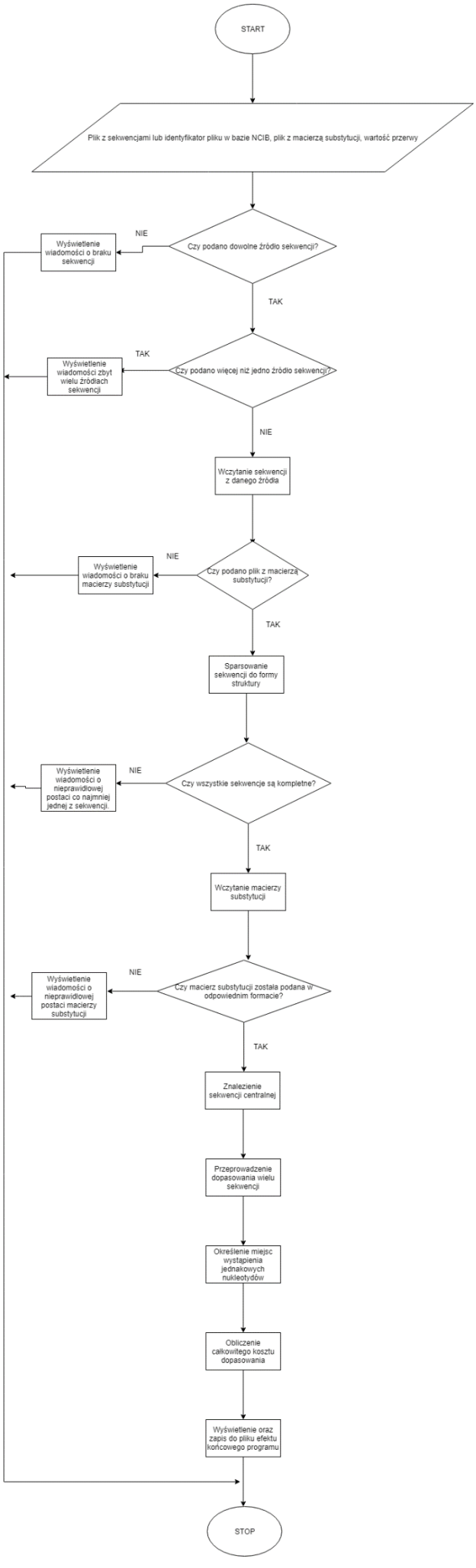
Jakub Cieplucha 230331

Wstęp do bioinformatyki

Zdanie 4

1. **Schemat blokowy algorytmu dopasowania wielu sekwencji.**


START
Plik z sekwencjami lub identyfikator pliku w bazie NCIB, plik z macierzą substytucji, wartość przerwy
Czy podano dowolne źródło sekwencji?
NIE
Wyświetlenie wiadomości o braku sekwencji
TAK
Czy podano więcej niż jedno źródło sekwencji?
TAK
Wyświetlenie wiadomości zbyt wielu źródłach sekwencji
NIE
Wczytanie sekwencji z danego źródła
Czy podano plik z macierzą substytucji?
NIE
Wyświetlenie wiadomości o braku macierzy substytucji
TAK
Sparsowanie sekwencji do formy struktury
Czy wszystkie sekwencje są kompletne?
NIE
Wyświetlenie wiadomości o nieprawidłowej postaci co najmniej jednej z sekwencji.
TAK
Wczytanie macierzy substytucji
Czy macierz substytucji została podana w odpowiednim formacie?
NIE
Wyświetlenie wiadomości o nieprawidłowej postaci macierzy substytucji
TAK
Znalezienie sekwencji centralnej
Przeprowadzenie dopasowania wielu sekwencji
Określenie miejsc wystąpienia jednakowych nukleotydów
Obliczenie całkowitego kosztu dopasowania
Wyświetlenie oraz zapis do pliku efektu końcowego programu
STOP

2. Analiza złożoności obliczeniowej czasowej i pamięciowej.

2.1 Funkcja multipleSequenceAligment

a)czasowa

O(2nm^3)

b)pamięciowa

n - rozmiar sekwencji wejściowych

m - rozmiar sekwencji po zmianach

S(n,m)= 4n(n+m+nm+5+n+m+nm+3) + 4m(n+m+nm+5+n+m+nm+3) + 2 =8n^2m+8n^2+8nm+32n+8nm^2+8m^2+8nm+32m=8n^2m+8nm^2+8n^2+8m^2+16nm+32n+32m+2

2.2 Funkcja progressiveMSA

a)czasowa

O(mn)

b)pamięciowa

nm – rozmiar macierzy

n + m – rozmiar sekwencji wejściowych

S(n,m)= n+m+nm+5

2.3 Razem

a) Czasowa: O(2nm^3)
b) Pamięciowa : S(n,m)= 8n^2m+8nm^2+8n^2+8m^2+17nm+33n+33m +7

3.Porównanie przykładowych par sekwencji ewolucyjnie powiązanych.

```
>> zadanieMSA('filename1','sequence.fasta','filename2','sequence1.fasta','subMatrixFile','msa.xlsx');
Complete cost : 24
X3333333.3 G_ACA_  4
X1111111.1 GCACAT  6
X2222222.2 TGAGA_  5
X4444444.4 GAACT_  5
             *
```

Rysunek nr 1. Dopasowanie wielu sekwencji dla przykładu pokazanego na wykładzie.

W celu określenia skuteczności opracowanej metody dopasowania wielu sekwencji, przetestowano ją na omówionym na wykładzie wzorcu. Uzyskano niższy o 6 punktów całkowity koszt dopasowania. Wynikać może to z faktu, iż w stworzonym rozwiązaniu nie stosuje się zachowania przerw, co nieuchronnie prowadzi do zwiększenia całkowitego kosztu dopasowania. Zamiast tego wpierw sekwencję centralną dopasowuje się po kolei do reszty sekwencji, po czym pozostałe sekwencje dopasowuje się do uprzednio zmienionej sekwencji centralnej.

Command Window

```
>> zadanieMSA('filename1','file1.fasta','filename2','file2.fasta','subMatrixFile','msa.xlsx');
Complete cost : 1422
AJ130824.1 A__TGA___CCAACATTCGTA___AA__AC  20
AJ314566.1 TCGTGAAAACCACCGTTGTTATTCAACTAC  30
Z12030.1   A__TGG___CCAACCTCCGAA___AA__AC  20
M28016.1   A__AGC___CCGA_ATGA_TA___TT__TC  18
               *    **    *    *        *

AJ130824.1 __G__C__A_T__CC__CCTACTAAAAA__  47
AJ314566.1 AAGAACCTAATGGCCAACCTCCGAAAAACC  60
Z12030.1   __C__C__A_C__CC__CCTCCTAAAAA__  47
M28016.1   _____C_TA_T__TC_GCCTAC_ACAA___  46
                *  *     *  *** * * **

AJ130824.1 __T__TGT__TAA____C_C_ACT__C_C_  74
AJ314566.1 CATCCTCTCCTAAAAATCGCTAATGACGCA  90
Z12030.1   __T__TGC__AAA____C_G_ACG__C_A_  74
M28016.1   __T__TCTCCGATCCGTC_C__CTAACAAA  83
             *  *     *     *        *

AJ130824.1 CTAATCGACCTTCCCGCCCCCTCAAATATC  120
AJ314566.1 CTAGTCGACCTCCCAGCACCCTCTAACATT  120
Z12030.1   CTAGTTGATCTCCCAGCTCCTTCAAACATT  120
M28016.1   CTAGGAGGCGTCCTTGCCCTATTAC_TATC  119
           ***   *   * *  ** *  *     **

AJ130824.1 TCTGCCTGATGAAACTTCGGCTCTCTATTG  150
AJ314566.1 TCAGTCTGATGAAACTTTGGCTCACTCCTA  150
Z12030.1   TCTGTTTGATGAAACTTTGGCTCCCTGCTA  150
M28016.1   CAT_CCTCATCCTAGCAATAATCCCCATCC  149
                * **   *       ** *

AJ130824.1 GGATTATG_CCTAATAATCCAG_ATCCTAA  178
AJ314566.1 GGCCTATG_TTTAGCCACCCAA_ATTCTTA  178
Z12030.1   GGGCTCTG_TCTAGCTGCCCAA_ATCCTGA  178
M28016.1   TCCATATATCCAAACAACAAAGCATAATAT  180
               * *     *       *  **  *

AJ130824.1 CTGGCTTATTCTTAGCCATACACTACACAT  210
AJ314566.1 CCGGGCTCTTCTTAGCCATACACTATACCT  210
Z12030.1   CAGGCCTCTTCCTAGCCATACATTACACCT  210
M28016.1   TTCGCCCA__CTAAGCCAATCACTTTAT_T  207
             *      *  *****  ** *  *  *

AJ130824.1 CAGACACAGCAACAGCATTCTCCTCAGTTA  240
AJ314566.1 CCGACATTTCAACAGCTTTTTCCTCTGTCT  240
Z12030.1   CCGATATCGCCACCGCCTTTTCCTCCGTTG  240
M28016.1   GACTCCTAGCCGCAG_AC_CTCCTCATTCT  238
                   *  * *     *****  *
```

```
                     *  * *     *****  *
AJ130824.1 CACATATTTGCCGAG_ACGTAAACTACGGC  269
AJ314566.1 GCCACATTTGCCGAG_ATGTTAGTTACGGC  269
Z12030.1   CCCACATCTGCCGTG_ATGTTAATTACGGC  269
M28016.1   AACCTGAAT_CGGAGGACAACCAGTAAGCT  269
             *     * * * * *       ** *
AJ130824.1 TGACTTATTCGTTACTTACACGCCAATGGA  300
AJ314566.1 TGACTCATTCGAAATATCCACGCCAACGGG  300
Z12030.1   TGACTCATCCGAAACATGCACGCTAACGGC  300
M28016.1   ACCCTTTTACCATCATTGGACAAGTA_GCA  299
             **  * *      *  **    * *
AJ130824.1 GCATCAATATTCTTTATTTGCTTATATATA  330
AJ314566.1 GCATCTTTCTTTTTTATTTGCATTTATATA  330
Z12030.1   GCATCCTTTTTCTTCATTTGCATTTATCTC  330
M28016.1   TCCGTACTATACTTCACAACAATCCTAATC  330
            *     * *  ** *      *     *
AJ130824.1 CATGTAGGCCGTGGAATCTATTACGGCTCA  360
AJ314566.1 CATATCGCCCGAGGACTTTATTATGGCTCT  360
Z12030.1   CACATCGGCCGAGGCTTGTACTACGGCTCC  360
M28016.1   CTAATACCAACTATC_TCCCTAATTGAAAA  359
           *   *           *     *  *
AJ130824.1 TATACTTACCTAGAAA_CCTGAAACATTGG  389
AJ314566.1 TACCTCTACAAAGAAA_CCTGAAATATTGG  389
Z12030.1   TACCTCTACAAAGAAA_CCTGAAACATTGG  389
M28016.1   CAAAATACTCAAATGGGCCTGAAACATTGG  390
            *         *     ******* *****
AJ130824.1 CATTATTCTATTA_TTCGCAGTTATGGCTA  419
AJ314566.1 GGTGGTACTTCTACTTCTCACT_ATAATAA  419
Z12030.1   AGTAATTCTCCT____C_CTTTTAA__CTA  413
M28016.1   CATTATTCTATTA_TTCGCAGTTATGGCTA  419
             *  * **  *    * *  * *     *
AJ130824.1 CAGCATTCATAGGCTATGTCCTCCCATGAG  450
AJ314566.1 CCGCCTTTGTAGGCTACGTCCTCCCATGAG  450
Z12030.1   T_G_ATA_ACAG_CT_T_T__TG___TG_G  438
M28016.1   CAGCATTCATAGGCTATGTCCTCCCATGAG  450
             *  *    ** **   *  *    ** *
AJ130824.1 GACAAATATCA  461
AJ314566.1 GACAAATATCA  461
Z12030.1   G_CTACG_TC_  458
M28016.1   GACAAATATCA  461
           * * *   **
```

Rysunek nr 2. Dopasowanie 4 sekwencji ewolucyjnie powiązanych.

Dopasowanie dla przykładów uprzednio sprawiających problemy

a)

```
>>  zadanieMSA('filename1','file1.fasta','subMatrixFile','msa.xlsx')
Complete cost : 29
A4.4       AGA___   3
A1.1       AAACGT   6
A2.2       CGT___   3
A3.3       ACA___   3

fx >>
```

Rysunek nr 3. Dopasowanie pierwszego z przykładów w ramach poprawy.

b)

```
>> zadanieMSA('filename1','file3.fasta','subMatrixFile','msa.xlsx')
Complete cost : 4
3          ACA  3
4          AGA  3
3          ACA  3
4          AGA  3
           * *
>>
```

Rysunek nr 4. Dopasowanie drugiego z przykładów w ramach poprawy.

c)

```
>> zadanieMSA('filename1','file4.fasta','subMatrixFile','msa.xlsx')
Complete cost : 24
A1.1       AAACGT  6
A2.2       ___CGT  3
A3.3       AAACGT  6
A4.4       ___CGT  3
              ***
>>
```

Rysunek nr 5. Dopasowanie trzeciego z przykładów w ramach poprawy.