Metody bioinformatyki (MBI)

Wykład 14 - wyszukiwanie motywów. Obliczenia na DNA.

Robert Nowak

2025L

Wyszukiwanie motywów

Wyszukiwanie motywów

Wejście: t sekwencji o długości n
Wyjście: znaleźć pozycje $\mathbf{s} = [s_0, s_1, ..., s_n]$, które są początkiem motywu o długości l
Przykład:

$t = 5, n = 18, l = 8$

$\mathbf{s} = (5, 0, 2, 9, 6)$

CGGGG<u>CTATCCAG</u>CTGGG
<u>CTATCCAG</u>ATCATCATCA
TT<u>CTATCCAG</u>AAAGTCAC
AATTTTTCT<u>CTATCCAG</u>T
AAGGCC<u>CTATCCAG</u>TGAT

Typowo: $t = 50..100, n = 500..1000, l = 8..30$

motywy nie są identyczne, ale bardzo podobne

Wyszukiwanie motywów (2)

Ocena dla $\mathbf{s}$ to $\sum_{j=0}^{l-1} \max P_j(\mathbf{s})$, gdzie $P(\mathbf{s})$ to profil dla $\mathbf{s}$, gdzie $P_j(\mathbf{s})$ jest j kolumną profilu Złożoność: $O(lt)$
Przykład:

CGGGGCTATCCAGCTGGG
CTATCCAGATCATCATCA
TTCTATCCAGAAAGTCAC
AATTTTTCTCTATCCAGT
AAGGCCCTATCCAGTGAT

$s = [1, 2, 3, 10, 0]$
ocena $= 2 + 3 + 2 + 3 + 4 + 2 + 2 + 4 = 22$

	G	G	G	G	C	T	A	T
	A	T	C	C	A	G	A	T
	T	A	T	C	C	A	G	A
	T	A	T	C	C	A	G	T
	A	A	G	G	C	C	C	T
A	2	3	0	0	1	2	2	1
C	0	0	1	3	4	1	1	0
G	1	1	2	2	0	1	2	0
T	2	1	2	0	0	1	0	4
	A	A	G	C	C	A	G	T

Wyszukiwanie motywów (3)

Algorytm pełnego przeglądania przestrzeni:
Dla wszystkich$(n - l + 1)^t$ pozycji początkowych **s** znaleźć to, o najwyższej ocenie
Złożoność (dla $n \gg l$): $O(ltn^t)$

Obserwacja: Maksymalna ocena $= l * t$ (wszystkie symbole identyczne)

można więc badać tylko pierwszych i pozycji $s_0, s_1, ..., s_i$ i jeżeli profil jest zły, to nie robić badań dla wszystkich $(n - l + 1)^{(t-i)}$ pozycji $s_{i+1}, ..., s_{t-1}$

function MotifSearch(X, t, n, l)
    $s \leftarrow [0, 0, ..., 0]$, $best \leftarrow 0$, $i \leftarrow 1$
    while $s \in \{[0, 0, ..., 0], ..., [n - l + 1, ..., n - l + 1]\}$ **do**
        if $i < t$ **then**
            optimistic = score(X,s,i) + (t-i)*l;
            if optimistic < best **then**
                pomiń poddrzewo
            else
                $i \leftarrow i + 1$
            end if
        else
            if score(X,s) > best **then**
                $best \leftarrow score(X, s)$, $motif \leftarrow s$
            end if
            $s \leftarrow next(s)$, $i \leftarrow 1$
        end if
    end while
    return motif
end function

Wyszukiwanie motywów (4) - mediana napisów

Mediana napisów, znaleźć napis v , $|v| = l$, który daje minimalną odległość dla t napisów $[X_0, X_1, ..., X_t]$, czyli $d_H(v, X) = \sum_{i=0}^{t-1} d_H(v, X_i)$

- napisów jest 4^l (dla alfabetu DNA)
- obliczenie $d_H(v, X)$ jest rzędu $O(nt)$ (dla każdego z t napisów znajdujemy pozycję, która jest najbardziej podobna do v (np. algorytm KMP)
- złożoność znajdowania mediany napisów $O(nt4^l) \ll O(ltn^t)$

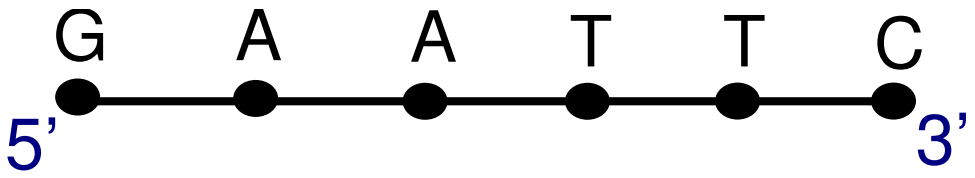
Mediana napisów jest motywem [a]

[a] dowód w N. Jones, P. Pevzner, An Introduction to Bioinformatics Algorithms, 2004

function MEDIANSEARCH(X, t, n, l)
    $v \leftarrow AA...A$, $|v| = l$, $best \leftarrow \infty$, $i \leftarrow 1$
    while $v \in \{AA...A, ..., TT...T\}$ **do**
        if $i < l$ **then**
            $prefix \leftarrow v[0..i]$, $optimistic \leftarrow distance(X, prefix)$
            if optimistic $>$ best **then**
                pomiń poddrzewo
            else
                $i \leftarrow i + 1$
            end if
        else
            if distance(X,v) $<$ best **then**
                $best \leftarrow distance(X, v)$, $median \leftarrow v$
            end if
            $v \leftarrow next(v)$, $i \leftarrow 1$
        end if
    end while
    return median
end function

Obliczenia molekularne

Pamięć na DNA



- synteza DNA - zapis
- sekwencjonowanie - odczyt

Właściwości:

	DNA	HDD	RAM
szybkość zapisu	$50 bps$	$100 Mbps$	$10 Gbps$
szybkość odczytu	$1 Mbps$	$200 Mbps$	$10 Gbps$
gęstość	$10^6 Gb/mm^3$	$10 GB/mm^3$	$1 GB/mm^3$
trwałość	10^6 lat	50 lat	1 rok

Obliczenia na DNA

- każda cząsteczka DNA może być traktowana jako 'procesor'
- mamy 10^{23} cząsteczek w 1 molu

Doświadczenia (laboratoryjne):

- problemy kombinatoryczne;
- przetwarzanie napisów;
- realizacja typowych elementów komputerów (bramki logiczne itp)

Dziękuję