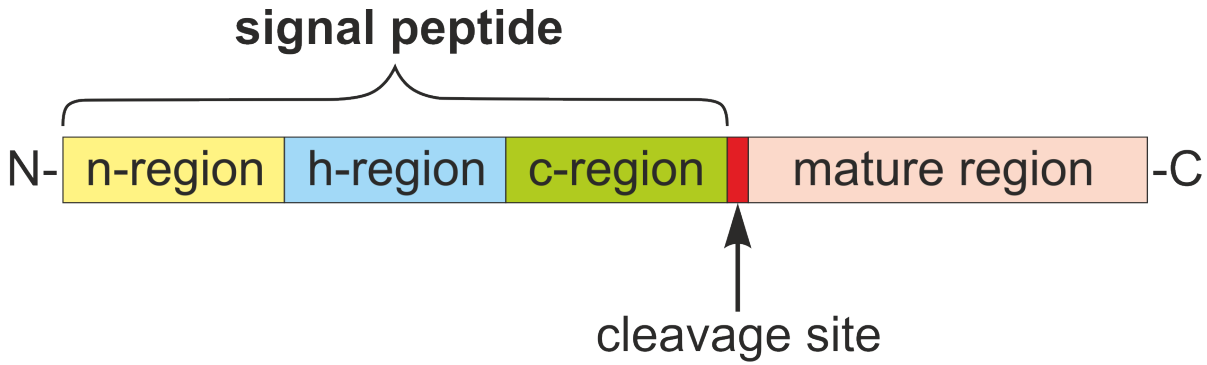
Wielowymiarowa przestrzeń atrybutów jest filtrowana z pomocą QuiPT (**Qui**ck **P**ermutation **T**est) łączącego zalety testów permutacyjnych (brak założeń) z szybkością wykonania.

W trakcie testu permutacyjnego 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 niepermutowanych danych).

Outline



- 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:

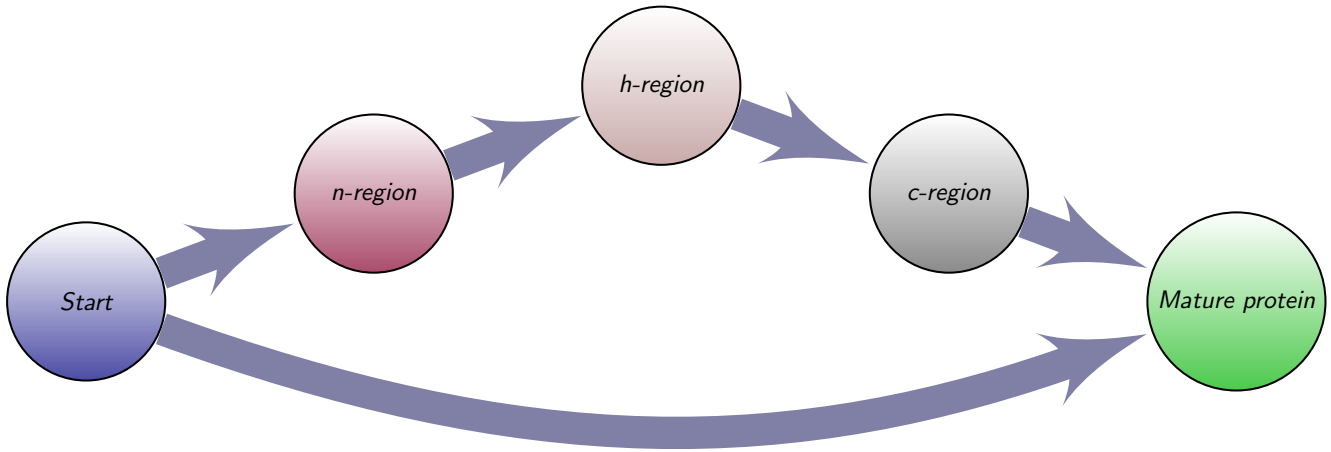
- signalP 4.1 (sieci neuronowe) (Petersen et al., 2011),
- PrediSi (Position Weight Matrix) (Hiller et al., 2004),
- Signal-3L (k-najbliszych sąsiadów) (Shen and Chou, 2007),
- Phobius (ukryte modele Markowa) (Käll et al., 2004).

Założenia modelu:

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

1. Pozyskanie eukariotycznych białek z bazy UniProtKB 2014_07 (po ocyszczeniu z nietypowych lub niedokładnie opisanych rekordów zbiór 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.



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.

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 ¡italic¿k-¡/italic¿tuple frequencies. *PLoS ONE*, 9(1):e84348.