

Wstęp do bioinformatyki

Laboratorium 4

Dopasowanie wielu sekwencji

Magdalena Trędak

236712

1. Schemat blokowy algorytmu dopasowania wielu sekwencji

Schematy blokowe algorytmów szukania sekwencji centralnej oraz dopasowania wielu sekwencji ze względu na duże rozmiary i umożliwienia poprawy ich czytelności zamieszczono w repozytorium jako pliki graficzne o nazwach: SchematBlokowySzukaniaSekwencjiCentralnej.png, SchematBlokowyDopasowaniaWieluSekwencji.png.

Do wygenerowania schematów użyto programu online znajdującego się na stronie:

<http://www.algorytm.org/narzedzia/edytor-schematow-blokowych.html> (data dostępu 16.05.19)

2. Analiza złożoności programu

a) czasowej

findCentralSequence – 7 przypisań poza pętlami, 11 przypisań w pętlach, każda o rozmiarze ilości możliwych dopasowań sekwencji. Dla liczby sekwencji k ilość tych dopasowań wynosi $x = k \cdot (k-1) / 2$. Wywoływane są funkcje scoringMatrix oraz traceBackMatrix, każda z nich x razy. Złożoność czasowa $x \cdot (n \cdot m) + x \cdot (n+m) \rightarrow$ co najwyżej rzędu $O(nm)$

alignAll – 2 przypisania poza pętlami, 4 wewnątrz nich. Dla x będącym ilością porównywanych sekwencji tworzona jest pętla dopasowująca kolejne sekwencje do sekwencji centralnej oraz powtarzająca poprzednie dopasowania. Ilość wywołań funkcji alignTwoSequences wynosi $n!$. Złożoność czasowa $x! \cdot (n \cdot m)$

3. Przykładowe dopasowanie dla sekwencji powiązanych ewolucyjnie

Porównano sekwencje mitochondrialnego cytochromu b poniższych organizmów:

X75581 - Antarctic minke whale

X75589.1 - sperm whale

X75584.1 - humpback whale

X75586.1 - pygmy right whale

X75583.1 - pygmy Bryde's whale

```
MSAAlignment — Notatnik
Plik Edycja Format Widok Pomoc
#Match: 0
#Mismatch: 1
#Gap: 2
#Length: 1140
#Central sequence: X75584.1
#Alignment cost: 492

X75583.1 ATGACCAACATCCGAAAAACACACCCACTAATAAAGATTGTCAACGATGCATTGTTGAT 60
X75589.1 ATGACCAACATCCGAAAAACACACCCACTAATAAAAAATCATTAAACATGCATTATCGAC 60
X75584.1 ATGACCAACATCCGAAAAACACACCCACTAATAAAAAATATCAACGACACATTCAATTGAT 60
X75586.1 ATGACCAACATCCGAAAAACACACCCACTAATAAAAAATATCAACGACGATTCAATTGAT 60
X75581.1 ATGACCAACATCCGAAAAACACACCCACTAATAAAAAATATCAACGACGATTCAATTGAT 60
*****
X75583.1 CTCCTACCCCATCAAACATTTCTCATGATGAAATTCGGCTCCCTACTCGGCTCTGC 120
X75589.1 CTCCTACCCCATCAAACATTTCTCATGATGAAATTCGGCTCCCTACTTGGACTCTGC 120
X75584.1 CTCCTACCCCATCAAACATTTCTCATGATGAAATTCGGCTCCCTACTCGGCTCTGC 120
X75586.1 CTCCTACCCCATCAAACATTTCTCATGATGAAATTCGGCTCCCTACTTGGCTCTGC 120
X75581.1 CTCCTACCCCATCAAACATTTCTCATGATGAAATTCGGCTCCCTACTCGGCTCTGC 120
*****
X75583.1 TTAATTACACAAATCCTAACAGGCCTATTCTAGCAATACACTACACACGACACACAACA 180
X75589.1 CTGATCATACAAATCCTAACAGGCCTATTCTAGCAATACACTACACACGACACACAACA 180
X75584.1 TTAATTATACAAATCCTAACAGGCCTATTCTAGCAATACACTACACACGACACACAACA 180
X75586.1 CTAATCATACAAATCCTAACAGGCCTATTCTAGCAATACATTACACGACGACACACAACA 180
X75581.1 TTAATTGTACAAATCCTAACAGGCCTATTCTAGCAATACACTACACACGACACACAACA 180
```

```
MSAAlignment — Notatnik
Plik Edycja Format Widok Pomoc
X75583.1 ACCGCTTCTCATCAGTTCACACATTTGCCGAGACGTAACCTACGGCTGAGTTATCCGA 240
X75589.1 ACCGCTTCTCATCAATCACACATTTGCCGAGACGTAACCTACGGCTGGACCATCCGA 240
X75584.1 ACCGCTTCTCATCAGTTCACACATTTGCCGAGACGTAACCTACGGCTGAGTTATCCGA 240
X75586.1 ACCGCTTCTCATCAGTTCACACATTTGCCGAGACGTAACCTACGGCTGAGTTATCCGA 240
X75581.1 ACCGCTTCTCATCAGTTCACACATTTGCCGAGACGTAACCTACGGCTGAGTTATCCGA 240
*****
X75583.1 TACCTACACGCAAAACGGAGCCTCCATATTCTTCATCTGTCTACGCTCAGATAGGACGA 300
X75589.1 TACCTACACGCAAAACGGAGCCTCCATATTTTCATTTGCCTTTACACTCAGATAGGACGA 300
X75584.1 TACCTACACGCAAAACGGAGCCTCCATATTCTTCATCTGCCTTACGCTCAGATAGGACGA 300
X75586.1 TATCTACATGCAAAACGGAGCCTCCATATTTTCATCTGCATCTACGCCACATAGGACGT 300
X75581.1 TATCTACATGCAAAACGGAGCCTCCATATTCTTTATTTGCCTTTACGCCACATAGGACGA 300
*****
X75583.1 GGCCTATACTACGGCTCCTATGCCTTTCCGAGAAACATGAAACATCGGAGTTATCCTACTA 360
X75589.1 GGCCTATACTACGGCTCCTATGCCTTTCCGAGAAACATGAAACATCGGAGTTATCCTACTA 360
X75584.1 GGCCTATACTACGGCTCCTATGCCTTTCCGAGAAACATGAAACATCGGAGTTATCCTACTA 360
X75586.1 GGCCTATACTACGGCTCCTATGCCTTTCCGAGAGACATGAAATATCGGAGTAATCTTATTA 360
X75581.1 GGCCTATACTACGGCTCCTATGCCTTTCCGAGAAACATGAAATATCGGAGTTATCCTACTG 360
*****
X75583.1 TTCACAGTTATAGCCACCGCATTTCATAGGCTACGTCCTACCTGAGGACAAATATCATTT 420
X75589.1 TTCACAGTTATAGCCACCGCATTTCATAGGCTACGTCCTACCTGAGGACAAATATCATTT 420
X75584.1 TTCACAGTTATAGCCACCGCATTTCATAGGCTACGTCCTACCTGAGGACAAATATCATTT 420
X75586.1 TTCACAGTTATAGCCACCGCATTTCATAGGCTACGTCCTACCTGAGGACAAATATCATTT 420
X75581.1 TTCACAGTTATAGCCACCGCATTTCATAGGCTACGTCCTACCTGAGGACAAATATCATTT 420
*****
```

MSAlignment — Notatnik

Plik Edycja Format Widok Pomoc

```
X75583.1 TGAGGCGCAACCGTCATACCAACCTCTTATCAGCAATCCCATACATTGGTACTACCCCTA 480
X75589.1 TGAGCGCGCAACCGTTATCACAACCTTCTATCAGCAATTCCTATATCGGCACCACCCTA 480
X75584.1 TGAGGCGCAACCGTCATACCAACCTTCTATCAGCAATCCCATACATTGGTACTACCCCTA 480
X75586.1 TGAGGCGCAACCGTCATACCAACCTCCTATCAGCAATCCCATATATTGGTACCACCCTA 480
X75581.1 TGAGGCGCAACCGTCATACCAACCTCCTATCAGCAATCCCATACATTGGTACCACCCTA 480
*****
X75583.1 GTCGAATGAATCTGGGGCGGTTTCTCTGTAGATAAAGCAACACTAACACGCTTTTTTGCC 540
X75589.1 GTAGAGTGAGTTTGAGGCGGTTTCTCCGTAGATAAGGCAACACTGACACGCTTCTTCACT 540
X75584.1 GTCGAATGAATCTGGGGCGGTTTTCGTAGACAAGCAACACTAACACGTTTCTTTGCT 540
X75586.1 GTTGAATGAATCTGGGGTGGCTTCTCCGTAGACAAGCGACACTAACTCGCTTCTTTGCT 540
X75581.1 GTTGAATGAATCTGAGGTGGCTTCTCTGTAGACAAGCAACATTAAACGCTTTTTTGCC 540
*****
X75583.1 TTCCACTTTATCTCCCTTTCATTATTCTAGCACTAGCAATGGTCCACCTCATTTTCCTC 600
X75589.1 CTCCACTTCATCTCCCTTTTATCACCTAACACTAACAACTAGTACATCTCCTATTTCTC 600
X75584.1 TTCCACTTCATCTCCCTTTCATTATTCTAGCACTAGCAATGGTCCACCTCATTTTCCTC 600
X75586.1 TTCCACTTCATCTCCCTTTCATTATTCTAGCGCTAGCAGCTGTTATCTCCTTTTCCTC 600
X75581.1 TTCCACTTCATCTCCCTTTCATTATCTAGCATTAGCAATGGTCCACCTCATTTTCCTC 600
*****
X75583.1 CACGAAACAGGATCCAATAACCCACAGGTATTCATCCAACATAGACAAAATCCCATTC 660
X75589.1 CATGAAACAGGATCCAACAACCCACAGGAATTCCTCCAACATAGACAAAATCCCATTC 660
X75584.1 CACGAAACAGGATCCAACAACCCACAGGCATCCCATCCAACATAGACAAAATCCCATTC 660
X75586.1 CACGAAACAGGATCTAACAACCCACAGGCATCCCATCCAACATAGACAAAATCCCATTC 660
X75581.1 CGCGAAACAGGATCCAATAACCCACAGGTATTCATCTGATATAGACAAAATCCCATTC 660
*****
```

MSAlignment — Notatnik

Plik Edycja Format Widok Pomoc

```
X75583.1 CACCCCTATTACACAACCTAAAGACATTCTAGGCGCCCTACTACTAATCCTAACCTACTA 720
X75589.1 CACCCCTACCACACAATCAAAGACACCATAGGTGCCCTACTACTAATCCTATCCCTACTT 720
X75584.1 CACCCCTACTACACAATCAAAGACACTCTAGGCGCCCTATTACTAATCCTAACCTACTA 720
X75586.1 CACCCCTACTACACAATTAAGACATCTGGGCGTCCCTACTACTAATCCTGACCTACTA 720
X75581.1 CACCCCTATTACACAATCAAAGACATTCTAGGCGCCCTACTACTAATCCTAACCTACTA 720
*****
X75583.1 ATGCTAACCCCTATTGCTACCCGACCTACTTGGAGACCCAGACAACACTACACTCCAGCAAAT 780
X75589.1 AACTAACCCCTGTTGCGACCCGACCTGCTAGGAGATCCCGACAACACTACACCCAGCAAAT 780
X75584.1 ATGTTAACCCCTATTGCGACCTGACCTGCTTGGAGACCCAGATAACTACACCCAGCAAAC 780
X75586.1 ATATTAACCTTATTACACCTGACCTGCTTGGAGACCCGACAACACTACACCCAGCAAAT 780
X75581.1 AACTAACCCCTATTGCGACCCGACCTGCTCGGAGACCCGACAACACTACACCCAGCAAAC 780
*****
X75583.1 CCACTCAGTACCCCAACACACATTAACCCAGAATGATATTTCTATTTCGATACGCAATC 840
X75589.1 CCACTAAATACCCCAACACACATCAAACCAAGATGGTATTTCTATTTCGCGTACGCCATC 840
X75584.1 CCACTCAGTACCCCAAGCACACATTAACCCAGAGTGATATTTCTATTTCGATACGCAATC 840
X75586.1 CCCCTCAGCACCCCAAGCACACATCAAGCCAGAATGATATTTCTATTTCGATATGCAATC 840
X75581.1 CCACTCAGTACCCCAAGCACACATTAACCCAGAATGATATTTCTATTTCGATACGCAATC 840
*****
X75583.1 CTACGATCAATTCCTAATAAATAGGCGGAGTCTTAGCCCTACTACTCTCAATCCTAATC 900
X75589.1 CTACGATCTGTCCTCAATAAATAGGAGGCGTCTTAGCCCTACTACTCTCAATCCTAATC 900
X75584.1 CTACGATCAATTCCTAATAAATAGGCGGAGTCTTAGCCCTGCTACTCTCAATCCTAATC 900
X75586.1 CTACGATCAATTCCTAATAAATAGGCGGAGTCTTAGCCCTATTATTTCCTAATCCTAATC 900
X75581.1 CTACGATCAATTCCTAATAAATAGGCGGAGTCTTAGCCCTATTACTCTCAATCCTAATC 900
*****
```


Command Window

```
Z96064.1 CTAGTAATACAAATCCTAACAGGCTTATTCTAGCATTACATTACACCTCTGACACCACC 180
Z96066.1 CTAGTAGCACAAATTATCACAGGACTATTTCTAGCAATACACTACACCTCTGATACCACA 180
Z96057.1 CTAATAATCCAAATCCTTACAGGCCTATTCTTAGCAATACACTACACCTCCGATACTACA 180
Z96062.1 CTAGTAATACAAATCCTCACAGGCCTATTCTTAGCAATACACTATACTCAGATACTATA 180
      *** **      ***** * ***** ***** ***** ** ***** ** ** *
```

```
Z96064.1 ACAGCGTTCTCATCCGTAACCCACATCTGCCGAGACGTAAACTACGGATGACTAATTCGT 240
Z96066.1 ACAGCATTTTCATCTGTAACCTCATATCTGCCGAGATGTAAACTACGGTTGATTGATTCGA 240
Z96057.1 ACAGCATTCACATCCGTAACCCACATCTGCCGAGACGTAAACTACGGCTGATTAATTCGA 240
Z96062.1 ACAGCATTTCTCATCCGTAACCCACATCTGCCGAGACGTAAACTACGGATGACTAATTCGA 240
      ***** **      ***** ***** ** ***** ***** ***** ***** ***** ** * *****
```

```
Z96064.1 TACCTCCATGCAAACGGGGCCTCCATATTCTTCATATGCCTATTCATACACGTAGGACGA 300
Z96066.1 TATCTTCATGCAAACGGAGCCTCAATATTCTTCATGTGTCTATTTATACACGTAGGACGA 300
Z96057.1 TATCTCCATGCAAACGGAGCCTCCATATTCTTCGTATGCTTGTTTATACACGTAGGACGA 300
Z96062.1 TATCTCCATGCAAACGGAGCCTCCATATTCTTCGTATGTTTGTTTCATACACGTAGGACGA 300
      ** ** ***** ***** ***** ***** * ** * ** ***** *****
```

Command Window

```
Z96064.1 GGAATCTACTACGGGTCCTATACCTTCATAGAAACCTGAAACATTGGCATCATTCTACTG 360
Z96066.1 GGAATCTACTACGGATCCTATACCTTTACAGAAACCTGAAATATTGGCATCATTCTACTG 360
Z96057.1 GGCATCTACTATGGATCTTACACCTTTACAGAAACATGAAACCTTGGTGTTATTCTACTA 360
Z96062.1 GGCATTTATTATGGATCCTACACCTTTACAGAAACATGAAACCTCGGCGTTATTCTATTG 360
      ** ** ** ** ** ** ** ** ** ** ** ** ** ***** * ***** ***** * ** * ***** *
```

```
Z96064.1 TTCGCCGTAATAGCAACTGCATTTATAGGATACGTTCTCCCATGGGGACAAATATCCTTC 420
Z96066.1 TTCGCCGTAATAGCAACTGCATTTATAGGATATGTGCTTCCCTGAGGACAAATATCCTTC 420
Z96057.1 TTTGCCGTAATAGCAACTGCATTTATAGGATATGTCTTCCATGAGGACAAATATCCTTC 420
Z96062.1 TTCGCCGTAATAGCAACTGCATTTATAGGATATGTCTTCCCATGGGGACAAATATCCTTC 420
      ** ***** ***** ***** ** ** ** ** ***** *****
```

```
Z96064.1 TGAGGGGCCACAGTCATTACAAATCTACTCTCAGCTATCCCTTACATAGGAACCTAATCTA 480
Z96066.1 TGAGGAGCCACAGTTATTACTAACCTTCTCTCAGCAGTTCCTTACATCGGTACTAATTTA 480
Z96057.1 TGAGGAGCCACAGTCATTACAAACCTCCTTTTCAGCTATTCCTTACATCGGTACTAATCTA 480
Z96062.1 TGAGGAGCTACAGTCATTACAAACCTACTATCAGCTATTCCTTATATCGGCACTAACCTA 480
      ***** ** ***** ***** ** ** ** ***** * ** ** ** ***** **
```

Command Window

```
Z96064.1 GTAGAATGGATCTGAGGTGGCTTTTCAGTAGACAAAGCCACTTTAACACGTTTCTTCGCC 540
Z96066.1 GTAGAATGAATCTGAGGGGGTTCTCAGTAGACAAAGCCACTCTTACACGCTTCTTCGCA 540
Z96057.1 GTAGAATGAATCTGAGGTGGGTTTTTCAGTAGACAAAGCCACCCTCACACGCTTCTTCGCA 540
Z96062.1 CTACAATGAATTTGAGGTGGGTTCTCAGTAGACAAAGCCACCCTTACACGCTTCTTCGCA 540
      ** **** ** ***** ** ** ***** ***** * ***** *****
```

```
Z96064.1 TTTCAC TTCATNCTACCN TTTATCATTGCAGCTCTAGCTATGGTCCACCTACTATTTCTT 600
Z96066.1 TTCCATTTTATCCTCCCATTCATCATTGCAGCCCTAGCTATAGTCCACCTACTTTTTCTC 600
Z96057.1 TTCCAC TTCATCCTTCCATTTATCATTGCAGCCCTAGCCATAGTACACCTACTATTCCCTC 600
Z96062.1 TTCCAC TTTATCCTCCCTTTATCATTGCAGCCCTAGCCATAGTACACCTTCTATTCCCTT 600
      ** ** ** ** ** ** ** ** ** ** ** ** ***** ***** ** ** ***** ** ** *
```

```
Z96064.1 CACGAAACAGGATCTAATAANCCCTCCGGAATAAACTCCGATGCAGACAAAATCCCCTTT 660
Z96066.1 CACGAAACTGGATCCAATAACCCACCGGAATCAACTCTGATGCAGACAAAATTCATTC 660
Z96057.1 CACGAAACAGGCTCTAATAACCCACAGGAATGAACTCTGACGCAGACAAAATCCCCTTC 660
Z96062.1 CACGAAACTGGCTCTAATAACCCACAGGAGTGAAGTCCGACGCAGACAAAATCCCATTC 660
      ***** ** ** ***** ** * ** * ***** ** ***** ***** ** *
```

```
Z96064.1 CACCCCTATTACACAATTAAAGATCTCCTAGGCATATTCCTAATATTAGCAGCCCTACTC 720
Z96066.1 CACCCGTACTACACTATAAAAGACTTCCTAGGGGCCCTACTCATATTAGTAGTTCTACTC 720
Z96057.1 CACCCCTACTATACAATCAAAGACCTCTTAGGCGCACTCCTCATACTACTAGTTCTCCTA 720
Z96062.1 CACCCCTACTACACAATTAAAGACCTTTTAGGTGTATTCTAATACTATTAATTCTCCTC 720
      ***** ** ** ** ** ***** * ***** * ** ** ** * ** **
```

```
Z96064.1 CTATTAGTCCTCTTCTCCCCAGACCTGTTAGGAGACCCCGACAACCTACACACCAGCCAAC 780
Z96066.1 CTACTTGTTTTATTTTCCCCAGATATTTTAGGCGACCCAGATAATTATACACCAGCTAAC 780
Z96057.1 CTATTAGTACTATTTTCCCCAGACTTATTAGGAGACCCCGACAACCTACACACCAGCCAAC 780
Z96062.1 CTATTAGTATTATTCTCCCCAGACCTATTAGGAGACCCAGACAACCTATACACCGGCTAAC 780
      *** * ** * ** * ***** * ***** ***** ** ** ** ***** ** ***
```

```
Z96064.1 CCCCTAAACACTCCCCCCCACATCAAACCAGAATGATATTTCTTATTTGCATACGCCATC 840
Z96066.1 CCCCTAAACACACCTCCTCACATCAAACCAGAGTGATACTTCTTATTCGCCTACGCTATC 840
Z96057.1 CCTCTAAATACCCCTCGCCACATTAACCAGAATGATACTTTCTATTTCGCATACGCTATC 840
Z96062.1 CCCCTAAACACCCCTCCCCACATCAAACCAGAGTGATACTTCTTATTCGCATACGCTATC 840
      ** ***** ** ** * ***** ***** ***** ** ***** ** ***** ***
```


Z96064.1 CTACGCTCTATTCCCTAACAAACTCGGAGGAGTACTAGCCCTAGTATTATCCATCTTAATC 900
Z96066.1 CTACGTTCTATCCCCAATAAACTCGGAGGAGTGTTAGCTTTAGTACTATCCATCCTTGTC 900
Z96057.1 CTACGCTCTATCCCCAATAAACTTGGAGGAGTACTAGCCCTAGTACTATCTATCCTAATC 900
Z96062.1 CTACGCTCTATCCCCAATAAACTTGGAGGCGTACTAGCCCTAGTATTATCCATCCTAATC 900
***** **

Z96064.1 CTGATCTTCTTACCCTTCATCCACACCTCCAAACAACGAAGCCTAATATTCCGCCCAATC 960
Z96066.1 CTAGCCTTACTACCTCTCATGCACACCTCAAACAACGAAGCCTTATGTTTCGCCCAATT 960
Z96057.1 CTAATTCTCATACCCTTAATCCACACCTCAAACAACGAAGCTTAATATTCCGCCCAATT 960
Z96062.1 CTGATTCTCATACCCCTCATCCACACATCAAACAACGAAGCTTAATGTTCCGCCCAATT 960
** * ***** **

Z96064.1 AGCCAAACCATATTTTGAATCTTAGCAGCAAACCTTCTTATCCTAACCTGAATCGGAGGA 1020
Z96066.1 AGCCAAATCCTATTTTGAATACTAACAGCTAATCTTCTAATTCTAACTTGAATTGGAGGA 1020
Z96057.1 AGCCAAACCATATTTCTGAATCCTAACAGCTAACCTACTCATCCTAACCTGAATCGGAGGA 1020
Z96062.1 AGTCAGGCCATATTTCTGAATCCTAACAGCTAACCTACTTATCCTAACCTGAATCGGAGGG 1020
** * ***** **

Z96064.1 CAACCTGTAGAACACCCCTTCATTATTATCGGCCAACTAGCCTCAATCAGCTATTTCTCC 1080
Z96066.1 CAACCAGTAGAACACCCCTTCATTATCATTGGACAAGTAGCCTCAATCAGCTACTTCTCT 1080
Z96057.1 CAACCCGTAGAACATCCCTTTATCATTATTGGCCAACCTAGCTTCAATCAGCTATTTCTCC 1080
Z96062.1 CAACCCGTAGAACACCCCTTTATCATCATCGGCCAACTAGCTTCAATCAGCTATTTCTCC 1080

Z96064.1 ATTATNTAATTCTAATACCAATCTCAGGCCTAATCTAAAACAAAATANTAAAAATAAAA 1140
Z96066.1 ATTATTTAATCCTAATACCAATCTCAGGACTAATCGAAAACAAAATACTAAACT---T 1137
Z96057.1 ATCATTTAATCCTAATACCAATCTCAGGACTAATTGAAAATAAAATACTAAATTAAC 1140
Z96062.1 ATTATTTAATTTAATACCAATCTCAGGACTAATTGAAAACAAAATACTAAATTAAC 1140
** * *****

Z96064.1 T 1141
Z96066.1 T 1138
Z96057.1 T 1141
Z96062.1 T 1141
*