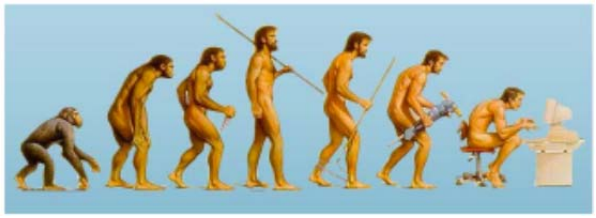
Ćwiczenie 07

Metody bioinformatyczne w genetyce molekularnej. Bazy danych sekwencji nukleotydowych. Rekordy baz danych. Genom człowieka w bazach danych.

Kornelia Polok

1. Metody bioinformatyczne w genetyce molekularnej

1.1. Bioinformatyka



Bioinformatyka zajmuje się opracowywaniem metod i narzędzi niezbędnych do analizy danych biologicznych. Stanowi ona połączenie statystyki, matematyki i informatyki. Bioinformatyka wykorzystywana jest w analizach *in silico*. Bioinformatyka stała się nieodłącznym elementem biologii molekularnej i genetyki. Metody bioinformatyczne wykorzystywane są do identyfikacji genów, polimorfizmu nukleotydów (SNP), ale także analizy dużych zbiorów danych jakie otrzymujemy z projektów sekwencjonowania genomów i transkryptomów. Większość badań dotyczy analizy sekwencji, składania genomów, poszukiwania genów, projektowania leków, modelowania i porównywania białek, analiza interakcji, analiza sprzężeń, analizy ewolucyjne i filogenetyczne.

Narzędzia, które umożliwiają analizy wykorzystują matematykę dyskretną, teorię mnogości, teorię systemów, teorię grafów, analizę matematyczną, statystykę i geometrię.

1.2. Dane wykorzystywane przez bioinformatykę

- Sekwencje DNA/RNA;
- SNPs
- Sekwencja, struktura i funkcja białek
- Dane dotyczące organizmów
- Sekwencje genomów
- Sekwencje transkryptomów
- Interakcje między cząsteczkami biologicznymi
- Szlaki metaboliczne

1.3. Typowe etapy analizy bioinformatycznej

- **Algorytmy**: zestaw operacji wykorzystywanych w złożonych obliczeniach i analizie danych. Przepis jak wykonać obliczenia. Najprostszy algorytm może dotyczyć zestawu operacji jakie należy wykonać aby dodać dwie liczby. Największym wyzwaniem jest podział zadania na poszczególne operacje.
- **Języki komputerowe**: umożliwiają zapis operacji za pomocą odpowiedniej gramatyki (syntax). W biologii obliczeniowej najczęściej używa się Python oraz R. Języki komputerowe zazwyczaj używają odpowiedniego zestawu operatorów.
- **Format plików**: standardowy sposób zapisu informacji. Dane biologiczne są najczęściej zapisywane jako zwykły tekst, co oznacza, że można je odczytać w prostym edytorze typu Notepad++. Pliki z sekwencjami występują w postaci FASTA i FASTQ. FASTA zawiera jedną lub kilka sekwencji, każda z unikalnym identyfikatorem (ID). Rozszerzenie nie jest stałe i może także występować z rozszerzeniem txt. FASTQ zawiera dodatkowe dane o jakości każdego nukleotydu.

2. Bazy danych sekwencji nukleotydowych

Liczba baz danych systematycznie zwiększa się. W 2019 roku było 148 baz danych, z czego 59 baz było nowych a 79 bazy były uzupełnione i unowocześnione. Bazy danych są corocznie opisywane w czasopiśmie Nucleic Acid Research w specjalnym numerze poświęconym bazom danych. Bazy pogrupowane są według typów:

- sekwencje nukleotydowe, struktury i regulacja transkrypcji;
- sekwencje aminokwasowe i struktury białek;
- szlaki metaboliczne i sygnałowe;
- genomy wirusów, bakterii, pierwotniaków i grzybów;
- genom człowieka i organizmów modelowych;
- zmienność genomu ludzkiego;
- bazy roślinne;
- inne.

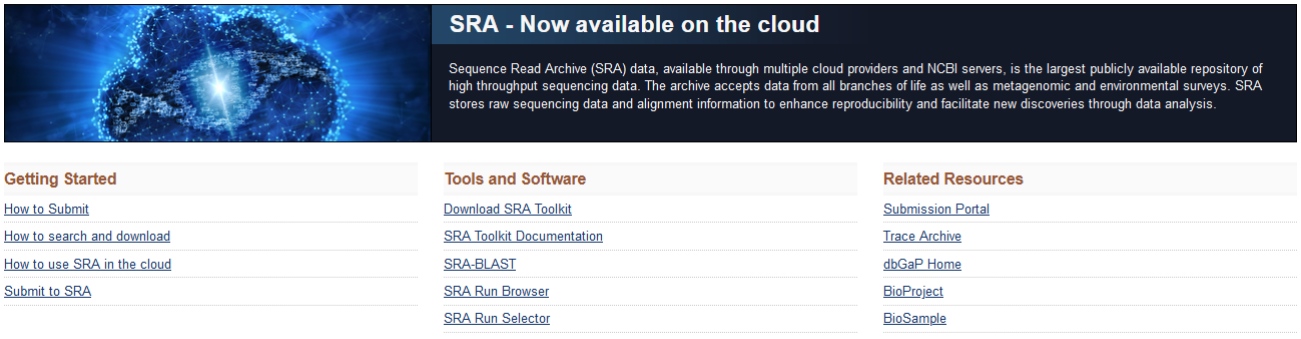
2.1. INSD: Międzynarodowa baza danych sekwencji nukleotydowych

INSD (International Nucleotide Sequence Database) jest międzynarodową inicjatywą, która skupia trzy bazy danych, będące repozytoriami sekwencji nukleotydowych wszystkich organizmów. Są to:

- DNA Data Bank of Japan
- EMBL: European Bioinformatics Institute
- GeneBank (National Centre for Biotechnology Information.

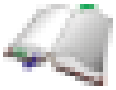
Bazy są synchronizowane w trybie dziennym tak aby każda zawierała tę samą informację. Natomiast dane nie są podawane według jednego standardu, co utrudnia krzyżowe korzystanie z baz nawet w zakresie liczby zdeponowanych danych. Pomimo koordynacji otrzymane wyniki podawane są w różniących się formatach. W przypadku dużych genomów, dane z EMBL mogą ułatwiać analizę, gdyż podawana jest sumaryczna strona z odnośnikami do informacji o genach, transkryptach etc.

Bazy współpracują z Archiwum zawierającym odczyty z sekwenatorów: **SRA** (Sequence Read Archive). Archiwum znajduje się na stronach NCBI i można pobierać dane lub parcować w chmurze. Wówczas konieczne jest zalogowanie się z kontem w Google oraz utworzenie wirtualnej maszyny (np. Oracle VM Virtual Box).



Rys. 2.1. Zrzut ekranu strony dostępowej do SRA.

2.2. Wtórne bazy danych



Wtórne bazy danych zawierają sekwencje, które zostały wstępnie opracowane. Dotyczą one określonego zagadnienia, grupy organizmów, ekspresji genów itd. Przykładem wtórnej bazy danych jest OMIM: Online Mendelian Inheritance in Man. Do takich baz należą także:

A. 1000 Genome Project (https://www.internationalgenome.org/)
B. HapMap (https://www.genome.gov/10001688/international-hapmap-project)
C. 23andMe (https://www.23andme.com/en-int/)

Korzystając z podanych stron internetowych proszę scharakteryzować każdą z wymienionych baz. Jaki był cel ich utworzenia? Jakich informacji wymienione bazy dostarczają?
(2 punkty/baza)

Czas wykonania: 15 minut.

3. Rekordy baz danych na przykładzie NCBi (GenBank)

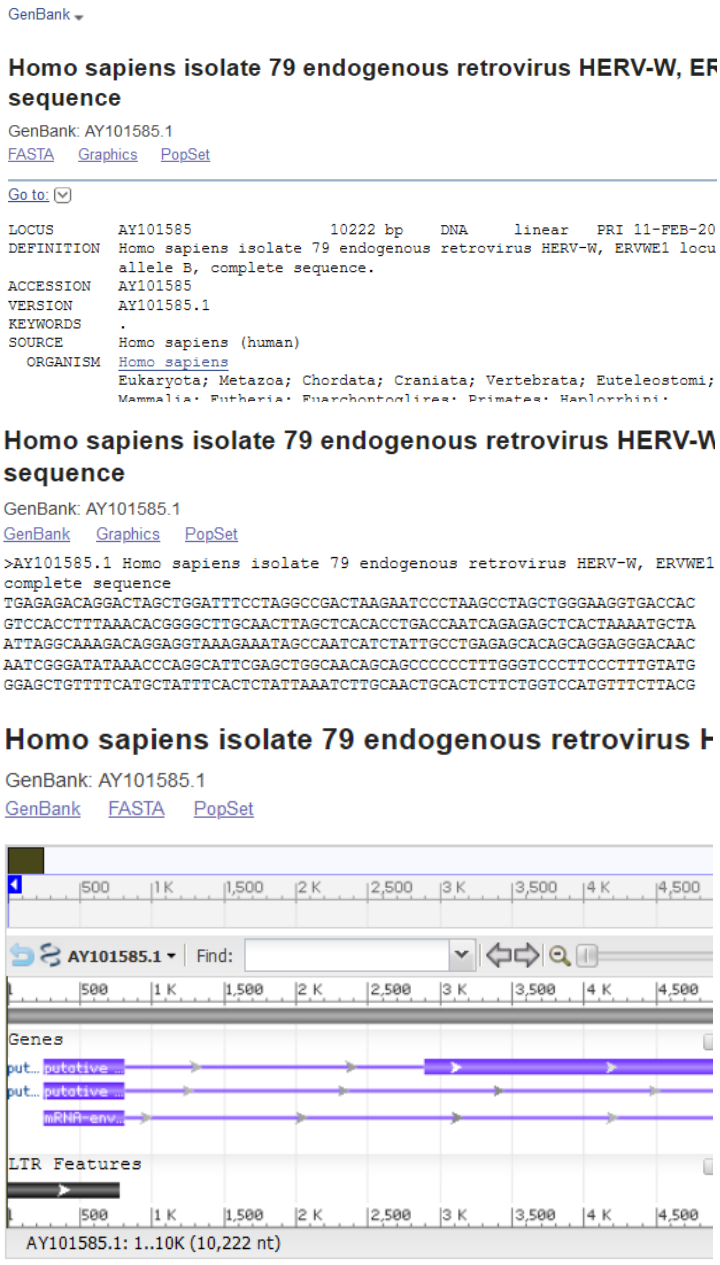
3.1. Struktura rekordu sekwencji nukleotydowej

3.1.1. Rekordy w bazie NCBI mogą być przedstawione jako:

- **GenBank**: podane są cechy sekwencji, skąd pochodzi, a także długość sekwencji, kolejność nukleotydów;
- **FASTA**: tylko sama sekwencja , ten widok jest wykorzystywany w analizach;
- **Graphics**: widok graficzny przedstawiający nici DNA, egzony, miejsca aktywne.

3.1.2. Informacje w widoku GenBank

- **LOCUS**: podany jest numer akcesyjny, który jest unikalny dla danego rekordu.
- **10222 bp**: długość sekwencji w parach zasad,
- **DNA**: cząsteczka, która została wykorzystana do sekwencjonowania. Zawsze podana jest wyjściowa cząsteczka.
- **DEFINITION**: nazwa genu, regiony



Rys. 3.1a. Widok rekordu w NCBI (od góry): genBank, Fasta, Graphic.

- **ACCESSION**: numer akcesyjny
- **VERSION**: wersja.
- **ORGANISM**: systematyka gatunku, od którego pochodzi sekwencja.

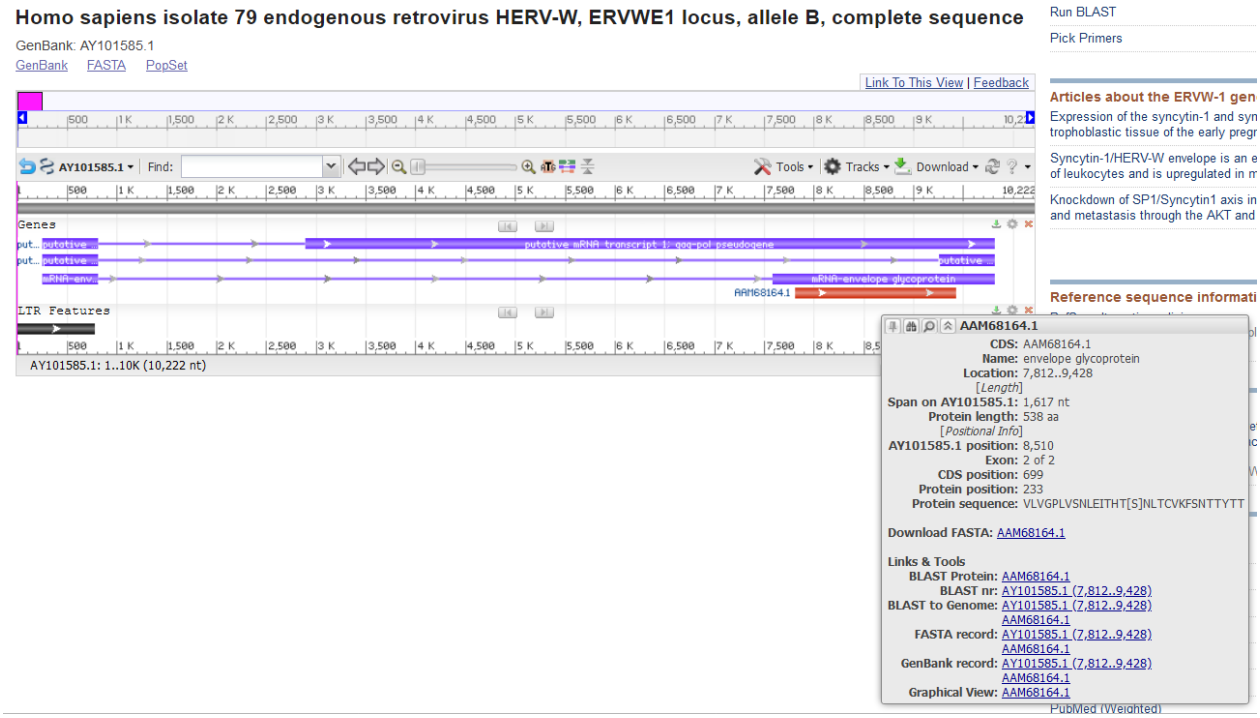
3.1.3. Informacje opisane jako Features

- **Source**: z jakiej tkanki otrzymano sekwencję, pozycja na mapie, jak wyizolowano.
- **Repeat region**: długość sekwencji powtórzonej.
- **mRNA**: jakie fragmenty tworzą mRNA, tutaj 248-811 oraz 2887-9813, instrukcja jak należy połączyć poszczególne elementy sekwencji.
- **CDC**: część kodująca wraz z translacją i odnośnikiem do białka.
- **ORIGIN**: sekwencja, w linijce po 60 nukleotydów: 6 x 10.

```
                d Italie, Lyon 69364 cedex 07, France
FEATURES             Location/Qualifiers
     source          1..10222
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /isolate="79"
                     /isolation_source="PBMC of indi
                     /db_xref="taxon:9606"
                     /chromosome="7"
                     /map="7q21-q22"
                     /sex="male"
                     /cell_type="PBMC"
                     /note="caucasian;
                     isolated by PCR; ERVWE1 locus,
                     endogenous virus: HERV-W"
     repeat_region   1..780
                     /rpt_type=long_terminal_repeat
     mRNA            join(248..811,2887..9813)
                     /note="putative mRNA transcript
     mRNA            join(248..811,7585..9813)
                     /product="envelope glycoprotein
                     /note="putative mRNA transcript
     mRNA            join(248..811,9248..9813)
                     /note="putative mRNA transcript
     CDS             7812..9428
                     /note="syncytin"
                     /codon_start=1
                     /product="envelope glycoprotein
                     /protein_id="AAM68164.1"
                     /translation="MALPYHIFLFTVLLPSF
                     GNIDAPSYRSLSKGTPTFTAHTHMPRNCYHS
                     TVCWTYFTQTGMSDGGGVQDQAREKHVKEVI
                     HTRLVSLFNTTLTGLHEVSAQNPTNCWICLP
                     LVGPLVSNLEITHTSNLTCVKFSNTTYTTNS
                     YRCLNGSSESMCFLSFLVPPMTIYTEQDLYS
                     TGIGGITTSTQFYYKLSQELNGDMERVADSL
                     RGGTCLFLGEECCYYVNQSGIVTEKVKEIRD
                     FLGPLAAIILLLLFGPCIFNLLVNFVSSRIE
                     RSDVNDIKGTPPEEISAAQPLLRPNSAGSS"
     repeat_region   9487..10222
                     /rpt_type=long_terminal_repeat
ORIGIN
        1 tgagagacag gactagctgg atttcctagg ccgactaag
       61 aggtgaccac gtccaccttt aaacacgggg cttgcaact
      121 gagagctcac taaaatgcta attaggcaaa gacaggagg
      181 attgcctgag agcacagcag gagggacaac aatcgggat
```

Rys. 3.1b. Fragment rekordu GenBank z opisem Features.

3.1.4. W widoku graficznym można odczytać dodatkowe informacje o elementach strukturalnych sekwencji. W tym celu należy „najechać myszą" na dany element. Po załadowaniu informacji można ją przypiąć do widoku. Pokazane są także egzony. Korzystając z suwaka można odczytać kodony.



Rys. 3.1c. Widok graficzny.

3.2. Analiza sekwencji nukleotydowej zdeponowanej w NCBI

Dla sekwencji o numerze akcesyjnym **AF135372.1** proszę podać następujące informacje:

A. Z jakiego gatunku pochodzi sekwencja?
B. Jaki gen został podany?
C. Z jakich tkanek/komórek wyizolowano kwas nukleinowy?
(1 punkt)
D. Jaki kwas nukleinowy był wykorzystany do otrzymania tej sekwencji?
E. Czy sekwencja zdeponowana w NCBI zawiera introny? Uzasadnij odpowiedź.
F. Podaj długość zdeponowanej sekwencji.
(1 punkt)
G. Proszę podać za pomocą pozycji nukleotydów lokalizację genu znajdującego się w obrębie tej sekwencji.
(1 punkt)
H. Proszę podać które fragmenty tej sekwencji tworzą mRNA.
(1 punkt)
I. Proszę przełączyć się na widok polipeptydu i podać jaki fragment tego polipeptydu zawiera motyw SNARE i z ilu aminokwasów składa się ten motyw?
(2 punkty)

Czas wykonania: 15 minut

Odpowiedzi

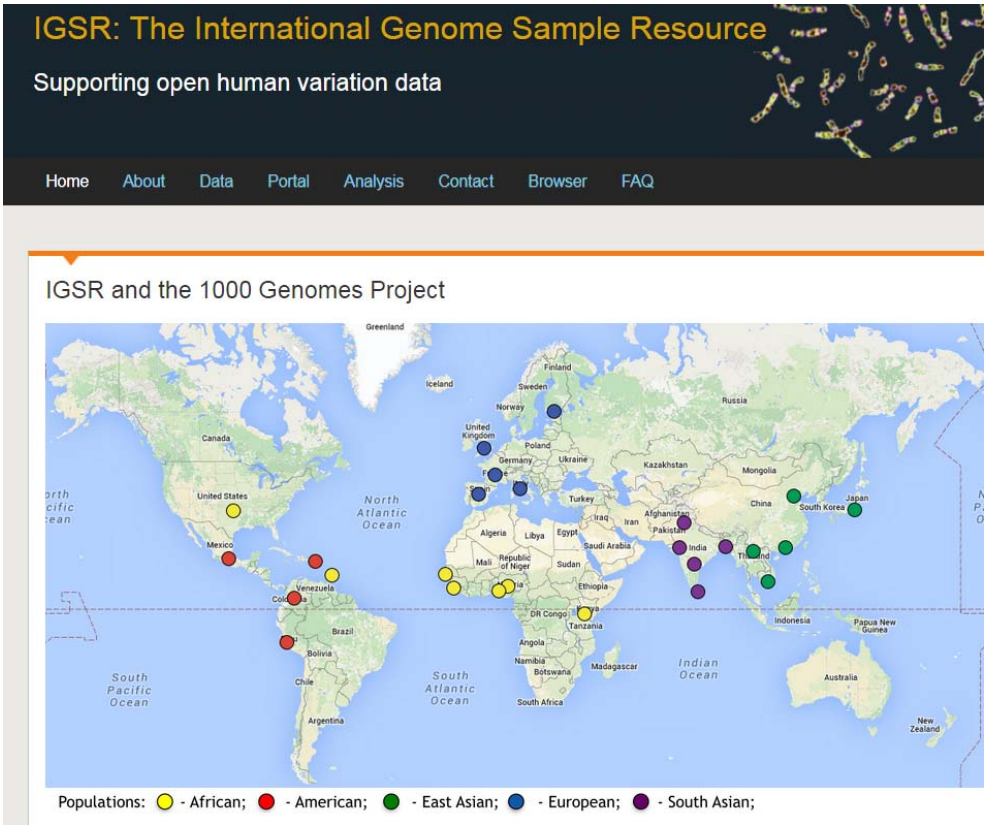
2. Bazy danych sekwencji nukleotydowych

2.2. Wtórne bazy danych

Wtórne bazy danych zawierają sekwencje, które zostały wstępnie opracowane. Dotyczą one określonego zagadnienia, grupy organizmów, ekspresji genów itd. Przykładem wtórnej bazy danych jest OMIM: Online Mendelian Inheritance in Man. Do takich baz należą także:

J. 1000 Genome Project (https://www.internationalgenome.org/home)

- Projekt prowadzono w latach 2008-2015, stanowi od największy katalog informacji o zmienności populacji ludzkich. Pierwotnym celem projektu była identyfikacja wariantów o częstości co najmniej 1% (polimorfizmy).
- Projekt wykorzystał dane 2504 osób z 26 populacji. Sekwencjonowanie o wysokiej rozdzielczości przeprowadzono dla 24 osób.



Rys. 2.2a. Zrzut ekranu strony dostępowej projektu 1000 Genome Project z pokazanymi lokalizacjami badanych prób.

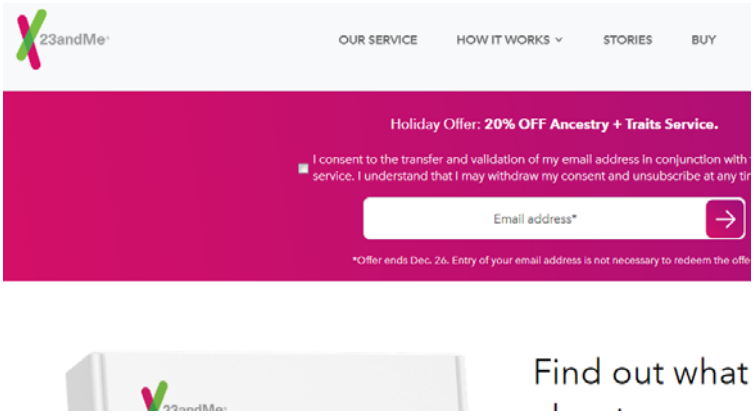
K. HapMap (https://www.genome.gov/10001688/international-hapmap-project)

- HapMap: International HapMap Project. Jego celem było poszukiwanie SNPs, które mogą być związane z chorobami poprzez mapowanie bloków haplotypów. SNPs, które zlokalizowane są w bliskiej odległości są sprzężone, co oznacza, że mogą one być dziedziczone wspólnie. Znalezienie charakterystycznych SNPs identyfikuje cały blok.
- Projekt umożliwił analizę podłoża chorób genetycznych jak również wpływu środowiska na zmienność.
- Obecnie jest to projekt zakończony.

Rys. 2.2b. Zrzut ekranu strony archiwalnej projektu HapMap.

L. 23andMe (https://www.23andme.com/en-int/)

- Prywatna baza danych, która gromadzi informacje z sekwencjonowania DNA osób poszukujących przodków. Analiza obejmuje pochodzenie DNA z 2000 lokalizacji na świecie, pochodzenie do 1000 lat wstecz oraz informację o udziale genów neandertalskich. Dane nie są dostępne publicznie.
- Umowa podpisana z bazą zakłada zgodę na prowadzenie badań na dostarczonym DNA, zgodę na przechowywanie DNA w okresie 1-10 lat.
- Strona nie podaje danych o firmie ani jej struktury.



Rys. 2.2c. Zrzut ekranu strony 23andMe.

3. Rekordy baz danych

3.2. Analiza sekwencji nukleotydowej zdeponowanej w NCBI

Dla sekwencji o numerze akcesyjnym **AF135372.1** proszę podać następujące informacje:

A. Z jakiego gatunku pochodzi sekwencja?

Homo sapiens

B. Jaki gen został podany?

Synaptobrevina 2, *VAMP2 (SYB2)*

C. Z jakich tkanek/komórek wyizolowano kwas nukleinowy?

Komórki krwi, leukocyty

D. Jaki kwas nukleinowy był wykorzystany do otrzymania tej sekwencji?

DNA, liniowy

E. Czy sekwencja zdeponowana w NCBI zawiera introny? Uzasadnij odpowiedź.

Tak, gdyż materiałem wyjściowym był genomowy DNA.

F. Podaj długość zdeponowanej sekwencji.

6885 bp.

Homo sapiens synaptobrevin 2 (VAMP2) gene, complete cds

GenBank: AF135372.1

FASTA Graphics

Go to:

```
LOCUS       AF135372               6885 bp    DNA     linear   PRI 21-SEP-2000
DEFINITION  Homo sapiens synaptobrevin 2 (VAMP2) gene, complete cds.
ACCESSION   AF135372
VERSION     AF135372.1
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 6885)
  AUTHORS   Zoraqi,G.K., Paradisi,S., Falbo,V. and Taruscio,D.
  TITLE     Genomic organization and assignment of VAMP2 to 17p12 by FISH
  JOURNAL   Cytogenet Cell Genet 89 (3-4), 199-203 (2000)
   PUBMED   10965122
REFERENCE   2  (bases 1 to 6885)
  AUTHORS   Taruscio,D., Zoraqi,K.G. and Falbo,V.
  TITLE     Direct Submission
  JOURNAL   Submitted (17-MAR-1999) Ultrastrutture, Istituto Superiore di
            Sanita, Viale Regina Elena, 299, Rome, RM 00161, Italy
FEATURES             Location/Qualifiers
     source          1..6885
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /chromosome="17"
                     /map="17p12"
                     /clone="pISSHG2b3A"
                     /cell_type="leukocyte"
                     /tissue_type="blood"
     gene            2701..6532
                     /gene="VAMP2"
                     /gene_synonym="SYB2"
     mRNA            join(2701..2796,3305..3425,3907..4065,4149..4200,
                     4802..6532)
                     /gene="VAMP2"
                     /gene_synonym="SYB2"
                     /product="synaptobrevin 2"
     CDS             join(2795..2796,3305..3425,3907..4065,4149..4200,
                     4802..4818)
                     /gene="VAMP2"
                     /gene_synonym="SYB2"
                     /note="SYB-2; VAMP-2; vesicle protein"
                     /codon_start=1
                     /product="synaptobrevin 2"
                     /protein_id="AAF15551.1"
                     /translation="MSATAATAPPAAPAGEGGPPAPPPNLTSNRRLQQTQAQVDEVVD
                     IMRVNVDKVLERDQKLSELDDRADALQAGASQFETSAAKLKRKYWWKNLKMMIILGVI
```

G. Proszę podać za pomocą pozycji nukleotydów lokalizację genu znajdującego się w obrębie tej sekwencji.

2701 do 6532 bp

H. Proszę podać które fragmenty tej sekwencji tworzą mRNA.

- 2701-2796
- 3305-3425
- 3907-4065
- 4149-4200
- 4802-6532

I. Proszę przełączyć się na widok polipeptydu i podać jaki fragment tego polipeptydu zawiera motyw SNARE i z ilu aminokwasów składa się ten motyw?

Motyw SNARE znajduje się między 30 a 92 aminokwasem, składa się z 62 aminokwasów.