Algorytmy kombinatoryczne w bioinformatyce, wykładowca prof. Marta Kasprzak

Materiały uzupełniające do wykładu 3: sekwencjonowanie cz. 2.

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 sekwencjonowaniu przez hybrydyzację istotnym problemem jest niejednoznaczność rozwiązania możliwego do złożenia na podstawie spektrum, czyli gdy więcej niż jedna sekwencja wynikowa równie dobrze reprezentuje wyniki eksperymentu hybrydyzacyjnego. Przykładowo, dla spektrum z przykładu ze slajdu 15 wykładu 2 można uzyskać dwie ścieżki Hamiltona w grafie Lysova lub dwie ścieżki Eulera w grafie Pevznera; bez dodatkowej informacji (nieobecnej w instancji problemu) nie sposób wskazać, która z sekwencji wynikowych jest tą właściwą, badaną sekwencją DNA. Problem identyfikacji sekwencji poprawnej z punktu widzenia biologicznego próbowano rozwiązać na różne sposoby, poprzez dostarczanie dodatkowej informacji umożliwiającej wyłonienie najbardziej prawdopodobnego rozwiązania. Wieloetapowe sekwencjonowanie przez hybrydyzację jest jednym z takich podejść, eksperyment hybrydyzacyjny identyczny jak w klasycznym podejściu SBH uzupełniany jest o serię pojedynczych reakcji hybrydyzacji przeprowadzanych pod konkretny graf uzyskany dla danego spektrum. Jeśli w grafie istnieją rozwidlenia, w których można obrać różne drogi i w efekcie uzyskać różne sekwencje wynikowe, rozwidlenia takie po kolei są rozwiązywane poprzez konstruowanie oligonukleotydów o długości większej niż l i zestawianie ich z badanym jednoniciowym DNA. Przykładowo, dwa łuki (TGTA, GTAC) i (TGTA, GTAG) przekładają się na rozwidlenie w wierzchołku TGTA; poprawną drogę może wskazać wynik hybrydyzacji badanej cząsteczki DNA z fragmentem komplementarnym do TGTAC (czyli GTACA), jeśli byłby on negatywny, to właściwym przejściem w grafie będzie (TGTA, GTAG) i łuk (TGTA, GTAC) może zostać usunięty.

SLAJD 3

Im większa długość oligonukleotydów zamieszczonych na mikromacierzy DNA, tym bardziej jednoznaczne rozwiązanie. Dłuższe słowa, którymi próbkujemy sekwencję, mają większą szansę wystąpić jednokrotnie w tej sekwencji (eliminacja błędów negatywnych wynikających z powtórzeń, mniej zapętleń w grafie), dłuższe oligonukleotydy współtworzą też dłuższe dwuniciowe kompleksy, czyli hybrydyzacja jest silniejsza (mniej błędów negatywnych eksperymentalnych). (Może wzrosnąć liczba błędów pozytywnych, ale błędy pozytywne w dużo mniejszym stopniu niż negatywne utrudniają znalezienie rozwiązania.) Jednak ze względu na ograniczenia technologii mikromacierzowej kompletna biblioteka oligonukleotydów nie mogła być generowana dla wystarczająco dużego l. Podejście wieloetapowe było bardzo pracochłonne, proponowane więc były inne podejścia, które zwiększały jednoznaczność rozwiązania przy niezwiększaniu rozmiaru biblioteki, omówione na kolejnych slajdach.

SLAJD 4

Nukleotyd uniwersalny był sztucznym tworem, który mógł być wplatany w łańcuchy nukleotydowe ze zwykłymi nukleotydami i mógł w dwuniciowych kompleksach występować naprzeciwko dowolnego nukleotydu („łączyć" się z nim). Zastosowanie go w bibliotece oligonukleotydów pozwoliło zwiększyć długość oligonukleotydów przy niezwiększaniu rozmiaru biblioteki. Weźmy dla przykładu bibliotekę tetranukleotydów (oligonukleotydów o długości 4), w klasycznej realizacji mamy ich 4^4. W realizacji z nukleotydami uniwersalnymi dokładnie tyle samo; dla schematu (2,2)-probes zamiast np. TTAG mielibyśmy TTUAUG, który to oligonukleotyd dostarcza informację o wystąpieniu w badanym łańcuchu DNA nadal tylko czterech nukleotydów, z tym że przeplatanych „spacjami": TT_A_G. Pozornie ilość informacji uzyskiwanej z takiej biblioteki pozostaje niezmieniona, jednak w rzeczywistości wydłużone oligonukleotydy pozwalają sięgnąć dalej i łatwiej rozwiązać rozgałęzienia w

grafie (podobny efekt obserwujemy obecnie na etapie asemblacji na podstawie sparowanych odczytów, tam też im większy odstęp pomiędzy odcinkami informacji, tym bardziej jednoznaczne rozwiązanie).

SLAJD 5

Autorzy nie konstruowali grafu w swoim algorytmie, tutaj dla wygody i porównania z wcześniejszym podejściem tak ta instancja została zwizualizowana. Teraz rozwiązaniem nie jest każda ścieżka Eulera, tylko taka, która na całej swojej długości nie powoduje konfliktów w nałożeniach oligonukleotydów (tzn. dwa niepasujące nukleotydy w tej samej kolumnie dopasowania, gdzie U pasuje do wszystkich). Taka ścieżka Eulera nie jest już znajdowana w czasie wielomianowym.

SLAJD 6

Nukleotydy uniwersalne z czasem okazały się niewystarczająco praktyczne w użyciu. Jeden z autorów tamtego podejścia kontynuował badania w tym kierunku i zaproponował podejście alternatywne, łatwiejsze w realizacji, które z informacyjnego punktu widzenia było prawie równoważne. W jednym polu mikromacierzy umieszczane były wszystkie oligonukleotydy pasujące do schematu danej „gapped probe", czyli np. TT_A_G, tylko w miejscach spacji zamiast nukleotydów uniwersalnych były wstawiane wszystkie kombinacje nukleotydów standardowych. Czyli w polu dla powyższej próbki byłyby umieszczone oligonukleotydy (w wielu kopiach): TTAAAG, TTAACG, TTCACG, TTTATG, itd. Świecący punkt na obrazie fluorescencyjnym mikromacierzy oznacza tutaj, że badany łańcuch DNA zawiera w sobie odcinek komplementarny do któregoś oligonukleotydu z powyższego zestawu, nie wiadomo, do którego, ale z pewnością można powiedzieć, że do któregoś w schemacie TT_A_G. Jednak takie podejście miało z kolei inną wadę, normalnie w polu mikromacierzy mielibyśmy X kopii tego samego oligonukleotydu, dla powyższego przykładu mamy 16 różnych oligonukleotydów w jednym polu w liczbie X/16 każdy. Siła hybrydyzacji i sygnał fluorescencyjny drastycznie maleją z liczbą spacji we wzorcu, co pozwala wstawić bardzo małą ich liczbę i przekłada się na zwiększenie liczby błędów eksperymentalnych. Dodatkowo siła wiązania w dwuniciowych kompleksach pary nukleotydów C i G (trzy wiązania wodorowe) nie jest taka sama jak pary A i T (dwa wiązania), co też wpływa na reakcję hybrydyzacji. Korekta tego podejścia (ostatni punkt na slajdzie) polegała na umieszczeniu w jednym polu mikromacierzy na danej pozycji zdegenerowanej tylko jednego z dwóch nukleotydów, albo A/T, albo C/G. Jednak wtedy liczba pól mikromacierzy wzrasta względem klasycznego podejścia.

SLAJD 7

Różna siła wiązania w dwuniciowych kompleksach (dupleksach) pary C/G i pary A/T powoduje błędy w hybrydyzacji także w klasycznej realizacji podejścia SBH. Bierze się to stąd, że inne są idealne warunki zajścia hybrydyzacji (głównie chodzi o temperaturę) pomiędzy parą jednoniciowych DNA, gdy mają one większy odsetek nukleotydów C/G vs. A/T (w pierwszym przypadku jest więcej pojedynczych wiązań wodorowych przypadających na odcinek o tej samej liczbie nukleotydów). Reakcja z mikromacierzą przeprowadzana jest w pewnej temperaturze, wspólnej dla całej mikromacierzy, która nie może być optymalna dla wszystkich oligonukleotydów z biblioteki. Wczesne modele przyjmowały uproszczone przybliżenie (z grubsza poprawne w pewnym zakresie długości oligonukleotydów), że każda para C/G wnosi do całościowej optymalnej temperatury zajścia hybrydyzacji cztery stopnie, każda para A/T dwa stopnie. (W rzeczywistości na optymalną temperaturę ma wpływ większa liczba czynników, m.in. najbliżsi sąsiedzi danego nukleotydu w nici.) Temperatura ta zwana jest także temperaturą topnienia dupleksów nukleotydowych, gdyż jest ona graniczna dla procesu łączenia/rozdzielania nici dupleksu. Propozycja zastosowania izotermicznej biblioteki oligonukleotydów uwzględnia tę okoliczność i polega na ujęciu w ramach jednej biblioteki wszystkich możliwych oligonukleotydów, które posiadają tę samą temperaturę wyliczoną wg powyższego przybliżenia. Będą się one wtedy różnić znacznie długością, najkrótsze oligonukleotydy z biblioteki będą dwa razy krótsze niż te najdłuższe.

SLAJD 9

Przykładowo, biblioteka o temperaturze 10 stopni będzie zawierała oligonukleotydy posiadające zawsze co najmniej jeden nukleotyd A/T. Dłuższy ciąg wystąpień samych nukleotydów C/G w badanym łańcuchu DNA nie zostanie pokryty żadnym oligonukleotydem z takiej biblioteki. Do zsekwencjonowania dowolnego łańcucha potrzebne są dwie biblioteki o „sąsiednich" temperaturach i użycie takiej pary bibliotek założone jest w izotermicznym podejściu SBH.

SLAJD 10

W problemie izotermicznego sekwencjonowania przez hybrydyzację, z racji użycia dwóch bibliotek o „sąsiednich" temperaturach, niektóre elementy spektrum będą zawierały się w innych (przesunięte do lewej lub prawej ich strony), w związku z czym sekwencję rekonstruować należy z założeniem, że sąsiednie elementy spektrum w uszeregowaniu nakładają się z przesunięciem (offsetem) o 0 znaków, o 1 znak, lub w razie wystąpienia błędów negatywnych o większą liczbę znaków.

SLAJD 11

Minimalizacja liczby błędów w rozwiązaniu z grubsza odpowiada maksymalizacji liczby elementów spektrum użytych do konstrukcji rozwiązania, czyli kryterium przyjętym w klasycznym SBH. Złożoność obliczeniowa odpowiednich wariantów problemu jest taka, jak w klasycznym SBH.

SLAJD 12

Algorytm dla wariantu bez błędów w spektrum stosuje transformację grafu z postaci, w której poszukiwana jest ścieżka Hamiltona do postaci, w której poszukiwana jest ścieżka Eulera (transformacja grafu liniowego do jego grafu oryginalnego), czyli coś co znamy z poprzedniego wykładu. Tym razem jednak nie jest to tak proste jak w przypadku podejść Lysova i in. i Pevznera. Graf konstruowany jest inaczej i nie jest z początku grafem liniowym, należy go doprowadzić do tej postaci i dopiero wtedy można przeprowadzić transformację. Wierzchołki grafu reprezentują elementy spektrum. Jeśli jeden element zawarty jest w innym, łączone są na zasadzie wyłączności: łukiem od krótszego do dłuższego, jeśli zawarty jest do lewej jego strony, od dłuższego do krótszego, jeśli do prawej, i nie są w grafie dopuszczone już żadne inne alternatywne połączenia dla tych dwóch wierzchołków (odpowiednio wychodzące bądź dochodzące). Pozostałe wierzchołki są łączone, jeśli ich etykiety mają tę samą długość i nakładają się na siebie z przesunięciem o 1 znak, pod warunkiem jednak, że takie połączenie nie generuje błędów w rozwiązaniu. Takie błędne połączenie zaznaczone jest na rysunku na czerwono, ACG i CGT nakładają się na siebie z przesunięciem o 1 znak, jednak takie połączenie dałoby w sekwencji wynikowej fragment ACGT, który przynależy do biblioteki o temperaturze 12 stopni (użytej w tym przykładzie) i w bezbłędnym wariancie problemu musiałby być obecny w spektrum (a nie jest). Skonstruowany graf, ze względu na późniejsze etapy algorytmu, nie może mieć żadnych łuków dochodzących do pierwszego elementu w rozwiązaniu i żadnych wychodzących z ostatniego elementu, a że nie znamy w ogólności końców rozwiązania, za takie przyjmujemy po kolei wszystkie pary i procedurę uruchamiamy $O(n^2)$ razy.

SLAJD 13

Niektóre łuki w utworzonym grafie na pewno nie wejdą w skład żadnego rozwiązania, a stoją na drodze do uczynienia grafu liniowym. Na rysunku zaznaczone są takie przypadki liniami przerywanymi. Gdyby któryś z tych łuków został wybrany, nie udałoby się poprowadzić ścieżki łączącej wszystkie elementy rozwiązania, dlatego wszystkie takie łuki usuwamy. Zastosowany wzorzec opisu oligonukleotydu $4°\langle\mathfrak{I}\text{-}2\rangle$ należy czytać jako każdy oligonukleotyd rozpoczynający się nukleotydem czterostopniowym (czyli C lub G), po którym następuje dalsza część o temperaturze $\mathfrak{I}$-2 (tu 8 stopni).

SLAJD 14

Jedyną przeszkodą, która stoi na drodze do uczynienia grafu liniowym i przeprowadzenia transformacji, jest brak łuku od 2°⟨ℑ-2⟩ do ⟨ℑ-2⟩2° w podgrafie z kroku 3 (o ile taki podgraf w ogóle występuje w danym grafie). Dodajemy więc taki łuk do grafu. W ogólności dodanie łuku może zmienić postać rozwiązania, gdyż umożliwiamy połączenie, którego w instancji wcześniej nie było. Tutaj jednak możemy zagwarantować, że tak dodany łuk nigdy nie zostanie użyty i pozostaniemy przy wielomianowej złożoności algorytmu. Mamy już graf liniowy i przeprowadzamy transformację w ten sposób, że każdy wierzchołek tego grafu staje się łukiem po transformacji i w nowym grafie istnieje przejście od łuku a do łuku b wtedy i tylko wtedy, gdy w grafie sprzed transformacji istnieje przejście od wierzchołka a do b. Możemy już poszukiwać ścieżki Eulera w nowym grafie. Podgraf z rysunku po lewej po transformacji staje się podgrafem z prawej strony, w którym należy zawsze pamiętać o tym, żeby nie wykorzystywać wykropkowanego połączenia. Jest to łatwo osiągalne, po wejściu po raz pierwszy do wierzchołka centralnego tego podgrafu od strony łuku 2°⟨ℑ-2⟩ należy wybrać któryś z łuków ⟨ℑ-2⟩4°, jeśli z kolei pierwszy raz wchodzimy od strony 4°⟨ℑ-2⟩, należy wyjść łukiem ⟨ℑ-2⟩2°.

SLAJD 15

Wcześniejsze techniki sekwencjonowania, laboratoryjne na żelu i SBH w różnych odmianach, zostały wyparte przez sekwencjonowanie nowej generacji, zautomatyzowane, wysokoprzepustowe i niewymagające etapu algorytmicznego. Pirosekwencjonowanie zaproponowane w 1996 r. stało się podstawą pierwszej technologii sekwencjonowania wysokoprzepustowego, tzw. sekwencjonowania 454 (od nazwy firmy 454 Life Sciences) uruchomionego w 2005 r. Obecnie już ta technologia nie jest wspierana.

SLAJD 16

Na rysunku jasność znaku ma obrazować mniejszą bądź większą emisję światła. Na podstawie obserwowanej w eksperymencie jasności można w przybliżeniu wywnioskować, ile nukleotydów danego rodzaju pod rząd przyłączyło się w danym kroku.

SLAJD 18

Pokazany fragment sekwencji i wiarygodności poszczególnych nukleotydów to fragment wyjścia z sekwenatora 454 z rzeczywistego eksperymentu. O ile pojedyncze nukleotydy danego rodzaju mają dość wysokie wartości wiarygodności odczytu, to niejasne są niskie wartości pierwszych nukleotydów A z serii sześciu — skoro emisja światła była na tyle duża, żeby przyjąć wystąpienie w tym miejscu sześciu albo pięciu nukleotydów pod rząd (gdyż szósty ma znikomą wiarygodność odczytu), to tym bardziej pewne staje się wystąpienie np. pierwszych trzech (tutaj wycenionych tylko na 18-19).

SLAJD 19

Odczytem jest nazywana pojedyncza sekwencja odczytana przez sekwenator. Odczyty z sekwenatora wchodzą na wejście algorytmów asemblacji (omawianych w ramach wykładu 5) wraz z wiarygodnościami, które mogą posłużyć do odsiania odczytów gorszej jakości lub do ich korekcji w słabszych miejscach. Obecnie najczęściej stosuje się protokoły sekwencjonowania z odczytami sparowanymi, gdzie na wyjściu sekwenatora podawane są pary odczytów, o których można powiedzieć, że są od siebie oddalone w badanym fragmencie DNA o odstęp określony w przybliżeniu jako pewien zakres wartości (np. dwa odczyty o długości 100 nukleotydów każdy oddalone od siebie o 50–150 nukleotydów), od czego są odstępstwa na skutek błędów eksperymentalnych.