**NCBI Entrez Direct / E-utilities**

Pakiet [NCBI Entrez Direct](https://www.ncbi.nlm.nih.gov/books/NBK179288/) umożliwia dostęp do baz danych serwisu NCBI przy użyciu wiersza poleceń.

**Zad. 1**

W terminalu wpisz polecenie z pakietu NCBI Entrez Direct: einfo -dbs

1. Ile baz danych jest obsługiwanych przez pakiet (połącz polecenie z odpowiednim poleceniem Linuxa)?

2. Wyświetl informacje o nukleotydowej bazie danych: einfo -db nucleotide

- Jak nazywa się format danych, który otrzymałe/aś?

- Ile sekwencji znajduje się w bazie nukleotydowej (`<Count>`)?

3. Ile artykułów znajduje się w bazie PubMed?

**Zad. 2**

W przeglądarce internetowej otwórz stronę NCBI, wybierz bazę `Protein` i przejdź do zaawansowanego wyszukiwania (`Advanced`). Utwórz zapytanie w celu znalezienia wszystkich białek kodowanych przez gen o nazwie `TNRC6A` pochodzących z człowieka i bazy danych RefSeq.

1. Podaj użyte zapytanie (pole `Search details`).

2. Ile rekordów znaleziono?

**Zad. 3**

W terminalu uruchom poniższe polecenie.

esearch -db protein -query "TNRC6A[Gene Name] AND Homo sapiens[Organism] AND refseq[Filter]"

1. Ile rekordów znaleziono?

2. Uruchom poniższe polecenia i odpowiedz do czego służy polecenie `xtract`.

esearch -db protein -query "TNRC6A[Gene Name] AND Homo sapiens[Organism] AND refseq[Filter]" | xtract -outline

esearch -db protein -query "TNRC6A[Gene Name] AND Homo sapiens[Organism] AND refseq[Filter]" | xtract -pattern ENTREZ\_DIRECT -element Count

**Zad. 4**

Uruchom poniższe polecenie.

esearch -db protein -query "TNRC6A[Gene Name] AND Homo sapiens[Organism] AND refseq[Filter]" | efetch -format fasta

1. Do czego służy polecenie `efetch`?

2. Zmodyfikuj polecenie, aby wyświetlić sekwencje w formacie GenBank (skorzystaj z `efetch -help`).

**Zad. 5**

Korzystając z poleceń `esearch` i `efetch` wyszukaj sekwencje białkowe w formacie FASTA, które mają w tytule rekordu wyraz `caspase` i pochodzą z *Bacillus subtilis*.

1. Ile białek znaleziono?

2. Podaj użyte polecenie.

**Zad. 6**

Korzystając z poleceń `esearch` i `efetch` przeszukaj nukleotydową bazę i wyświetl w formacie GenBank wszystkie cząsteczki mRNA ludzkiego genu o nazwie TNRC6A pochodzące z bazy RefSeq. Jeżeli nie masz pewności jak utworzyć zapytania do bazy NCBI, przećwicz je najpierw w przeglądarce internetowej.

1. Ile sekwencji znaleziono?

2. Podaj użyte polecenie.

**Zad. 7**

Przy użyciu poleceń Linuxa zmodyfikuj polecenie z poprzedniego zadania, aby odpowiedzieć na następujące pytania:

1. Na którym chromosomie znajdują się znalezione geny?

2. Jaka jest łączna liczba egzonów we wszystkich znalezionych sekwencjach?

3. Wyświetl linie rekordów zaczynające się od `LOCUS` i uszereguj je ze względu na malejącą długość sekwencji.

LOCUS XM\_024450231 8606 bp mRNA linear PRI 28-FEB-2021

LOCUS XM\_017023145 8537 bp mRNA linear PRI 28-FEB-2021

LOCUS NM\_001351850 8506 bp mRNA linear PRI 19-FEB-2021

…

4. Wyświetl listę niepowtarzających się identyfikatorów do bazy PubMed.

PUBMED 11950943

PUBMED 12831532

PUBMED 13130130

…

**Zad. 8**

Uruchom poniższe dwa polecenia:

esearch -db nucleotide -query "TNRC6A[Gene Name] AND Homo sapiens[Organism] AND refseq[Filter] AND mrna[Filter]" | efetch -format docsum

esearch -db nucleotide -query "TNRC6A[Gene Name] AND Homo sapiens[Organism] AND refseq[Filter] AND mrna[Filter]" | efetch -format docsum | xtract -outline

Następnie zmodyfikuj drugie polecenie, aby uzyskać poniższe wyniki:

NM\_014494 8491 mRNA linear human 2021/04/15

XM\_017023152 6771 mRNA linear human 2021/02/28

XM\_024450233 6828 mRNA linear human 2021/02/28

…

**Zad. 9**

Przy pomocy narzędzi `esearch`, `efetch`, `xtract` i `sort` utwórz jedno polecenie, które wyszuka w bazie `gene` wszystkie geny o nazwie BRCA2 u naczelnych, tak aby wyświetlić poniższą listę (tj. identyfikator, nazwa genu, organizm) uszeregowaną ze względu na nazwę organizmu.

105726195 BRCA2 Aotus nancymaae

100397509 BRCA2 Callithrix jacchus

103267329 BRCA2 Carlito syrichta

108310783 BRCA2 Cebus imitator

105587897 BRCA2 Cercocebus atys

**Zad. 10**

Skorzystaj z polecenia `efetch` wyświetlające sekwencje FASTA o numerach dostępu: NP\_476567 i NP\_476565 (`efetch -h`).

**Zad. 11**

Wyświetl abstrakty artykułów bazy PubMed dotyczących schizofrenii i opublikowanych w ciągu ostatnich 30 dni. Podaj użyte polecenie. Wskazówka: Ograniczenie wyników ze względu na czas opublikowania umożliwi polecenie [efilter](https://www.ncbi.nlm.nih.gov/books/NBK179288/#chapter6.Searching_and_Filtering).

1. Ile artykułów znaleziono?

2. Podaj użyte polecenie.

**Zad. 12**

Korzystając z narzędzia [elink](https://www.ncbi.nlm.nih.gov/books/NBK179288/#chapter6.Writing_Commands_on_Multiple_Li) wyszukaj wszystkie sekwencje białkowe, o których mowa w artykułach o schizofrenii z ostatnich 30 dni. Podaj użyte polecenie oraz liczbę sekwencji.

**Zad. 13**

Podaj polecenie `efetch`, które wyświetli abstrakty trzech artykułów o identyfikatorach PubMed: 24102982, 21171099, 17150207.

**Zad. 14 (Python dla chętnych)**

Pobierz plik <http://www.combio.pl/files/vertebrates.txt>. Napisz skrypt, który dla każdego organizmu z pliku wyszuka (korzystając z Entrez Direct) sekwencje białkowe genu TNRC6A z bazy RefSeq. Znalezione sekwencje w obrębie organizmu powinny zostać zapisane w osobnym pliku tekstowym w formacie FASTA. Na przykład, sekwencje białkowe TNRC6A dla organizmu *Mus musculus* powinny zostać zapisane w pliku `mus\_musculus.fasta`. Uwzględnij w skrypcie sugestię NCBI, aby nie przekraczać trzech zapytań do bazy w ciągu 1 sekundy.