Imię i Nazwisko	Przedmiot	Data oddania	Ocena
Paulina Roszkowska Aleksandra Szum	Algorytmy i struktury danych	9 stycznia 2020	
Sprawozdanie II			
Temat sprawozdania	Odległość Levenshteina		

1 Opis ćwiczenia

Celem ćwiczenia było obliczenie odległości Levenshteina w sekwencji DNA. Zadanie wymagało dodatkowo sprawdzenia nierówności trójkąta, a także zaproponowanie liczby insercji, delecji oraz transpozycji przeprowadzających jedną sekwencję w inną.

Najpierw wygenerowano sekwencje DNA, a następnie wykonano trzykrotnie mutacje. Następnie wyznaczono odległość Levenshteina wygenerowanej sekwencji DNA z sekwencjami po mutacjach i wyciągnięto wnioski.

2 Wstęp teoretyczny

Odległością Levenshteina nazywa się miarę odmienności ciągów znaków, zdefiniowaną następująco: działaniem prostym (wstawienie nowego znaku do napisu, usunięcie znaku z napisu, zamianę znaku w napisie na inny znak) i odległością pomiędzy dwoma napisami. Miara ta znajduje zastosowanie w przetwarzaniu informacji, danych w postaci ciągów symboli: w maszynowym rozpoznawaniu mowy, analizie DNA czy rozpoznawaniu plagiatów.

Sekwencja DNA to kolejność nukleotydów w cząsteczce DNA. Oznaczana jest za pomocą skrótów od zasad wchodzących w skład nukleotydów.

Zmianę w sekwencji DNA nazywamy mutacją. Mutacje można podzielić na:

- delecję czyli zmianę w materiale genetycznym polegającą na utracie jego fragmentu.
- insercję czyli najczęściej spontaniczną mutację genu polegającą na wstawieniu krótkiej sekwencji DNA w obrębie pojedynczego genu albo wstawieniu dłuższego fragmentu chromosomu.
- transpozycję, która powoduje mutacje i może zmieniać ilość DNA w genomie.

Nierównością trójkąta nazywa się twierdzenie matematyczne które mówi, że dla dowolnego trójkąta miara każdego boku musi być mniejsza lub równa sumie miar dwóch pozostałych, ale większa lub równa od różnicy ich miar. W przestrzeni metrycznej zapisuje się ją za pomocą metryki:

$$d(x,z) \leqslant d(x,y) + d(y,z) \qquad \forall x,y,z \in X \tag{1}$$

co oznacza, że odległość między x, a z jest nie większa, niż suma odległości od x do y oraz y do z.

3 Obliczenia numeryczne

Kod znajduje się na stronie `https://github.com/aleksandraszum/AiSD_projekty/tree/master/report2`.

4 Analiza wyników

Oryginalna sekwencja przedstawiała się następująco:

– agaggtgtactgggaaagtcgggattgtccatgatgcggtggggtagggc.

Przeprowadzone mutacje przedstawiały się następująco:

1. mutacja: agaggtgtcctgggaaaggaggcattatccaaggtacggtggggtaggat
2. mutacja: agaggcgtacgggcattttcaagattgtccatgatgcggtggagtagcgc
3. mutacja: agaggtgtactaggaaaatcgtgattatgcatgctgaggcggggtcgagc.

4.1 Odległość Levensteina między oryginalną sekwencją, a jej mutacjami

Wyznaczone przez funkcję macierze odległości znajdują się w rozdziale 7. Czerwona krzywa przedstawia zmiany edycyjne analizowanych ciągów znaków.

W wierszach macierzy jest określony koszt wstawiania nowego znaku. W kolumnie macierzy określone są koszty usunięcia znaku. Na przekątnej macierzy określony jest koszt zastąpienia danego znaku. Przykładowo, gdyby analizować dwa identyczne ciągi znaków, na przekątnej powinna znaleźć się liczba 0. W związku z powyższym, w ostatnim elemencie macierzy Levensteina określona jest minimalna liczba kroków, która umożliwia przekształcenie jednego ciągu znaków w drugi.

Zgodnie z oczekiwaniami, odległości między oryginalną sekwencją oraz jej mutacją wynosi 10, co związane jest z faktem, że dokonano 10 mutacji.

4.1.1 Zaproponowana liczba insercji, delecji oraz substytucji

Zaproponowana przez Autorki sprawozdania liczba insercji, delecji oraz transpozycji przedstawia się następująco:

mutacja 1: insercja: 0, delecja: 0, substytucja: 10, gdzie tranzycja: 5, transwersja: 5.

mutacja 2: insercja: 0, delecja: 0, substytucja: 10, gdzie tranzycja: 4, transwersja: 6.

mutacja 3: insercja: 0, delecja: 0, substytucja: 10, gdzie tranzycja: 5, transwersja: 5.

4.1.2 Odległość Levenshteina między mutacjami

Wyznaczono macierze odległości Levenshteina między mutacjami. W celu zapewnienia czytelności sprawozdania, nie umieszczono pełnych macierzy w sprawozdaniu. W tabeli 1 zawarto odległości między danymi mutacjami.

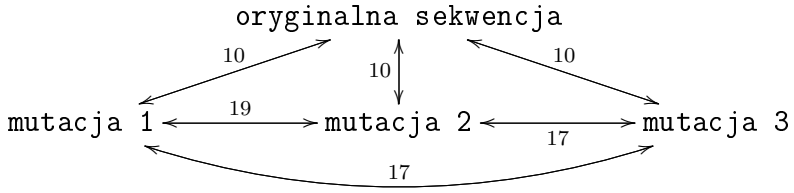
Tabela 1: Wyznaczona ogledłość Levenshteina między mutacjami

Porównane mutacje	Odległość Levenshteina
Mutacja1 i Mutacja2	19
Mutacja1 i Mutacja3	17
Mutacja2 i Mutacja3	17

4.1.3 Nierówność trójkąta

Odległość Levensteina należy do metryki, zatem spełnia zasady nierówności trójkąta. Można ją wyrazić za pomocą wzoru 1.

Poniższy schemat przedstawia odległości między poszczególnymi ciągami znaków:



Powyższy schemat udowadnia zachodzenie zasady nierówności trójkąta. Podstawiając do wzoru 1 dowolną trójkę odległości z powyższego schematu uzyska się spełnienie nierówności.

5 Wnioski

Odległość Levenshteina daje możliwość wyznaczenia minimalnej liczby kroków związanych ze zmianą jednego ciągu znaków w inny ciąg znaków.

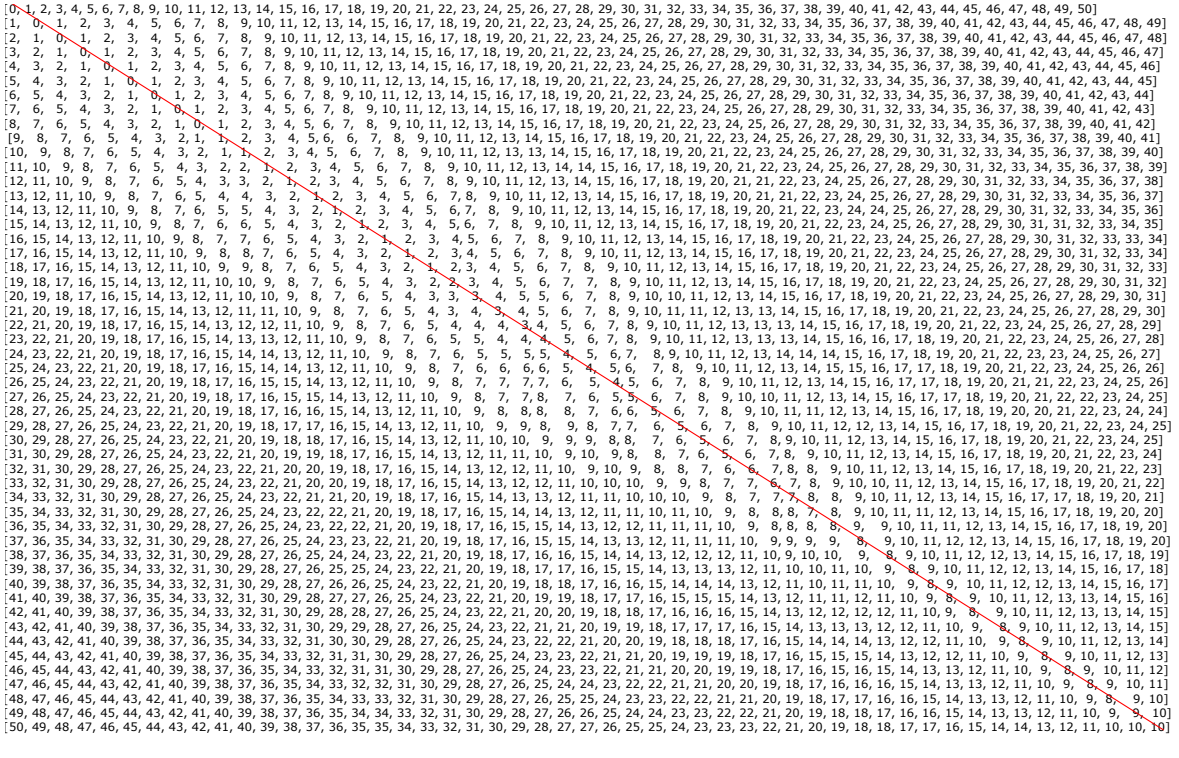
Po dodaniu 10 mutacji do wygenerowanej wcześniej sekwencji DNA i analizie ciągu znaków z wykorzystaniem algorytmu wyznaczania odległości Levenshteina, uzyskano odległości Levenstheina wynoszące 10. Był to wynik oczekiwany, związany z liczbą zaaplikowanych mutacji.

Odległość Levenshteina jest metryką i spełnia nierówność trójkąta, co zostało udowodnione w trakcie analizy.

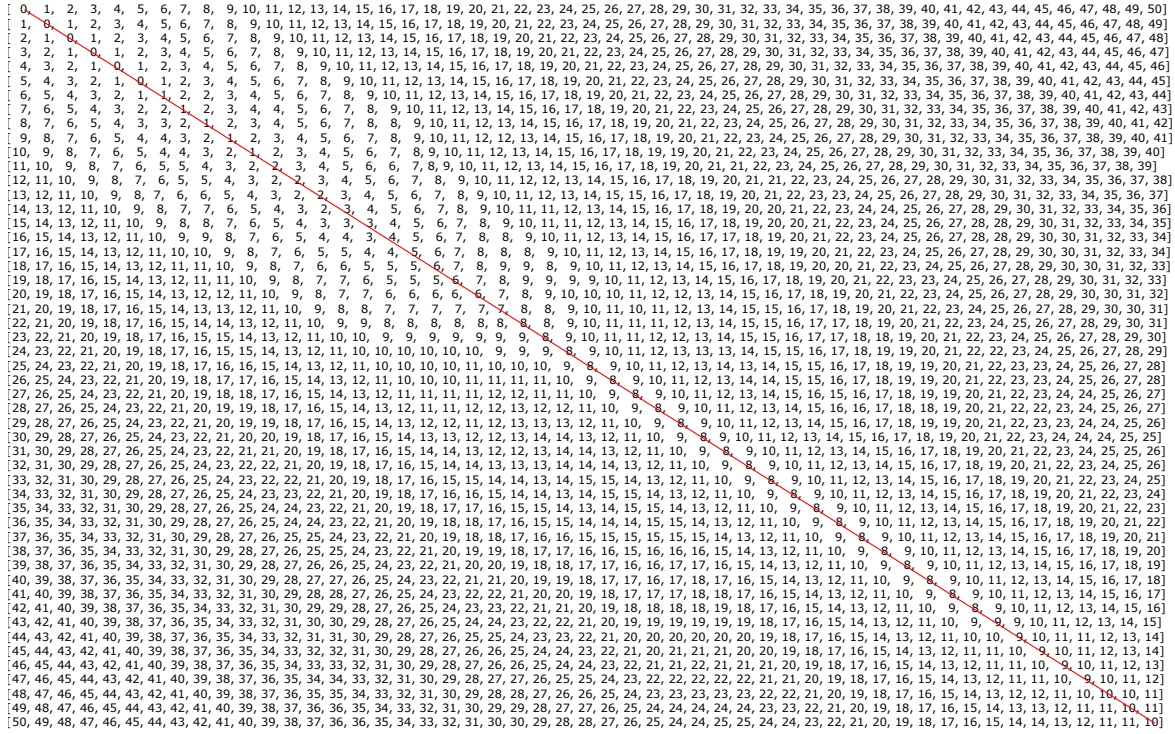
6 Bibliografia

1. Malinowski, P., *Algorytmy i Struktury Danych, Wykład 7*, Biostatystyka 2019/2020
2. Devopedia, *Levenshtein Distance*, dostęp online: `https://devopedia.org/levenshtein-distance`
3. *Levenshtein Distance and the Triangle Inequality* , dostęp online: `http://richardminerich.com/2012/09/levenshtein-distance-and-the-triangle-inequality/`
4. *Levenshtein Distance and the Triangle Inequality*, dostęp online: `https://jeremykun.com/tag/levenshtein-distance/`
5. *Portal Algorytm*, dostęp online: `http://www.algorytm.org/`
6. *Opis mutacji na Wikipedii*, dostęp online: `https://pl.wikipedia.org/wiki/Mutacja`
7. *Portal Edunauka*, dostęp online: `http://www.edunauka.pl/biomolmutacja.php`

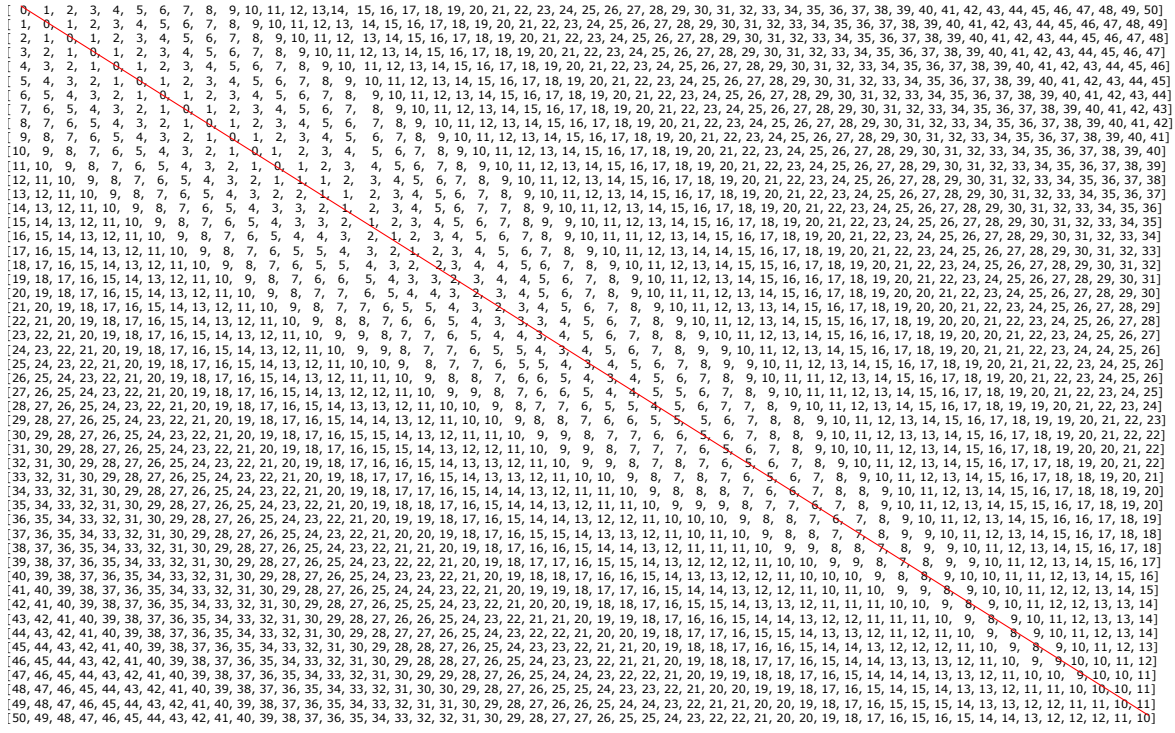
7 Załącznik



Rysunek 1: Macierz odległości Levensteina dla oryginalnej sekwencji i mutacji 1.



Rysunek 2: Macierz odległości Levensteina dla oryginalnej sekwencji i mutacji 2.



Rysunek 3: Macierz odległości Levensteina dla oryginalnej sekwencji i mutacji 3.