Algorytmy kombinatoryczne w bioinformatyce, wykładowca prof. Marta Kasprzak

Materiały uzupełniające do wykładu 6: mapowanie.

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

W ramach tego wykładu skupimy się na punkcie pierwszym i czwartym.

SLAJD 3

Mapowanie jako etap procesu składania sekwencji genomowej tym różni się od pierwszych dwóch etapów (sekwencjonowania i asemblacji), że nie porządkuje sekwencji przez porównanie ich ciągów nukleotydowych, tylko na podstawie innych wyników eksperymentalnych. Robi się tak między innymi dlatego, że czasami uporządkować chcemy odcinki genomu, a potem na podstawie tego wyłuskać niektóre z nich do dalszych badań, kiedy jeszcze ich sekwencji nukleotydowych nie znamy. Obecnie etap mapowania często jest pomijany w procesie składania sekwencji genomowej, gdyż za pomocą masowego sekwencjonowania nowej generacji i późniejszej asemblacji z użyciem nowoczesnych komputerów rozpoznawane są nawet całe genomy.

SLAJD 4

W podejściu mapowania przez hybrydyzację wykorzystuje się znaną nam już z sekwencjonowania skłonność jednoniciowych fragmentów DNA do wiązania się w dwuniciowe kompleksy (hybrydyzację) z komplementarnymi do nich odcinkami drugiej nici DNA. Jeśli zestawimy w ten sposób krótką jednoniciową próbkę, która z dużym prawdopodobieństwem ma unikalną w skali badanego wycinka genomu sekwencję nukleotydów, ze zbiorem dłuższych jednoniciowych fragmentów DNA, to informacja o tym, że dwa lub więcej fragmentów związało się z daną próbką daje podstawy przypuszczać, że pokrywają one to samo miejsce w genomie, zatem nakładają się na siebie. Eksperyment z wieloma różnymi unikalnymi (przypuszczalnie) próbkami pozwala uszeregować względem siebie cały zbiór fragmentów.

SLAJD 5

Jest to wariant teoretyczny problemu, gdzie zakłada się brak błędów hybrydyzacji oraz unikalność próbek w ramach badanego wycinka genomu.

SLAJD 6

Własność *consecutive ones* w wierszach macierzy oznacza, że da się tak poprzestawiać kolumny macierzy, żeby uzyskać dla wszystkich wierszy efekt występowania jedynek w co najwyżej jednym spójnym bloku pod rząd. Taką własność ma każda macierz reprezentująca bezbłędną informację o hybrydyzacji w tym wariancie problemu (dla tego przykładu jest nią macierz ze slajdu 5). Na slajdzie pokazana jest macierz z odpowiednio poprzestawianymi kolumnami i zwizualizowane na jej podstawie uporządkowanie względem siebie fragmentów, które wprost można odczytać w macierzy z ciągów jedynek. Dwa wyróżnione punkty odpowiadają dwóm niebieskim elementom macierzy. Kolejność kolumn odpowiadających próbkom (b, d, a, c) to przypuszczalna kolejność występowania próbek w badanym wycinku genomu. Informacja z macierzy nie daje nam podstaw do ustalenia, gdzie dokładnie kończą się poszczególne fragmenty, rysunek jest pewnym przybliżeniem. Wiadomo na przykład, że fragmenty 1 i 2 pokrywają dwie pierwsze próbki i że fragment 2 rozciąga się dalej na prawo poza fragmentem 1, ale fragment 1 może zaczynać się w genomie zarówno przed początkiem fragmentu 2, jak i po nim. Odtworzenie właściwej kolejności kolumn dla macierzy posiadającej własność *consecutive ones* w wierszach jest problemem łatwym obliczeniowo (algorytm Bootha i Luekera). Uporządkowanie fragmentów można też (choć nie zawsze) ustalić na drodze rozwiązania problemu poszukiwania ścieżki Hamiltona w grafie interwałowym (problem ten jest rozwiązywalny w

czasie wielomianowym w grafach interwałowych). Graf interwałowy pokazany na slajdzie skonstruowany został dla macierzy ze slajdu 5, wierzchołki odpowiadają fragmentom, a krawędzie łączą dwa wierzchołki, jeśli odpowiednie dwa fragmenty nakładają się (współdzielą próbkę). Ścieżka Hamiltona w takim grafie, o ile istnieje, odpowiada z grubsza jednemu z możliwych rozwiązań tego problemu. W przykładowym grafie są dwie ścieżki Hamiltona, (1, 3, 2, 4) i (3, 1, 2, 4), pierwsza z nich może odpowiadać rozwiązaniu jak na slajdzie, druga sekwencji kolumn (a, d, b, c). Instancję problemu, dla której istnieje rozwiązanie, ale nie istnieje ścieżka Hamiltona w utworzonym na jej podstawie grafie interwałowym, można uzyskać poprzez zamianę 1 na 0 w przykładowej macierzy w polu [1, d].

SLAJD 7

Chimera oznacza tutaj rodzaj błędu, kiedy dwa jednoniciowe odległe fragmenty sklejają się ze sobą przypadkowo w jeden kompleks, przełoży się to na wiersz macierzy z dwoma rozłącznymi blokami jedynek w optymalnym ustawieniu. Błąd negatywny to brak sygnału hybrydyzacji, choć powinien wystąpić (0 w macierzy w miejsce 1), błąd pozytywny to błędny sygnał hybrydyzacji (1 w miejsce 0). Problem mapowania przez hybrydyzację z dopuszczeniem błędów eksperymentalnych jest już trudny obliczeniowo. Zaprezentowana przykładowa macierz jest macierzą ze slajdu 5 z dwoma błędami (na niebiesko), choć tylko jeden z nich wprowadza zaburzenie we własności *consecutive ones*.

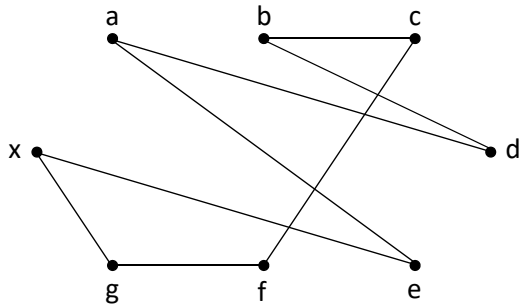
SLAJD 8

Autorzy sprowadzają problem do poszukiwania cyklu komiwojażera, a że w pierwotnym problemie mamy uszeregowanie liniowe kolumn, domykają je, wprowadzając dodatkową sztuczną kolumnę x wypełnioną samymi zerami. Kolumny odpowiadają wierzchołkom w grafie pełnym, odległość pomiędzy nimi to odległość Hamminga pomiędzy kolumnami. Cykl komiwojażera minimalizuje liczbę niezgodności pomiędzy sąsiednimi kolumnami, a tego spodziewamy się w optymalnym uszeregowaniu. Cykl przerywany jest w wierzchołku x i otrzymujemy wynikowe uszeregowanie kolumn (czytane od dowolnej strony), na rysunku takie samo jak na slajdach wcześniejszych. Podejście jest heurystyczne już na etapie transformacji problemu, gdyż optymalny cykl utworzony z uwzględnieniem odległości do wektora x (niezależnie od jego zawartości) nie musi przekładać się na optymalną ścieżkę po przerwaniu cyklu w tym miejscu.

SLAJD 9

Odległość Hamminga wyliczona dla kolumn przykładowej macierzy oraz jeden z optymalnych cykli komiwojażera:

	b	c	d	e	f	g	x
a	2	3	1	0	2	2	2
b	1	1	2	2	2	4	
c	2	3	3	3	3		
d	1	3	3	3			
e	2	2	2				
f	0	2					
g	2						



Wynikowe uszeregowanie kolumn macierzy po przerwaniu cyklu w miejscu x: (e, a, d, b, c, f, g).

Macierz z kolumnami w powyższym ustawieniu:

0	0	0	1	1	1	1
1	1	1	1	0	1	1
0	0	1	1	1	0	0
1	1	1	1	1	0	0

SLAJD 11

W problemie założono brak błędów eksperymentalnych, natomiast próbki już nie muszą być unikalne. Podany przykład jest niewygodny w rozwiązywaniu z tego względu, że jest mało próbek, często powtarzają się w badanym wycinku „genomu" i dają w ten sposób przypadkowe podobieństwo wektorów odpowiadających odległym fragmentom (np. fragmenty 1 i 5).

SLAJD 12

Przykład zastosowania heurystyki zachłannej do rozwiązania problemu. W pary łączone są albo pojedyncze fragmenty, albo fragment z już złączoną wcześniej grupą – „kontigiem" (zapisaną w przykładzie jako indeksy w nawiasie), albo dwie grupy. Heurystyka zwróciła rozwiązanie trochę inne niż oczekiwane (tu w szarym kolorze dla porównania), należy je czytać: fragmenty 6 i 7 z jednego końca cząsteczki, pozostałe pogrupowane w kolejności przyłączania, czyli (((1+5)+4)+2)+3).

SLAJD 13

Mapowanie z użyciem enzymów restrykcyjnych (mapowanie restrykcyjne) wykorzystuje enzymy do przecięcia kopii badanego wycinka genomu w różnych miejscach, tak by wynikowe fragmenty pochodzące z różnych kopii nakładały się na siebie. Enzymy restrykcyjne rozpoznają specyficzne dla siebie miejsca w dwuniciowym DNA, obejmujące zaledwie kilka sąsiednich par zasad, i przecinają łańcuch (trawią) w rozpoznanym miejscu lub obok niego. Enzymy rozpoznające rzadziej występujące miejsca restrykcyjne (zwykle te dłuższe) potną genom rzadziej, inne częściej, konkretny rodzaj enzymu dobierany jest do długości badanego DNA i oczekiwanej liczby fragmentów, jakie się pragnie uzyskać. Przykładowo, enzym Alu I rozpoznaje dwuniciowy fragment AGCT/ AGCT (tzn. AGCT w jednej nici i jej odwrotnie komplementarny odpowiednik o tej samej sekwencji w drugiej nici) występujący często, enzym Not I rozpoznaje znacznie rzadziej występujący fragment GCGGCCGC/GCGGCCGC. Trawienie przeprowadzane w celu późniejszego mapowania fragmentów odbywa się najczęściej z użyciem dwóch enzymów restrykcyjnych lub jednego, za to w różnych przedziałach czasowych. Długość tak uzyskanych fragmentów mierzona jest za pomocą elektroforezy żelowej i multizbiór lub multizbiory tych długości stanowią instancję problemu. Mapowanie restrykcyjne tym różni się od mapowania przez hybrydyzację, że pozbawione jest jakiejkolwiek informacji o nakładaniu się fragmentów.

SLAJD 14

Multizbiór *A* zawiera długości fragmentów powstałych w wyniku trawienia kopii badanej cząsteczki pierwszym enzymem. Większość kopii zostanie przecięta we wszystkich miejscach rozpoznawanych przez ten enzym (na rysunku zaznaczone na niebiesko) i efekt będzie możliwy do zaobserwowania na obrazie z elektroforezy. Druga reakcja przeprowadzana jest z kopiami cząsteczki i drugim enzymem, dokona on cięcia ponownie w niemal wszystkich miejscach dla niego charakterystycznych (na rysunku na czerwono) i długości tych fragmentów utworzą multizbiór *B*. Multizbiór *AB* jest efektem trzeciej reakcji, gdzie trawienie kopii cząsteczki dokonywane jest równocześnie dwoma enzymami. Elementy w obrębie każdego z tych multizbiorów należy uporządkować w ten sposób, żeby dla wszystkich multizbiorów uzyskać całkowicie zgodne położenie w obrębie cząsteczki zidentyfikowanych miejsc restrykcyjnych.

SLAJD 15

Błędami w tym problemie (i w kolejnych wariantach mapowania restrykcyjnego) są głównie błędy pomiaru długości fragmentów (elektroforeza umożliwia pomiar z pewnym przybliżeniem) i błędy negatywne, czyli braki niektórych elementów w multizbiorach. Długości fragmentów uzyskiwanych w rzeczywistych eksperymentach są znacznie większe niż w tym prostym przykładzie, mają zatem zwykle widoczną różnicę w długości i pewna korekta błędów jej pomiaru jest możliwa.

SLAJD 16

Gdybyśmy przeprowadzili reakcję trawienia jednym enzymem tak jak w problemie powyższym, uzyskalibyśmy rozłączne fragmenty bez informacji umożliwiającej ich uporządkowanie. Tutaj używany jest jeden tylko enzym, ale przeprowadzana jest seria reakcji w różnych przedziałach czasowych. Im dłuższy czas reakcji, w której DNA poddawane jest trawieniu, tym większa szansa na odszukanie i przecięcie miejsc restrykcyjnych rozpoznawanych przez enzym. Jeśli reakcja będzie trwała wystarczająco długo, cząsteczki enzymu odszukają zdecydowaną większość miejsc restrykcyjnych we wszystkich kopiach badanego wycinka DNA i dokonają trawienia. Jeśli reakcja będzie trwała skrajnie krótko, tylko niektóre miejsca w niektórych kopiach zdążą zostać odszukane, w efekcie zaobserwujemy kopie nieprzecięte w ani jednym miejscu, kopie przecięte w jednym, ale różne w różnych miejscach, także małą liczbę kopii przeciętą w większej liczbie miejsc naraz. Kilka reakcji przeprowadzonych w różnym czasie – najkrótszym, najdłuższym i pośrednich – daje fragmenty rozpięte nie tylko pomiędzy sąsiednimi miejscami restrykcyjnymi w badanym DNA, ale pomiędzy (w optymalnej sytuacji) każdą parą miejsc restrykcyjnych, wliczając w to także oba końce cząsteczki. Ze względu na współdzielenie tych samych fragmentów przez zbiory wynikowe „sąsiednich" reakcji, długości fragmentów nie są rozdzielane na rozłączne multizbiory, tylko na wyjściu podawany jest jeden multizbiór z wszystkimi długościami. W idealnym przypadku multizbiór *A* zawiera dokładnie $\binom{k+2}{2}$ elementów, gdzie k jest liczbą miejsc restrykcyjnych.

SLAJD 17

Prawie zawsze bezbłędny multizbiór *A* umożliwia dopasowanie do niego jednej tylko mapy miejsc restrykcyjnych, czyli takiego położenia miejsc, żeby odległości pomiędzy wszystkimi parami punktów (miejsca restrykcyjne plus końce cząsteczki) pokryły się idealnie z wartościami z *A*. Literalnie mamy dwa rozwiązania na każdą mapę, gdyż można ją czytać od lewej strony do prawej lub odwrotnie. Warianty problemu z rozmaicie ograniczanymi modelami błędów są trudne obliczeniowo.

SLAJD 18

Ponieważ eksperyment z problemu częściowego trawienia jest kłopotliwy w przeprowadzeniu i produkuje instancje z dużym odsetkiem błędów negatywnych w stosunku do idealnego multizbioru *A*, zaproponowany został uproszczony wariant tego eksperymentu. Okazało się, że informacja płynąca jedynie z dwóch reakcji, najkrótszej i najdłuższej, często jest wystarczająca nawet do jednoznacznej rekonstrukcji mapy. W bezbłędnym wariancie problemu multizbiór *A* zawiera długości fragmentów będących rezultatem trawienia badanej cząsteczki w dokładnie jednym miejscu (cząsteczki nieprzecięte są łatwe do odsiania na etapie elektroforezy na podstawie długości, a przecięte w wielu miejscach da się odsiać na podstawie małej liczby wystąpień), multizbiór *B* we wszystkich miejscach naraz. Elementy multizbioru *A* łatwo dobierane są w pary sumujące się długością do długości całej cząsteczki, problem stanowi takie ich uporządkowanie, aby w sumie złożyły się na mapę zgodną z multizbiorem *B*. To wystarczyło, żeby problem okazał się silnie NP-trudny.

SLAJD 20

Odpowiadające sobie fragmenty genomów pochodzące od zbliżonych genetycznie osobników charakteryzują się podobnym profilem cięć enzymami restrykcyjnymi (lub, w ogólności, podobnymi rezultatami różnych innych eksperymentów biologicznych). Taki genetyczny „odcisk palca" jest

przydatny na przykład w identyfikacji materiału genetycznego lub do konstrukcji drzew filogenetycznych oddających podobieństwo różnych organizmów. W omawianym tutaj podejściu porównanie dwóch próbek materiału biologicznego dokonywane jest na podstawie zbiorów długości fragmentów uzyskanych w trawieniu enzymami wraz z przypisanym im identyfikatorem danego enzymu. Fragmenty pozyskiwane są nieco inaczej niż w poprzednim zastosowaniu, tutaj wszystkie kopie trawionej cząsteczki mają znakowany (fluorescencyjnie bądź radioaktywnie) jeden, ten sam koniec, co umożliwia wyłuskanie ze wszystkich pociętych fragmentów tych, które zaczynają się w tym końcu (pozostałe nas nie interesują), a to z kolei umożliwia łatwe odłożenie miejsc restrykcyjnych na osi reprezentującej mapę.

SLAJD 21

Jak takie sekwencje A i B mogą wyglądać, można zobaczyć na slajdzie 25 wraz z ich cząsteczkami źródłowymi.

SLAJD 22

Przekształcenie jednej sekwencji w drugą odbywa się trochę na wzór przekształcenia sekwencji w algorytmie programowania dynamicznego dla problemu dopasowania globalnego dwóch sekwencji (wykład 4).

SLAJD 23

Fragment algorytmu w szarym kolorze nie jest tak naprawdę potrzebny, gdyż realizowany jest następną instrukcją.

SLAJD 24

Rozwiązanie w wypełnionej macierzy odczytujemy poprzez odtwarzanie decyzji, które doprowadziły do wypełnienia skrajnie prawego pola ostatniego wiersza macierzy. Cofamy się od tego pola do pola [0,0], tutaj ścieżka ta została zaznaczona na żółto. Dopasowanie profili cięć odpowiadające tej ścieżce pokazane jest na slajdzie 25.