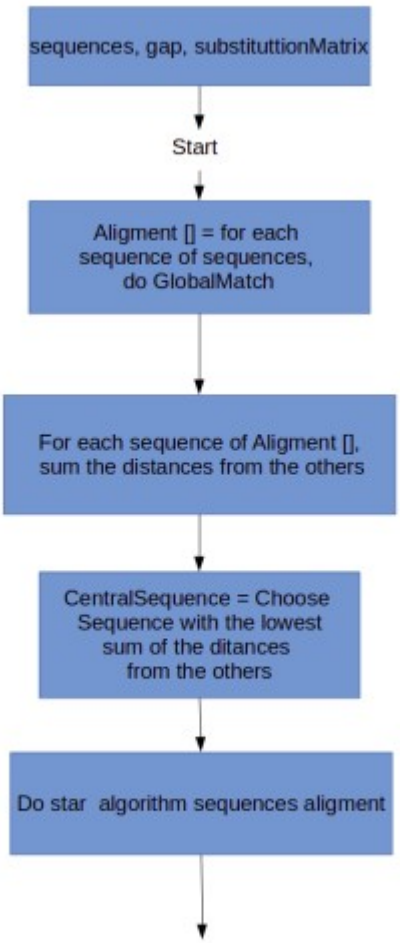
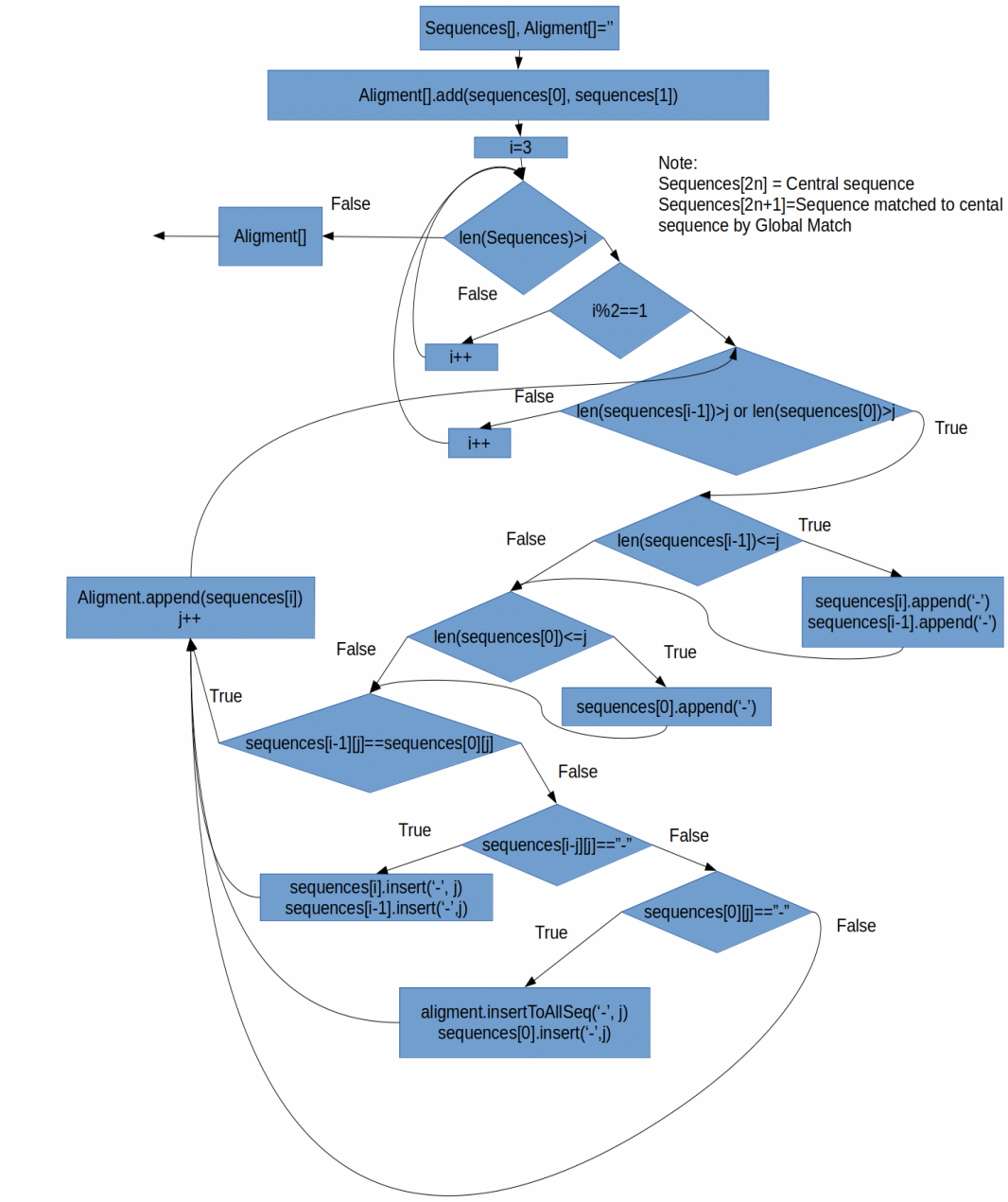
Anna Lasota
236727
Zadanie 4 (MSA)

1. Ogólny schemat blokowy algorytmu dopasowania wielu sekwencji:



Rysunek nr 1, Ogólny schemat blokowy algorytmu dopasowania wielu sekwencji.

Schemat dopasowania sekwencji za pomocą algorytmu na gwiazdę przedstawiono na rysunku nr 2



Rysunek nr 2, Schemat dopasowanie wielu sekwencji za pomocą algorytmu na gwiazdę

2. Analiza złożoności obliczeniowej czasowej i pamięciowej:

2.1 Dopasowanie globalne wielu sekwencji:
- czasowa: $O(m*n*o^2)$
- pamięciowa: $O(m*n*o^2)$

Gdzie m i n to długości sekwencji, a o to ilość dopasowywanych sekwencji

2.2 Zsumowanie odległości do pozostałych sekwencji dla każdej sekwencji:
- czasowa: $O(o^2)$
- pamięciowa: (o)

2.3 Wybranie sekwencji o najmniejszej sumie odległości:
Wielkości pomijalne

2.4 Dopasowanie wielu sekwencji przy pomocy algorytmu na gwiazdę:
- czasowa: O(m*n*o)
- pamięciowa: O(o)

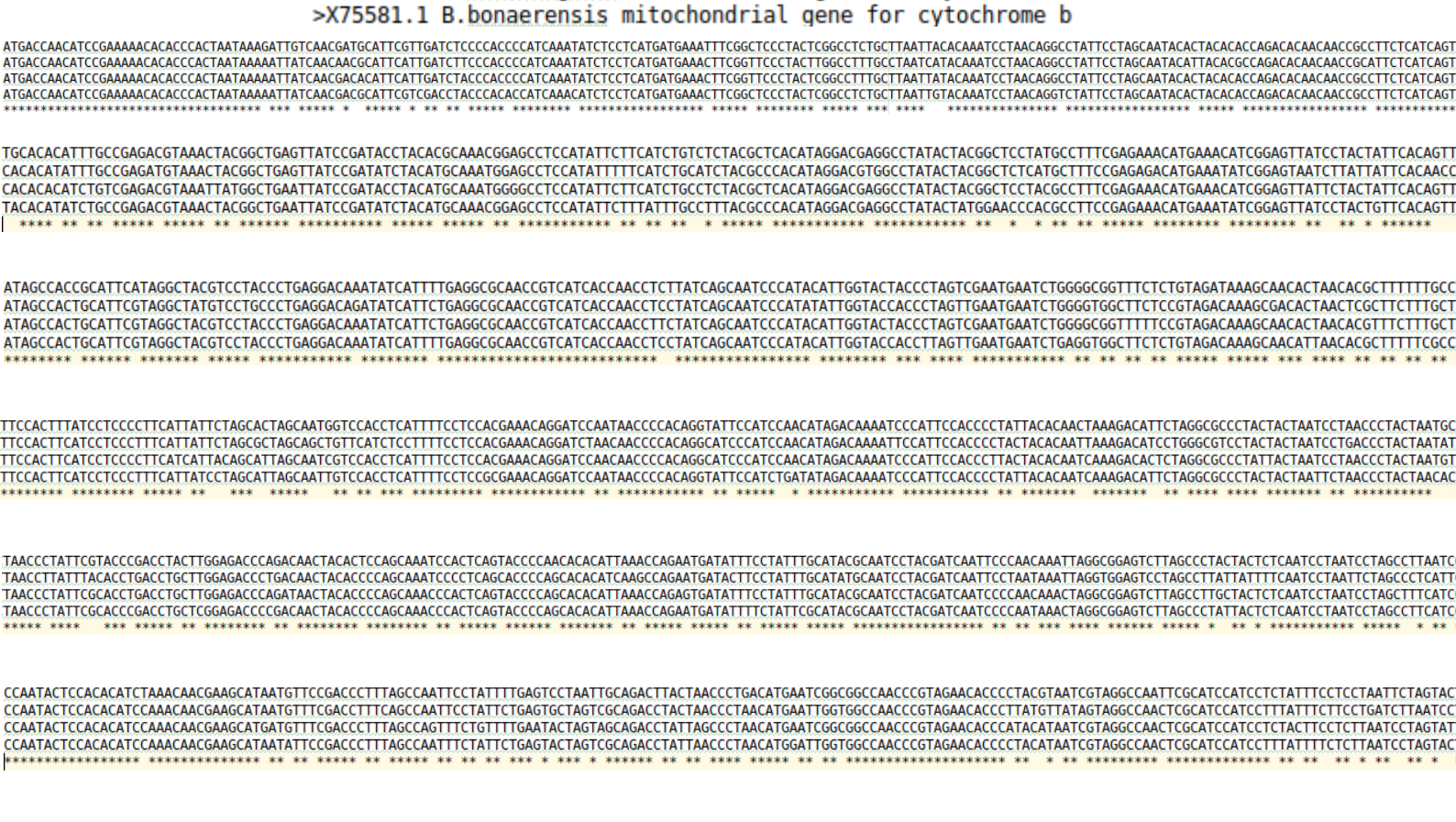
2.5 Całokształt:
- czasowa: $O(m*n*o^2)$
- pamięciowa: $O(m*n*o^2)$

3. Przykładowe dopasowanie dla czterech powiązanych ewolucyjnie sekwencji

Wybrano cztery sekwencje kodujące cytochrom b, zwierząt takich jak:
- walenik mały (*Caperea marginata), GenBank: X75586.1*
- *długopłetwiec oceaniczny (Megaptera novaeangliae), GenBank: X75584.1*
- *płetwal Bryde'a (Balaenoptera edeni), GenBank: X75583.1*
- *płetwal antarktyczny (Balaenoptera bonaerensis), GenBank: X75581.1*

Wyniki:

```
>X75583.1 B.edeni mitochondrial gene for cytochrome b
>X75586.1 C.marginata mitochondrial gene for cytochrome b
>X75584.1 M.novaeangliae mitochondrial gene for cytochrome b
>X75581.1 B.bonaerensis mitochondrial gene for cytochrome b
```



```
TAATACCAGTAACTAGTCTTATCGAGAATAAACTTATAAAATGAAGA
TAATGCCAGTAACCAGTCTTATCGAAAATAAACTTATAAAATGAAGA
TAATACCAATAACTAGTCTTATCGAGAACAAACTTATAAAATGAAGA
TAATACCAGTAGCTAGCCTTATCGAGAACAAGCTTATAAAATGAAGA
**** *** ** * ** ******** ** ** ***************
```

4 Wnioski:

- Na podstawie uzyskanych wyników, możemy stwierdzić, iż gen kodujący cytochrom b jest dobrze zachowany, w procesie ewolucji nie doszło do znacznych zmian w jego sekwencji. Świadczy o tym mała ilość kolumn niezakonserwowanych i bardzo mała mała przerwa w wydruku gwiazdek (maksymalna przerwa w wydruku gwiazdek = 3).
- Zmiany powodujące brak konserwacji kolumny dotyczą przeważnie substytucji jednego nukleotydu, co może również świadczyć o bardzo silnej konserwacji badanego genu.
- Algorytm dopasowania sekwencji na gwiazdę możemy uznać za dużo bardziej wydajny (złożoność czasowa: $O(m*n*o^2)$) od tradycyjnego MSA (złożoność obliczeniowa $O(n^N)$, N - liczba sekwencji)