Algorytmy kombinatoryczne w bioinformatyce

wykład 3: sekwencjonowanie cz. 2

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

Wieloetapowe SBH

[V.T. Phan i S. Skiena, *Bioinformatics* 17 (2001)]

- Podejście wieloetapowe do spektrum z dowolnymi błędami
- Tradycyjny eksperyment hybrydyzacyjny jest uzupełniony o serię porównań krótkich próbek (dłuższych niż l) z sekwencją oryginalną
- Do grafu konstruowanego jak w metodzie Lysova i in. wprowadza się modyfikacje oparte na dodatkowych eksperymentach. Dodatkowe „zapytania” przeprowadza się ze wzrastającą długością próbek, począwszy od l+1, do momentu usunięcia wszystkich niejednoznacznych przejść w grafie (rozgałęzień)

SBH z nukleotydami uniwersalnymi

- Wieloetapowe podejście do sekwencjonowania przez hybrydyzację umożliwia jednoznaczne rozwiązanie problemu, niestety dużym nakładem pracy
- Zwiększanie długości oligonukleotydów z biblioteki sprzyja jednoznacznemu rozwiązaniu, wiąże się jednak z wykładniczym wzrostem rozmiaru biblioteki
- Oligonukleotydy o długości 10 pozwalają stosunkowo jednoznacznie zrekonstruować sekwencje o długości ok. 500 nukleotydów, dla l=8 są to sekwencje o długości 100–200
- Statystycznie rzecz ujmując, zastosowanie nukleotydów uniwersalnych umożliwia przesunięcie granicy jednoznacznej rekonstrukcji przy zachowaniu niezmienionej liczebności biblioteki

SBH z nukleotydami uniwersalnymi

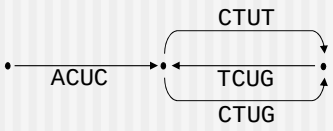
[F.P. Preparata i E. Upfal, zgłoszenie patentowe USA (2001)]

- Użycie *gapped probes* w miejsce klasycznej biblioteki oligonukleotydów – wprowadzenie nukleotydu uniwersalnego U hybrydyzującego w zamierzeniu z każdym innym nukleotydem
- Nowa biblioteka w porównaniu do klasycznej o tej samej liczebności zawiera dłuższe oligonukleotydy, co może umożliwić jednoznaczne odtworzenie dłuższych sekwencji
- Metoda algorytmiczna zakłada spektrum pozbawione błędów
- Oligonukleotydy opisane jako (s,r)-*probes* skonstruowano wg schematu: s standardowych nukleotydów + r wystąpień wzoru „$s-1$ nukleotydów uniwersalnych + 1 nukleotyd standardowy”; (3,2)-*probe* = XXXUUXUUX

SBH z nukleotydami uniwersalnymi

- Jeśli zwizualizujemy instancję problemu w postaci grafu na wzór metody Pevznera, poszukiwać w nim będziemy szczególnej ścieżki Eulera

ACTCTGG, (2,1)-*probes* = {ACUC, CTUG, CTUT, TCUG}




```
ACUC
 CTUT
  TCUG
   CTUG
--------
ACTCTGG
```

- Istnieją dwie ścieżki Eulera w grafie, ale tylko jedna nie produkuje konfliktów w nałożeniach oligonukleotydów

SBH z nukleotydami zdegenerowanymi

[F.P. Preparata i J.S. Oliver, *J. Comput. Biol.* 11 (2004)]

- Zamiast sztucznego nukleotydu uniwersalnego hybrydyzującego w teorii z każdym innym nukleotydem zastosowano nukleotydy „zdegenerowane", będące mieszaniną nukleotydów A, C, G i T
- W każdym polu generowanej mikromacierzy znajdą się oligonukleotydy różniące się między sobą na pozycjach wystąpienia nukleotydów zdegenerowanych
- Liczba nukleotydów zdegenerowanych w oligonukleotydzie jest mocno ograniczona ze względu na zanik sygnału hybrydyzacji
- Ponieważ sygnał hybrydyzacji zależy dodatkowo od siły wiązania nici komplementarnych, wprowadzono bardziej szczegółowy model nukleotydów „na wpół zdegenerowanych", osobno dla nukleotydów „słabych" (A/T) i „mocnych" (C/G)

Izotermiczne SBH

[J. Błażewicz i in., zgłoszenie patentowe PL (1999)]

- W izotermicznej wersji sekwencjonowania DNA eksperyment hybrydyzacyjny jest przeprowadzany z *izotermicznymi bibliotekami oligonukleotydów*. Biblioteka izotermiczna zawiera wszystkie oligonukleotydy o jednakowej *temperaturze topnienia dupleksów*, lecz o różnej długości
- Podejście izotermiczne do problemu sekwencjonowania daje szansę na uniknięcie znacznej liczby błędów eksperymentalnych w etapie biochemicznym
- Uproszczony model z wczesnych prac zakłada, że każda para C/G wnosi 4° do temperatury dupleksu, a para A/T wnosi 2°

Izotermiczne SBH

[J. Błażewicz i in., zgłoszenie patentowe PL (1999)] – cd.

- Izotermiczna biblioteka oligonukleotydów o temperaturze $\mathcal{T}$ zawiera wszystkie oligonukleotydy spełniające zależności:

$$w_A x_A + w_C x_C + w_G x_G + w_T x_T = \mathcal{T},$$

$$w_A = w_T,$$

$$w_C = w_G,$$

oraz $2w_A = w_C$.

Zakłada się, że $w_A = w_T = 2°$ i $w_C = w_G = 4°$.

Przykład: $\mathcal{T}_{\mathtt{ACATT}} = 2°+4°+2°+2°+2° = 12°$,
$\mathcal{T}_{\mathtt{GCC}} = 4°+4°+4° = 12°$.

Izotermiczne SBH

- Jedna biblioteka izotermiczna nie wystarcza do pokrycia dowolnej sekwencji DNA

  Przykład: Biblioteka o temperaturze niepodzielnej przez 4 nie pokryje sekwencji złożonej wyłącznie z C i G. Biblioteka o temperaturze podzielnej przez 4 może nie pokryć pojedynczego A/T otoczonego przez C/G.

- Każdą sekwencję DNA można pokryć elementami dwóch bibliotek o temperaturach różniących się dwoma stopniami

Izotermiczne SBH

badana sekwencja: CCTACGT

temperatury bibliotek: 10° i 12°

spektrum bez błędów: {ACG, ACGT, CCT, CCTA, CGT, CTAC, TACG}

błędy negatywne: ACG, CCTA, TACG

błędy pozytywne: ACC, TTTG

spektrum z błędami: {ACC, ACGT, CCT, CGT, CTAC, TTTG}

Rozwiązanie:

```
CCTACGT
CCT
 CTAC
   ACGT
    CGT
```

Izotermiczne SBH

- Sformułowanie problemu izotermicznego sekwencjonowania DNA przez hybrydyzację z błędami – wersja optymalizacyjna
  Instancja: Zbiór oligonukleotydów o temperaturach $\mathcal{T}$ lub $\mathcal{T}$+2, długość sekwencji oryginalnej n.
  Odpowiedź: Sekwencja o długości $\leq n$ odpowiadająca minimalnej liczbie błędów negatywnych i pozytywnych.
- Problem izotermicznego sekwencjonowania jest silnie NP-trudny, nawet po ograniczeniu błędów do tylko negatywnych lub tylko pozytywnych. Problem bez błędów jest rozwiązywalny w czasie wielomianowym

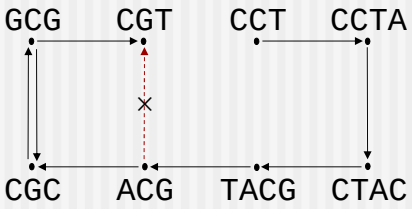
Izotermiczne SBH bez błędów

[J. Błażewicz i M. Kasprzak, *Discrete Appl. Math.* 154 (2006)]

- Algorytm rozwiązujący w czasie wielomianowym problem izotermicznego sekwencjonowania przez hybrydyzację bez błędów w instancji

CCTACGCGT
S = {ACG, CCT, CCTA, CGC, CGT, CTAC, GCG, TACG}

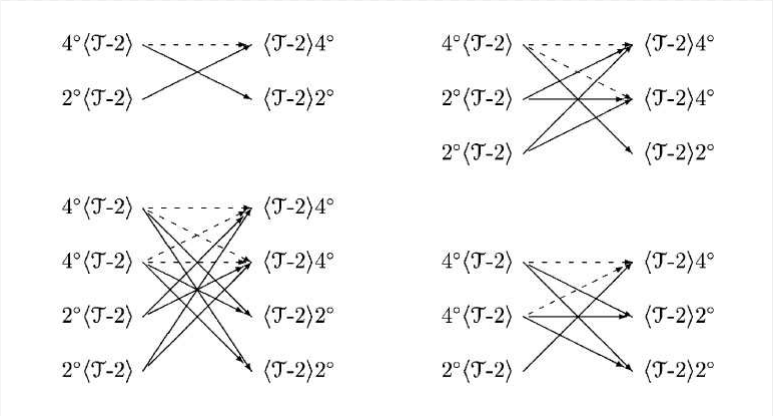
- Krok 1: konstrukcja grafu



Izotermiczne SBH bez błędów

[J. Błażewicz i M. Kasprzak, *Discrete Appl. Math.* 154 (2006)] − cd.

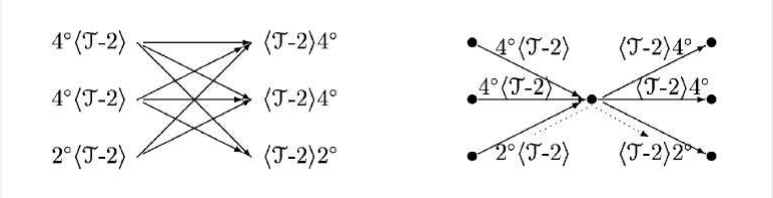
- Krok 2: usunięcie nadmiarowych łuków



Izotermiczne SBH bez błędów

[J. Błażewicz i M. Kasprzak, *Discrete Appl. Math.* 154 (2006)] − cd.

- Krok 3: wstawienie łuków tymczasowych



- Krok 4: transformacja uzyskanego grafu liniowego w graf oryginalny, poszukiwanie ścieżki Eulera zmodyfikowanym algorytmem wielomianowym

Sekwencjonowanie nowej generacji

[M. Ronaghi i in., *Anal. Biochem.* 242 (1996)]

- *Pirosekwencjonowanie* (ang. *pyrosequencing*) to biochemiczna metoda sekwencjonowania DNA polegająca na odczycie fragmentu pojedynczej nici DNA poprzez syntezę wzdłuż tego fragmentu jego komplementarnego odpowiednika
- W trakcie tej syntezy, za każdym razem gdy kolejny nukleotyd A, C, G lub T jest dodawany do nici komplementarnej, uruchamiana jest seria reakcji chemicznych kończących się emisją światła
- Proces ten jest w pełni zautomatyzowany, wydajny i stosunkowo tani. Ograniczeniem była kiedyś długość sekwencjonowanych fragmentów: w artykule 4 cykle A/G/C/T (15 nukleotydów; dwa lata później już 40 cykli)

Sekwencjonowanie nowej generacji

[M. Ronaghi i in., *Anal. Biochem.* 242 (1996)] – cd.

- Ilość światła zależy od liczby nukleotydów tego samego typu dodanych w jednym kroku. Najmniejsza jasność towarzyszy dodaniu pojedynczego nukleotydu
- W jednym kroku dodawane są nukleotydy jednego typu. Po każdym kroku reszta nieprzyłączonych nukleotydów jest wypłukiwana z roztworu i wprowadzane są kolejne innego typu

Sekwencjonowanie nowej generacji

[M. Margulies i in., *Nature* 437 (2005)]

- Firma 454 Life Sciences (później w Roche) zastosowała pirosekwencjonowanie do asemblacji DNA. Cały proces sekwencjonowania odbywa się maszynowo, metody algorytmiczne stosuje się dopiero na etapie asemblacji
- Sekwenator 454 umożliwia odczytanie milionów nukleotydów w jednym przebiegu maszyny. Jest to historycznie pierwsza metoda szybkiego sekwencjonowania
- W wyniku sekwencjonowania otrzymywany jest zbiór sekwencji (o długości nawet ok. 700 nukleotydów) o wysokiej jakości odczytu i stosunkowo dużym pokryciu badanego fragmentu DNA (ok. 30–40 sekwencji na jedną pozycję we fragmencie). Jako dane uzupełniające otrzymuje się „wiarygodność" dla każdego odczytanego nukleotydu

Sekwencjonowanie nowej generacji

[M. Margulies i in., *Nature* 437 (2005)] – cd.

- Wiarygodność nukleotydu wyprowadzana jest z ilości światła emitowanego podczas jego dołączania
- Wiarygodność reprezentowana jest przez liczbę całkowitą z przedziału 0–40. Największe wartości są zarezerwowane dla pojedynczych nukleotydów odczytanych bezproblemowo
- Dla ciągłej sekwencji nukleotydów tego samego typu wiarygodność procesu ich odczytu spada

A	G	C	A	A	T	T	T	A	A	A	A	A	A	T	T
30	30	31	31	26	23	22	03	19	19	18	16	11	03	27	24

Sekwencjonowanie nowej generacji

- Inne popularne technologie sekwencjonowania drugiej generacji: Solexa/Illumina (Illumina), SOLiD (Applied Biosystems)
- Wysoka jakość odczytywanych sekwencji oraz duże pokrycie nimi badanego fragmentu sprzyja sekwencjonowaniu znacznych fragmentów DNA, nawet całych genomów mniejszych organizmów (zwłaszcza na podstawie sparowanych odczytów). Aby maksymalnie wykorzystać naturę tych danych, a także załączony do nich współczynnik wiarygodności nukleotydów, opracowywane były nowe algorytmy specjalizowane
- Rozwijane są technologie produkcji długich odczytów (PacBio, Oxford Nanopore, ponad 10/100 tys. nukleotydów, nawet miliony), gdzie największym wyzwaniem jest redukcja błędów sekwencjonowania