

Instrukcja do ćwiczenia nr 2 z MBI

Adnotacja DNA

Wiktor Kuśmirek

Robert Nowak

2025L

1 Wstęp

Nić DNA składa się z regionów kodujących i niekodujących. W regionach kodujących (eksonach) zawarta jest informacja o sekwencji aminokwasów w cząsteczce białka, regiony niekodujące (introny) rozdzielają eksony, ich rola w organizmie człowieka nie jest jeszcze do końca zbadana.

Przedmiotem niniejszego ćwiczenia jest adnotacja DNA, czyli identyfikacja regionów kodujących nici DNA i przypisanie im pełnionych w organizmie człowieka funkcji. Kolejnymi etapami ćwiczenia będą:

- zamaskowanie sekwencji repetytywnych - proces maskowania sekwencji repetytywnych w wejściowej sekwencji DNA pozwoli na uzyskanie poprawniejszych wyników adnotacji
- mapowaniu znanych sekwencji białek i mRNA na sekwencje DNA (białka i mRNA są zakodowane w eksonach, zmapowanie pozwoli na oznaczenie sekwencji kodujących w wejściowym DNA)
- oznaczeniu sekwencji kodujących i niekodujących w wejściowej sekwencji DNA
- ekstrakcji sekwencji do dalszych analiz (geny)
- przypisaniu funkcji zidentyfikowanemu genom

2 Adnotacja DNA

Instrukcja pokazuje poszczególne kroki na Ubuntu 24.04 w wersji 64 bitowej. Wymagania wstępne to instalacja aplikacji docker.

```
sudo apt-get install docker
```

2.1 Sekwencje kontigów i aplikacje

Pobierz sekwencje kontigów badanego w ćwiczeniu organizmu oraz potrzebne do ćwiczenia aplikacje

```
sudo docker pull wkusmirek/repeatmasker
sudo docker pull wkusmirek/maker
```

```
wget path_to_genome
wget path_to_proteins
wget path_to_mRNA_transcripts
```

Odnosniki do danych (pliki FASTA) są dostępne na stronie przedmiotu MBI.

2.2 Przygotowanie danych

Rozpakuj pobrane pliki FASTA, np:

```
gzip -d *
```

Następnie z pliku zawierającego wyniki assemblingu (plik ze skafoldami) wybierz jedną sekwencję DNA. Proszę wybrać identyfikator sekwencji o indeksie $\text{numer_indeksu} \bmod 150$. Jeżeli w numerze indeksu występują jakieś litery - pomijamy je. Rozważamy numer indeksu jednej osoby z zespołu. Przykładowo dla numeru 01027972 wybieramy identyfikator sekwencji z linii o indeksie 22 ($1027972 \bmod 150 = 22$). Proszę zamieścić w sprawozdaniu informacje o numerze indeksu oraz identyfikatorze wybranej sekwencji. Plik z mapowaniem indeksów na identyfikatory sekwencji dostępny jest na stronie przedmiotu. Po wybraniu identyfikatora sekwencji wybierz odpowiednią sekwencję z pliku ze skafoldami i zapisz ją w oddzielnym pliku o nazwie 'single_scaffold.fa'.

2.3 Maskowanie genomu

Wygeneruj zmodyfikowaną sekwencję DNA, która będzie miała zamaskowane (oznaczone symbolami 'N') podsekwencje powtarzające się. Wykorzystaj program RepeatMasker:

```
sudo docker run -it --rm -v /tmp:/tmp
-w /tmp wkusmirek/repeatmasker RepeatMasker
--species arabidopsis /tmp/single_scaffold.fa
```

Zapoznaj się z wygenerowanym plikiem oraz z logami aplikacji z konsoli. Porównaj pliki 'single_scaffold.fa' i plik 'single_scaffold.fa.masked'.

Odpowiedz w sprawozdaniu na następujące pytania:

- Ile nukleotydów zostało zamaskowanych?
- Czy zamaskowane nukleotydy były pojedynczymi nukleotydami, czy ciągami nukleotydów?
- Kolejnym etapem ćwiczenia będzie zmapowanie sekwencji mRNA i białek na genom z zamaskowanymi sekwencjami repetytywnymi. W jaki sposób maskowanie sekwencji repetytywnych może wpłynąć na wynik mapowania?

2.4 Mapowanie znanych sekwencji i adnotacja strukturalna

Adnotacja strukturalna dostarcza dodatkowych danych ułatwiających zidentyfikowanie struktur genowych. Główną metodą adnotacji jest zmapowanie białek i mRNA na genom. Wykorzystaj program Maker:

- wygeneruj pliki konfiguracyjne:

```
sudo docker run --rm -v /tmp:/tmp -w /tmp wkusmirek/maker maker -CTL
```

- zmodyfikuj plik 'maker_opts.ctl', ustawiając odpowiednie ścieżki do plików dla opcji: 'genome', 'est', 'protein'. Ścieżka dla 'genome' powinna wskazywać na ścieżkę do zamaskowanego pliku 'single_scaffold.fa.masked'.
- uruchom aplikację Maker

```
sudo docker run --rm -v /tmp:/tmp -w /tmp wkusmirek/maker maker
```

Zapoznaj się z wygenerowanym plikiem .gff. Przykładowa ścieżka do pliku .gff to:

```
/tmp/single_scaffold.fa.maker.output/single_scaffold.fa_datastore  
/F8/34/HDID_contig0001624/HDID_contig0001624.gff
```

Ścieżka ta może nieznacznie różnić się w zależności od użytej sekwencji DNA. Wklej 10 pierwszych wierszy z tego pliku do sprawozdania.

Odpowiedz w sprawozdaniu na następujące pytania:

- Jakie informacje można odczytać z wygenerowanego pliku .gff?
- Oblicz ilość wygenerowanych zdarzeń typu *expressed_sequence_match* i *protein_match*. Co oznaczają wymienione typy zdarzeń?

2.5 Adotacja funkcjonalna

Znajdź w pliku .gff wiersz opisujący fragment genu zawierający w opisie znacznik 'expressed_sequence_match'. Przykładowy wiersz może wyglądać w następujący sposób:

```
HDID_scaffold0000001 blastn expressed_sequence_match 17669 18426 40 -.  
ID=HDID_scaffold0000001:hit:57:3.2.0.0;Name=HDID_0000675501-mRNA-1
```

Jeśli wiersz ze znacznikiem 'expressed_sequence_match' nie jest obecny w pliku gff, to powtórz całe ćwiczenie dla innej sekwencji wejściowej, która została wybrana na początku ćwiczenia. Sekwencje wybierz w sposób losowy, opisz w sprawozdaniu fakt braku 'expressed_sequence_match' oraz identyfikator nowej sekwencji.

Następnie z sekwencji FASTA, zapisanych na końcu pliku .gff, wyodrębnij sekwencję fragmentu genu. Dla zaprezentowanego powyżej wiersza należy z sekwencji o identyfikatorze HDID_scaffold0000001 wyodrębnić sekwencję nukleotydów o indeksach 17669 – 18426.

Dla otrzymanej sekwencji przypisz funkcje zidentyfikowanym genom: użyj algorytmu BLASTX z bazy NCBI (https://blast.ncbi.nlm.nih.gov/Blast.cgi?LINK_LOC=blasthome&

PAGE_TYPE=BlastSearch&PROGRAM=blastx) i wyszukaj opisy dla wybranych sekwencji genów. Wklej wyodrębnioną sekwencję DNA, nie zmieniaj żadnych opcji i kliknij przycisk "BLAST".

U.S. National Library of Medicine | NCBI National Center for Biotechnology Information

BLAST » blastx

Translated BLAST: blastx

blastn | blastp | **blastx** | tblastn | tblastx

BLASTx search protein databases using a translated nucleotide query. [more...](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

AAATCCACATGAAAGTACCTCTACCTTCAGAGAGGTGAAAGCTCGATCCACATTGGTTGAG
TCTAAGGCTGATGTTTCAATGAAGTGAAGACAGATATCTTCAGCGAATGGCTTGGCTTCGCTGAG
AACCGTACGACATGGCGCAGATCACATTATTTCCAAATATGGTGAATATTTGATCAGCAT
TATTTCTAAGTTCTTCAACCATGCTTTCATTTTCAAGAGTATGCTAGTATATCATATACA
ACGAGGCTTCAACGACACCGGATATATCTTATGAAACGACACTTATTATAGAAATATCTCAAT
TTACGCTGACGTTATTGCTCTGATCTTCTTGGCCGGCTGATCCCAATTTGGATTTTTATAACC

Clear Query subrange

From

To

Or, upload file

Browse... No file selected.

Genetic code

Standard (1)

Job Title

Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Database

Non-redundant protein sequences (nr)

Organism

Optional

Enter organism name or id--completions will be suggested

Exclude

Optional

☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Entrez Query

Optional

Enter an Entrez query to limit search

BLAST

Search database Non-redundant protein sequences (nr) using Blastx (search protein databases using a translated nucleotide query)

☐ Show results in a new window

[Algorithm parameters](#)

przkładowy wynik

	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	ras protein Rab-11A [Hymenolepis microstoma]	293	356	94%	2e-97	77%	CD330947.1
☐	rab [Echinococcus multilocularis]	276	340	94%	4e-91	73%	CD336404.1
☐	Ras-related protein Rab-11A [Echinococcus granulosus]	275	338	94%	2e-90	72%	XP_024345837.1
☐	Ras family protein [Toxotrypanus	258	258	81%	2e-83	64%	KH448718.1
☐	ras protein Rab-11B [Toxotrypanus	256	256	80%	2e-83	65%	CD056810.1
☐	Putative Small GTPase [Rhizopus microsporus]	253	253	78%	2e-82	66%	CD073387.1
☐	PREDICTED: ras-related protein Rab-11B-like [Aureothidium quercusindica]	253	253	76%	1e-81	67%	XP_003388803.1
☐	ras protein Rab-11A [Hymenolepis microstoma]	252	300	99%	9e-81	62%	CD330941.1

Dodaj do sprawozdania nazwy organizmów, które mają 5 najbardziej podobnych sekwencji do badanej, procent podobieństwa oraz istotność wyniku (E-value).

Odpowiedz w sprawozdaniu na następujące pytania:

- Co oznacza oraz jak interpretować wartość E-value?
- Zinterpretuj liste uzyskanych organizmów (w ćwiczeniu pracujemy na genomie tasiemca szczurzego *Hymenolepis diminuta*).

3 Zadanie implementacyjne

Proszę zapoznać się z tematyką konwersji sekwencji aminokwasów na mRNA. Proszę zaimplementować prosty skrypt umożliwiający odczytać zawartość pliku w formacie FASTA

z sekwencjami aminokwasów. Następnie aplikacja konwertuje sekwencje aminokwasów na sekwencje RNA i zapisuje ją oddzielnego pliku w formacie FASTA. Proszę wykorzystać biblioteki dedykowane do przetwarzania danych genomowych, np.:

- SeqAn;
- Biopython;
- BioJava.

Uwaga: Jeden aminokwas może być kodowany przez więcej niż jedną trójkę nukleotydów - w takim wypadku należy na wyjściu podać dowolną trójkę kodującą ten aminokwas.