

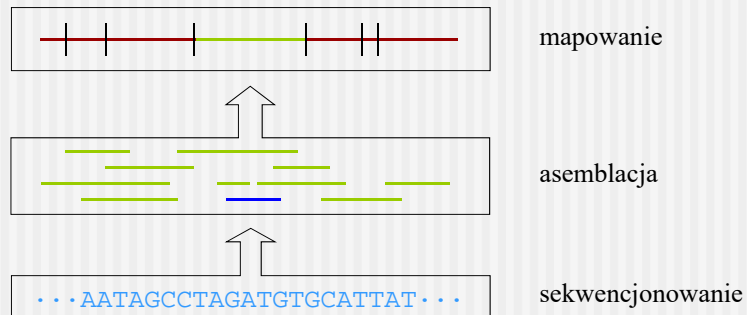
# Algorytmy kombinatoryczne w bioinformatyce

wykład 2: sekwencjonowanie cz. 1

prof. dr hab. inż. **Marta Kasprzak**  
Instytut Informatyki, Politechnika Poznańska

## Poznawanie sekwencji genomowej

- Poznawanie sekwencji genomów na trzech poziomach



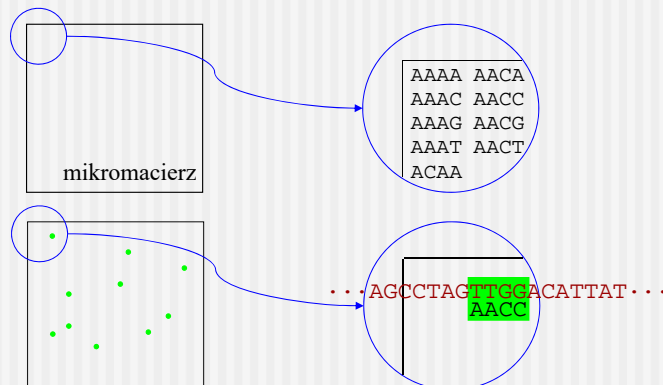
## Sekwencjonowanie

- Sekwencjonowanie DNA jest procesem wyznaczenia sekwencji nukleotydów pewnego fragmentu DNA. Fragmenty, w zależności od metody, osiągają długość od kilkudziesięciu do tysiąca nukleotydów (więcej w sekwencjonowaniu trzeciej generacji)
- Metody laboratoryjne sekwencjonowania na żelu nie wymagają metod obliczeniowych, są za to nieodporne na błędy eksperymentalne
- Sekwencjonowanie przez hybrydyzację (SBH) korzysta z wyników eksperymentu hybrydyzacyjnego, czyli zbioru oligonukleotydów (*spektrum*) wchodzących (przeważnie) w skład badanego fragmentu DNA. Długość badanego fragmentu można zmierzyć w procesie elektroforezy. W klasycznym modelu przyjmuje się, że żadne dodatkowe informacje o elementach spektrum nie są dostępne

3

## Sekwencjonowanie przez hybrydyzację

- SBH z biblioteką oligonukleotydów o stałej długości



4

## Sekwencjonowanie przez hybrydyzację

badana sekwencja: CCGACGT  
długość oligonukleotydów: 3  
spektrum bez błędów:  
{ACG, CCG, CGA, CGT, GAC}

Rozwiązanie:

CCGACGT

CCG

CGA

GAC

ACG

CGT

5

## Sekwencjonowanie przez hybrydyzację

Błędy eksperymentu hybrydyzacyjnego:

- Błąd negatywny — gdy oligonukleotyd obecny w sekwencji nie zostanie wykryty
  - (a) oligonukleotyd występuje więcej niż raz w badanej sekwencji
  - (b) oligonukleotyd komplementarny na skutek nieoptymalnych warunków eksperymentu nie połączy się z badaną sekwencją
- Błąd pozytywny — gdy zostanie wskazany oligonukleotyd nieobecny w sekwencji
  - (c) oligonukleotyd niezupełnie komplementarny połączy się z badaną sekwencją
  - (d) zaszumiony obraz fluorescencyjny mikromacierzy

Spektrum zazwyczaj zawiera oba rodzaje błędów

6

## Sekwencjonowanie przez hybrydyzację

Przykłady błędów eksperymentu hybrydyzacyjnego

	sekwencja	spektrum
(a)	TTACATTA	{ACA, ATT, CAT, TAC, TTA}
(b)	TTACATTC	{ACA, ATT, TAC, TTA, TTC}
(c)	TTACAT	{ACA, CAT, TAC, TTA, TTT}
(d)	TTACAT	{ACA, CAT, GAG, TAC, TTA}

7

## Sekwencjonowanie przez hybrydyzację

badana sekwencja: CCGACGT

długość oligonukleotydów: 3

spektrum bez błędów:

{ACG, CCG, CGA, CGT, GAC}

błąd negatywny: CGA

błędy pozytywne: AAT, TTG

spektrum z błędami:

{AAT, ACG, CCG, CGT, GAC, TTG}

Rozwiązanie:

CCGACGT

CCG

GAC

ACG

CGT

8

## Sekwencjonowanie przez hybrydyzację

- Sformułowanie problemu klasycznego sekwencjonowania DNA przez hybrydyzację z błędami — wersja optymalizacyjna  
Instancja: Zbiór oligonukleotydów o jednakowej długości  $l$ , długość sekwencji oryginalnej  $n$ .  
Odpowiedź: Sekwencja o długości  $\leq n$  zawierająca maksymalną liczbę oligonukleotydów ze zbioru.
- Problem sekwencjonowania jest silnie NP-trudny, nawet po ograniczeniu błędów do tylko negatywnych lub tylko pozytywnych
- Nawet przy założeniu wiedzy o istnieniu unikalnego rozwiązania w instancji oba podproblemy pozostają trudne obliczeniowo

9

## Sekwencjonowanie przez hybrydyzację

- Prezentowane modele opisują warianty klasycznego problemu sekwencjonowania przez hybrydyzację, z założeniem braku lub obecności błędów w instancji
- Wspólnym celem omawianych podejść jest znalezienie ścieżki w grafie skierowanym zbudowanym na podstawie instancji
- Etykiety przypisane wierzchołkom pozwalają jednoznacznie przekształcić ścieżkę w grafie na sekwencję nukleotydów

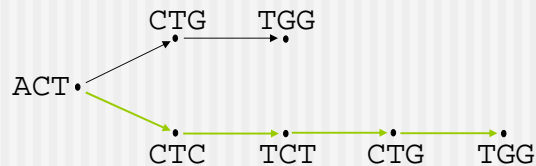
10

## Sekwencjonowanie przez hybrydyzację

[W. Bains i G.C. Smith, *J. Theor. Biol.* 135 (1988)]

- Pierwszy algorytm rekonstruujący sekwencję oryginalną na podstawie spektrum
- Założenie o braku błędów w instancji
- Konstrukcja drzewa, w którym oligonukleotydy odpowiadają wierzchołkom (przeszukiwanie z nawrotami)

ACTCTGG,  $S = \{ACT, CTC, CTG, TCT, TGG\}$



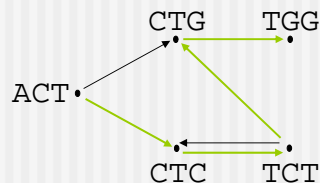
11

## Sekwencjonowanie przez hybrydyzację

[Y.P. Lysov i in., *Doklady Akademii Nauk SSSR* 303 (1988)]

- Pierwszy algorytm odwołujący się do znanego problemu z teorii grafów
- Założenie o braku błędów w instancji
- Poszukiwanie ścieżki Hamiltona w grafie skierowanym, w którym oligonukleotydy odpowiadają wierzchołkom

ACTCTGG,  $S = \{ACT, CTC, CTG, TCT, TGG\}$



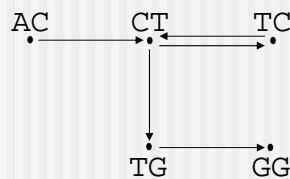
12

## Sekwencjonowanie przez hybrydyzację

[P.A. Pevzner, *J. Biomol. Struct. Dyn.* 7 (1989)]

- Pierwszy algorytm rozwiązujący problem sekwencjonowania bez błędów w instancji w czasie wielomianowym
- Poszukiwanie ścieżki Eulera w grafie skierowanym, w którym oligonukleotydy odpowiadają łukom

ACTCTGG,  $S = \{ACT, CTC, CTG, TCT, TGG\}$



13

## Grafy dla SBH bez błędów

[J. Błażewicz i in., *Discrete Appl. Math.* 98 (1999)]

- Pevzner w swoim artykule nie wyjaśnił, dlaczego możliwa była transformacja grafu, w którym poszukiwana jest ścieżka Hamiltona (problem w ogólności silnie NP-trudny) w graf, w którym poszukiwana jest ścieżka Eulera (problem obliczeniowo łatwy)
- Grafy z metody Lysova i in. (tzw. grafy DNA) należą do klasy *grafów liniowych* (krawędziowych) skierowanych, dla których problem poszukiwania ścieżki Hamiltona rozwiązywany jest w czasie wielomianowym. Ich grafami oryginalnymi są grafy z metody Pevznera
- Wierzchołki grafu liniowego  $G=(V,A)$  reprezentują łuki grafu oryginalnego  $H=(U,V)$ , łuki w  $G$  łączą wierzchołek  $x$  z  $y$  jeśli w  $H$  koniec łuku  $x$  pokrywa się z początkiem łuku  $y$

14

## Grafy dla SBH bez błędów

### ■ Zadanie

Dla następującego spektrum bez błędów:

$S = \{AGTA, AGTG, GTAC, GTAG, GTGT, TACC, TAGT, TGTA\}$   
należy skonstruować grafy metodami Lysova i in. oraz Pevznera,  
znaleźć w nich odpowiednie ścieżki reprezentujące rozwiązania  
i odczytać możliwe sekwencje wynikowe.

Przeprowadzić transformację grafu liniowego z metody Lysova i in.  
do jego grafu oryginalnego.

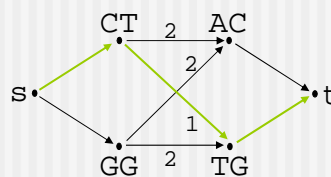
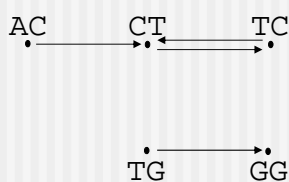
15

## SBH z błędami negatywnymi

[P.A. Pevzner, *J. Biomol. Struct. Dyn.* 7 (1989)]

- Założenie o obecności błędów negatywnych w instancji
- Algorytm heurystyczny wielomianowy
- Poszukiwanie brakujących oligonukleotydów jako przepływu o wartości  $m-1$  i minimalnym koszcie w sieci opartej na grafie dwudzielnym  $K_{m,m}$

ACTCTGG,  $S = \{ACT, CTC, TCT, TGG\}$



16



## SBH z błędami negatywnymi

[P.A. Pevzner, *J. Biomol. Struct. Dyn.* 7 (1989)] – cd.

- Niepowodzenie, gdy przepływ o minimalnym koszcie okaże się mieć koszt inny niż liczba brakujących oligonukleotydów
- Niepowodzenie, gdy uzupełniony graf okaże się niespójny
- Ewentualne niepowodzenie, gdy wierzchołek z brakującymi incydentnymi łukami nie zostanie wstawiony do grafu dwudzielnego

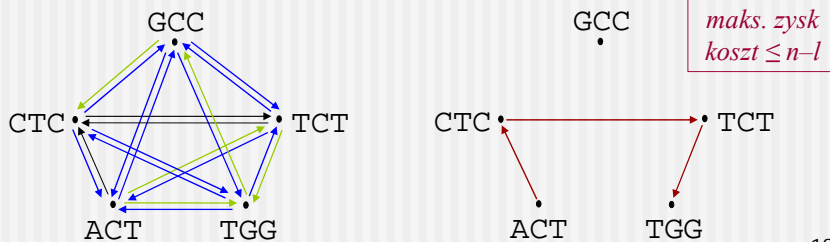
17

## SBH z błędami negatywnymi i pozytywnymi

[J. Błażewicz i in., *J. Comput. Biol.* 6 (1999)]

- Pierwszy algorytm rozwiązujący problem z dowolnymi błędami negatywnymi i pozytywnymi w instancji
- Sformułowanie wariantu problemu selektywnego komiwojażera w grafie skierowanym pełnym

ACTCTGG,  $S = \{\text{ACT}, \text{CTC}, \text{GCC}, \text{TCT}, \text{TGG}\}$



18