n-gramy w analizie sekwencji biologicznych

Michał Burdukiewicz[1], Piotr Sobczyk[2]

[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.

	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

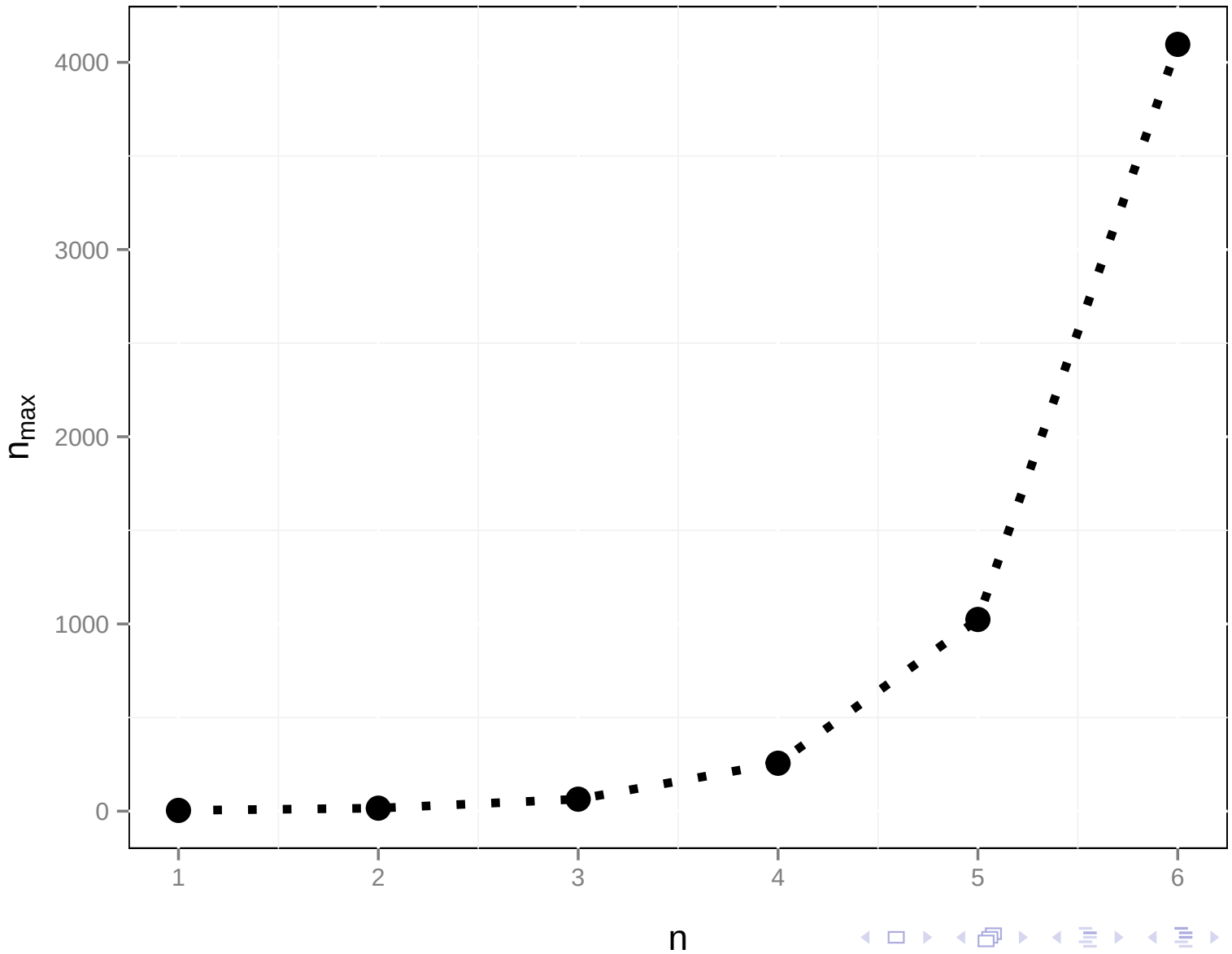
Przykładowe sekwencje. S - sekwencje, P - pozycja nukleotydu.

	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


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

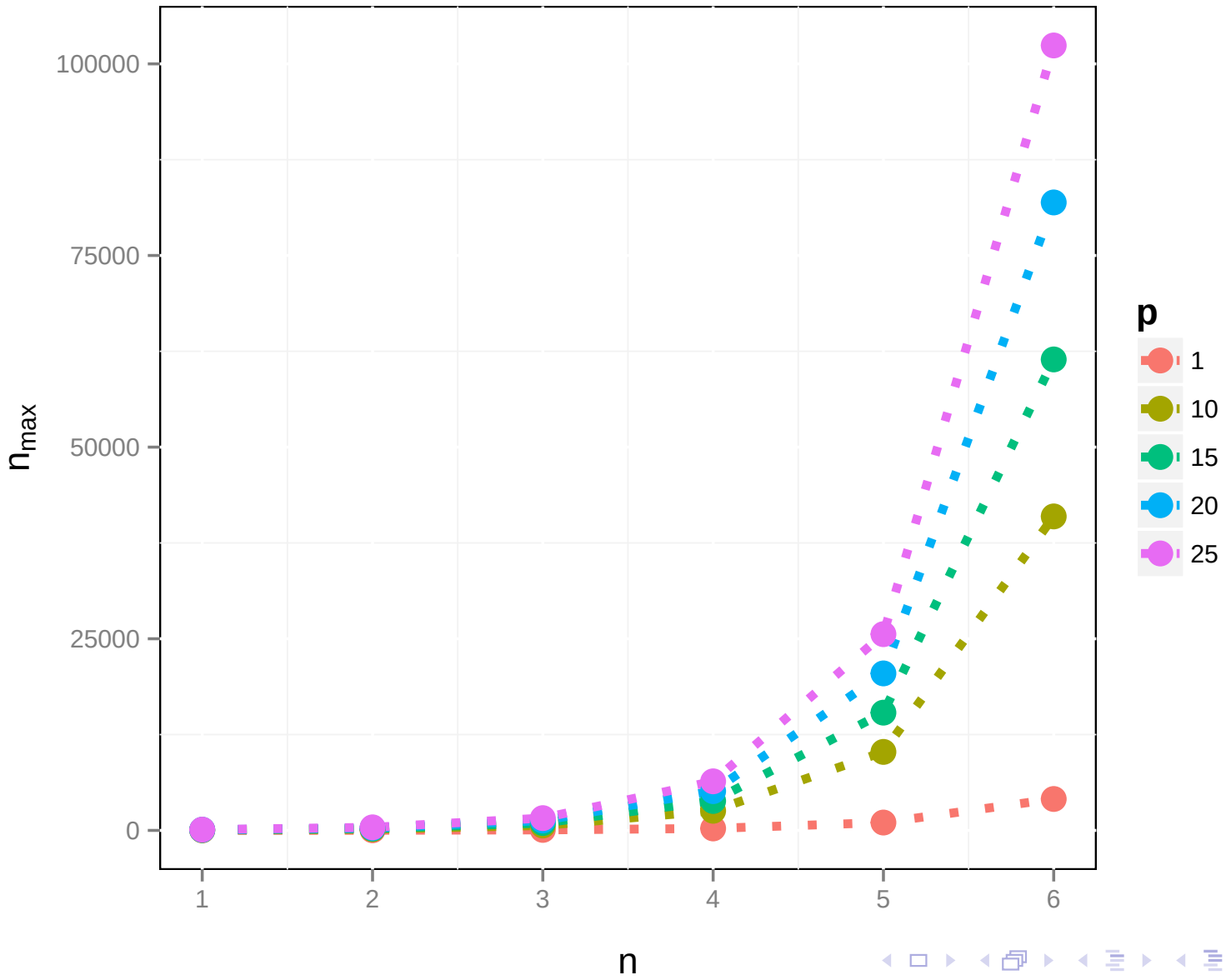
Przykładowe sekwencje. S - sekwencje, P - pozycja nukleotydu.

	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).

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

$n_{\max}$: 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).