Algorytmy kombinatoryczne w bioinformatyce, wykładowca prof. Marta Kasprzak

Materiały uzupełniające do wykładu 2: sekwencjonowanie cz. 1.

Na części slajdów podane są namiary na artykuł źródłowy opisujący daną metodę. Zachęcam do dalszej lektury osoby, które chciałyby pogłębić swoją wiedzę nt. danej metody.

SLAJD 2

Wyznaczanie sekwencji genomowej organizmów można podzielić na trzy etapy w zależności od długości analizowanej sekwencji: sekwencjonowanie — rozpoznawanie sekwencji nukleotydów we fragmencie genomu o długości zwykle od kilkudziesięciu do kilkuset nukleotydów (mowa o sekwencjonowaniu tradycyjnym i nowej generacji krótkimi odczytami), asemblacja — składanie zsekwencjonowanych fragmentów w dłuższe odcinki (nawet setki milionów nukleotydów), oraz mapowanie — uszeregowanie zasemblowanych sekwencji w obrębie chromosomu lub genomu. Obecnie etap mapowania może być pomijany, gdy produktem asemblacji staje się cały genom. Wyznaczanie sekwencji jest pierwszym krokiem prowadzącym do poznania i zrozumienia genomów i zakodowanych w nich funkcji, zajmuje zatem ważną pozycję w prowadzonych współcześnie badaniach.

SLAJD 3

Laboratoryjne metody wykrywania sekwencji DNA poprzez elektroforezę w żelu – metody Maxama i Gilberta (1977) oraz Sangera i in. (1977) – używane były powszechnie przed wynalezieniem technik wysokoprzepustowych (sekwenatorów nowej generacji). Pozwalały wtedy na zsekwencjonowanie cząsteczek DNA o długości do ok. 100 nukleotydów. Metoda sekwencjonowania przez hybrydyzację akceptowała cząsteczki parokrotnie dłuższe, wymagała za to wspomagania algorytmicznego. Ona również została wyparta przez sekwenatory nowej generacji. Jest dla nas interesująca z tego względu, że po etapie laboratoryjnym wchodzi etap obliczeniowy, dane wyjściowe z eksperymentu biochemicznego muszą zostać złożone algorytmem w celu uzyskania sekwencji wynikowej. Chociaż metoda ta nie jest już wykorzystywana, modele obliczeniowe dla niej opracowane zostały zaadaptowane na etapie asemblacji i są obecnie w powszechnym użyciu.

SLAJD 4

W eksperymencie hybrydyzacyjnym przeprowadzanym w pierwszej fazie procesu sekwencjonowania przez hybrydyzację, w klasycznej jego realizacji, celem jest wykrycie wszystkich oligonukleotydów o zadanej długości (w praktycznych realizacjach 8–12 nukleotydów) składających się na badany łańcuch DNA (kilkaset nukleotydów). W tym celu generowana jest biblioteka oligonukleotydów składająca się ze wszystkich możliwych jednoniciowych fragmentów DNA o zadanej długości (każdy z nich w wielu kopiach). Dla oligonukleotydów o długości l biblioteka taka będzie składała się z 4^l różnych elementów. Biblioteka jest syntetyzowana w sposób systematyczny na płytce o specjalnie przygotowanym podłożu w celu utworzenia mikromacierzy DNA. Kopie tego samego oligonukleotydu umieszczane są w jednym polu mikromacierzy i współrzędne pola dla każdego oligonukleotydu są znane. Każde pole ma wymiary około 25 µm $\times$ 25 µm, rozmiar biblioteki wpływa na rozmiar mikromacierzy i jest to przyczyną, dla której l nie mogło być większe. Następnie mikromacierz wprowadzana jest do roztworu z jedną z nici badanego łańcucha DNA powieloną w reakcji PCR i oznakowaną fluorescencyjnie lub radioaktywnie. Dzięki sprzyjającym warunkom (temperatura, inne parametry roztworu) zajdzie reakcja hybrydyzacji pomiędzy jednoniciowymi krótkimi fragmentami DNA przytwierdzonymi do mikromacierzy i długimi, wolno pływającymi w roztworze jednoniciowymi kopiami badanego DNA — komplementarne fragmenty połączą się w dwuniciowe kompleksy. W ogólności, im dłuższe są takie komplementarne fragmenty, tym większa siła ich przyciągania i większa liczba kopii utworzonych dwuniciowych kompleksów. Po reakcji mikromacierz jest przepłukiwana w celu usunięcia niezwiązanych kopii badanego DNA, pozostaną jedynie te związane z

oligonukleotydami przytwierdzonymi do mikromacierzy. Obraz fluorescencyjny (radioaktywny) mikromacierzy wykaże świecące punkty, tym jaśniejsze, im więcej kopii oligonukleotydu w danym polu przyłączyło się do kopii badanego DNA. Najbardziej jasne punkty z obrazu odpowiadają z grubsza oligonukleotydom przyłączonym do badanego łańcucha na całej ich długości. Wiele kopii oligonukleotydu w każdym polu zwiększa szansę na poprawne wyniki eksperymentalne poprzez statystyczną eliminację niewłaściwych lub niepełnych połączeń. Na podstawie współrzędnych najjaśniejszych punktów ustalany jest zbiór odpowiadających im oligonukleotydów, co do których możemy założyć, że ich odwrotnie komplementarne odpowiedniki wchodzą w skład badanego fragmentu DNA. Te odwrotnie komplementarne odpowiedniki zapisane jako sekwencje nad alfabetem {A, C, G, T} tworzą zbiór (nazywany spektrum) przekazywany do etapu obliczeniowego. Przykładowo, kopie badanej sekwencji CCGACGT przyłączyłyby się do następujących oligonukleotydów z biblioteki o parametrze l=3: ACG, CGG, CGT, GTC, TCG, co dałoby spektrum postaci {ACG, CCG, CGA, CGT, GAC}.

SLAJD 5

Obliczeniowa część procesu sekwencjonowania polega na rekonstrukcji sekwencji nukleotydów badanego łańcucha DNA na podstawie spektrum. W klasycznym ujęciu problemu spektrum jest zbiorem bez powtórzeń. Gdy zakłada się wyidealizowany wariant problemu bez jakichkolwiek błędów eksperymentalnych, o spektrum można powiedzieć, że zawiera wszystkie podciągi badanej sekwencji o długości l i nic więcej. Wtedy długość badanej sekwencji (oznaczana n) nie jest potrzebna do wygenerowania rozwiązania, należy użyć każdy element spektrum dokładnie raz, sąsiednie elementy muszą pokrywać się sufiksem/prefiksem o długości l-1, co w efekcie da zawsze sekwencję o długości n, gdyż bezbłędne (idealne) spektrum zawiera n-l+1 elementów. Jednak nawet w wariancie bez błędów z danego spektrum można uzyskać tym sposobem więcej niż jedną sekwencję wynikową, z informatycznego punktu widzenia wszystkie takie rozwiązania są równie dobre, ale z biologicznego nie, gdyż tylko jedna z tych sekwencji ma znaczenie biologiczne. Ustalenie, które z wielu rozwiązań jest tym właściwym wymaga dostarczenia dodatkowej informacji, np. poprzez przeprowadzenie dodatkowego eksperymentu z badaną cząsteczką.

SLAJD 6

Taki wyidealizowany model eksperymentu hybrydyzacyjnego nie jest jednak realizowalny w praktyce. Jak każdy eksperyment biologiczny, także i ten nie jest wolny od błędów. W sekwencjonowaniu przez hybrydyzację wyróżniamy dwa główne typy błędów: błędy negatywne, czyli brakująca informacja w spektrum, oraz błędy pozytywne, czyli niewłaściwa informacja obecna w spektrum.

SLAJD 7

Wiersze a–d zawierają odpowiednio przykłady błędów opisanych literami a–d na poprzednim slajdzie. W klasycznym modelu SBH (tak jak pierwotnie był ujęty w literaturze) przyjmuje się, że eksperyment z mikromacierzą daje tylko informację o zawieraniu danego oligonukleotydu w badanej sekwencji i że nie można stwierdzić, ile razy w niej wystąpił (jasność punktu na obrazie fluorescencyjnym przekracza pewien założony próg bądź nie). Dlatego nawet w razie braku błędów typowo eksperymentalnych może w spektrum pojawić się błąd wynikający ze składu badanego łańcucha DNA (wiersz a). Taki model jest założony w tym wykładzie. W późniejszych pracach przyjęto, że do pewnego stopnia można określić, ile razy dany oligonukleotyd wystąpił w sekwencji (np. model 1-2-wiele) na podstawie tego, że spośród punktów o jasności przekraczającej założony próg niektóre świecą wyraźnie mocniej niż inne.

SLAJD 8

W wariancie problemu zakładającym błędy negatywne w spektrum należy dopuścić, że sąsiednie słowa w rekonstruowanym uszeregowaniu nie muszą już nakładać się sufiksem/prefiksem o długości l-1, mogą krótszymi odcinkami, w skrajnym przypadku wcale (gdy pomiędzy nimi wystąpi l-1 błędów negatywnych pod rząd). Nadal jednak nakładające się odcinki muszą się idealnie pokrywać. Z kolei

założenie obecności błędów pozytywnych daje podstawę do pominięcia w rekonstrukcji niektórych elementów spektrum. Założenie obu rodzajów błędów naraz, czyli model najbardziej ogólny i pasujący do rzeczywistości, to dopuszczenie obu tych sytuacji. Teraz już znajomość długości badanej sekwencji n staje się konieczna, żeby dopasować liczbę odrzucanych słów do oczekiwanej długości rozwiązania.

SLAJD 9

Bez kryterium optymalizacji problem sekwencjonowania przez hybrydyzację z błędami negatywnymi i pozytywnymi mógłby przełożyć się na rozwiązanie złożone z jednego elementu spektrum albo z serii elementów skonkatenowanych ze sobą do długości n. Nie o takie rozwiązanie nam chodzi, a o rozwiązanie w największym stopniu wykorzystujące informację ze spektrum. Taki efekt uzyskamy, stosując kryterium maksymalizacji liczby elementów spektrum użytych do budowy rozwiązania o długości nie większej niż n (rozwiązanie może być krótsze ze względu na błędy negatywne na końcach sekwencji).

SLAJD 11

Czasami da się ustalić, który element spektrum powinien wystąpić jako pierwszy w rozwiązaniu — na podstawie startera używanego wcześniej w reakcji PCR albo sekwencji rozpoznawanej przez enzym restrykcyjny użyty wcześniej do wycięcia badanego fragmentu DNA. W ogólności jednak, gdy tej wiedzy nie posiadamy, rozrysowane drzewko należy uzupełnić o dodatkowy poziom „zerowy", tutaj pominięty ze względu na brak miejsca. W tym algorytmie do bieżącego elementu dokładany jest kolejny jeszcze nieużyty w danej ścieżce, o ile nakłada się na bieżący na l-1 znakach. Algorytm kończy pracę, jeśli osiągnie ścieżkę o długości n-l+1. Sekwencja nukleotydowa będąca rozwiązaniem problemu odczytywana jest poprzez łączenie etykiet w wynikowej ścieżce z założeniem odpowiedniego pokrycia sufiksów/prefiksów (podobnie w innych omawianych algorytmach).

SLAJD 12

W tym grafie wierzchołki u i v łączymy łukiem (u,v), gdy sufiks o długości l-1 etykiety (oligonukleotydu) wierzchołka u pokrywa się z prefiksem etykiety wierzchołka v. Każda ścieżka Hamiltona w takim grafie to możliwe rozwiązanie problemu (po przetłumaczeniu na sekwencję nukleotydową), gdyż każdy element spektrum jest wtedy użyty dokładnie raz i sąsiednie elementy w rozwiązaniu nakładają się na l-1 znakach.

SLAJD 13

Graf ten tworzymy w ten sposób, że każdy element spektrum staje się łukiem poprowadzonym od prefiksu tego elementu o długości l-1 do jego sufiksu o tej samej długości. Jeśli wierzchołek jest już obecny w grafie, przy wstawianiu następnych łuków wykorzystujemy go i nie tworzymy jego kopii. W tym grafie mogą być wierzchołki niepołączone łukiem, chociaż ich sufiksy/prefiksy nakładają się na l-2 znakach (bo łuków jest tylko tyle, ile elementów spektrum). Każda ścieżka Eulera jest tu rozwiązaniem problemu sekwencjonowania (po przetłumaczeniu na sekwencję nukleotydową).

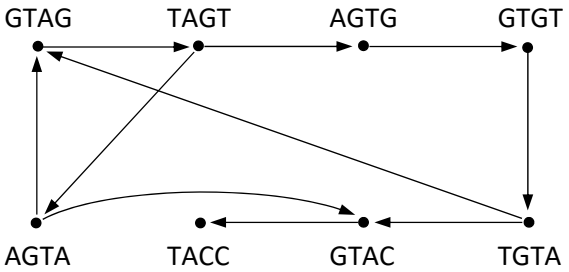
SLAJD 14

Zamiana problemu poszukiwania ścieżki Hamiltona w grafie na problem poszukiwania ścieżki Eulera, bez straty na dokładności rozwiązania, nie jest przypadkiem, obok którego można przejść obojętnie. Gdyby taka wielomianowa transformacja z zachowaniem poprawności rozwiązania była możliwa dla dowolnego grafu, problem ścieżki Hamiltona stałby się rozwiązywalny w czasie wielomianowym, a że należy do klasy problemów NP-trudnych (jego wariant decyzyjny do klasy problemów NP-zupełnych), udowodnione zostałoby, że P=NP (czyli wszystkie problemy NP-zupełne byłyby rozwiązywalne w czasie wielomianowym). Okazało się jednak, że transformacja ta działa tylko dla pewnej podklasy grafów skierowanych, obejmującej m.in. wyszczególnione tu grafy liniowe. Do grafów liniowych (zwanych także krawędziowymi) często odwołujemy się poprzez nawiązanie do ich grafowych odpowiedników (nazwanych tu grafami oryginalnymi), gdyż wierzchołek grafu liniowego reprezentuje

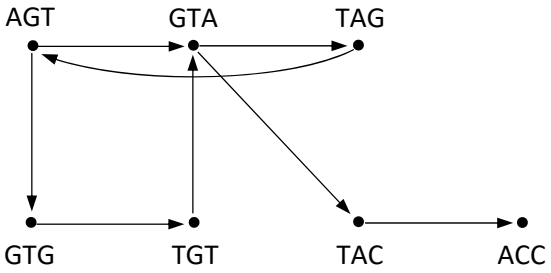
krawędź (łuk) tego drugiego grafu z zachowaniem zasady, że bezpośrednie przejście w grafie liniowym pomiędzy wierzchołkami jest obecne wtedy i tylko wtedy, gdy w jego grafie oryginalnym jest obecne bezpośrednie przejście pomiędzy odpowiednimi krawędziami (łukami). Grafy liniowe są także definiowane w oderwaniu od ich grafów oryginalnych, stosowną definicję zawiera opis drugiego zadania laboratoryjnego. W ten sposób można łatwo stwierdzić, czy dowolny graf skierowany jest grafem liniowym, zatem czy można rozwiązać w nim problem ścieżki Hamiltona w czasie wielomianowym. Graf z metody Lysova i in. jest grafem liniowym grafu z metody Pevznera skonstruowanego dla tego samego spektrum.

SLAJD 15

Graf Lysova dla podanego spektrum:

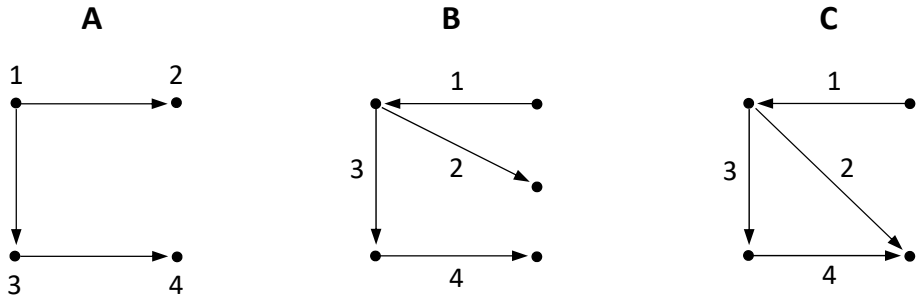


W grafie tym są dwie ścieżki Hamiltona, które przekładają się na dwa rozwiązania problemu sekwencjonowania: AGTAGTGTACC, AGTGTAGTACC. Graf Pevznera dla tego samego spektrum:



Przykładowo, łuk (AGT, GTA) odpowiada elementowi spektrum AGTA. W tym grafie są dwie ścieżki Eulera, które przekładają się na dwie takie same sekwencje jak powyżej.

Gdy wierzchołki są zaetykietowane, transformacja z jednej postaci grafu do drugiej (w dowolną stronę) staje się bardziej czytelna. Ale grafy bez etykiet też w prosty sposób można przetransformować w dowolną stronę. Graf liniowy z grafu oryginalnego tworzymy na podstawie definicji z poprzedniego slajdu: dla każdego łuku grafu oryginalnego tworzymy osobny wierzchołek w grafie liniowym i odtwarzamy połączenia na zasadzie 1-1. Transformacja w drugą stronę odbywa się przez odwrócenie tych operacji: dla każdego wierzchołka grafu liniowego tworzymy łuk w grafie oryginalnym, dbając o to, żeby tak je zaczepić w wierzchołkach, aby odtworzyć wszystkie połączenia z grafu liniowego. Transformacja bez etykiet nie zawsze jest jednoznaczna, dla pewnego grafu liniowego można uzyskać różne postaci grafu oryginalnego. Przykładowo, dla poniższego grafu liniowego A można uzyskać w wyniku transformacji zarówno graf B, jak i C. Obie transformacje są poprawne, ponieważ zachowane są wszystkie połączenia 1-1, czyli w grafie A istnieje łuk (*x*,*y*) wtedy i tylko wtedy, gdy w grafie B/C koniec łuku *x* pokrywa się z początkiem łuku *y*.



SLAJD 16

W grafie z metody Pevznera (slajd 13) zabraknie pewnych łuków, gdy spektrum będzie zawierać błędy negatywne (graf po lewej). Łuki te w dużej mierze można uzupełnić, stosując metodę Pevznera odwołującą się do poszukiwania przepływu w sieci. Sieć ta oparta jest na grafie dwudzielnym $K_{m,m}$, w której następniki s to wierzchołki z grafu po lewej, które mają więcej łuków wchodzących niż wychodzących (gdy różnica pomiędzy tymi łukami jest większa niż 1, wierzchołek w sieci jest powielany odpowiednią liczbę razy); z kolei poprzedniki t to wierzchołki, które mają więcej łuków wychodzących niż wchodzących (również powielanych, gdy różnica jest większa niż 1). Łuki pomiędzy tymi wierzchołkami zaetykietowane zostają wartością minimalnego offsetu (przesunięcia) ich etykiet w dokładnym ich nałożeniu (np. ATTC i TCAG mają minimalny offset 2) i jest to równocześnie liczba łuków, które należałoby wstawić do grafu po lewej, żeby połączyć daną parę wierzchołków (np. do połączenia ATTC z TCAG potrzebne są dwa łuki: od ATTC do TTCA i od TTCA do TCAG). Wartości te to koszty jednostkowego przepływu w sieci, koszty pozostałych łuków ustawiane są na 0, pojemności wszystkich łuków ustawiane są na 1. Należy wyznaczyć przepływ od s do t o wartości m–1 i o minimalnym koszcie, który wskaże dokładnie m–1 łuków z grafu $K_{m,m}$ (czyli warunek na istnienie ścieżki Eulera w grafie mówiący o stopniach wierzchołków zostanie spełniony), a sumaryczny koszt przepływu oznaczać będzie liczbę łuków wstawianych do grafu po lewej. W podanym przykładzie zielona ścieżka to wyznaczony przepływ o koszcie 1, który wskazuje łuk (CT, TG) jako ten, którym należy uzupełnić graf po lewej. W tym akurat przypadku brakujący łuk zrekonstruowany został poprawnie (por. slajd 13).

SLAJD 17

Podejście to okazało się heurystyką, gdyż nie zawsze zadziała zgodnie z intencją autora. Trzeci punkt mówi o sytuacji, gdy w wierzchołku zabraknie równocześnie łuku wchodzącego i wychodzącego. Nie zostanie on wtedy uwzględniony w sieci przepływowej. Nie zawsze musi się to zakończyć niepowodzeniem, gdyż czasami takie brakujące łuki zostaną odtworzone jako część połączenia innej pary wierzchołków.

SLAJD 18

W problemie komiwojażera (w ujęciu grafowym) poszukiwany jest cykl Hamiltona o minimalnym koszcie, gdzie koszt nadawany jest krawędziom. Problem selektywnego komiwojażera pozwala pominąć część wierzchołków (miast) w cyklu. Wierzchołki mają przypisany zysk z ich odwiedzenia i poszukiwany jest cykl maksymalizujący sumaryczny zysk przy ograniczeniu na jego długość. Odpowiada to sytuacji, gdy komiwojażer chce zmaksymalizować swój oczekiwany zarobek przy ograniczonych środkach na podróż. Wariant tego problemu w grafie skierowanym, o którym mowa w metodzie, polega na poszukiwaniu ścieżki zamiast cyklu, każdemu wierzchołkowi nadawany jest zysk równy 1, a koszt w łukach to minimalny offset pomiędzy etykietami wierzchołków (elementami spektrum). Na rysunku po lewej kolory łuków oznaczają różne wartości kosztu: czarny to 1, zielony to 2, niebieski to 3. Poszukiwana jest ścieżka o koszcie nie większym niż n-l i o maksymalnym zysku, czyli w efekcie sekwencja nukleotydowa nie dłuższa niż n włączająca maksymalną liczbę elementów spektrum. Optymalne rozwiązanie zostało przedstawione na rysunku po prawej. W rozwiązaniu

uwzględniony został błąd pozytywny z racji pominięcia wierzchołka GCC i błąd negatywny przez wybór połączenia (TCT, TGG) o koszcie 2.