Algorytmy kombinatoryczne w bioinformatyce, wykładowca prof. Marta Kasprzak

Materiały uzupełniające do wykładu 4: dopasowanie sekwencji, 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

Dopasowanie sekwencji, dla dwóch sekwencji, polega na takim zestawieniu znaków jednej i drugiej sekwencji, że maksymalizowana jest całkowita wartość takiego dopasowania rozumiana jako suma wartości dla poszczególnych par zestawionych znaków. Można także zestawić znak ze spacją wstawioną w jedną bądź drugą sekwencję. W najprostszym schemacie punktacji wartość dla pary znaków określana jest jako +1 za zgodność pary znaków, -1 za niezgodność takiej pary i -1 za zestawienie znaku ze spacją. Można dopasowywać sekwencje na całych ich długościach (dopasowanie globalne) lub tylko dla ich fragmentów (dopasowanie lokalne, dopasowanie semiglobalne). Przykładowo, dla powyższej punktacji optymalne dopasowanie globalne dla pary sekwencji nukleotydowych CCGATACGT i CGATTACGAA to:

CCGAT-ACGT-
C-GATTACGAA

o wartości 3. Sekwencje dopasowuje się celem określenia stopnia podobieństwa sekwencji (w sensie wartości dopasowania), wyszukania różnic pomiędzy sekwencjami (gdy porównywane są sekwencje podobne) albo wyszukania fragmentów podobnych (gdy sekwencje są raczej różne). Różnice pomiędzy w miarę podobnymi sekwencjami pochodzenia biologicznego (nukleotydowymi, aminokwasowymi) wynikać mogą ze zmian na drodze ewolucji lub z niedokładnego rozpoznania sekwencji. W biologii obliczeniowej niezgodność pomiędzy parą nukleotydów bądź aminokwasów określa się mianem substytucji (zamiany), zestawienie znaku ze spacją mianem insercji/delecji.

SLAJD 3

Odległość Levenshteina dla dopasowania pary sekwencji jak wyżej wynosi 4. Gdy dopasowujemy sekwencje globalnie, wartość dopasowania i odległość Levenshteina są w pewnym sensie komplementarne, równie dobrze określają podobieństwo/różnicę pomiędzy sekwencjami. Jednak gdy długość dopasowania nie jest ustalona, wartość dopasowania sprawdza się lepiej, gdyż premiuje zgodne odcinki dopasowania.

SLAJD 4

Dla różnych schematów punktacji otrzymamy różne optymalne dopasowania sekwencji, zatem ustalenie tej punktacji ma istotny wpływ na wynik i zależy od celu, który chcemy osiągnąć. Gdy porównywane sekwencje są w miarę podobne i chcemy, żeby różnice miały większy wpływ na ocenę, możemy przyznać im większą liczbę punktów ujemnych. Jeśli spodziewamy się przewagi jednego rodzaju błędów eksperymentalnych nad innymi (wynikającej np. z metody użytej do odczytu sekwencji), możemy to uwzględnić, różnicując punktację za substytucję i insercję/delecję. Gdy porównywane sekwencje są krótkie, możemy zechcieć bardziej ukarać niedopasowania niż w przypadku sekwencji długich. Znaczenie ma też rozmiar alfabetu (4 dla sekwencji nukleotydowych, 20 dla aminokwasowych), w sekwencjach nukleotydowych łatwiej o zgodność pary znaków, chociażby przypadkową.

SLAJD 5

Prawdopodobnie najpowszechniej stosowany algorytm w biologii obliczeniowej (w różnych wariantach), w tej wersji realizujący dopasowanie globalne pary sekwencji. Algorytmy z tego i kolejnego slajdu były już przedstawione na przedmiocie „Wprowadzenie do bioinformatyki", tu przypomniane dla kompletności wykładu. W przykładzie przyjęto punktację -1 za wstawienie spacji (oznaczona literą g od *gap*), -1 za niezgodność pary znaków i +1 za zgodność (funkcja porównania

znaków $s(i,j)$). Po wypełnieniu całej macierzy o n+1 wierszach i m+1 kolumnach, gdzie n i m to długości sekwencji, wynikowa wartość dopasowania znajduje się w skrajnie prawym polu ostatniego wiersza macierzy. Samo dopasowanie (zestawienie par znaków) odczytuje się, począwszy od tego pola, poprzez cofanie się aż do pola [0,0] na podstawie tego, skąd wyprowadzona została wartość danego pola. Każde pole może być wyliczone na podstawie trzech pól sąsiadujących z nim: lewego, z lewego górnego narożnika, lub górnego, w zależności od tego, które da wartość maksymalną dla wzoru na wypełnienie macierzy programowania dynamicznego. Częstą sytuacją jest uzyskanie wartości maksymalnej dla więcej niż jednego pola sąsiadującego, wtedy przy cofaniu się można wybrać dowolne z takich pól. W istocie macierz ta zawiera wszystkie rozwiązania optymalne (nawet wykładniczą ich liczbę) i moglibyśmy je odtworzyć, podążając wszystkimi możliwymi ścieżkami od pola [n,m]. W praktyce jednak generuje się jedną taką ścieżkę, na rysunku oznaczoną kolorem czerwonym. Odpowiada jej dopasowanie sekwencji pokazane na dole slajdu po lewej. Przejście w macierzy ścieżką z góry na dół oznacza zatrzymanie się na znaku sekwencji niebieskiej i dodanie nowego znaku sekwencji zielonej, czyli wstawienie spacji w sekwencję niebieską. Przejście w poziomie na prawo oznacza odpowiednio wstawienie spacji w sekwencję zieloną. Przejście po skosie to zestawienie kolejnego znaku sekwencji zielonej z kolejnym znakiem niebieskiej. Przykład alternatywnej ścieżki: na slajdzie wartość w polu [3,2] macierzy mogła zostać uzyskana z wyliczenia M[2,1]+s(3,2) (co daje 0–1, bo znaki C i G są różne) albo M[2,2]+g (co również daje 0–1). Wybrana została ta druga opcja, ale równie dobrze mogła zostać wybrana pierwsza, co zmieniłoby początek dopasowania na ATC z A–G.

SLAJD 6

Problem dopasowania lokalnego tym różni się od globalnego, że nie musimy uwzględniać całych sekwencji. Wybierany jest podciąg jednej sekwencji i podciąg drugiej dające maksymalną wartość dopasowania i reszta sekwencji nie jest obarczana karą za niedopasowanie. Zmienia to wzór na wypełnienie macierzy w taki sposób, że w przypadku wejścia w zakres ujemnych wartości dopasowania możemy takie dopasowanie porzucić i rozpocząć od początku, wybierając 0 jako wartość w danym polu. Po wypełnieniu macierzy optymalne dopasowanie lokalne odtwarzamy, wybierając maksymalną wartość w macierzy (to wartość tego dopasowania) i identyfikując ścieżkę wyprowadzania wartości pól, aż do osiągnięcia pola z wartością 0. W macierzy na slajdzie jest kilka wartości 3, każda z nich mogłaby stać się początkiem rekonstruowanego rozwiązania optymalnego. Ścieżka zaznaczona na czerwono odpowiada dopasowaniu ukazanemu w lewej dolnej części slajdu.

SLAJD 7

Dopasowanie semiglobalne pary sekwencji stanowi podstawę procesu rekonstrukcji sekwencji genomowej na etapie asemblacji. Polega na tym, że oceniane są nachodzące na siebie fragmenty obu sekwencji na całej długości takiego nałożenia, ale nieoceniane są odcinki sekwencji wystające poza nałożenie, czyli te kolumny dopasowania, gdzie występuje tylko jedna sekwencja. Na slajdzie przykład nałożenia sufiksu jednej sekwencji z prefiksem drugiej, oceniany jest tylko kolorowy fragment. Innym przykładem mogą być dwie sekwencje, z których jedna zawiera się w drugiej: dla AAATCGCCAA i ATGC ocenione zostanie dopasowanie całej sekwencji krótszej z fragmentem dłuższej ATCGC, nieobarczone karą pozostaną wystające końce sekwencji dłuższej AA oraz CAA. Wszystkie warianty dopasowania semiglobalnego: sufiks pierwszej sekwencji z prefiksem drugiej, odpowiednio prefiks z sufiksem, zawieranie pierwszej w drugiej i odwrotnie, przeliczane są poprzez jednokrotne wypełnienie macierzy z użyciem podanego wzoru. Mamy tu inną przykładową punktację za spację (-2 punkty). Po wypełnieniu macierzy optymalną wartość dopasowania wyszukuje się w całym ostatnim wierszu i w ostatniej kolumnie (tu: 2 i jest tylko jedna taka wartość) i rozwiązanie odczytuje się, począwszy od tego pola, odtwarzając drogę wypełniania pól macierzy aż do dotarcia do pierwszego wiersza lub pierwszej kolumny macierzy. Ścieżka prowadząca od ostatniego wiersza do pierwszej kolumny (jak w przykładzie) oznacza dopasowanie sufiksu sekwencji zielonej z prefiksem niebieskiej. Ścieżka od ostatniej kolumny do pierwszego wiersza oznaczałaby dopasowanie sufiksu sekwencji niebieskiej z prefiksem zielonej. Ścieżka od ostatniego wiersza do pierwszego wiersza to zawieranie

sekwencji zielonej w niebieskiej, od ostatniej kolumny do pierwszej kolumny to zawieranie sekwencji niebieskiej w zielonej.

SLAJD 8

Optymalne rozwiązanie dla dopasowania globalnego plasuje się w okolicy przekątnej macierzy, jeśli sekwencje są dość do siebie podobne. Na slajdzie 5 widać, że im dalej od przekątnej, tym niższe ujemne wartości dopasowania. Zainteresowanie użytkowników często ogranicza się do poznania optymalnego dopasowania dla sekwencji w miarę do siebie podobnych i wtedy wystarcza wypełnienie okolicy przekątnej o pewnej szerokości. Wypełniane pola można też ograniczyć dla dopasowania semiglobalnego, jeśli zakładamy pewne warunki odnośnie nakładania się sekwencji (np. tylko prefiks z sufiksem lub odwrotnie i tylko długie nałożenia).

SLAJD 9

Algorytm Crochemore'a i in. jest interesującym przykładem na to, że można ograniczyć złożoność algorytmu programowania dynamicznego dla dopasowania pary sekwencji bez straty na dokładności rozwiązania. Podana tu złożoność może być wyraźnie mniejsza od $O(n^2)$, jeśli sekwencje są dość długie i po części skomponowane z powtarzających się fragmentów.

SLAJD 10

Faktoryzację LZ przeprowadza się osobno dla każdej sekwencji (używa się w tym celu drzew prefiksowych, na slajdzie po prawej) w ten sposób, że sekwencję czyta się od lewej do prawej i odcina słowa na takiej zasadzie, że każde następne odcięte słowo ma być słowem rozpoznanym wcześniej wydłużonym o jeden znak. Z początku nie są rozpoznane żadne słowa, odcinane są zatem pojedyncze znaki. Dla sekwencji TAGACTAC odcinane słowa po kolei to: T, A, G, AC, TA, C. Nawet jeśli sekwencja nie jest za bardzo repetytywna (powtarzalna), sama długość sekwencji sprawi, że będą odcinane coraz dłuższe słowa. Repetytywność sekwencji słowa te znacznie wydłuża. Zysk na obliczeniach przy wypełnianiu macierzy bierze się z obserwacji, że kolejne słowa to zawsze słowo występujące wcześniej plus jeden znak (wyjątkiem może być koniec sekwencji), a w odniesieniu do bloków macierzy, występujący dalej blok obejmuje porównania dokonane dla bloków wcześniejszych i nowe jest tylko jedno pole. Należy zwrócić uwagę na istotną różnicę z podejściem tradycyjnym: tam wartości wpisywane do macierzy były bezwzględne, nie można takich wartości po prostu skopiować kawałek dalej. Tutaj wykorzystać należy względny przyrost wartości dopasowania w obrębie bloku, dokonane wcześniej porównania takich samych podciągów znaków.

SLAJD 12

Tradycyjny model punktacji w algorytmie programowania dynamicznego ocenia tak samo oba dopasowania:

```
TGTACAT        TGTACAT
T-T-C-T        TG---AT
```

podczas gdy to drugie jest bardziej prawdopodobne ewolucyjnie — jedno usunięcie (bądź wstawienie) trzech nukleotydów naraz — a to pierwsze to trzy osobne operacje. Model kar afinicznych uwzględnia preferencję biologów i traktuje łagodniej serię spacji występujących naraz. W takim modelu inną karą może być obarczona pierwsza spacja z rzędu i inną kolejne, wymaga on jednak wyliczania trzech macierzy programowania dynamicznego zamiast jednej.

SLAJD 13

Różne znaki w sekwencjach aminokwasowych czasami odpowiadają aminokwasom, które są do siebie podobne pod względem funkcji (wpływ na zwijanie łańcucha białkowego) i w trakcie ewolucji bywały zastępowane jeden drugim. Zatem nie można interpretować zerojedynkowo zgodności pary znaków w takich sekwencjach, ważny jest stopień ich podobieństwa/zastępowalności. Aspekt ten uwzględniony jest przez zastosowanie macierzy substytucji, która podaje wartość podobieństwa $s(i,j)$ dla każdej pary znaków obliczoną na podstawie analizy statystycznej grup mniej lub bardziej

podobnych sekwencji aminokwasowych. Przykłady takich macierzy są pod podanym odnośnikiem, np. macierz BLOSUM62; tamże, pierwsze 20 pozycji to kody IUPAC aminokwasów, B i Z to grupy dwuelementowe aminokwasów (odpowiednio D i N, E i Q), a X to dowolny aminokwas. Więcej o macierzach substytucji było na przedmiocie „Wprowadzenie do bioinformatyki".

SLAJD 14

Dopasowanie globalne wielu sekwencji (tu czterech) wygląda np. tak:

```
T-AT-AGATTGA
GGATTAG-T-GT
TGATTA-ATTGT
AGAT-AGATAGA
-------------
TGAT-AGATTGA
```

gdzie pod kreską podana jest sekwencja konsensusowa reprezentująca to dopasowanie (tu wygenerowana z zastosowaniem reguły względnej większości: daną kolumnę dopasowania reprezentuje znak, w tym spacja, który występuje w tej kolumnie największą liczbę razy). Przykład dopasowania lokalnego wielu sekwencji i odpowiadająca mu sekwencja konsensusowa:

```
   TATCACCTTGA
GGATTCGCCGT
      TCATCAATTGT
AGATCA-CAGA
-----------
   TCACC
```

Obliczanie dopasowania sekwencji parami i składanie wyniku w całość może być traktowane jedynie jako podejście heurystyczne: miejsca wstawienia spacji i przyporządkowanie znaków ustalone na podstawie dopasowania par sekwencji P i R oraz R i Q mogą się kłócić z optymalnym dopasowaniem pary sekwencji P i Q. Na slajdach 15 i 16 przedstawione są podejścia heurystyczne bazujące na informacji o dopasowaniu sekwencji parami.

SLAJD 18

W algorytmie sekwencje wejściowe dekomponowane są na serie nakładających się krótszych k-merów, gdzie k jest parametrem metody, z których budowany jest graf Pevznera jak w problemie sekwencjonowania przez hybrydyzację (k-mery to tamtejsze elementy spektrum), przykład na slajdzie 19. Dodatkowo przypisywane są wagi łukom mówiące, w ilu sekwencjach wejściowych dany k-mer wystąpił. Graf jest w ogólności dość zapętlony, aby łatwiej było wyznaczyć sekwencję reprezentującą rozwiązanie (sekwencję konsensusową) graf jest w sposób heurystyczny transformowany do postaci acyklicznej. Dzieje się to przez powielanie niektórych wierzchołków i rozdzielanie łuków z nimi incydentnych aż do osiągnięcia postaci acyklicznej. Wtedy już w sposób prosty wyznaczana jest najdłuższa ścieżka w grafie w sensie wag łuków składowych. Ścieżka tłumaczona jest na sekwencję wynikową (konsensusową) jak w oryginalnej metodzie Pevznera.

SLAJD 20

Przykładowy graf, po transformacji do postaci acyklicznej, zawiera dwie ścieżki o maksymalnej wadze wynoszącej tu 8. Na każdej z nich można zbudować wynikowe dopasowanie globalne sekwencji wejściowych. Wysoka (w ogólności) waga łuków składowych ścieżki pozwala przyporządkować danym miejscom większą liczbę sekwencji wejściowych.

SLAJD 21

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 22

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 24

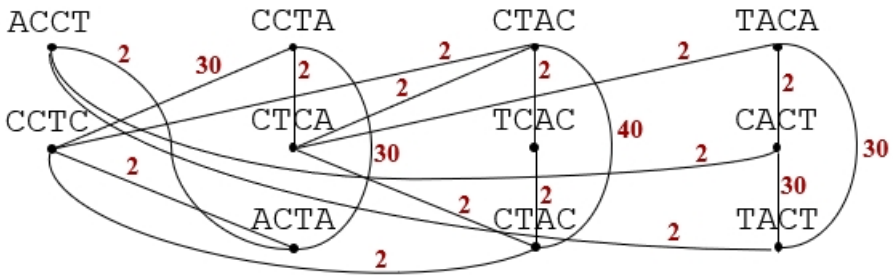
W różnych algorytmach służących wyszukiwaniu motywów, które odwołują się do teorii grafów, kliki (lub gęste podgrafy zbliżone do klik) w grafie nieskierowanym interpretowane są jako potencjalne miejsca wystąpienia motywu. Kliki takie, w zależności od konkretnego algorytmu i sekwencji wejściowych, bywają łatwiejsze bądź trudniejsze do wyodrębnienia. Tutaj graf nierzadko jest dość zagmatwany i zawiera wiele struktur zbliżonych do klik, odwołanie do wag krawędzi bardzo pomaga w identyfikacji tych właściwych struktur odpowiadających poszukiwanym motywom.

SLAJD 25

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). 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 26

Graf dla tej przykładowej instancji wygląda następująco:



Dla zwiększenia czytelności wierzchołki są tutaj porozmieszczane w rzędach odpowiadających sekwencjom wejściowym. 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. Kliki w takim grafie będą więc 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.