Sprawozdanie 1
Teoretyczne opracowanie metody heurystycznej

Mateusz Babiaczyk, Bartosz Nawrotek

2018-04-08

1 Wprowadzenie

Sekwencjonowanie DNA polega na odczytywaniu sekwencji, czyli kolejności par nukleotydowych w cząsteczce DNA. Obejmuje ona dowolną metodę lub technologię, która jest stosowana do określenia porządku czterech zasad: adeniny, guaniny, cytozyny i tyminy w nici DNA.
Do rozwiązania problemu zastosowano hybrydę algorytmu genetycznego.

2 Opis algorytmu

2.1 Kodowanie

Osobniki są kodowane jako wektory liczb całkowitych w których indeks oznacza kolejność występowania oligonukleotydu w sekwencji:

$$X = [x_1, x_2, \ldots, x_m], \tag{1}$$

gdzie:

1. m - wielkość zbioru dostępnych oligonukleotydów,
2. x_i dla $i \in \langle 1, m \rangle$ - indeks oligonukleotydu w liście dostępnych oligonukleotydów.

Jako rozwiązanie zostanie traktowana sekwencja zbudowana z oligonukleotydów począwszy od x_1 nie przekraczająca długości n.

2.2 Generowanie populacji początkowej

Tworzone jest s - osobników. Osobniki są permutacjami liczb całkowitych.

2.3 Funkcja oceny

Jako minimalizowaną funkcję oceny osobnika przyjęto następującą postać addytywną:

$$f_{min}(X) = 1,5 \sum_{i=1}^{i=k-1} [len(x_i, x_{i+1}) - l + 1] + \sum_{i=k}^{i=m-1} [len(x_i, x_{i+1}) - l + 1] \tag{2}$$

gdzie:

1. k - taka liczba całkowita, dla której długość sekwencji $[x_1, x_2, \ldots, x_k]$ będącej złożeniem k pierwszych oligonukleotydów osobnika X jest mniejsza lub równa n,
2. $len(x, y)$ jest długością sekwencji otrzymanej z połączenia oligonukleotydów x oraz y,
3. l - długość oligonukleotydu.

Przyjmując funkcję kosztu powyższej postaci zwrócono szczególną uwagę na pierwsze k oligonukleotydów z osobnika, które zawierają szukane rozwiązanie. Druga część sumy ma na celu nielosowe uporządkowanie pozostałych oligonukleotydów, które będą wykorzystane w operacji krzyżowania.

2.4 Wybór osobników do krzyżowania

Losuje się c osobników z populacji. Dla trójek kolejnych osobników wybiera się z nich tego o najniższej funkcji oceny, następnie jest on dodawany do populacji rodzicielskiej.

2.5 Krzyżowanie

Krzyżowanie będzie odbywało się metodą podobną do *krzyżowania z zachowaniem porządku*. Losowane są dwa punkty. Do pierwszego potomka kopiowany jest fragment pomiędzy punktami krzyżowania z pierwszego rodzica. Fragmenty uzupełniane są od drugiego punktu krzyżowania, następnie uzupełniany jest początek chromosomu. Przyjęto tę metodę, aby rozwiązania po krzyżowaniu nadal należały do zbioru rozwiązań dopuszczalnych. Dodatkową zaletą tej metody jest występowanie stosunkowo niewielkich zmian w kolejności oligonukleotydów, co potencjalnie zmniejszy ryzyko utraty ciągów dokładnych dopasowań. W naszym rozwiązaniu jednak zdecydowano się na modyfikację fazy uzupełniania chromosomu w stosunku do *krzyżowania z zachowaniem porządku*. Pomija się oligonukleotydy, które w potomku już występują. Uzupełnia się rozwiązanie o brakujące oligonukleotydy dobierając miejsca z najmniejszym pogorszeniem funkcji celu.

2.6 Mutacje

Osobniki wybierane są losowo zarówno z populacji przodków, jak i potomków. Następnie, dla danego osobnika wybiera się oligonukleotyd najsłabiej pasujący do sąsiednich oligonukleotydów, który zamieniany jest miejscem z oligonukleotydem sąsiednim w taki sposób aby minimalizować funkcję kosztu. Zabieg ten ma na celu przyspieszenie zbiegania rozwiązań do optimum. Prawdopodobieństwo występowania mutacji jest równe u razy wielkość instancji.

2.7 Wybór nowej populacji

Wybór nowej populacji polegać będzie na zachowaniu d chromosomów z populacji rodziców oraz wszystkich chromosomów z populacji potomków. Gdzie d jest stałą zapewniającą brak zmiany wielkości populacji.

2.8 Dobór parametrów algorytmu

Dobór parametrów algorytmu: s, c, u wyznaczony będzie eksperymentalnie przez algorytm najszybszego spadku gradientu.

2.9 Błędy jednego typu

W przypadku występowania tylko błędów negatywnych w problemie, dodatkową informacją jest to, iż długość sekwencji zbudowanej z olinukleotydów będzie mniejsza lub równa n. Wtedy wystarczyło by usunąć w naszej metodzie możliwość wychodzenia poza tą długość.
W przypadku występowania tylko błędów pozytywnych, można by było odrzucać nukleotydy które kompletnie nie pasują do naszego ciągu, ale takie których brak nic nie zmieni. Prawdopodobnie takie nukleotydy byłyby właśnie tymi błędami.