**Bioinformatyka - sprawozdanie 1**

Filip Śnieguła 122458

Tomasz Peliński 122474

**I. Opis metody**

**I.1 Metaheurystyka**

Wybraliśmy algorytm genetyczny.

**I.2 Przedprzetwarzanie:**

Po pobraniu danych wejściowych tworzymy graf pełny, w którym wierzchołkami są oligonukleotydy, a wagi łuków są równe przesunięciu jakie należy wykonać, żeby dwa nukleotydy się nałożyły.

Przed przystąpieniem do głównej pętli algorytmu upraszczamy zbiór oligonukleotydów wejściowych w taki sposób, że:

a) jeśli dla danego olinukleotydu (o1) istnieje tylko jeden następujący oligonukleotyd (o2) oddalony o 1 i równocześnie o1 jest jedynym oligonukleotycem poprzedzającym o2 o 1 to tworzymy nowy oligonukleotyd składający się ze wszszystkich zasad o1 i ostatniej zasady o2, ustawiamy jego długość na długość o1 + 1, jego indeks początkowy na indeks o1, indeks końcowy na o2 i zastępujemy nim o1 i o2 w grafie;

b) jeśli dla danego olinukleotydu (o1) istnieje tylko jeden poprzedzający oligonukleotyd (o2) oddalony o 1 i równocześnie o1 jest jedynym oligonukleotydem następującym o2 o 1 to tworzymy nowy oligonukleotyd składający się ze wszszystkich zasad o2 i ostatniej zasady o1, ustawiamy jego długość na długość o1 + 1, jego indeks początkowy na indeks o1, indeks końcowy na o2 i zastępujemy nim o1 i o2 w grafie.

Do głównej części algorytmu przekazujemy te oligonukleotydy, których indeks zgadza się z ich indeksem początkowym.

**I.3 Główna część algorytmu**

Algorytm składa się z czterech głównych części: generowania populacji rozwiązań początkowych oraz wykonywanych cyklicznie mutacji i krzyżowania rozwiązań (skutkujące zwiększeniem liczności populacji rozwiązań) i turnieju, który po dwóch poprzednich etapach przywraca populację do początkowego rozmiaru, z zadaną liczbą iteracji.

**I.3.1 Generowanie populacji początkowej:** Tworzona jest zadana liczba rozwiązań, do których oligonukleotydy dodawane są w losowej kolejności.

**I.3.2 Mutacja:** dla każdego rozwiązania z populacji losowana jest wartość, która porównywana jest z parametrem prawdopodobieństwa ewolucji i, jeśli jest mniejsza, przeprowadzana jest ewolucja tego rozwiązania. Polega ona na przesunięciu losowo wybranego oligonukleotydu w rozwiązaniu o wylosowaną liczbę pozycji, przy czym jeżeli oligonukleotyd znalazłby się przed początkiem rozwiązania ustawiany jest na jego początku, zaś jeśli znalazłaby się za jego końcem przesunięcie w ogóle nie zachodzi.

**I.3.3 Krzyżowanie:** dla każdego rozwiązania z populacji losowana jest wartość, która porównywana jest z parametrem prawdopodobieństwa krzyżowania i, jeśli jest mniejsza, przeprowadzana jest krzyżowanie tego rozwiązania z losowo wybranym innym rozwiązaniem. Podczas krzyżowania oligonukletody pierwszego rozwiązania są rozdzielone na dwa zbiorów.

Operacje z pierwszej połowy pierwszego rozwiązania według liczby słów są dodawane do pierwszej połowy rozwiązania wynikowego według kolejności w pierwszym rozwiązaniu, a te z drugiej połowy według kolejności w rozwiązaniu drugim.

**I.3.4 Turniej:** Przed rozpoczęciem turnieju ustalana jest wartość funkcji celu dla wszystkich rozwiązań - liczba słów mieszczących się w zadanej długości łańcucha wynikowego.

W trakcie turnieju z grupy w kolejnych w rozwiązań, gdzie w = aktualny rozmiar populacji / rozmiar populacji początkowej (zaokrąglając w górę), wybierane jest to o największej wartości funkcji celu i dodawane do populacji wynikowej tej iteracji.

**II. Przypadki o określonych błędach**

**II.1 Błędy tylko pozytywne**

W przypadku wyłącznie takich błędów, dokładnie n-l+1 oligonukleotydów jest prawidłowych i wejdzie do rozwiązania. Na etapie wstępnego przetwarzania, dla każdego oligonukleotydu nie znajdzie się następnik/poprzednik jedynie w przypadku gdy jest to element końcowy/początkowy lub błędny. Jeśli będzie tylko jeden element bez następnika/poprzednika, ale dopasowany z drugiej strony to możemy przyjąć go jako ostatni/pierwszy, bez możliwości dalszego przesuwania. Element, który nie daje się dopasować z żadnej ze stron jest na pewno błędny i możemy go z góry odrzucić.

**II.2 Błędy tylko negatywne**

Tu również po ilości słów możemy znaleźć liczbę błędów, jest to liczba, która brakuje do n-l+1. Z góry wiemy, że wszystkie oligonukleotydy muszą zmieścić się w rozwiązaniu. Suma odległości większych niż. W przypadku ilości błędów mniejszej od l, możemy z góry odrzucić wszystkie połączenia o większej wadze.