Ćwiczenie nr 2 z MBI, adnotacja DNA

Jakub Skałecki, Kinga Kimnes

4 grudnia 2019

1 Przygotowanie danych

W ramach doświadczenia wykorzystano, uzyskany wskutek assemblingu, genom tasiemca (hymenolepis diminuta), zawierający całościowo 3035 kontigów (scaffoldów). W przeprowadzonym badaniu wykorzystany został kontig fizyczny o identyfikatorze:

```
> HDID_scaffold0000073 length=149394,
```

spełniający warunek odznaczania się długością powyżej 100kbp (149394kbp).

2 Maskowanie genomu

W dalszej kolejności przeprowadzono maskowanie genomu, polegające na wygenerowaniu zmodyfikowanej sekwencji DNA, zawierającej zamaskowane powatrzające się sekwencje. W tym celu wykorzystano program RepeatMasker, sprawdzający fragmenty DNA pod kątem powtórzeń, jak i sekwencji o niskiej złożoności. Dane wyjściowe stanowią: szczegółowa adnotacja powtórzeń obecnych w sekwencji zapytań wraz ze zemodyfikowaną wersją sekwencji, w której zamaskowano wszelkie powtórzenia (wstawiając w miejscach tych literę N). Działanie programu oparte jest o algorytm Smitha-Watermana-Gotoh'a Liczba uzyskanych, zamaskowanych nukleotydów wyniosła w naszej próbie 2296.

3 Mapowanie znanych sekwencji i adnotacja strukturalna

W kolejnym etapie dokonano adnotacji strukturalnej, dostarczającej danych usprawniających identyfikację struktur genowych. Dzięki zastosowaniu programu Make, w efekcie przyrównania białek i mRNA do genomu (ich zmapowania), uzyskano adnotacje genów w formacie GFF3. Poniżej zamieszczono 10 pierwszych linii, wygenerowanego przez program Maker, pliku gff:

```
##gff-version 3
HDID_scaffold0000073 .      contig 1      149394 .      .      .      ID=HDID_scaffold0000073
    ;Name=HDID_scaffold0000073
###
HDID_scaffold0000073 repeatmasker  match  32910  32976  263    +      .      ID=
    HDID_scaffold0000073:hit:0:1.3.0.0;Name=species:NONAUT-5|genus:LTR%2FGypsy;Target=species:
    NONAUT-5|genus:LTR%2FGypsy 1346 1412 +
HDID_scaffold0000073 repeatmasker  match_part    32910  32976  263    +      .      ID=
    HDID_scaffold0000073:hsp:0:1.3.0.0;Parent=HDID_scaffold0000073:hit:0:1.3.0.0;Target=species
    :NONAUT-5|genus:LTR%252FGypsy 1346 1412 +
HDID_scaffold0000073 repeatmasker  match  33324  33816  718    +      .      ID=
    HDID_scaffold0000073:hit:1:1.3.0.0;Name=species:NONAUT-3|genus:LTR%2FGypsy;Target=species:
    NONAUT-3|genus:LTR%2FGypsy 1606 2102 +
HDID_scaffold0000073 repeatmasker  match_part    33324  33816  718    +      .      ID=
    HDID_scaffold0000073:hsp:1:1.3.0.0;Parent=HDID_scaffold0000073:hit:1:1.3.0.0;Target=species
    :NONAUT-3|genus:LTR%252FGypsy 1606 2102 +
HDID_scaffold0000073 repeatmasker  match  33918  34592  492    +      .      ID=
    HDID_scaffold0000073:hit:2:1.3.0.0;Name=species:Gypsy2_MH-I|genus:LTR%2FGypsy;Target=
    species:Gypsy2_MH-I|genus:LTR%2FGypsy 2392 3335 +
HDID_scaffold0000073 repeatmasker  match_part    33918  34592  492    +      .      ID=
    HDID_scaffold0000073:hsp:2:1.3.0.0;Parent=HDID_scaffold0000073:hit:2:1.3.0.0;Target=species
    :Gypsy2_MH-I|genus:LTR%252FGypsy 2392 3335 +
```

```
HDID_scaffold0000073 repeatmasker match 34932 35128 262 + . ID=
    HDID_scaffold0000073:hit:3:1.3.0.0;Name=species:GYPSY1-I_CB|genus:LTR%2FGypsy;Target=
    species:GYPSY1-I_CB|genus:LTR%2FGypsy 3101 3300 +
```

Przykładowa adnotacja typu `expressed_sequence_match` w pliku .gff (wygenerowanym przez program Maker):

```
HDID_scaffold0000073 blastn expressed_sequence_match 1868 12755 887 + . ID
    =HDID_scaffold0000073:hit:11:3.2.0.0;Name=HDID_0000755601-mRNA-1
```

Jej źródłemi jest program blastn, służący do porównywania sekwencji z bazą danych sekwencji nukleotydowych. W naszym przypadku bazą danych jest plik z sekwencją mRNA dla tasiemca (sekwencja dopasowana została do scaffoldu o identyfikatorze `HDID_0000755601-mRNA-1`).

Nukleotydy tej sekwencji:

```
ATGGCTCAAGAAGATGATGATCGTAATCTTCTTATGATCTTAATTGATCTTACCCCAGTG
TGGTGGGGTACTTATGCTCACGAGTTTCTTATTCTTCCCACATTCATCGAAAATATTCTC
GCTTTCGCCAATAGTCATTTGGCTCTATCACCCCTTAATGAGGTAGCAATTGTTGGCGTA
ACTCCCGAGAAAACCGAATTTTTATGGCCTTCTCTAAACCCCATTGATGAAATTGAATGC
CACAATGGTCAGTATGAACCGTTCTCGATTATCGGTCAAACAGTACGAAAGAAAGTCAAC
CAGATGATTATGTCCTGTGAGTCGACGAGATGTACGGTCGCTTTCGCTAATGCGATTAAT
AACGCTCTTTGTTACTTCATTCGTCGCTGTCGAGAAATGCGTCCTACTTTTGCTTACACT
CGAATCGATTCAAATACATTGATGGAGGATGACATTCATAGTCTGCTGAAGGACAATTTC
CATGCTCGGATTCTTGTCGTTCGAGCGGCTGAGGATGATTCTTCTCAATACCTTTCTCTC
ATGAATGCTGTCTTTACCGCTCAAAAGATGGGAGTGCTGATCGACGCTTGTATTATACCC
CCTACTCGAATGTCTTACTCCACCGAAGATCATCTTCGTCAATCATCCACAACTCTCCAG
GAGTCTGGACATTCTAGTCTCTTCCAACAAGCTGCGGAACTGACTGGAGGTATTTATCTT
CGTATTCCGAGACCTGCAGGACTTTTACAGTACCTCCTTTGTGTTTTTCTCCCGCATGCA
GGTCTCAGATCTCAAATGATCCTTCCAGATTCTAATGGGGGCTCATCTGCAGGGGTGGAT
TTTCGTTCGGCCTGTTTTTGTCACCACAAGATGGTCGATTTAGCTTATGTTTGCTCTGTC
TGCCTGTCTGTATTCTGTGAATTCTCTCCTATCTGTTCAACTTGTCAAACTCCTTTTCGC
ATTCCCACTTCTGTAACTCTACCTCCTGAGAAGCTCAAAAATGTAAAATAAATATCTTCT
GTTTCTCAATTGCTCTTCGGCTTTTTTGATATTCCGTTTGGTTACTTTTGTTCTAATCTT
ACCTTGGTTGTTTCAGCGCGAAGATGGATTCAGATGATATCTTGCCATCAAAACGCATTA
AGGTATTAAAATAGTTCTTTTCTTTAAGATGAAGAACGTGATTTAAGTCTTGTTGAATAA
TTAGTTTATTTCTTCACTCCTTCCACATTTGTAGGTTGATGATACCCAAATCGGTTCCAA
GGGTTGGGACTTGATGGTGAGTTATCGCTTAATAAGTGGATTTGGAACTTTGTTCCTAAT
TATTTTGCTCTGTTGTGGACGGGAGCTGTCGATTTTCTCCTCGTCTGGGGATTTTGATTT
TGATTAATGCTGAGCCCTAAGCACAGATTAGTTAGTTAGTCTGTTAGTCTGTTAATGTAA
GGTAAAATCATCTAAGCTCAACTAATATGTCAAAAACTGAGGAGATGTGGTATCCAAGGC
TGCAAATTCAAATTTTGCTTTCAAGTTCATTGGTGAAAATCGTCTTCCAAGATGTTTATA
TTCGAGGATAAAAGCACTTCCATTTTTTCTAACTAGTTCATTTGCGCTAAACGAAAATCA
TAGCTCAATGCCTATGATTTCTCTGCTACTATGGGAGTAAGACGGGGGCCACCACTCAAG
ATTTCTTTTGCTAAAAATGAATTAGAGCTGGATTGTTAAATATGCTGCAATATACAATTA
GCGCTTCCCATTGCTAAATGAAATGTCCACTGTATTAAACTGCTAAAGTCTGGGATATCT
AGAATTATCATCCCGTAGCTGCTAAATTGCTTTTCCCAGTGCATTTTAATTTAAAAGACG
AGAAAGTAACTGTAACGTCTATTAGATACACTGCGTAAGGTGGTACAAAGCGTACTTCTT
GCCCCAAATATCGAGTAGGTGTTAATTTTCTTTCATAGAATATAGCTTTCGGTGGAATTT
TTTTCTCAAGGTCTCTATTGGCCGACTGATTAACTGATATCGTTGGCCATTTGATTTCGG
AATATTATCTTCTAATGGCCACGACAGCAGGATAATATTAGCGTACATGATTTTAACATA
ACACGGTTCAATAAGGTTATTTCTATGTGAGCATTAACCGTTAATATTTTGATTGAATAT
ACTTGTTAGGTTTTTATTATGACTCGATGGGGTTGACATGTATAATCGTAGGCTCTTTTT
GAATTGAATAGTTTTGTGGGAGCGTTTTTATTCGCGACTTGCAGTTGAATAGTTTACATC
GAGGAGTGTTTTACGAAAAATGTGTTTATTATTCTTTCGTAATAATCCCCGTTATACTCAG
CGATAATGTAAATTTAAGTTGTCACCTTTTTTTTCAACACAGCATTTTTAATGTGTAATGA
GTACGAGGTCTGTGGAAGCTTCACGTTCGACTATTTAGCTAGATGTCTTCTGGTTAGAAT
TTATTAATATCTACGTTATCCCCGCAGGATAGTACTGATGGGCTTATGATCCCGATAGTT
AACTCGTTAATTGCCTTCCATTAACAGAGAAAATTAGGTTGGACCGACGGCAAGGGTCTA
GGTGCTCAAGGTCAGGGTCGTCTCGAACCCATATCAGCCAAGGTGCCGAAGAATCGTCAA
GGTTTGGGTGCCGATGCCGCCTCAGGATTTCGCGCTCTCGACATAGGCGAGCTGCCAAAT
```

GGTCCGAATGCTTGGTCCGACGACAGGACCGATCTGGTCCCCGCCTCAGCTAATGATGAC
CTCTTCTATCGATTCTCCCTTTGGTTCCCCCCGGGATCAAACTTGGAGATTCCAGATTCG
CCTGAGAGGTGTTTGGGAGCACTGAGATCGGCGATTTTGCCGTTGGATTCAGGGTCTGCT
CATGGTCCTCCAGTTGACTCGATGGAGAGTCAGTCTCGATATTGCTCGGAGGAATTACTC
TTTGAGATGCTGTTTTATAAGGTAAGTCCTGGAAACCGTATGGAAAATCGCTCCGAAATC
TAAATTATGTGTTACTCAATTCTACAACGAGGTGAAGATGAGTTGATTAAGAAAATTCCA
TATTGTTTTTTTGTTCTGTTCTCTTCATTCTCCCCAGTGGTTTGGTTTGGAGTTGAAGGT
GTTTAATCATCGCATTTCTCTCCCGAATTTGTTTACATCTGAAGATTTTCACTTAACCTT
TTGTTAGGTTCAATTTATGTATGTTTTTTTCTTCCTTGCTAAAGAATTCTTCTCCATTTC
GTCTATGAATTTCTAACAGTTTGGAAGTGGGCTTAAGTAATACGCCGTAGATATGAGTCC
CTATGGTAGCTTTTCAATAGATTATTTGATGTTGCTCACATAAATACGATTTTGCTGCCC
GCTTTCATCATCTGTCAAATATACAGAAAAATGGACATCACTAATAGTCCCTAAATGAAA
TATGGTGCAAAATTTTATTTGAAAAGCACGAATCTAGTTTCGCAGGTCAAATTCTAAGCC
TCTTACATGAGATATTTCCCTCGTCTAATTAGTGACCTAGCGCTTCCTTGCCTTCTTTAT
TGATTCGCCTTAGCAACAATTTTGTGTGTTTTCTAGTGGCCTGAAAGAAGTTTCCCTAGA
CTTTAAATCAGTCCCAAAGTAAAGTTCTCTCTGTTGAAACTTGGTCACTGGATTTCCCAG
AGCATTCGCCTGGGCAAGCTTTTCGGCCCAGTGAAAATACAGTTCTAAGATTAGTCTTAT
CGCAACATTCAAAACAGTCTTCGTCTTGGATACAGGAAGCATGAAAAAAGAAATTGGCTA
CGCAACCATTTCTTTAAATTTATTTTTGCGCTGCCGGATGTGTAGTTTCATTTTATTATT
CCAACTACAAGGTTGCTTAGCTACATGAAGACATAAGACTTCCAGAACTAAATCAAATTT
GGAAAGAATGCCAAAATGTAGCCTCGCATATCTTTCCTGAATGCTATGCATCTATTGACC
AATTGGTATGTCTTATTTAATATTTAAGTAGATGGACTTTTTATTATCACAATATTTTTG
TTGAAAGACCATAATGAAAGACTGGCTCCCTAAAGATCTAAATTAACCCCGCCTTATCTT
GAATGGCAATTTTGAACTCTAAAATGAATTGGAACTTCGCTGGATGCTACTTGGCTATAC
AAGTGTGGTTCGAAAATGTTTGTCTACACATGAGAAAAAGTAGTATGCTATTAGAAACTC
AGTGAATTGGAAATCCTTCCCCATCTATCCTTCTGCATCATCGAGGCACCAGTATTATTA
TTTCCTTAATATCGTTTTTTATTCCTTTTAGAACTCCTTGGATAAGCGCTCTCGTGAAGC
CATTATGAGTGCCCGTTTGCGGTCGAATCCTTACGAAGATATTCGCAGTGGAATATTCAT
GAATAGGTTAGTAGTTTTTATCCGGCTAATTTAATATCGCATAGAGAGGAGAGGGAAATT
TGATGATTCCGGTCATTAGTGGTAAAGACTTCTCTTTTTCCCTTAAGACCTTCCAATTAA
ATGTTTGGGCCTATGTGTCCGCTTGTCAAACTAATCTAACATATCCTAAAAATCCTAAAC
CGGTTTTAAACCAGCCTTTCCCAAGGCTTATCTTAGAAGCTTTTTGATGTTCCCACTCCT
CTCTTTGTCAACTTTTTTTCTAATATAATTGGTCATATTGTTCACTTAAGCCGCCTTTAT
AAGCATCGAGGCCATGATCTCTGACTGTAGACGTGAATAGTTGCCGTTGAGCCGCCGACG
CATACGCAAATTTATAAATAAACGAAATTTTCAGTGTATAATCTCCATTTACGTTCTCAT
TTACTTATTAGAGCTGCGATGAAAATGGCCAATATGGACTCCTTATTCATGGGATTGTTC
TCCGGGAAACCTCCCCACAATGTAATTACTTATACACCTTTAAGTAGTATTGTCTTACAT
ATAATTTCGTCATTCTTCCCCATCCCCTTGTTCCTTTTAGGAAATTCTCCACTTTGCAGA
CATCTGTGCTGGTCCAGGAGGTTTTTCTGAGTACTTGACATGGCGTCGTGGATGTCCCCT
AGGTTCAAAATCGCATCATTTTCAATCTATTCGAGGCTATGGCATGACGCTCAAGGGGGA
GTGTGACTTCAAGTTAAGGAAATTCATTGCTGGACCGATGGAGAATTTTCGAGCTTACTA
TGGCACCGCAGATGATGGTGATATTACGAAGTGGTGCAACTTGGCCTCATTTGCGAAGAC
TATCATTTCTGCGACATATGGAAAGGGAGTCCACCTTGTTATGGCTGATGGCGTGAGTGA
CAAATAGATTTCTTTATCGTGTTTTAATAATCGTAATGTAAGGCTAGATCATCACTGAGA
AAATACTGCAGTCTGTTTGGCTCGGACCATAGAAAACTCCCAGAATTTTAATTACTGAAC
ATTTGAATCAGAAGGCTGTATCAAAGATACGATTTTGACAGGTTTTTGGCTTTCAGATCA
GGCGATTTCTATCATAAAACCAATTATTTATTATAGAATTGAAGTTACAATTGGCACTGT
AATTAACGAAACATTACTTAGGAGTGTTTCGTTCAATTTAGGCTGAGCTTTAAAAGGGCA
TCCGATTGATAGATAACGATTTTCAAGGGTGAAAGTTTATATTTGTCTACATTTTCGTTA
TCTCGTACTAATCTTTAGAGATTGGGTGGTGGCGTCACCCCACTCTTTGGAGTTAGGGGG
CTGTTAATCGTCTTAACCTTCGATTTACAATGCATTGATCCGTCGATTATGTTGAATGAT
TTATAATACATTTTAAGATACGTTTTGTCTTCCAACAGCCAAAATGCTTGGCACCATATT
GCGAACTAATGTAGCAGCTCACCTGTGTTCACTCAAGGTGAACCTAACAATTTTCTAATT
CAATCACTAAAATCACTTGGATGAATAATGGACGAAGACCGAATCATCGTCGCATCGATG
ACAACTTATTCTTTGGCTTTATTAGGTTGTTTGGCTCAGAATAGGGCAGCACATGCTGCT
TGTAAATAATAAGTGACATACGGACATGGGCATAACCTAAGAGATAGTTATTGCCCTCTC
CTGAGATTTTCACTCACGCCTGTCCAAAGAATAATCTTGATTGATTAAACCTCTCACCTT
ATTTAACGGTCCGTGGTTAGGATGTATGTTGCATGATCCGAATAGATCTTTATTAAGGCG

CATCATAGATAAAGTAAGTAAATATGAAGGTGCATAAAGTCTATAACATATATGAAAAAT
GGAAATAGGGACCAAGAAACATAAAAATTATAAATGGTCAGAGCAGGTGGCCATTCCCCA
TTATATCCGTCTAGCTTTACCACGAGCAATGGCAACCGATAACTATGCAGCCTGATCAAG
CCAATGGCCACGCTGAGCTCTCAGTTCCGCTTCCCCATTGGTTCACACTGCGTATAGGCA
TGTGCTTCGGTTTGATCCGGCAAGGTTACGGGCATACTTTTAACAGGCCTGGGGCATATG
AGGGAGAGCTTTATTGTCAGCAACACTGGTTGACTTTTGTAGAGATTTGGGCTGATATCG
AGATTCAAGTATATACCAAATAATGACATACTCTTAATGGATCTTGATTTCCGTTGATCT
CAATATTACACTTTTGTTCTCTTAAGGGTTTCGATGTGTCTGATGGCTACAATCTGCAAG
AAGTGAAATCGAAGCACATCTACCTTTGTCAATGCCTCTGTGCCCTTACTATTCTACGCC
CAGGTAAATCACTTTCCCATAAATGTTCAACAAACACTTTTTCAGGCGGTCGCTTCGTTA
CAAAGTTGTTCGACACCTTCACCGATTTCACTGTTGATTTACTTTGGCTTATGAGTCATG
TTTTCAGGAAGATTTACATAGTCAAGCCCATCACCAGTAGACCAGCAAATTCTGAACGGT
AAGCGTCTAGCTCCGAAGTCGATTTTCTAGATATCTTGTCTGTGATGGTTTGATTTCTCC
CAACGATTGCATGGCTGGCTGTCCACCGGCTCCTGAATCCTCTAATAAAATTAAAAAGAA
AAATACGGAGGAATTTACTCGTCGACAGTTTAAGCAGAAGAAAATCCAGAAACCCGTAGT
AATGCCTTCTCAAGCCTCCAAAGATAGTCATGAGGATCAATTTCACATAGATACAACTTC
GGCGGTTGGAGTTCTAATTGGACATTTTCTGGCAGCTGGAGAAGAACTACATAAATTGGA
TGGAAACGATTCTGTGGACCTTCTTCGATTAGCTCGAGATGAGATTCTACATGGCGAAAA
CTCCCAGATCTCGGAAAGCTTCCCGGAGTTCATTACAAGGGTTAATGAGTTGCAAGTGGT
TTTGCTTTTTGTTTTTAGCTACTGATTTCATTTTAAGTAGTTGTGTTGTGTTGTTTAATT
ATGTTATGTATTTTTTTCTCTTGCTTTAGATTGATAAAACGCCAAAGTTTGTACCTCTCA
AAGATGATTGTTTTCGCTGACGATGCAACTAAGAGTGACGATTCACAAGGTGACATTGCA
AAAGCGTGTTTGGAGAAGTGGCAGGTAAAATGTCATTGAATAGGTTTTATTCTTTTTAAT
TTTAATAATAACTCCCCATCATAAACAGATCCATAAAAGTGGATCTGAGGTTTTATTTGA
AAATTAAATTGCACATAAATCTGCAGTTTCCTTTATATTATCTATAATCCTGGACTTTGT
ATGGTAAATCTTGTTGATCTTTGGTGTTTTTCGTTGTCATTGATACCCTCCTGAGATCGT
TTCAAGTTTTACCAATGTGTCCATGCATTTCCTTGTTGGGAGATTAGAAGCTGATGGTCG
ATAGAATCAAATTTTTGATGTTTTAGCAGAATTCAGAAGGATTCTATGGGAGGAATTTTT
CTTTTCGAGAGTTACATTTTTCGATAACTCTTGTGCTTGTTCCTAAGACTACCCTTTCCT
TAAAGGTAGATTACTTGCTAACTTGGTAATCAGTCGATCCCCTTTACCGAACTCTTATAT
AAATAACCATACATTTGTAATAGAAAAAAAACAGCTGTCTGAGCTATAAGTCTTCTCCTT
ACTTCTCGGAGTATCAGCCCCCATAGCCTCTAGCCCATACCGTTGAAGTTTGTATGATCC
GATGTAAAACTCTTACTCCTGCTTTTTTGATCCCCCAATTTGTGAATTCGAAACAGTTGT
TAGAAGTTTTTACTTAACATTAATCCCATCAAAGATTTCATTTGGCACTTTGGCAAATTT
ATTTTGACCCATACTTTGTGTGCAATTTCTTTGTTTATCATCTTAAAATAAATCTATCCA
TGTTTCCTCAAGATAAGCTTCGAGAAAATCATCTATAATGGCTATACTCCTAACTGTAAA
AACTGCTTTAAGAGCCAGACAAAATTAGGTCTAGAAATCCCTTTGCAAAACGAAGGATCT
TCCTAAAGGTCATCAGAATACTACGGTCTTTCTCGAATGGGAAACCTAAGAGTTGGGTTG
AGACCAAGCTTATTTTAGACAACAGGATGAAGGTATAATTGTGAGTTGGGAAGGCAGAGG
TATCTGTTCTGAGTGGGTGGAGTCCGGGTTTGGTTAAAGAGGGCTCATCGAGGTACGACA
TAGGAGTTTACACACTTCGCGTTCTGTTTTTTGATAAGCAATTGATTAAATACATCAATC
GACAATTAGGTGTAAGCCTCCTCGAGTGTATTAGAAGTTGGCTTGTAGCCTTGCTCGCTA
TTAAGAATTAGTGTTATATGCACATCTAATCGATTCTTGGTGCTTATCGATTTTAGATCC
CTCATATCAAGCGGGGGATTCAAAGCTGGCCTCTGTGGTCGGAGAATATCTCTCCTGCTC
TACGTGAAATTATCCCCGTGCGTACTCATTCTCTACTTATAATTTTATCAGTAGATCGTG
TATTTGTAAATTGTTTTTTGTCATTTCACTCATCAGGATCCCTCGAATTTCACAAAAGGAT
CTCTTCCTAAGCATCTGGAGCGTCCAAAAATAATTGAGACTTTTAAAACGGCAGATCTTG
AACGATCGTTATTCCAACCCTCTTCATTTTATGCCCTAGTTACTTCTGGTCTTCATAGTG
TAACCTCGGAACAACCCAGTGATCCAATGATCCTGCTATCTCAAGGAAATGATCGGATAT
TTATATGGGAGGAGGGGAGATTCCTTCGACTTGAATCAAAAATTCAAATTCAACTTCCTG
CGGGTTGTCTACTTTGGGTTATTAATGTTAATATTTACTCTAATGTGCGTTAATATTTTT
GTAGCGCTACCTTTTTACGTAGACATGATTGATCACATTTTACTTTTTTACAGAAAGGAC
GTCGATATCGCGCGTTAATGATTCTAGATGCAGCATTTATCTACGGAATAGATATTCAAA
ATTTACCCCTTGTCGAGAGAATGGCACATATTAGAGCACTATGCAACACCTTGGATTTCC
CGGAAACTGATTGCGCTAAAGTTATCTGTCCACCGTGTAGGCCTCTGACTAGCATGCCGG
ATTTTATTAATGGGTAATTAAAGGAAGATTGCTCAGTTATTTTCCTGGTCTTCTTCCTTT
ATTCAATCTCGTTTTCTCCTGTAGTTTGAAGGAGCTTCCATGTAAAGACTATCCAGATGG
TCACATTATGTTTCACGCCACCAAATCTGGAATGACTTGTGCTCCAAAGAGTCTTCTGCT

```
CGTTCAACCTTTATCCTGTGAGTAAGGAGATTCGCTGTGTATGTTTGATAAGTATTTGAA
TATTGTGCATGCCCATCCAATCATAGATCACAAGATCGGTGCTTATTCAGTGACGCAGAT
GACTAACCGTACTGTCTAATTTTGGTCAACAAGAGTGGTGGTGAATGTAAATTTCTGAAA
GACTCCCTGCAAAAATTGATTGTTAATGAGCTTGAGAAGTTATCATAATATCTGGATCTC
TTTTTAAAATTTTAGCTCCTTGGACAATGGGAGTAAGTCGTAGCACTGGTCATGTATACT
ATTTGAATTCGAAGGATCACATATCGAGCTACGAAGTTCCAGCTGGTATTTGTTTACCAT
TCAGGTATGATTATAATTTGCTATTGTTTAACTCGATAAAAAGAGACATCTATGTAGGTT
TAATAGTGAAGTAAGCTAATTTGGGTATGAATGAGGAATCAGGTATAACTGGAAATTAAA
TATCGGATCTTCTTAGCCTAATTGCAGGGTTTGACCCCCAGATCTATGAGTGGTAGTTCA
AATTTGTGATGTTCAGAGCTGATAGGCCAACTGATCAAATTGAAAAGCCGGTCATACCAT
CTTATAACTTCCGTCTCTTCTTACAACTAGGGGATACTAGACAAAGGACTTAATAGTCTA
TTAATAACATTTATGTGGAGCGACAGTGGTTCGAAGTCTCTGACTGAGATAATAGTACCC
TAGACGCAGTAGTAGTATGTTAGCAAGTCAGATAGAAAACGTCTTCTAATTATAAAATGT
GGGATCATGGGAAACGTCAACTTACATGCCTGTAAAGTTAGAGATCGTGAGATCTATGCC
CATCAGTGCACATCCAGTGACGTGGATGGCCAACCGCCATGGTTGATTCTGTTCAACAAG
AGACGTGGTAGAATTTTTCCCGTAAGAGCAACTCCAGGATACCTCTAAGTCAAGTTTAAT
GAACACTTCTAATGGCTGTAAGAAGTTACTAAACATGGATCTTGTCGGGAATAGTCCGCA
GGTTAATCATTTCGGGAAACCACAAATAGCAATCGCTTGTTTTCGTAAAGAAAGAGATCG
CAAAATCCTATCTTGTTTTCTATAGATAAAGGTGTTTCTTTAACATTACAGCAAAACAAA
GTTCTACCAGGTTCCTTGGACCACTGCACATGACGGCGATTTAGACGTGAAGAGCCTTCT
GGTGTGGTTAAGAAAGCATCCACAGTA
```

4 Adnotacja funkcjonalna

Badany w ramach ćwiczenia genom należy do tasiemca szczurzego (`hymenolepis diminuta`) Dla wybranej sekwencji genów zastosowano algorytm BLASTX, który pozwolił na przeprowadzenie analizy podobieństwa uzyskanej w punkcie powyżej sekwencji nukleotydów do tych przechowywanych przez NCBI.

W wyniku działania algorytmu (rysunek 4), znaleziono 4 sekwencje o procencie podobieństwa powyżej 90 (100% dla dwóch sekwencji oraz ponad 94% dla 2 pozostałych – w tym także dla badanej sekwencji tasiemca). Istotność wyniku `E-value` jest na poziomie około `3*e-178`, zatem prawdopodobieństwo przypadkowego dopasowania sekwencji źródłowej do znalezionych jest znikome.

Dwie pierwsze pozycje na liście wskazują, iż badana sekwencja nukleotydów faktycznie stanowi poprawny scaffold tego organizmu (procent podobieństwa =100, E-value=0). Ponadto zaobserwować można także duże podobieństwo sekwencji do `Hymenolepis microstoma` - znany jako tasiemiec gryzoni oraz `Rodentolepis nana` - tasiemca karłowatego.

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	unnamed protein product [Hymenolepis diminuta]	702	702	9%	0.0	100.00%	VUZ57475.1
☑	unnamed protein product [Hymenolepis diminuta]	699	2438	35%	0.0	100.00%	VDL59872.1
☑	cap specific mRNA [Hymenolepis microstoma]	587	1786	29%	3e-178	94.61%	CDS26052.2
☑	unnamed protein product [Rodentolepis nana]	649	2148	35%	0.0	94.15%	VDO06395.1
☑	unnamed protein product [Hymenolepis diminuta]	384	1675	25%	1e-121	80.00%	VUZ57477.1

Rysunek 1: Wyniki algorytmu BLASTX z bazy NCBI