Algorytm rozwiązujący problem sekwencjonowania przez hybrydyzację

Marceli Jerzyński 136725
Korneliusz Szymański 136813

23 czerwca 2020

1 Opis problemu

Algorytm rozwiązuje problem sekwencjonowania łańcuchów DNA, z wykorzystaniem standardowych bibliotek oligonukleotydów o stałej długości i przy założeniu obecności odpowiednio błędów negatywnych lub pozytywnych w danych wejściowych.

2 Dane wejściowe

Na wejściu algorytmu podawane jest spektrum z błędami negatywnymi oraz pozytywnymi, dalej oznaczane jako **S** (Jest to zbiór słów nad alfabetem ['A','C','T','G']), długość słowa w spektrum, oznaczane jako **l**, oraz długość sekwencji oryginalnej **n**.

3 Dane wyjściowe

Algorytm zwraca oryginalną sekwencję nukleotydową.

4 Założenia

Oryginalny problem jest NP-trudny, dlatego też zaprezentowany poniżej algorytm jest heurystyką, co oznacza że dane wyjściowe mogą zawierać błędy, jednak złożoność obliczeniowa rozwiązania jest wielomianowa a nie wykładnicza, co znacząco zmniejsza czas trwania obliczeń.

5 Podejście teoretyczne

Aby rozwiązać problem sekwencjonowania przez hybrydyzację, można stworzyć graf skierowany, którego wierzchołkami będą słowa z **S**. Pomiędzy każdą parą

wierzchołków powinny zostać poprowadzone dwa łuki (A $\rightarrow$ B oraz B $\rightarrow$ A). Koszt danego łuku, jest to minimalne przesunięcie pomiędzy etykietami wierzchołków. (np ACT $\rightarrow$ CTC = 1, CTC $\rightarrow$ ACT = 3, GCC $\rightarrow$ CTC = 2). Za każde odwiedzenie wierzchołka otrzymuje się 1 punkt, za każde przejście przez łuk należy dodać jego wartość do sumy kosztów. Poszukiwana ścieżka musi mieć koszt nie większy niż **n - l**, oraz musi maksymalizować zysk. Problem zatem sprowadza się do problemu komiwojażera, który nie musi odwiedzić wszystkich wierzchołków (ze względu na błędy pozytywne), nie musi też odnaleźć on cyklu a jedynie ścieżkę (długość ścieżki nie może być jednak dłuższa niż n).

6 Algorytm mrówkowy

Aby osiągnąć optymalne rozwiązanie, zastosowany zostanie algorytm mrówkowy. Algorytm ten jest heurystyką rozwiązującą problem komiwojażera, zainspirowaną zachowaniem mrówek, szukających pożywienia. Mrówki w prawdziwym świecie poruszają się, sugerując się zapachem feromonów innych mrówek, które wcześniej znalazły pożywienie. Jeśli zapach jest słabszy, oznacza to, że mrówki, które szły tą drogą zostawiły go dawno temu, co implikuje mniejsze prawdopodobieństwo znalezienie pożywienia. Natomiast jeśli zapach jest silny, oznacza to, że duża ilość mrówek znalazła niedawno pożywienie. Należy również wziąć pod uwagę czynnik długości ścieżki $\rightarrow$ jeśli ścieżka jest dłuższa, feromony mają więcej czasu żeby wyparować, więc ostatecznie jest ich mniej.

7 Algorytm właściwy

1. Na wejściu zostały podane spektrum z błędami oraz długość sekwencji oryginalnej.

2. Na podstawie spektrum został utworzony graf, opisany w sekcji **Podejście teoretyczne**.

3. Na początku działania algorytmu dla każdego wierzchołka obliczana jest heurystyka, dzięki której, po normalizacji jesteśmy w stanie obliczyć prawdopodobieństwo, że dany wierzchołek jest początkiem sekwencji

$$H(i) = in_i - out_i$$

$H(i)$ - heurystyka, z której można obliczyć prawdopodobieństwo, że i-ty wierzchołek jest początkiem sekwencji

in_i - suma kosztów łuków wchodzących do i-tego wierzchołka

out_i - suma kosztów łuków wychodzących z i-tego wierzchołka

Zgodnie z tym prawdopodobieństwem, mrówki są rozmieszczane na wierzchołkach i to z nich rozpoczną poszukiwania ścieżki optymalnej

4. Każda mrówka wybiera krawędź którą podąży, zgodnie z prawdopodobieństwem obliczonym z poniższego wzoru:

$$p(ij) = \frac{\tau_{ij}^{\alpha} \cdot \eta_{ij}^{\beta}}{\sum \tau_{il}^{\alpha} \cdot \eta_{il}^{\beta}}$$

$p(i)$ - prawdopodobieństwo, że mrówka znajdująca się w wierzchołku i-tym pójdzie łukiem do wierzchołka j-tego

τ_{ij} - natężenie feromonów na krawędzi ij

α - parametr sterujący intensywnością feromonów (waga powyższego kryterium)

η_{ij} - stosunek zysku zdobytego przez przejście do j-wierzchołka (w wersji algorytmu z błędami poz. i neg. jest to 1) do kosztu łuku ij :

$$\frac{1}{koszt_{ij}}$$

β - parametr sterujący znaczeniem stosunku η_{ij}

Mrówka rozważa tylko te wierzchołki, dla których całkowity koszt przebytej ścieżki nie przekracza $\mathbf{n} - \mathbf{l}$

5. Gdy ilość odwiedzonych odwiedzonych wierzchołków jest równa $\mathbf{n}$, lub gdy całkowity koszt przebytej ścieżki jest tak duży, że nie pozwala mrówce przejść dalej do żadnego wierzchołka, mrówka kończy swoją podróż

6. Gdy wszystkie mrówki zakończą poszukiwanie spektrum, następuje parowanie na wszystkich krawędziach feromonów zgodnie ze wzorem:

$$\tau_{ij} = \rho * \tau_{ij}$$

τ_{ij} - ilość feromonów na krawędzi ij
ρ - współczynnik parowania feromonów

7. Następnie wszystkie mrówki zostawiają na wszystkich krawędziach które przeszły swoje feromony, zgodnie ze wzorem:

$$\forall_{ij \in P} \tau ij = \tau ij + \frac{zysk_P^2}{koszt_P}$$

ij - krawędź z wierzchołka i do wierzchołka j
P - ścieżka którą dana mrówka przeszła
τ_{ij} - ilość feromonów na krawędzi ij
$zysk_P$ - zysk, jaki mrówka osiągnęła przechodząc przez ścieżkę P
$koszt_P$ - koszt, jaki mrówka poniosła przechodząc przez ścieżkę P

8 Parametry

Parametry potrzebne do algorytmu mrówkowego zostaną wyznaczone empirycznie, jednak na podstawie poprzednich doświadczeń z tym algorytmem jesteśmy w stanie wyestymować wartości, które następnie będziemy zmieniać, aby osiągnąć najlepsze rezultaty:

ilość mrówek w jednej iteracji: $20 * \mathbf{S}.lenght$
ilość iteracji: 50
ρ: 0.7
τ_{ij} początkowe : 1
α : 1
β: 5

9 Rozwiązanie problemu przy braku błędów pozytywnych

Przy dodatkowej wiedzy, że instancja zawiera tylko błędy negatywne, należy znacznie zwiększyć zysk wynikający z odwiedzenia wierzchołka, co zwiększy prawdopodobieństwo odwiedzenia wszystkich wierzchołków. Jako że błędy w podanej instancji będą tylko negatywne, wiemy że zbiór **S** podany na początku jest podzbiorem spektrum idealnego, co oznacza, że wszystkie podane nukleotydy się w nim znajdują.

10 Rozwiązanie problemu przy braku błędów negatywnych

Przy dodatkowej wiedzy, że instancja zawiera tylko błędy pozytywne, należy zwiększyć koszt łuków, które w oryginalnym algorytmie jest większy niż 1 do liczby bliskiej ∞, tak, by prawdopodobieństwo pójścia tą drogą było równe 0.

11 Potencjalne udoskonalenia algorytmu

- Prawdopodobnie może się okazać, że tworzenie grafu pełnego może powodować zbyt długi czas oczekiwania na wyniki. W takim wypadku warto założyć, że prawdopodobieństwo wystąpienia x błędów pozytywnych pod rząd jest równe 0 i pozostawić tylko te łuki, których koszt jest mniejszy niż x. Parametr x zostałby w takim wypadku wyznaczony empirycznie.
- Może się okazać, że nasza heurystyka wyznaczająca prawdp. rozpoczęcia ścieżki w danym wierzchołku jest niepoprawna. W takim wypadku być może zmiana wartości H_{ij} co iterację, sugerując się uzyskiwanymi wynikami podniesie jakość algorytmu i zmniejszy to procent błędów

- W trakcie implementacji wzory na pozostawianą ilość feromonów oraz obliczenia η_{ij} mogą się okazać niepoprawne i w ostatecznej wersji mogą zostać zmienione na bardziej odpowiednie do tego problemu
- Ze względu na błędy negatywne polegające na tym, że dane słowo w **S** może występować raz, a w oryginalnej sekwencji może się ono pojawić wielokrotnie, mrówki będą miały możliwość powrotu do raz już odwiedzonego wierzchołka. Może się jednak okazać, że mrówki zbyt chętnie będą wykorzystywały tą możliwość. W takim wypadku zaimplementowany zostanie system, w którym wierzchołek, który już raz pojawił się na ścieżce, może pojawić się na niej ponownie, jednak albo zysk uzyskany z jego odwiedzenia zmniejszy się, albo koszt odwiedzenia danego wierzchołka zostanie zwiększony . Parametr zmniejszenia zysku zostanie wyznaczony empirycznie, jeśli taki system w ogóle zostanie zaimplementowany.