Uniwersytet Przyrodniczy we Wrocławiu
Wydział Biologii i Hodowli Zwierząt

Pracownia Informatyczna

biologiczne bazy danych

LISTA 2

Szymon Armata
117360

Bioinformatyka, rok III, grupa II

1. Formaty danych

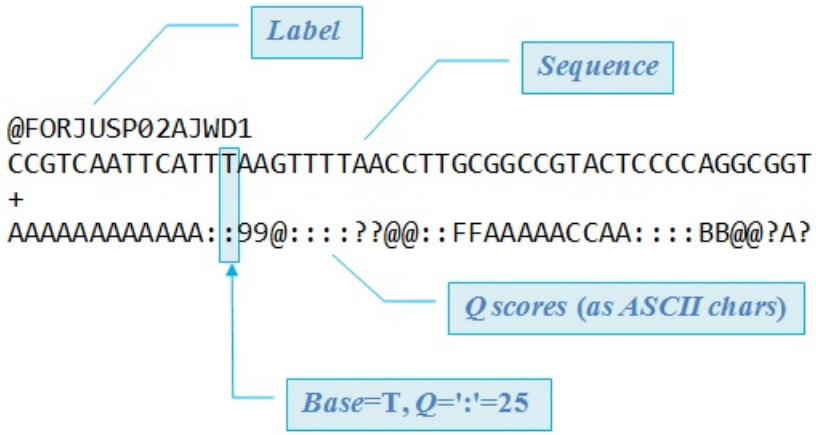
1.1. FASTA

FASTA – format zapisu sekwencji kwasów nukleinowych oraz białek używany w bioinformatyce. Nukleotydy (dla DNA i RNA) oraz aminokwasy (dla białek) oznaczone są jednoliterowymi skrótami. Format FASTA uwzględnia również możliwość dodawania opisów i komentarzy do sekwencji. Dane zapisane w formacie FASTA składają się z pojedynczej linii tekstu zawierającej opis sekwencji oraz z kolejnych linii zawierających samą sekwencję. Linia z opisem rozpoczyna się od znaku "większe niż" (">"). Pierwsze słowo po tym znaku służy jako identyfikator sekwencji. Dalej w tej samej linii umieszczany jest opis. W kolejnych liniach znajduje się ciąg znaków składający się na sekwencję. Przykładowa sekwencje w formacie FASTA wygląda tak:

```
>NG_008679.1:5001-38170 Homo sapiens paired box 6 (PAX6)
ACCCTCTTTTCTTATCATTGACATTTAAACTCTGGGGCAGGTCCTCGCGTAGAACGCGGCTGTCAGATCT
GCCACTTCCCCTGCCGAGCGGCGGTGAGAAGTGTGGGAACCGGCGCTGCCAGGCTCACCTGCCTCCCCGC
CCTCCGCTCCCAGGTAACCGCCCGGGCTCCGGCCCCGGCCCGGCTCGGGGCCCGCGGGGCCTCTCCGCTG
CCAGCGACTGCTGTCCCCAAATCAAAGCCCGCCCCAAGTGGCCCCGGGGCTTGATTTTTGCTTTTAAAAG
GAGGCATACAAAGATGGAAGCGAGTTACTGAGGGAGGGATAGGAAGGGGGGTGGAGGAGGGACTTGTCTT
TGCCGAGTGTGCTCTTCTGCAAAAGTAGCAAAATGTTCCACTCCTAAGAGTGGACTTCCAGTCCGGCCCT
GAGCTGGGAGTAGGGGGCGGGAGTCTGCTGCTGCTGTCTGCTAAAGCCACTCGCGACCGCGAAAAATGCA
GGAGGTGGGGACGCACTTTGCATCCAGACCTCCTCTGCATCGCAGTTCACGACATCCACGCTTGGGAAAG
TCCGTACCCGCGCCTGGAGCGCTTAAAGACACCCTGCCGCGGGTCGGGCGAGGTGCAGCAGAAGTTTCCC
GCGGTTGCAAAGTGCAGATGGCTGGACCGCAACAAAGTCTAGAGATGGGGTTCGTTTCTCAGAAAGACGC
```

1.2. FASTQ

FASTQ - format do przechowywania danych z sekwencjonowania; 4 linie na fragment:
1. Linia zaczyna się od znaku @, następnie identyfikator sekwencji i opcjonalnie opis
2. Linia zawiera wyłącznie sekwencję
3. Linia zaczyna się od znaku + i może opcjonalnie zawierać identyfikator i opis
4. Linia zawiera symbolicznie zapisaną ocenę jakości dla każdego nukleotydu. Musi zawierać dokładnie tyle samo znaków co sekwencja. Skala jakości (! najniższa ~ najwyższa):



1.3. VCF (Variant Call Format)

Format Variant Call Format (VCF) określa format pliku tekstowego używanego w bioinformatyce do przechowywania zmienności sekwencji genów. Format ten został opracowany wraz z pojawieniem się dużych projektów genotypowania i sekwencjonowania DNA, takich jak 1000 Genomes Project. Istniejące formaty dla danych genetycznych, takie jak General Feature Format (GFF), przechowywały wszystkie dane genetyczne, z których znaczna część jest zbędna, ponieważ będzie współdzielona przez wszystkie genomy. Dzięki zastosowaniu formatu wywoływania wariantów tylko warianty muszą być przechowywane wraz z genomem referencyjnym.

- metadane

- nagłówek

VCF

```
##fileformat=VCFv4.2
##contig=<ID=2,length=51304566>
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
#CHROM  POS ID  REF ALT QUAL FILTER  INFO FORMAT     SAMPLE1     SAMPLE2     SAMPLE3     SAMPLE4     SAMPLE5     SAMPLE6     SAMPLE7
2  81170  .  C  T   .  .    AC=9;AN=7424   GT:DP:GQ  0/0:4:12    0/0:3:9     0/1:1:3     0/1:9:24    1/0:4:12    0/0:5:15    0/0:4:12
2  81171  .  G  A   .  .    AC=6;AN=7446   GT:DP:GQ  0/1:4:12    0/0:3:9     0/0:1:3     0/0:9:24    0/1:4:12    0/1:5:15    0/0:4:12
2  81182  .  A  G   .  .    AC=5;AN=7506   GT:DP:GQ  0/0:5:15    0/0:4:12    0/0:5:15    0/0:9:24    0/0:4:12    0/0:4:12    0/0:4:12
2  81204  .  T  G   .  .    AC=2;AN=7542   GT:DP:GQ  1/0:5:15    0/0:9:27    0/0:10:30   0/0:15:39   0/0:9:27    1/0:13:39   0/1:14:42
```

BCF

```
2  81170  .  C  T   .  .    AC=9;AN=7424   GT:0/0:0/0:0/1:0/1:1/0:0/0:0/0    DP:4:3:1:9:4:5:4          GQ:12: 9: 3:24:12:15:12
2  81171  .  G  A   .  .    AC=6;AN=7446   GT:0/1:0/0:0/0:0/0:0/1:0/1:0/0    DP:4:3:1:9:4:5:4          GQ:12: 9: 3:24:12:15:12
2  81182  .  A  G   .  .    AC=5;AN=7506   GT:0/0:0/0:0/0:0/0:0/0:0/0:0/0    DP:5:4:5:9:4:4:4          GQ:15:12:15:24:12:12:12
2  81204  .  T  G   .  .    AC=2;AN=7542   GT:1/0:0/0:0/0:0/0:0/0:1/0:0/1    DP:5:9:10:15:9:13:14      GQ:15:27:30:39:27:39:42
```

2. Baza Ensembl

Otwórz przeglądarkę internetową i korzystając z bazy danych Ensembl znajdź:
a) wszystkie sekwencje białkowe dostępne dla człowieka
b) sekwencje ncRNA dla dingo
c) adnotację genomową w formacie GTF dla szczura wędrownego
Otwórz terminal. Przejdź do katalogu nazwanego dwiema literami Twojego imienia oraz nazwiska (np. Adam Mickiewicz posiada katalog o nazwie „AdMi”). Pobieranie znalezionych danych wykonaj w wierszu poleceń. W terminalu wpisz odpowiednie polecenie wraz z adresem linku prowadzącym do danych. Przyjrzyj się nazwom pobranych plików.

```
a) wget ftp://ftp.ensembl.org/pub/release-101/fasta/
homo_sapiens/pep/Homo_sapiens.GRCh38.pep.all.fa.gz

b) wget ftp://ftp.ensembl.org/pub/release-101/fasta/
canis_lupus_dingo/ncrna/Canis_lupus_dingo.ASM325472v1.
ncrna.fa.gz

c) wget ftp://ftp.ensembl.org/pub/release-101/gtf/
rattus_norvegicus/Rattus_norvegicus.Rnor_6.0.101.gtf.gz
```

3. Praca na plikach z bazy Ensembl

Korzystając z plików z poprzedniego zadania:

3.1. wyświetl zawartość pliku przechowującego sekwencje niekodujące

```
less Canis_lupus_dingo.ASM325472v1.ncrna.fa.gz
```

3.2. rozpakuj plik zawierający sekwencje białkowe bez usuwania oryginalnego pliku

```
gunzip -k Homo_sapiens.GRCh38.pep.all.fa.gz
```

3.3. sprawdź czy znany jest gen o nazwie Pcmt1-201 dla szczura wędrownego Na którym chromosomie się znajduje?

```
grep 'Pcmt1-201' Rattus_norvegicus.Rnor_6.0.101.gtf
| head -n 1 | cut -f 1
```

4. Transkryptomy

Wiele nowoczesnych programów do mierzenia poziomu ekspresji genów w oparciu o dane RNA-seq potrzebuje jako plik wejściowy wszystkich znanych sekwencji transkryptów dla badanego organizmu.

4.1. Utwórz katalog o nazwie „transkryptom_hsapiens”

```
mkdir transkryptom_hsapiens
```

4.2. Z bazy Ensembl pobierz do niego transkrypty kodujące białko i transkrypty niekodujące u człowieka

```
cd transkryptom_hsapiens

wget http://ftp.ensembl.org/pub/release-105/fasta/
homo_sapiens/cdna/Homo_sapiens.GRCh38.cdna.all.fa.gz

wget http://ftp.ensembl.org/pub/release-105/fasta/
homo_sapiens/ncrna/Homo_sapiens.GRCh38.ncrna.fa.gz
```

4.3. Połącz oba pliki w jeden i spakuj go

```
zcat Homo_sapiens.GRCh38.cdna.all.fa.gz
Homo_sapiens.GRCh38.ncrna.fa.gz
| gzip > combined.fastq.gz
```

5. Polimorfizmy

Plik o nazwie „homo_sapiens_variant_annot.txt" jest wynikiem adnotacji funkcjonalnej 4 polimorfizmów:

5.1. jakie to polimorfizmy?

1_65568_A/G

3_319781_A/G

4_125638_A/T

5_145726_C/T

5.2. jakie są ich pozycje w genomie?

65568

319781

125638

145726

5.3. czy są zlokalizowane w genach?

protein_coding

6. Koordynaty polimorfizmów

Z pliku o nazwie „bos.txt” wyodrębnij koordynaty polimorfizmów i zapisz je w nowym pliku.

```
cut -f2 bos.txt > coordinates.txt
```