



BIOINFORMATYKA

SEKWENCJONOWANIE PRZEZ HYBRYDYZACJĘ - DEFINICJA PROBLEMU

PIOTR WAWRZYNIAK

INSTYTUT INFORMATYKI, POLITECHNIKA POZNAŃSKA

PIOTR.WAWRZYNIAK@CS.PUT.POZNAN.PL

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

DANE WEJŚCIOWE

Długość badanego fragmentu DNA

Długość sond oligonukleotydowych

Początkowy fragment długości k badanego DNA

Spektrum dopasowanych oligonukleotydów

INSTANCJA

N: długość szukanego łańcucha DNA

K: długość sond oligonukleotydowych

S_0 : początkowe słowo szukanego łańcucha DNA

Spektrum: $\{S_1, S_2, \dots, S_{n-k}\}$ - zbiór $n-k$ słów S_i będących fragmentami szukanego DNA

Słowo S_i : ciąg o długości k

$$n_{i+1}n_{i+2}\dots n_{i+k}$$

nad alfabetem $\{A, C, G, T\}$

INSTANCJA Problem podstawowy i błędy

$n=9$ $k=3$

TCGTACGAC

TCG

CGT

GTA

TAC

ACG

CGA

GAC



S_0 : TCG

Spektrum

Bez błędów

NNN:

ACG

CGA

CGT

GAC

GTA

TAC

TCG

S_0 : TCG

Spektrum

Błędy pozytywne

NNN:

ACG

CGA

CGT

GAC

GGC

GTA

TAC

TCG

S_0 : TCG

Spektrum

Błędy negatywne

NNN:

ACG

CGA

~~CGT~~

GAC

GTA

TAC

TCG

INSTANCJA Chipy binarne

$n=9$ $k=2$

TCGTACGAC

TCG

CGT

GTA

TAC

ACG

CGA

GAC



S_0 : TCG

Spektrum

słabe – silne

W: AT – S: CG

ZZN:

SSA

SST

SWA

SWC

WSG -> ACG TCG

WWC

S_0 : TCG

Spektrum

puryny – pirymidyny

R: AG – Y: CT

PPN:

RRC

RYA

RYG

YRA

YRC

YRT

YYG

INSTANCJA Chipy alternatywne

$n=9$ $k=3$

TCGTACGAC

TCGTA

CGTAC

GTACG

TACGA

ACGAC



S_0 : TCGTA

Spektrum

NXNXN:

AXGXC

CXTXC

GXAXG

TXCXA

TCGTACGAC

TCGT

CGTA

GTAC

TACG

ACGA

CGAC



S_0 : TCGT

Spektrum

NXNN:

AXGA

CXAC

CXTA

GXAC

TXCG

INSTANCJA Informacja o położeniu

$n=9$ $k=3$

TCGTACGACCCA

TCG

CGT

GTA

TAC

ACG

CGA

GAC

ACC

CCC

CCA



S_0 : TCG

Spektrum

Bez błędów położenia

NNN:

ACC : 3

ACG : 2

CCA : 4

CCC : 4

CGA : 2

CGT : 0

GAC : 3

GTA : 1

TAC : 1

TCG : 0

S_0 : TCG

Spektrum

Z błędami położenia

NNN:

ACC : ~~3~~ 2

ACG : 2

CCA : ~~4~~ 3

CCC : 4

CGA : 2

CGT : ~~0~~ 1

GAC : ~~3~~ 1

GTA : 1

TAC : 1

TCG : 0

INSTANCJA Informacja o powtórzeniach

n=9 k=3

TCTCTCTGGGGGG

TCT
CTC
TCT
CTC
TCT
CTG
TGG
GGG
GGG
GGG



S_0 : TCG

Spektrum

Bez błędów

NNN:

CTC : 2 -> 2
CTG : 1 -> 1
GGG : 4 -> 3
TCT : 3 -> 2
TGG : 1 -> 1

S_0 : TCG

Spektrum

Z błędami pozytywnymi

NNN:

CTA : 1 -> 1
CTC : 23 -> 2
CTG : 1 -> 1
GGG : 4 -> 3
TCT : 35 -> 23
TGG : 1 -> 1

S_0 : TCG

Spektrum

Z błędami negatywnymi

NNN:

CTC : 21 -> 21
CTG : 1 -> 1
GGG : 4 -> 3
TCT : 32 -> 2
~~TGG : 1 -> 1~~

A decorative graphic consisting of blue circuit-like lines with small circles at the ends, extending horizontally from the left and right sides of the central black box.

CEL

REKONSTRUKCJA SEKWENCJI DNA PODDANEJ
EKSPERYMENTOWI HYBRYDYZACJI

ODPOWIEDŹ

- Superciąg R , powstały przez uporządkowanie wszystkich słów S_i

- S_y następuje po S_x , gdy:

$$\forall i > 1 : n_{y+i} = n_{x+i-1}$$

- Superciąg R zawiera wszystkie słowa ze spektrum

TWORZENIE SUPERCIAĞU

TCG

ACG

CGA

CGT

GAC

GTA

TAC

PRZESZUKIWANIE DRZEWA W GŁĄB

Korzeń – sekwencja startowa

Wierzchołki – kolejne rozwiązania

Dzieci – rozwiązania wydłużone o jeden nukleotyd

Łuki – relacja następowania słów

Szukana sekwencja – ścieżka o maksymalnej głębokości

Złożoność pamięciowa zależna od głębokości ścieżki

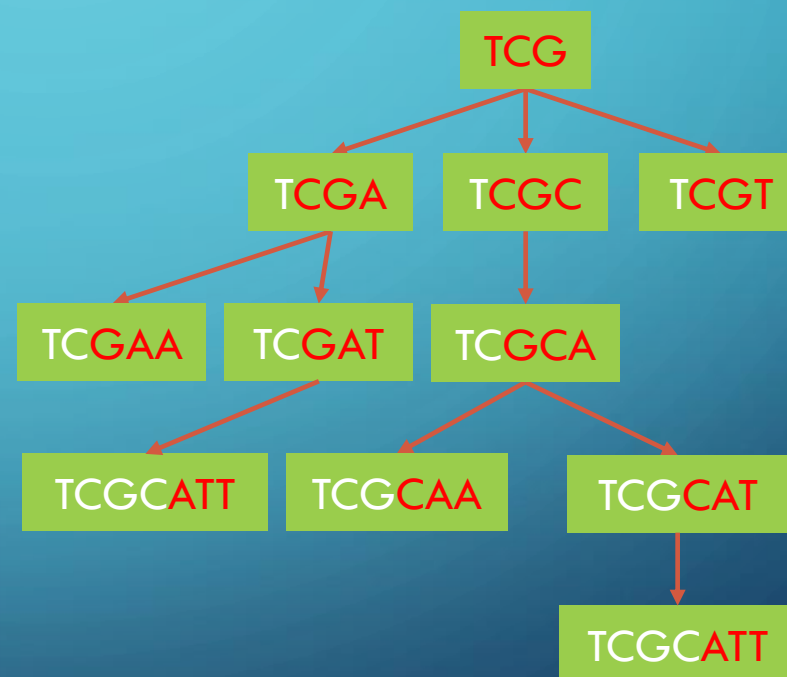
Złożoność obliczeniowa $O(|V| + |E|)$

Problem?

V – cała przestrzeń rozwiązań 4^n

Spektrum

TCG; ATT CAA CAT CGA CGC CGT
GAA GAT GCA TCG



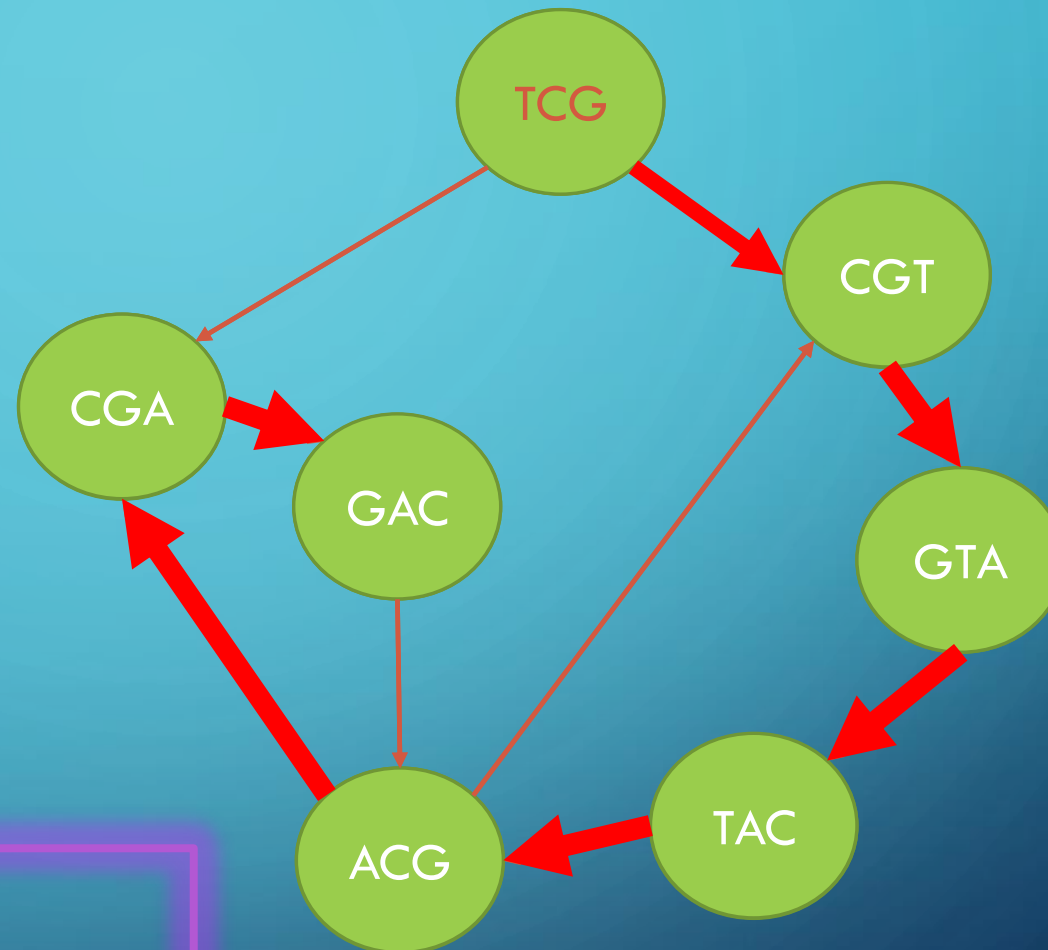
Spektrum: TCG; ACG CGA CGT GAC GTA TAC TCG

ŚCIEŻKA HAMILTONA

Wierzchołki – słowa ze spektrum

Łuki – relacja następowania słów

Szukana sekwencja – ścieżka
przechodząca dokładnie raz przez
wszystkie wierzchołki – ścieżka
Hamiltona



Problem silnie NP-zupełny

Złożoność?

Proste przeszukiwanie $O(|V|!)$ (znany wierzchołek początkowy)

Algorytm Bellman, Held & Karp $O(2^{|V|} |V|^2)$

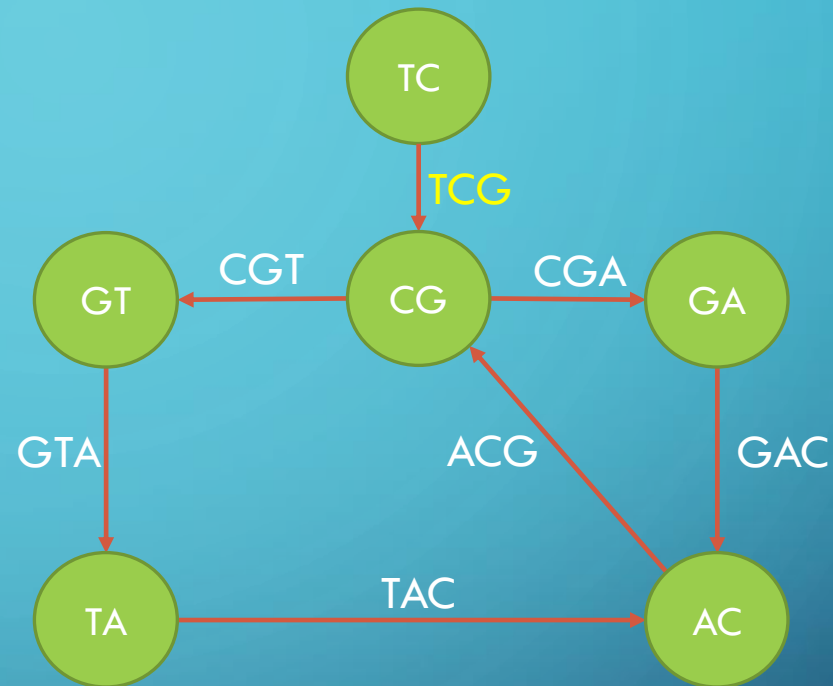
Spektrum: TCG; ACG CGA CGT GAC GTA TAC TCG

ŚCIEŻKA EURELA

Łuki – słowa ze spektrum

Wierzchołki – prefiksy i sufiksy słów ze spektrum

Szukana sekwencja – ścieżka przechodząca dokładnie raz przez wszystkie krawędzie – ścieżka Eureka



Złożoność wielomianowa

$O(|V| + |E|)$

Problem?

Błędy w spektrum uniemożliwiają praktyczne zastosowania