Bioinformatyka – Sprawozdanie

Autorzy:

Omar Shelbayeh

Jakub Raczkiewicz

Temat projektu: Sekwencjonowanie DNA

Cel: Zaprojektowanie i zaimplementowanie algorytmu heurystycznego do sekwencjonowania DNA którego złożoność obliczeniowa jest wielomianowa.

Problem ogólny:

Celem algorytmu jest sekwencjonowanie DNA. Jako parametry wejściowe zostaje mu podana żądana długość sekwencji oraz słowa składające się z liter {A, C, T, G}. Każde słowo ma długość „L". Znalezienie perfekcyjnego dopasowania jest bardzo kosztowne w złożoności obliczeniowej gdyż O(n) = n! Z tego powodu trzeba wymyślić i zaimplementować heurystyczną metodykę znajdywania sekwencji z podanych słów. W naszym podejściu pierwszym krokiem będzie zbudowanie macierzy sąsiedztwa. Macierz ta będzie wypełniona dopasowaniami tj. liczbami w ilu znakach słowo jest dopasowane do innego.

Przykład:

{ACTGCTG}

{CTGAGTC}

Dopasowanie dla tych słów będzie wynosiło 3 gdyż litery zaznaczone na zielono się pokrywają. Tak wypełniona macierz będzie naszą przestrzenią przeszukiwań. Kolejnym krokiem będzie implementacja algorytmu mrówkowego. Każda mrówka na podstawie feromonu czyli w naszym przypadku jakości dopasowania będzie podążać za śladem aż nie dojdzie do momentu gdy nie może już przejść dalej. Każda mrówka będzie posiadała listę Visited aby móc odznaczać wierzchołki w których już była. Powoduje to że nie dojedzie do cykli a mrówka zawsze skończy swoją drogę w skończonej ilości kroków.

Błędy

W tym problemie charakteryzujemy 2 typy błędów. Pozytywne i negatywne.

Błędy pozytywne charakteryzują się tym że w bazie słów wystąpiły takie które się nie pojawiły w podstawowej sekwencji. Powoduje to że takie słowa są ślepymi zaułkami dla

mrówki. Każda mrówka która dojdzie do błędu pozytywnego nie będzie miała już gdzie iść dlatego wykrycie takiego wierzchołka nie będzie trudne.

Błędy negatywne polegają na utracie pewnych informacji. W skrajnych przypadkach spowoduje to przerwanie całej sekwencji i w konsekwencji 2 łańcuchy wynikowe. Wiadomością że znalezione zostało najlepsze dopasowanie jest fakt iż jakość dopasowania jest mniejsza lub równa N – L + 1 gdzie N oznacza ilość słów a L Długość słów. Gdy takowe dopasowanie mrówka znajdzie i nie będzie miała już gdzie iść będzie to dopasowanie perfekcyjne.

Możliwości do optymalizacji:

Na wstępie widać że przy błędach pozytywnych można zauważyć że przepatrywanie wierzchołka do przodu można zoptymalizować. Każdy wierzchołek powinien mieć w sobie parametr dopasowanie wierzchołków. Wartość ta reprezentowałaby ilość wierzchołków które są do niego dopasowane chociaż w jednej literce. Gdy takowych nie ma można założyć że jest to błąd pozytywny i nie brać go pod uwagę. Spowoduje to jeszcze lepsze wykorzystanie każdej mrówki i uniemożliwienie im wchodzenia w ślepe zaułki.

Kolejną możliwością jest wprowadzenie parametru odcięcia. Byłoby to minimalne dopasowanie. Wynosiłoby np. 3. Każdy wierzchołek który jest dopasowany w stopniu mniejszym niż 3 nie byłby jego sąsiadem. Unikałoby to w jeszcze lepszy sposób ślepych zaułków dla mrówek. Zagrożeniem takiego podejścia jest niestety fakt iż jeżeli z sekwencji zniknie 4 kolejne słowa a długość L = 7 to dopasowanie na tym utraci.