Sprawozdanie z laboratorium: Bioinformatyka (szablon)

Część I: Analiza teoretyczna

25 marca 2017

Prowadzący: prof. dr hab. inż. Marta Kasprzak

Autorzy:	**Damian Jurga**	inf.....	I2	jasiu@serwer.domena.poczta.pl
	Grzegorz Miebs	inf122453	I2	grzegorz.miebs@student.put.poznan.pl

Zajęcia środowe, 11:45.

Oświadczamy, że niniejsze sprawozdanie zostało przygotowane wyłącznie przez powyższych autorów, a wszystkie elementy pochodzące z innych źródeł zostały odpowiednio zaznaczone i są cytowane w bibliografii.

1 Wstęp

Celem tego sprawozdania jest przedstawienie teoretycznego opracowania metody heurystycznej rozwiązującej problem sekwencjonowania łańcuchów DNA z błędami pozytywnymi oraz negatywnymi w czasie wielomianowym. Alogrytm mając dany na wejściu zbiór S słów o długości l nad alfabetem $\{A, C, G, T\}$, długość n sekwencji oryginalnej, powinien zwrócić sekwencję o długości nie większej niż n zawierającą maksymalną liczbę słów z S.

2 Algorytm

2.1 Opis

Do rozwiązania tego problemu zbudujemy graf, którego wierzchołkami będą słowa ze zbioru S, a wartości łuku między wierzchołkami będą równe przesunięciu między tymi słowami. Przykładowo łuk z wierzchołka ACCGT do wierzchołka CCGTC będzie miał wartość 1 a do wierzchołka GTCGT wartość 3. Na skonstruowanym w ten sposób grafie rozwiązujemy problem komiwojażera maksymalizujący liczbę odwiedzonych wierzchołków przy ograniczeniu na sumę wartości wykorzystnych łuków, która nie może być większa od $n - l$, gdyż w przeciwnym razie na wyjściu otrzymamy sekwencję dłuższą niż oryginalna. Aby ograniczyć ponowne odwiedzanie tych samych wierzchołków, będziemy zwiększać wartość łuków prowadzących do odwiedzonych już wierzchołków o pewną stałą C. Do rozwiązania problemu komiwojażera posłużymy się przeszukiwaniem wiązkowym oraz algorytmem wspinaczki.

2.2 Lista kroków

1. Zbudowanie grafu
2. Zaczynamy z losowego wierzchołka
3. Znajdujemy k najbliższych miast
4. Do każdej z k dotyczasowych ścieżek liczymy odległość po dodaniu każdego z wierzchołków z uwzględnieniem kary za powtórne odwiedzenie tego samego wierzhołka.Wybieramy k najkrótszych ścieżek
5. Powtarzamy krok 4 aż koszt ścieżek przekroczy krytyczną wartość $n - l$
6. Dla każdej z k wygenerowanych ścieżek sprawdzamy które przestawienie parami pozwoli maksymalnie zmniejszyć koszt ścieżki i wykonujemy je
7. Powarzamy krok 6 aż do uzyskania lokalnego optimum
8. Ponownie wykonujemy przeszukiwanie wiązkowe

2.3 Złożoność obliczeniowa

Złożoność pierwszego przeszukiwania wiązkowego jest równa $O(k * |S|^2)$
Złożoność algorytmu wspinaczkowego (jeszcze nie wiem)
Złożoność drugiego przeszukiwania wiązkowego $O(k * |S|^2)$