



BIOINFORMATYKA

SEKWENCJONOWANIE PRZEZ HYBRYDYZACJĘ - ALGORYTMY HEURYSTYCZNE

PIOTR WAWRZYNIAK

INSTYTUT INFORMATYKI, POLITECHNIKA POZNAŃSKA

PIOTR.WAWRZYNIAK@CS.PUT.POZNAN.PL

WWW.PIOTR.E.WAWRZYNIAK.DOCTORATE.PUT.POZNAN.PL

INSTANCJA

ACGT

ACGTAACCTG

ACGT
CGTA
GTAA
TAAC
AACC
ACCT
CCTG



AACC
ACCT
ACGT
CCTG
CGTA
GTAA
TAAC



AACC
~~ACCT~~
ACGT
CCTG
CGTA
GTAA
~~TAAC~~
AACT
CTGG



AACC
AACT
ACGT
CCTG
CTGG
CGTA
GTAA

ALGORYTM ZACHŁANNY

Błażewicz J., Formanowicz P., Kasprzak M.,
Markiewicz W. T., Węglarz J., 1999.

DNA sequencing with positive and negative errors.

J. Comput. Biol. 6 113–123.

Startujemy z początkowego oligonukleotydu

Do tworzonej sekwencji dobieramy
oligonukleotyd z najlepszym dopasowaniem

Dopasowanie oceniamy poprzez liczbę
nakładających się nukleotydów, oraz potencjalne
dopasowanie kolejnego nukleotydu

Dokładanie oligonukleotydów kończymy gdy
następny dodany zwiększył by długość wyniku
powyżej dopuszczalnej wartości

ALGORYTM ZACHŁANNY

ACGT, AACC, AACT, CCTG, CTGG, CGTA, GTAA; $n=10$

ACGT $n=4$

AACC(4) + CCTG(2) 6
AACT(4) + CTGG(2) 6
CCTG(4) + CTGG(1) 5
CTGG(4) + GTAA(3) 7
CGTA(1) + GTAA(1) 2 ✓
GTAA(2) + AACC(2) 4

ACGTA $n=5$

AACC(3) + CCTG(2) 5
AACT(3) + CTGG(2) 5
CCTG(4) + CTGG(1) 5
CTGG(4) + GTAA(3) 7
GTAA(1) + AACC(2) 3 ✓

ACGTAA $n=6$

AACC(2) + CCTG(2) 4 ✓
AACT(2) + CTGG(2) 4
CCTG(4) + CTGG(1) 5
CTGG(4) + AACC(4) 8

ACGTAACC $n=8$

AACT(4) $8+4 > 10$ x
CCTG(2) $8+2 = 10$ ✓
CTGG(3) $8+3 > 10$ x

ACGTAACCTG

ACGTAACCTGG

PRZESZUKIWANIE TABU

Błażewicz J., Glover F., Kasprzak M., 2004.

DNA sequencing — tabu and scatter search combined.

INFORMS J. Comput. 16, 232–240.

Stworzenie początkowego rozwiązanie poprzez algorytm zachłanny

Ruchy – dodanie oligonukleotydu, usunięcie lub przesunięcie oligonukleotydu lub klastra (oligonukleotydy z pełnym dopasowaniem)

Lista tabu – lista wstawionych, usuniętych lub przesuniętych oligonukleotydów, pamiętana określona ilość iteracji

Funkcje oceny:

globalna: liczba oligonukleotydów w rozwiązaniu

zagęszczania: stosunek oligonukleotydów do nukleotydów w rozwiązaniu

Naprzemienne ruchy:

zagęszczania – funkcja oceny zagęszczania

wydłużania – globalna funkcja oceny

Przy braku poprawy nowe początkowe rozwiązanie stworzone przez algorytm zachłanny, ale z dopasowaniami oligonukleotydów ograniczonymi do tych z już znalezionych rozwiązań

PRZESZUKIWANIE TABU

ACGT, AACC, AACT, CCTG, CTGG, CGTA, GTAA; **ACGTAACTGG**

Zagęszczanie

Tabu: -

ACGTAACTGG 0,50

AACT	0,40
ACGT	0,44
CTGG	0,50
GTAA	0,40
ACGTAA	0,33

Zagęszczanie

Tabu: CTGG

ACGTAACT 0,50

AACT	0,50
ACGT	0,43
GTAA	0,37
ACGTAA	1,00

Zagęszczanie

Tabu: AACT, CTGG

ACGTAA 0,50

ACGT	0,40
GTAA	0,40
ACGTAA	0/0

PRZESZUKIWANIE TABU

ACGTAA; AACC, AACT, CCTG, CTGG

Wydłużanie

Tabu: AACT, CTGG

ACGTAA 3

AACC 4

CCTG 4

Wydłużanie

Tabu: AACC, AACT, CTGG

ACGTAACC 4

CCTG 5

Wydłużanie

Tabu: CCTG, AACC, AACT, CTGG

ACGTAACCTG 5

-

ALGORYTM GENETYCZNY

Bui T. N., Youssef W. A, 2004.

An enhanced genetic algorithm for DNA sequencing by hybridization with positive and negative errors.

Lect. Notes Comput. Sci. 3103, 908–919.

Łączenie oligonukleotydów z dopasowanymi $k-1$ nukleotydami (jeśli jest to jedyne możliwe dopasowanie)

Osobnik w algorytmie to permutacja uzyskanych sekwencji z początkowego kroku

Dopasowanie osobnika – liczba nakładających się nukleotydów pomiędzy składowymi sekwencjami oraz różnica długość wynikowej od oczekiwanej

Krzyżowanie – podział na 3 miejscach i wybór różnych części od rodziców. Wymagane usuwanie duplikacji i dodawanie brakujących części

Mutacja – zamiana miejscami losowo wybranych fragmentów

Potomkowie lepsi niż jeden z rodziców zastępują najgorsze osobniki

Algorytm kończy się gdy nie potrafi już polepszyć wyniku

INSTANCJA

n=12 k=3

TCAGTACGACAA

TCA

CAG

AGT

~~GTA~~

TAC

ACG

CGA

~~GAC~~

ACA

CAA

AAT

GTT

Spektrum

AAT

ACA

ACG

AGT

CAA

CAG

CGA

GTT

TAC

TCA

Łańcuchy

1 ACA

2 ACGA

3 CAAT

4 CAGTT

5 TAC

6 TCA

ALGORYTM GENETYCZNY

1 ACA, 2 ACGA, 3 CAAT, 4 CAGTT, 5 TAC, 6 TCA

Początkowa losowa populacja

[1452]	ACAGTTACGA	5-2= 3
[21354]	ACGACAATACAGTT	5-2= 3
[35146]	CAATACAGTTCA	6-0= 6
[45136]	CAGTTACAATCA	6-0= 6
[51463]	TACAGTTCAAT	7-1= 6
[64521]	TCAGTTACGACA	6-0= 6

ALGORYTM GENETYCZNY

1 ACA, 2 ACGA, 3 CAAT, 4 CAGTT, 5 TAC, 6 TCA

Krzyżowanie

[5 14 6 3]	TACAGTTCAAT	6
[64 5 2 1]	TCAGTTACGACA	6

[5561]	[5361]	TACAATCACA	3-2= 1
[641423]	[641523]	TCAGTTACATACGACAAT	4-6=-2
[64563]	[64513]	TCAGTTACAAT	7-1= 6
[51421]	[51423]	TACAGTTACGACAAT	4-6=-2

ALGORYTM GENETYCZNY

1 ACA, 2 ACGA, 3 CAAT, 4 CAGTT, 5 TAC, 6 TCA

Mutacja

[64513]	TCAGTTACAAT		6
[24513]	ACGACAGTTACAAT	$7-2=$	5
[62513]	TCACGATAACAAT	$5-0=$	5
[64213]	TCAGTTACGACAAT	$5-2=$	3
[64523]	TCAGTTGACAAT	$5-0=$	5
[64512]	TCAGTTACACGA	$6-0=$	6