Sekwencjonowanie łańcuchów DNA metodą hybrydyzacji

Wyszukiwanie rozwiązań przy pomocy alogrytmu mrówkowego

Michał Zajączkowski 106655
Karol Lisiecki 106458

30.04.2014

Spis treści

1 Opis problemu **3**

2 Opis metody heurystycznej **3**

2.1 Opis ogólny 3

2.2 Budowa grafu 3

2.3 Rozwiązanie i jego ocena 4

2.4 Wybór wierzchołka początkowego 4

2.5 Lokalna ocena heurystyczna 4

2.6 Atrakcyjność krawędzi 5

2.7 Metoda uaktualniania feromonów 5

3 Złożoność obliczeniowa **5**

4 Optymalizacje z wykorzystaniem wiedzy o rodzaju błędów **6**

4.1 Sekwencjonowanie łańcuchów DNA z błędami negatywnymi 6

4.2 Sekwencjonowanie łańcuchów DNA z błędami pozytywnymi 6

1 Opis problemu

Sekwencjonowanie jest techniką odczytywania sekwencji, czyli kolejności par nukleotydowych w cząsteczce DNA. [1]

Matematycznie problem ten przedstawia się następująco:
Problem: Sekwencjonowanie łańcuchów DNA z błędami negatywnymi i pozytywnymi
Instancja: Zbiór S słów o jednakowej długości l nad alfabetem {A, C, G, T}, długość n sekwencji
Odpowiedź: Sekwencja o długości nie większej niż n zawierająca maksymalną liczbę słów z S.

Problem ten jest NP-trudny. Niżej opisany zostanne heurystyczna metoda o złożoności wielomianowej, pozwalająca na wyszukiwanie przybliżonych rozwiązań

2 Opis metody heurystycznej

2.1 Opis ogólny

Metoda oparta będzie o algorytm grafowy. Algorytm, który chcemy zastosować, to algorytm mrówkowy (ACO). Żeby go zastosować poza grafem potrzebna jest jeszcze koniecznie heurystyka lokalna opisana w punkcie 2.5.

ACO jest metaheurystyką iteracyjną. W każdej iteracji przeprowadzana jest pewna liczba przejść po grafie - "przejść mrówek", z których to każde generuje pewne rozwiązanie. Następnie rozwiązania te są oceniane i lepsze rozwiązania mocniej niż gorsze wpływają na proces uczenia się kolonii, składując na krawędziach feromony - pewną względną wartość oceniającą wybór tej krawędzi. Wiedza ta jest wykorzystywana w kolejnych iteracjach. Ponadto w każdej iteracji uwzględniany jest proces stopniowego ulatniania się tej wiedzy - przez to dobre rozwiązania lokalne są mocniej utrwalane, a te gorsze zapominane. Mechanizm ten zwany jest odparowywaniem feromonów. Taki algorytm ma większe szanse na wygenerowanie dobrego rozwiązania niż zwykłe heurystyki, ponieważ ma mniejsze tendencje do wpadania w minima lokalne.

2.2 Budowa grafu

Konstrujemy pełny skierowany graf G z wagami na krawędziach, którego wierzchołki odpowiadają olinukleotydom (słowom ze zbioru S). Waga krawędzi (s_1, s_2) odpowiada długości najdłuższego sufiksu słowa s_1 pokrywającego się z prefiksem słowa s_2 i może przyjmować wartości z przedziału *(0; l-1)*.

[1] `http://pl.wikipedia.org/wiki/Sekwencjonowanie_DNA`

2.3 Rozwiązanie i jego ocena

Odpowiednikiem rozwiązania problemu w reprezentacji grafowej jest ścieżka o długości co najwyżej n, mogąca zawierać wielokrotne wystąpenia tego samego wierzchołka. Na podstawie znalezionej ścieżki odtwarzana jest oryginalna sekwencja. Rozwiązanie jest dopuszczalne jeżeli ścieżka nie przekracza długością n. Ponieważ dążymy do maksymalizacji liczby wybranych słów, rozwiązanie jest tym lepsze im więcej unikalnych wierzchołków zawiera.

2.4 Wybór wierzchołka początkowego

Możliwa jest ocena wierzchołka pod względęm prawdopodobieństwa jego wystapienia na pierwszej pozycji rozwiązania. Początkowy wierzchołek powinien charakteryzować się niskimi wagami krawędzi wejściowych oraz wysokimi wagami krawędzi wyjściowych. Dla każdego wierzchołka obliczamy różnicę sumy wag jego krawędzi wejściowch i sumy wag krawędzi wyjściowych. Aby uwzględnić tę właściwość, ale zapewnić również odporność na błędy, korzystamy z metody ruletki [2]

Dla każdej iteracji algorytmu wybierany jest zbiór wierzchołków o rozmiarze równym liczbie mrówek, każda mrówka startuje w jednym z wierzchołków.

2.5 Lokalna ocena heurystyczna

Lokalna ocena heurystyczna krawędzi dla k-tej mrówki xy jest dana wzorem:

$$\eta_{xy} = w_{xy} - vis_y^k \cdot \gamma \tag{1}$$

gdzie:
w_{xy} - waga krawędzi
vis_y^k - liczba dotychczasowych odwiedzeń węzła y przez k-tą mrówkę
γ - parametr regulujący ujemną wagę liczby odwiedzeń danego węzła

Jak wspomniano w punkcie 2.2, waga krawędzi zależy liczby dopasowanych zasad incydentnych węzłów. Oznacza to, iż wybranie krawędzi o wyższej wadze w mniejszym stopniu wydłuża aktualną sekwencję niż krawędzi o niższej wadze. To z kolei pociąga za sobą fakt, że wybranie takiej krawędzi umożliwia potencjalnie odwiedzenie większej liczby węzłów, co wpływa bezpośrednio na ocenę rozwiązania. Odejmowanie liczby odwiedzeń z pewnym parametrem ma za zadanie dodatkowe preferowanie nowych węzłów nad te już odwiedzone.

[2] `http://pl.wikipedia.org/wiki/Algorytm_genetyczny#Metody_selekcji`

2.6 Atrakcyjność krawędzi

Wzór na prawdopodobieństwo wybrania krawędzi xy dla k-tej mrówki będącej aktualnie w węźle x jest dane wzorem:

$$p_{xy}^k = \frac{(\tau_{xy}^\alpha)(\eta_{xy}^\beta)}{\sum_{y \in \text{allowed}_y}(\tau_{xy}^\alpha)(\eta_{xy}^\beta)} \tag{2}$$

gdzie:
$allowed_y$ - zbiór docelowych węzłów dopuszczalnych w danym kroku,
τ_{xy} - ilość feromonów na krawędzi xy,
η_{xy} - lokalna ocena heurystyczna wyboru krawędzi xy,
α - parametr regulujący udział feromonów w ocenie krawędzi,
β - parametr regulujący udział heurystyki w ocenie krawędzi.

W celu wylosowania krawędzi mając te prawdopodobieństwa można wykorzystać algorytm ruletki.

2.7 Metoda uaktualniania feromonów

Po każdej iteracji wartość feromonów na każdej krawędzi jest uaktualniana wg. wzoru:

$$\tau'_{xy} = (1-\rho) \cdot \tau_{xy} + \sum_k \frac{Q \cdot max(R - r_k, 0)}{L_k} \tag{3}$$

gdzie:
ρ - parametr odparowywania feromonów, $\rho \in (0..1)$
Q - parametr wzmocnienia
R - liczba najlepszych rozwiązań danej iteracji, które mają być uwzględniane we wzmocnieniu
r_k - miejsce rozwiązania w rankingu najlepszych rozwiązań w danej itercji
L_k - ocena rozwiązania k-tej mrówki w danej iteracji

Taka procedura zapewnia, że lepsze rozwiązania będą lepiej zapamiętywane niż te gorsze.

3 Złożoność obliczeniowa

Na złożoność wpływa liczba iteracji, która może być regulowana bezpośrednio, lub przez zadany czas wykoniania. Oprócz tego kluczowymi czynnikami jest rozmiar instancji wejściowej oraz liczba mrówek. Złożoność czasowa algorytmu dana jest wzorem:

$$O(m,n,i) = i \cdot (m \cdot n^2 + m \cdot log(m) + m \cdot n + n^2) \tag{4}$$

W każdej iteracji i należy:

1. dla każdej mrówki przejść po średnio n węzłów i każdym wybrać z pośród n kandydatów - $m \cdot n^2$
2. posortować rozwiązania - $m \cdot log(m)$
3. dla każdej mrówki zsumować jej cząstkowe przyrosty dla średnio n krawędzi - $m \cdot n$
4. uaktualnić n^2 krawędzi

4 Optymalizacje z wykorzystaniem wiedzy o rodzaju błędów

4.1 Sekwencjonowanie łańcuchów DNA z błędami negatywnymi

Dla instancji w których występują jedynie błedy negatywne, każdy wierzchołek w grafie G musi być odwiedzony co najmniej raz. Można zakończyć szukanie rozwiązania po znalezieniu ścieżki zawierającej wszystkie wierzchołki.

4.2 Sekwencjonowanie łańcuchów DNA z błędami pozytywnymi

W przypadku występowanie jedynie błędów pozytywnych każdy wierzchołek w grafie G odwiedzany będzie co najwyżej raz. W przedstawionej metodzie heurystycznej uprości to funkcje oceny do postaci:

$$\eta_{xy} = w_{xy} \tag{5}$$

Ponieważ nie występują błedy negatywne (nie ma brakujacych słów, czyli wierzchołków w grafie), więc nie ma potrzeby tworzenia grafu pełnego - wystarczy połączyć ze sobą wierzchołki nachodzące na siebie na długości l-1, dzięki czemu znacznie zmniejszy się rozmiar grafu.