Algorytmy kombinatoryczne w bioinformatyce, wykładowca prof. Marta Kasprzak

Materiały uzupełniające do wykładu 7: drzewa filogenetyczne.

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

Drzewa filogenetyczne służą odzwierciedleniu powiązań ewolucyjnych organizmów, które ulokowane są w liściach drzewa. Na slajdzie przykładowe drzewo filogenetyczne sporządzone dla organizmów niższego poziomu, głównie bakterii. Im bliżej dwa organizmy są ulokowane w drzewie, tym są bardziej podobne ewolucyjnie. Podobieństwo może być określane różnymi miarami, np. przez dopasowanie globalne wybranych fragmentów genomów, ale też na różne inne sposoby, na podstawie cech organizmów (fenotypu) czy wyników wybranego eksperymentu laboratoryjnego (np. porównanie profili cięć enzymami restrykcyjnymi). Wagi przypisane krawędziom drzewa są tutaj wartością procentową oddającą wiarygodność topologii danej gałęzi drzewa, taką informację podaje się dla drzewa, które powstało ze złożenia wielu propozycji (drzewa konsensusowego); tu pominięto rozgałęzienia o wagach mniejszych od 50.

SLAJD 3

Drzewo filogenetyczne przedstawiane jest czasem w postaci ukorzenionej, gdy chcemy zobrazować kierunek ewolucji. Ewolucja biegnie od korzenia do liścia i w takim przypadku drzewo jest traktowane jak skierowane rozchodzące się. Węzły odpowiadają wtedy punktom pośrednim ewolucji, hipotetycznym przodkom, z których wyewoluowały organizmy współczesne. Krawędzie można uzupełnić o informację oddającą odległość ewolucyjną lub czas, jaki upłynął pomiędzy węzłami incydentnymi, informacja ta może być wyrażona długością krawędzi lub wagą.

SLAJD 4

Drzewa budowane są na podstawie danych znakowych lub numerycznych (danym numerycznym poświęcone są slajdy 10-17). Cechy obiektu reprezentowane w postaci wektora znaków gromadzone są w macierzy znakowej, mogą wśród nich znaleźć się także wartości liczbowe, jeśli nie są interpretowane numerycznie, tylko jako możliwe skończone stany pewnej cechy. Przykładowa macierz znakowa na slajdzie opracowana została dla różnych gatunków ryb i cechy opisują ich wygląd (legenda pod tabelką). Obiekty w wierszach są tym bardziej podobne ewolucyjnie, im więcej mają takich samych wartości w odpowiednich kolumnach.

SLAJD 5

W drzewie utworzonym na podstawie macierzy znakowej wszystkie obiekty z wejścia znajdą się w liściach, zatem liściom można przypisać wektory stanów cech z macierzy, ale węzłom wewnętrznym drzewa także przypisuje się wektory stanów cech, tylko często inne niż obecne w macierzy. Wektory węzłów wewnętrznych opisują stany pośrednie ewolucji, wektory w gałęzi wychodzącej z pewnego węzła „ewoluują" z wektora tego węzła. Przypadek ewolucji równoległej w drzewie zachodzi wtedy, gdy jakaś cecha niezależnie wyewoluowała dwa lub więcej razy w ten sam stan. Czyli istnieje jakaś gałąź drzewa, w której korzeń v tej gałęzi ma na danej pozycji wektora stanów cech pewną wartość X, a inna wartość Y występuje na tej samej pozycji wektora w węzłach co najmniej dwóch gałęzi ukorzenionych w następnikach v. Przypadek ewolucji wstecznej ma miejsce wtedy, gdy jakaś cecha wyewoluowała ze stanu X w stan Y, a potem powróciła do stanu X, czyli w drzewie istnieje ścieżka od korzenia do liścia, która obejmuje zmianę X-Y-X, niekoniecznie w sąsiednich węzłach. Zakłada się, że idealna (bezbłędna) macierz znakowa umożliwia utworzenie drzewa pozbawionego takich przypadków. Obie takie sytuacje można łącznie opisać jako przypadki, kiedy pewien stan jakiejś cechy występuje w wektorach węzłów, które nie tworzą jednego spójnego poddrzewa (i wtedy pozbywamy się nawiązań do drzewa ukorzenionego – skierowanego). O stanach mówimy, że są uporządkowane,

gdy w trakcie ewolucji mogą zmieniać się tylko zgodnie z ustalonym porządkiem (np. liczba zębów może tylko maleć).

SLAJD 6

Wersja problemu o znikomej przydatności praktycznej ze względu na skrajnie małą liczbę cech (i założenie o braku błędów), ale z interesującą metodą odwołującą się do znanych nam już grafów liniowych. Przy okazji sekwencjonowania mowa była o grafach liniowych skierowanych, tu mamy nieskierowane. Wierzchołki grafu liniowego nieskierowanego G odpowiadają krawędziom grafu oryginalnego H i dwa wierzchołki w G są połączone krawędzią wtedy i tylko wtedy, gdy odpowiadające im krawędzie w H są incydentne. Graf przecięcia stanów (ang. *state intersection graph*, SIG) jest tak konstruowany, że wierzchołki odpowiadają stanom (wartościom z macierzy znakowej) i są połączone krawędzią, jeśli jakiś obiekt posiada oba stany. Acykliczność grafu SIG zapewnia, że w wynikowym drzewie filogenetycznym nie nastąpi w żadnym węźle konflikt wartości na tej samej pozycji wektora stanów cech przy zachowaniu zasady, że każdy stan każdej cechy formuje w nim spójne poddrzewo. Żeby się o tym przekonać naocznie, proszę wziąć macierz ze slajdu, zamienić w niej ostatni wiersz na [x_1, y_1] (SIG zawierałby wtedy cykl) i spróbować skonstruować dla niej (własnym sposobem, z pominięciem SIG i transformacji do grafu liniowego) idealne drzewo filogenetyczne; można też usunąć celem uproszczenia niepotrzebne w tym przykładzie wiersze D i E.

SLAJD 7

Przekształcenie w graf liniowy oznacza zamianę krawędzi na wierzchołki z zachowaniem połączeń: jeśli krawędzie były incydentne, po przekształceniu wierzchołki łączone są krawędzią. Jakiekolwiek drzewo rozpinające w G może stanowić szkielet rozwiązania, wynikowe drzewo filogenetyczne uzyskujemy z niego poprzez proste operacje: obiekty z wejścia mają być w liściach drzewa, więc jeśli liść drzewa rozpinającego ma przypisane stany któregoś obiektu i tylko jednego, podpinamy tam ten obiekt (tak możemy tutaj zrobić z C i D); jeśli ma przypisane stany więcej niż jednego obiektu, wyprowadzamy z niego odpowiednią liczbę liści (tu A i E); jeśli jest węzłem wewnętrznym, wyprowadzamy z niego liść lub liście (tu B i F). Uzyskaliśmy idealne drzewo filogenetyczne, gdyż wszystkie stany: x_1, x_2, x_3, y_1, y_2, y_3, przypisane są do węzłów składających się na spójne poddrzewo (liście A, B, E, F posiadają pary stanów jak w macierzy). Chętni mogą spróbować sprawdzić poprawność drzewa zbudowanego na innym drzewie rozpinającym grafu G.

SLAJD 8

Kolejny problem z ograniczoną liczbą cech i założeniem o braku błędów. Graf SIG konstruowany jest analogicznie jak w poprzedniej metodzie, cechom przypisane są różne kolory. Każdy obiekt z wejścia w tym grafie będzie reprezentowany kliką na trzech wierzchołkach oraz żadne dwa stany tej samej cechy nie będą połączone krawędzią. Triangulacja grafu to podział cięciwami wszystkich cykli grafu o więcej niż trzech wierzchołkach i nieposiadających cięciw na cykle o trzech wierzchołkach. C-triangulacja odnosi się do kolorowania wierzchołkowego grafu i jest to taka triangulacja, że zachowuje poprawne pokolorowanie po dodaniu cięciw. Przykładowy graf SIG jest poprawnie pokolorowany (wierzchołki połączone krawędzią mają zawsze różne kolory) i taki pozostanie po triangulacji poprzez dodanie krawędzi (y_2, z_3). Jeśli można dla grafu SIG przeprowadzić c-triangulację, można z uzyskanej struktury wyprowadzić drzewo filogenetyczne. Każda klika na trzech wierzchołkach w grafie po triangulacji odpowiada węzłowi drzewa, kliki wcześniej obecne to liście, kliki nowe to węzły wewnętrzne (można je zaetykietować poprawnymi wektorami stanów cech) i dwa węzły w drzewie połączone są krawędzią, jeśli w SIG po triangulacji odpowiednie kliki współdzielą krawędź. W przykładzie po triangulacji dochodzą nam dwie nowe kliki, zarysowane na niebiesko i zielono, przekładają się one na dwa węzły wewnętrzne drzewa o tych samych kolorach, do których podpinamy liście. Ponownie uzyskujemy drzewo, w którym wszystkie stany wszystkich cech formują spójne poddrzewa.

SLAJD 10

Więcej o różnych miarach odległości filogenetycznej przy okazji slajdu 13. W praktyce rzadko kiedy mamy do czynienia z idealną (addytywną) macierzą numeryczną.

SLAJD 11

Tutaj wagi w krawędziach drzewa oznaczają odległość filogenetyczną pomiędzy węzłami. Wszystkie odległości pomiędzy obiektami w drzewie (zsumowane wagi na ścieżce pomiędzy parą liści) równają się odległościom z macierzy. Warunek na addytywność macierzy spełniony jest po przypisaniu obiektów do etykiet: A-i, B-l, C-j, D-k; d_{ij} oznacza odległość pomiędzy obiektem i i obiektem j.

SLAJD 13

Wartość dopasowania sekwencji (nukleotydowych, aminokwasowych) należy tu rozumieć jako odległość Levenshteina, gdyż im większa wartość w macierzy, tym obiekty są dalej od siebie w drzewie. Gdy z jakiegoś względu nie pasuje nam porównanie globalne sekwencji, możemy odnieść się do ich fragmentów (np. porównując motywy) lub zastosować analizę k-merową – każdą sekwencję opisujemy multizbiorem jej podciągów o długości k i na tej podstawie określamy ich podobieństwo. Odległość filogenetyczna może być też wyrażana funkcją liczby motywów wykrytych w sekwencjach, liczby obecnych w nich wspólnych miejsc restrykcyjnych albo markerów, innych wartości. Albo może to być odległość Hamminga pomiędzy ciągami znaków przypisanymi do obiektów (np. wektorami z macierzy znakowej). Odległość można także określić na podstawie różnic strukturalnych pomiędzy cząsteczkami produkowanymi przez organizmy (np. białkami). W końcu może nią być jakakolwiek inna charakterystyka wyrażana w liczbach.

SLAJD 14

Faktoryzacja LZ była już wykorzystywana w metodzie omawianej w ramach wykładu 4. Tutaj metoda wykorzystuje tylko część informacji płynącej z faktoryzacji, nie interesują nas odcinane słowa, a tylko ich liczba uzyskana dla sekwencji. Mieliśmy już okazję zauważyć, że z długością sekwencji odcinane są coraz dłuższe słowa; także im sekwencje bardziej repetytywne, tym dłuższe są odcinane kolejne słowa w porównaniu do sekwencji losowej. Nawet jeśli sekwencje nukleotydowe S i Q są do siebie niepodobne, sekwencja będąca rezultatem ich skonkatenowania (w dowolnej kolejności) będzie miała mniejszą liczbę słów po faktoryzacji niż suma takich wartości uzyskanych osobno dla S i Q. Jeśli z kolei S i Q wykazują podobieństwo, wartość złożoności LZ dla skonkatenowanych sekwencji będzie jeszcze mniejsza. Im większa różnica pomiędzy sumą słów uzyskanych w faktoryzacji osobno S i Q i liczbą słów uzyskanych dla SQ (lub QS), tym bardziej S i Q możemy uznać za podobne (przykład na slajdzie 15).

SLAJD 16

Autorzy w artykule zaproponowali pięć miar odległości filogenetycznej d(S,Q) i nie zdecydowali się na wskazanie najlepszej.

SLAJD 17

Przykład przeliczony dla wszystkich miar. Wartości są różnego rzędu, ale dla tego przykładu niezależnie od miary uzyskujemy taką samą relację w zbiorze: najbardziej podobna jest para Q i R, potem R i S, najbardziej odległa para S i Q.

SLAJD 18

Drzewa wynikowe generowane przez różne metody dla tej samej macierzy wejściowej mogą mieć podobne topologicznie fragmenty, jeśli bazują one na informacji oddającej wyraźną relację pomiędzy obiektami. Zapewne będą też różnić się w niektórych miejscach. Przyjętym podejściem jest użycie kilku różnych metod (albo jednej metody ale dla zmienianego wejścia) w celu wygenerowania hipotetycznych drzew i uzyskanie na ich podstawie jednego drzewa konsensusowego.

SLAJD 20

Opisana na kolejnych slajdach metoda jest odmienna od typowych z dwóch względów. Po pierwsze, autorzy zauważyli, że ograniczenie wyjścia w problemie generowania drzewa konsensusowego do jednego tylko drzewa jest niekiedy zbyt silne, drzewa na wejściu mogą się bardzo różnić i jedno drzewo wyjściowe może okazać się dość przypadkowym zlepkiem, nieoddającym wiarygodnej relacji ewolucyjnej pomiędzy obiektami. Po drugie, nie skupiali się na dopasowywaniu fragmentów struktur drzew, uprościli tę informację do podziałów zbioru obiektów na dwa podzbiory i w ten sposób ułatwili sobie obliczenia z zyskiem w postaci ewentualnej mniejszej liczby drzew na wyjściu.

SLAJD 21

Przykład dwóch różnych podziałów zbioru obiektów uzyskanych dla tego samego drzewa na slajdzie 22. Większy przykład od slajdu 26.

SLAJD 22

Dla tego jednego drzewa wejściowego uzyskujemy dwa możliwe podziały S_1 i S_2 (bo inne spowodowałyby utworzenie podzbioru o mniej niż dwóch elementach, co nie jest dopuszczone), każdy o wadze 1 (jednokrotne wystąpienie w zbiorze drzew na wejściu). Jeśli za próg wiarygodności przyjmiemy 1, to oba podziały zostaną uwzględnione w grafie kompatybilności (na rysunku). Dwa podziały nie są sprzeczne, jeśli mogą razem współistnieć w jakimś drzewie (niekoniecznie drzewie z wejścia), a to ma miejsce wtedy, gdy podzbiory obiektów z obu podziałów można tak sparować (po jednym z każdego podziału w parze), że jeden z pary mieści się w całości w tym drugim. W podanym przykładzie podzbiory obiektów można sparować następująco: {A, B} z {A, B, C} w jednej parze i {D, E} z {C, D, E} w drugiej, i w każdej parze jeden z podzbiorów mieści się w całości w drugim.

SLAJD 23

Wszystkie podziały uzyskane dla pewnego drzewa z wejścia z pewnością nie będą sprzeczne, zatem w grafie kompatybilności, o ile przekroczą próg wiarygodności, będą połączone krawędziami każdy z każdym (utworzą klikę). Ale możemy też zaobserwować kliki, które nie odpowiadają podziałom drzew z wejścia, mówi nam to wtedy, że takie podziały można zestawić w jedno spójne drzewo i mogłoby to być drzewo konsensusowe na wyjściu. Ponieważ na wyjściu chcemy uwzględnić całą informację z wejścia uznaną za wiarygodną, wszystkie podziały z grafu kompatybilności muszą zostać użyte, dlatego interesuje nas pokrycie wszystkich wierzchołków grafu. Klika-rdzeń pasuje do całej reszty, dlatego wejdzie w skład każdego drzewa konsensusowego.

SLAJD 24

Problem pokrycia grafu minimalną liczbą klik jest równoważny problemowi pokolorowania wierzchołków dopełnienia tego grafu minimalną liczbą kolorów (przy zachowaniu zasady, że wierzchołki o tym samym kolorze nie są połączone krawędzią) i autorzy zdecydowali się rozwiązać w swojej metodzie ten drugi problem (heurystyką zachłanną).

SLAJD 27

Gdy przyjmiemy 3 za próg wiarygodności, graf kompatybilności jest bardzo mały (na rysunku), składa się z kliki-rdzenia i po jej odjęciu już nic nie pozostaje. Dlatego nie wyszukujemy innych klik i podajemy na wyjściu jedno drzewo konsensusowe zgodne z oboma podziałami.

SLAJD 28

Dla progu wiarygodności 2 uzyskujemy bardziej interesujący przykład. Nie ma tu kliki-rdzenia, pozostałe kliki określimy poprzez rozwiązanie problemu pokolorowania grafu po prawej.

SLAJD 29

Podejście zachłanne dało nam rezultat jak na rysunku. Wierzchołki kolorujemy w kolejności malejącej wagi, tu kolejność pokrywa się z rosnącymi indeksami wierzchołków. Wierzchołki o tym samym kolorze odpowiadają klice w pierwotnym grafie. Dla każdej znalezionej kliki tworzone jest osobne drzewo konsensusowe na wyjściu, nie można dwóch takich klik połączyć w jednym drzewie, gdyż obejmują sprzeczne podziały. Gdyby dla tego przykładu istniała niepusta klika-rdzeń, współtworzyłaby każde z tych drzew.

SLAJD 30

Optymalne rozwiązanie składa się z mniejszej liczby drzew konsensusowych niż to uzyskane heurystyką zachłanną.