Przedmiot obieralny Bioinformatyka, wykładowca prof. Marta Kasprzak

Materiały uzupełniające do wykładu 5: poszukiwanie motywów

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

Region promotorowy to odcinek DNA poprzedzający sekwencję kodującą gen, służący do regulacji procesu transkrypcji tego genu. Regiony takie w genomach różnych organizmów mogą zawierać podobne fragmenty istotne funkcjonalnie. Często możemy spodziewać się, że takie podobne fragmenty (motywy) występują dla podobnych organizmów w zbliżonej odległości od początku danego genu i/lub w zbliżonej kolejności występowania w genomie. Także w sekwencjach aminokwasowych białek z różnych organizmów możemy spodziewać się podobnych fragmentów, jeśli białka pełnią podobną funkcję w organizmach, gdyż mają wtedy podobną strukturę przestrzenną, która z kolei wynika z sekwencji.

SLAJD 3

Wyszukanie pojedynczego motywu może być zrealizowane jak w problemie dopasowania lokalnego sekwencji, jeśli motyw jest najwyżej punktowanym podobnym podciągiem sekwencji. Wyszukanie serii motywów oddzielonych niepodobnymi odcinkami realizowane jest innymi podejściami.

SLAJD 4

Przykład prezentuje osiem białek o podobnej strukturze, co widać na schematycznym nałożeniu przestrzennym tych cząsteczek (właściwie ich części odpowiadających odcinkom sekwencji aminokwasowych, w tabelce poniżej). Odcinki sekwencji w tabelce mogą być traktowane jako rozłączne motywy obecne w ośmiu sekwencjach aminokwasowych. Tutaj szczególnie widać różnicę w porównaniu do dopasowania sekwencji nukleotydowych, gdyż kolejne wystąpienia motywów różnią się wyraźnie od siebie (niektóre aminokwasy mogą być zastąpione innymi bez dużej różnicy w strukturze przestrzennej cząsteczki wynikowej). Ta różnica jest w tym przykładzie pogłębiona przez to, że motywy zostały wyznaczone na podstawie dopasowania struktur, a nie osobnym podejściem na drodze analizy sekwencji.

SLAJD 6

Sequence logo służy zobrazowaniu efektu dopasowania wielu sekwencji w postaci zbliżonej do sekwencji konsensusowej, tym razem jednak na danej pozycji w sekwencji obecne są wszystkie litery z danej kolumny dopasowania, zróżnicowane rozmiarem zależnym od udziału w kolumnie. Zastosowanie *sequence logo* w problemie poszukiwania motywów, jako sposobu na ich wyłuskanie z sekwencji dopasowywanych globalnie, sprawdzi się tylko w nielicznych przypadkach, gdy sekwencje są zbliżonej długości, wystąpienia motywów mają podobną lokalizację w sekwencjach i stanowią na tyle solidny szkielet dopasowania, że zostaną zestawione razem w dopasowaniu globalnym. Ale i wtedy podejście to może zawieść, por. slajd 13, tak więc *sequence logo* można traktować najwyżej jako podpowiedź. W przykładzie wyodrębnić można np. motyw GCGAT (na pozycjach 3-7), który we wszystkich sekwencjach poza jedną ma wystąpienia różniące się od niego na co najwyżej jednej pozycji, a więc może zostać uznany za właściwy dla tego zbioru. Albo motyw GTATTA (pozycje 12-17), którego wystąpienia w pięciu sekwencjach odbiegają od niego na co najwyżej jednej pozycji, a w trzech sekwencjach na dwóch pozycjach. Po skróceniu tego motywu do GTATT otrzymujemy siedem wystąpień odbiegających od niego na co najwyżej jednej pozycji. Logo wygenerowane aplikacją ze strony weblogo.berkeley.edu.

SLAJD 7

Metody ze slajdów 8-13 służą w ogólności rozwiązaniu problemu dopasowania globalnego wielu sekwencji, ale mogą być również zastosowane do problemu poszukiwania serii motywów (jak w przykładach na slajdach). Metody ze slajdów 14-21 są dedykowane temu drugiemu problemowi.

SLAJD 8

Strategie dopasowania wielu sekwencji zostały omówione w ramach wykładu 4. Bazują one na dopasowaniach par sekwencji, obliczanych dla wszystkich par, ale w konstrukcji rozwiązania wykorzystują tylko wybrane dopasowania i pozostałe pomijają (inaczej niż omawiany tu algorytm).

SLAJD 9

Wagi krawędzi pozwalają wyłuskać bardziej zgodne miejsca, które, o ile sąsiadują, mogą zostać zagregowane do motywów występujących w instancji. W modelu grafu bez wag można się ograniczyć do wstawiania tylko tych krawędzi, które oddają zgodność pary znaków.

SLAJD 10

Przykład jest bardzo mały i małe są też kliki (rozpięte na trzech sekwencjach), dlatego też drobne niezgodności sprawiają tu dużą różnicę. W przykładzie można wyodrębnić np. motyw AC o identycznych wystąpieniach we wszystkich sekwencjach lub, wydłużając go i dopuszczając niezgodności, motyw ACA/ACT, albo też inny motyw o wystąpieniach ACCT/CCT/ACT.

SLAJD 11

Składowe grafu z metody Kececioglu są pod koniec kompresowane do wierzchołków, czego nie pokazano na rysunku, przy czym wierzchołki z pierwotnego grafu w ramach jednej składowej mogą być zaetykietowane różnymi znakami. W modelu grafu bez wag, jeśli dopuścilibyśmy tylko krawędzie oddające zgodność pary znaków, składowe obejmowałyby tylko wierzchołki o tych samych znakach i wierzchołek po kompresji odpowiadający tej składowej mógłby zostać zaetykietowany jednym znakiem. Graf z metody Raphaela i in. przypomina ten drugi rodzaj grafu, a w przykładzie na slajdzie 12 pokazany jest po kompresji. W grafie po kompresji (inaczej niż u Kececioglu) łuki mają wagę równą liczbie wystąpień danej relacji poprzedzania w sekwencjach wejściowych. Wagi te są uwzględniane na etapie usuwania niepożądanych struktur z grafu.

SLAJD 12

W tym przykładzie sekwencje wejściowe są identyczne jak na slajdzie 10. Krótkie ścieżki o wysokich wagach mogą zostać uznane za odpowiadające motywom w zbiorze. Najwyższą wagą jest tutaj 3, na tych łukach można zbudować motywy, wydłużając je zgodnie z preferencjami.

SLAJD 13

Algorytm Zhanga i Watermana służy dopasowaniu wielu sekwencji i omówiony został w ramach wykładu 4. Tu inny przykład, bardziej skomplikowany (graf dla instancji ze slajdu 10 jest na slajdach z wykładu 4). W algorytmie graf poddawany jest heurystycznej transformacji do postaci acyklicznej, tu pokazany graf przed transformacją, jednak w obu postaciach za ścieżki reprezentujące wiarygodne dopasowanie uznawane są te o wysokich wagach. Identyfikując takie krótkie ścieżki w tym grafie, otrzymujemy zestaw potencjalnych motywów. Tutejszy przykład celowo został tak skonstruowany, żeby pokazać niejednoznaczność interpretacji. Jeśli uznamy, że motyw powinien być zbudowany na łukach o wadze przekraczającej połowę liczebności zbioru wejściowego (tu 4, gdyż mamy siedem sekwencji), w kręgu naszego zainteresowania będą łuki (AC, CC), (CG, GT) i (TA, AC). Te połączenia podświetlone zostały w zestawieniu po prawej stronie na niebiesko. Można zauważyć, że podciąg ACC występuje w pięciu sekwencjach wejściowych, ale czasem na odległych pozycjach, więc nie do końca nas satysfakcjonuje, jeśli zależy nam na podobnych pozycjach wystąpień danego motywu w różnych sekwencjach lub na podobnej kolejności wystąpień serii motywów w sekwencjach. Połączenie podciągów TAC i ACC w TACC, naturalne ze względu na ich sąsiedztwo w grafie, występuje

tylko jednokrotnie w tej instancji. Jednak w bardziej rzeczywistych instancjach, gdzie sekwencje są dłuższe i dłuższe też mogą być k-mery, wynik zwykle jest bardziej jednoznaczny. Załączone *sequence logo* pokazuje, jak może odbiegać od rzeczywistej oceny. Da się w nim co prawda wyróżnić poprawny motyw CGT, ale i równie wyrazisty podciąg CGTA, w istocie występujący jednokrotnie w instancji; CGTA nie może też zostać uznany za poprawny wzorzec motywu, od którego dopuszczone są drobne odchylenia, gdyż równoczesne wystąpienia T i A albo G i A na odpowiednich pozycjach w sekwencjach wejściowych są bardzo rzadkie. Wyraźnie odcinający się w *sequence logo* początkowy podciąg AGC także występuje tylko raz na tej pozycji, a sekwencje zawierające A na pierwszej pozycji i G na drugiej są niemal rozłączne.

SLAJD 14

Rola i postać macierzy substytucji dla aminokwasów przedstawione zostały w ramach wykładu 4.

SLAJD 15

Próg podobieństwa aminokwasów został podniesiony względem propozycji autorów z oryginalnego artykułu celem uproszczenia przykładu. Wagi łuków przeliczane są w ten sposób, że do początkowej wagi wyznaczonej jak w oryginalnym algorytmie Zhanga i Watermana dodawane są wagi łuków podobnych przeskalowane przez współczynnik obliczany na podstawie macierzy substytucji, szczegóły na slajdzie 16. Najsilniejszy motyw w tej instancji odpowiada k-merom ulokowanym na końcu sekwencji: LYS/LFS. Za drugi można by uznać YVL/YIL, a że oba motywy się nakładają, można je połączyć w jeden pomimo nieco słabszego środka tej ścieżki.

SLAJD 17

Chociaż i w tej metodzie, i we wcześniejszych omawianych w ramach tego wykładu kliki (lub gęste podgrafy zbliżone do klik) w grafie nieskierowanym (ważonym bądź nie) interpretowane są jako potencjalne miejsca wystąpienia motywu, różnica jest taka, że tu nie wyszukuje się całych składowych. Tutaj graf nierzadko jest spójny i dość zagmatwany, zawiera wiele struktur zbliżonych do klik, odwołanie do wag krawędzi bardzo pomaga w identyfikacji tych właściwych struktur, odpowiadających poszukiwanym motywom.

SLAJD 18

Metoda jest heurystyką. Wierzchołki konstruowanego grafu odpowiadają wszystkim podciągom o długości l, które można wyodrębnić w sekwencjach wejściowych i jeśli jakiś l-mer powtarza się w tym zbiorze, w grafie występuje taką samą liczbę razy (inaczej niż u Zhanga i Watermana). Na slajdzie 19 wierzchołki są dla wygody porozmieszczane w rzędach, gdzie każdy rząd odpowiada innej sekwencji wejściowej; powtórzenia wierzchołków z takimi samymi etykietami mogą wystąpić w różnych rzędach (jak w przykładzie), ale także w obrębie tego samego rzędu. Krawędzie wstawiane są tylko pomiędzy rzędami, nigdy w obrębie tego samego rzędu, gdyż interesuje nas wykrycie podobieństw pomiędzy sekwencjami, nie w obrębie tej samej sekwencji. Odległość Hamminga dla dwóch sekwencji (w ogólności wektorów) o tej samej długości jest liczbą pozycji, na których te sekwencje różnią się. Zdegenerowane pozycje to takie, na których wystąpienie motywu w sekwencji różni się od wzorcowego motywu (możliwe, że nieobecnego w sekwencjach, można na niego patrzeć jak na sekwencję konsensusową), dlatego dopuszczone są różnice pomiędzy dwoma wystąpieniami motywu w liczbie $2d$. Wartości l i d są parametrami ustalanymi przez użytkownika, k to wyliczona odległość Hamminga.

SLAJD 19

Kliki w takim grafie będą obejmowały co najwyżej po jednym wierzchołku z każdej sekwencji. Struktury zbliżone do klik (gęste podgrafy) mogą już obejmować więcej wierzchołków z tej samej sekwencji, jednak mając na uwadze cel, który chcemy osiągnąć, powinniśmy poszukiwać takie struktury rozpięte pomiędzy rzędami, nie w obrębie tego samego rzędu. W przykładowym grafie jest wiele struktur zbliżonych do klik rozpiętych na wszystkich trzech sekwencjach, głównie dlatego, że

pełna klika ma tutaj tylko trzy krawędzie. Jeśli ograniczymy się tylko do pełnych klik, i tak jest ich trochę w tym grafie, ale wagi pozwalają nam wyłuskać te bardziej znaczące: {CCTA, CCTC, ACTA} i {TACA, CACT, TACT} o wadze 62 oraz trzecią o wadze 44 w wariancie {CTAC, TCAC, CTAC} lub {CTAC, CTCA, CTAC}. Można z tych klik wyprowadzić sekwencje konsensusowe, odpowiednio CCTA, TACT i CTAC. Ta ostatnia akurat nie spełnia naszego oczekiwania na odstępstwo każdego wystąpienia od wzorca na jednej pozycji, bo wzorzec pokrył się z jednym z wystąpień; możemy to odstępstwo zignorować albo uznać, że ostatni podciąg reprezentuje motyw obecny tylko w dwóch sekwencjach. Te trzy sekwencje konsensusowe nakładają się idealnie na siebie, po złożeniu w całość mogą reprezentować dłuższy motyw CCTACT. Wygląda na wiarygodny, gdyż jego odległość Levenshteina do odpowiednich fragmentów wszystkich sekwencji wejściowych wynosi 1. Gdybyśmy chcieli ściśle trzymać się wskazania, że odstępstwo wzorca wyznaczonego na podstawie kliki od wszystkich jego wystąpień musi wynosić najwyżej d, można by zastąpić CTAC bardziej sztuczną reprezentacją, niebędącą konsensusem, czyli CCAC lub TTAC dla kliki {CTAC, TCAC, CTAC} czy CTAA lub CTCC dla kliki {CTAC, CTCA, CTAC}. Wtedy jednak te trzy wzorce (dwa poprzednie i jeden z powyższych czterech wariantów) przestają nam się idealnie nakładać, nadal można skleić je w jeden motyw, ale z niedokładnym ich nałożeniem. Problem wyznaczania dopasowania wielu sekwencji, z którym mamy tu do czynienia, jest trudny obliczeniowo, na tym etapie trudność sprawia wybranie i połączenie wzorców, które reprezentować mają kolejne nakładające się podciągi rozwiązania. Chociaż cała metoda jest heurystyczna, autorzy do realizacji tego etapu obliczeń zaimplementowali algorytm dokładny.

SLAJD 21

Motywy wprowadzone zostały do instancji ze zmianami, dlatego zostały tu zaprezentowane w postaci *sequence logo* (kolumna 1 tabeli). Częstość ich wystąpienia w sekwencjach podana została w kolumnie 2. Slajdy zapisane w PDF dają niezbyt czytelny rysunek, rysunek oryginalny w artykule: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-19 .