Laboratorium 5

Czas na oddanie finalnego raportu: 6 tygodni

Nazwa pliku: imie_nazwisko_5_bio.pdf

Typ ćwiczenia: jednotygodniowe

Cel: Celem ćwiczenia jest zapoznanie studentów z bazami danych wykorzystywanymi w bioinformatyce.

Zadania

Każdego roku w styczniu czasopismo naukowe NAR publikuje artykuł zawierający aktualną listę biologicznych baz danych. Znajdź ostatni artykuł i zrzut ekranu z pierwszej strony tej publikacji [**task 1**].

1. W treści artykułu znajdź odnośnik do wersji on-line tej bazy i podaj ten link [**task 2**].
2. W sekcji "International Nucleotide Sequence Database Collaboration" tej bazy danych znajdź rekord dotyczący bazy NCBI GenBank i odnajdź informację dotyczącą tego, jak często jest ona aktualizowana [**task 3**].

Zapoznaj się z publikacją: [Artykuł naukowy](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8022358).

1. W jednej z tabel znajdź bazę danych związaną z czerniakiem. Podaj jej nazwę [**task 4**].
2. W tabeli "Other databases" znajdź bazę HLA-ADR i z jej zasobów odczytaj dla ilu zdeponowanych alleli (wariantów genów lub innych odcinków DNA) aspiryna podnosi ryzyko chorób autoimmunologicznych, a dla ilu obniża [**task 5**]. Podpowiedź — risk (podnosi ryzyko), protective (obniża ryzyko); każdy wariant to oddzielny wiersz w tabeli; skorzystaj z danych widniejących pod linkiem "HLA and Adverse Drug Reaction report".

Wejdź do bazy [NCBI](https://www.ncbi.nlm.nih.gov/genbank/).

1. Znajdź wszystkie rekordy zdeponowane pod nazwiskiem "Kuzdraliński" i umieść tutaj listę gatunków mikroorganizmów, których one dotyczą (widoczne są w ramce po prawej stronie) [**task 6**].
2. Wyświetl tylko rekordy dotyczące drożdży z gatunku *Pichia kudriavzevii*, a następnie eksportuj je do jednego pliku w formacie FASTA posortowanego według długości sekwencji i w odpowiedzi do zadania umieść całą zawartość tego pliku [**task 7**].
3. Wróć do poprzedniego widoku i wejdź tylko do listy rekordów dla gatunku *Rhizopus stolonifer* (zdeponowanych przez tego samego autora), posortuj je w oknie przeglądarki według długości sekwencji i wejść do dowolnego rekordu o największej długości (jest ich kilka). W odpowiedzi do tego zadania podaj jego Accesion Number [**task 8**]. Następnie znajdź ostatniego autora tego rekordu w bazie Nauka polska (załóżmy hipotetycznie, że poszukujesz współpracy – baza OPI zawiera sporo ciekawych rekordów dotyczących polskiej nauki) i podaj dane bibliograficzne ostatniej publikacji tego autora z tej bazy [**task 9**]. Następnie za pomocą bazy danych np. *scholar.google.com* (lub innej) sprawdź, czy rzeczywiście jest to ostatnia publikacja tego autora, jeśli nie — podaj dane bibliograficzne jednej z ostatnich publikacji w formacie APA [**task 10**]. Tutaj nauczysz się, że baza OPI jest trochę nieaktualna, a dodatkowo – w jaki sposób pobierać cytowania z *scholar.google.com* (bardzo przydatna umiejętność do budowania profesjonalnych raportów).

W jednej z baz danych NCBI zdeponowany jest kompletny ludzki genom: ncbi.nlm.nih.gov.

1. Podaj, na ile obecnie ocenia się liczbę genów kodujących białka w ludzkim genomie (tabela, na przecięciu "GRCh38.p14" i "protein-coding") [**task 11**]. Posługując się wyszukiwarką Google, znajdź dane z około 1999-2001 roku na temat naszych wyobrażeń dotyczących liczby genów w ludzkim genomie [**task 12**]. To, co wiemy o świecie, szybko się zmienia dzięki odkryciom naukowym.

Jednostkami, w jakich zapisywana jest długość nici DNA to kilo par zasad – kpz (ang. kilo base – kb), mega par zasad – mpz (ang. mega base – Mb) i giga par zasad – gpz (ang. giga base – Gb).

- 1 kb to 1000 par zasad (nt).
- 1 Mb to 1 000 000 par zasad (nt).
- 1 Gb to 1 000 000 000 par zasad (nt).

Inne skróty i jednostki to:

- nt — nukleotyd
- bp — bas pair, czyli para zasad azotowych (oznacza się tak często nukleotyd).

Scaffold to pojęcie związane z sekwencjonowaniem DNA, które odnosi się do połączenia kilku krótkich sekwencji DNA w jedną dłuższą sekwencję. Scaffoldy są tworzone w procesie asemblacji sekwencji krótkich fragmentów DNA, które zostały uzyskane w procesie sekwencjonowania. Asemblacja jest procesem polegającym na scaleniu kilku krótkich sekwencji DNA w jedną dłuższą sekwencję, która będzie odpowiadać sekwencji genomu. Scaffoldy są zazwyczaj używane jako punkt wyjścia do dalszej analizy genomu.

W bazie danych NCBI posługując się wyszukiwarką znajdź rekord dla chromosomu 22 z GRCh38.p14 i podaj jakiej wielkości jest ten chromosom (podpowiedź: ponad 50 Mb; odrzuć wszystkie rekordy, które mają "scaffold" w tytule) [**task 13**].

1. Z rekordu tego przejdź następnie do bazy BioProject, tam kliknij w sekcji Project Data, w tabeli Assembly Details w kolumnie Chrs pozycję "25". Wygenerowamą listę uszereguj według wielkości (Sequence Length). W jakim porządku ustawione są ludzkie chromosomy [**task 14**]?
2. Przenieś listę do pliku "Summary", a jego treść umieść jako odpowiedź do tego zadania [**task 15**]

Afiliacja w publikacji naukowej jest to informacja o instytucji, w której autor publikacji jest zatrudniony lub związany zawodowo. Afiliacja jest zazwyczaj umieszczana na początku artykułu naukowego i wskazuje, w jakim instytucji autor był zaangażowany w prowadzenie badań i/lub pisanie publikacji. Instytucja ta może być uniwersytetem, instytutem badawczym, firmą lub jakimkolwiek innym miejscem, w którym autor prowadzi swoją działalność naukową lub zawodową. Afiliacja także może wskazywać na kraj, w którym instytucja jest zlokalizowana.

W bazie danych PubMed wyszukaj wszystkie publikacje afiliowane na Polskę. Podaj ich liczbę [**task 16**].

1. W polu "results by year" możesz eksportować plik `csv` z liczbą publikacji z każdego roku. Skopiuj jako odpowiedź do zadania dane od roku 2015 do roku 2022 z rozdzielczością co do roku [**task 22**].

2. Podaj liczbę publikacji afiliowanych na Polskę, które dotyczą nowotworu jelita ("intestinal cancer" w tytule) [**task 17**].

Zapoznaj się z informacjami dotyczącymi markera genetycznego typu SNP o oznaczeniu rs53576.

1. Znajdź rekord dotyczący tego markera genetycznego w tej bazie: NCBI SNP. Jaki jest MAF dla tego markera w populacji ludzkiej [**task 18**]? Co to jest MAF (Minor allele frequency) [**task 19**]?
2. Znajdź rekord tego markera w bazie danych ENSEMBL i odpowiedz na pytanie: w jakiej populacji (odnośnik: "Population genetics") allel związany z fenotypem predysponującym do bycia bardziej empatycznym występuje w największej frekwencji tj. 81% [**task 20**]? Znajdź również sekwencję DNA okalającą ten marker i umieść ją jako odpowiedz na zadanie (tzw. flanking sequence) [**task 21**].

NRRL (Northern Regional Research Laboratory) jest kolekcją mikroorganizmów, która jest jednym z największych źródeł kultur mikroorganizmów na świecie. Została założona w 1944 roku w Peoria w Illinois, w Stanach Zjednoczonych. NRRL jest częścią amerykańskiej USDA i zawiera ponad 100 000 kultur mikroorganizmów, w tym grzybów, bakterii, wirusów i prątków.

Inne znane kolekcje mikroorganizmów to:

- ATCC (American Type Culture Collection) - jedna z największych i najbardziej znanych na świecie kolekcji mikroorganizmów, która zawiera ponad 100 000 kultur mikroorganizmów z ponad 140 państw.
- DSMZ (Deutsche Sammlung von Mikroorganismen und Zellkulturen) - niemiecka kolekcja mikroorganizmów, która zawiera ponad 100 000 kultur mikroorganizmów z ponad 120 krajów.
- JCM (Japan Collection of Microorganisms) - japońska kolekcja mikroorganizmów, która zawiera ponad 40 000 kultur mikroorganizmów z ponad 50 krajów.

Kolekcje mikroorganizmów są często wykorzystywane do badań naukowych, w szczególności w dziedzinach takich jak biologia, medycyna, chemia i inżynieria genetyczna. Są one używane do identyfikacji, klasyfikacji i opracowywania nowych technik badawczych dla różnych mikroorganizmów. Kolekcje te stanowią również ważne źródło materiału do produkcji leków, szczepionek i innych produktów biotechnologicznych. Ponadto kolekcje mikroorganizmów służą jako referencyjne bazy danych dla identyfikacji mikroorganizmów i ich monitorowania w różnych środowiskach.

W bazie danych EBI znajdź rekord dla całego genomu mikroorganizmu NRRL 26941 (szczep o numerze 26941 pochodzący z kolekcji NRRL). Co to za gatunek [**task 22**]? Jakiej wielkości posiada genom [**task 23**]? Ile wynosi Contig N50 i co oznacza ten parametr [**task 24**].

Enzym trawienny to białko, które jest niezbędne do rozkładania złożonych związków organicznych, takich jak białka, węglowodany i tłuszcze, w prostsze związki, które nasz organizm może wykorzystać jako źródło energii i składników odżywczych.

Niektóre przykłady ludzkich enzymów trawiennych to:

- Pepsyna — enzym występujący w soku żołądkowym, który rozkłada białka.
- Amylaza — enzym, który rozkłada skrobię na prostsze związki, takie jak glukoza.
- Lipaza — enzym, który rozkłada tłuszcze na kwasy tłuszczowe i glicerol.
- Laktaza — enzym, który rozkłada laktozę (cukier mleczny) na glukozę i galaktozę.

Znajdź w bazie danych UNIPROT rekord reprezentujący dowolny ludzki enzym trawienny (wybierz

któryś). W odpowiedzi do zadania wpisz nazwę enzymu i numer rekordu [**task 25**].

1. Wypisz kofaktory wybranego przez ciebie enzymu i napisz co to jest kofaktor dla enzymu [**task 26**]. Jeśli wybrany enzym nie posiada kofaktorów, wyszukaj inny, który posiada.
2. Z jakich baz danych pochodzi predykcja struktury tego białka [**task 27**]?

Numer EC (Enzyme Commission) jest systematycznym identyfikatorem białek w klasyfikacji biblioteki Enzyme. Każdy białko jest przypisane do określonej klasy, podklasy i typu reakcji według kodu EC, który składa się z kilku liczb oddzielonych kropkami. Na przykład, EC 3.2.1.20 oznacza trzecią klasę enzymów (hydrolazy), podklasę 2 (hydrolazy zasadowe), typ 1 (hydrolazy zasadowe złożone z dwóch podjednostek) i 20-ty enzym w tej podklasie.

W biologii molekularnej termin "ligand" jest używany do opisania cząsteczki, która wiąże się z białkiem za pomocą interakcji chemicznych, takich jak wiązania jonowe, wiązania wodorowe lub wiązania kowalencyjne. Te interakcje powodują, że ligand i białko tworzą kompleks, który może modyfikować funkcję białka. Wiele białek w organizmach żywych pełni funkcje receptorów, które są zaprojektowane do wchodzenia w interakcje z ligandami, takimi jak hormony, czynniki wzrostu i neurotransmitery.

W bazie danych RCSB wyszukaj rekordy, które posiadają numer EC ten sam co enzym, który wytypowałeś w poleceniu powyżej. Dlaczego wyszukiwarka pokazuje rekordy również z innych gatunków niż człowiek [**task 28**]?

W bazie danych Protein Atlas przejdź do odnośnika "Brain".

1. Jaka liczba genów ulega ekspresji tylko w ludzkim mózgu [**task 29**]?
2. Kliknij liczbę genów ulegających ekspresji tylko w mózgu, a następnie znajdź tam rekord dla genu DRD3. Co to jest za białko [**task 30**]? Jakie markery genetyczne z nim związane możesz znaleźć (np. tutaj: https://www.snpedia.com/). Podaj przynajmniej jeden przykład i powiązany z nim fenotyp [**task 31**].

W bazie danych NCBI VIRUSES znajdź sekwencję referencyjną genomu wirusa SARS-CoV-2 oznaczoną jako Wuhan-Hu-1. Jaką ma wielkość i z jakiego rodzaju kw. nukleinowego jest zbudowany ten genom [**task 32**]?

1. Znajdź datę pozyskania tej próbki (collection date) [**task 33**]
2. Znajdź i skopiuj sekwencję aminokwasową słynnego białka kolca, w które celowane są szczepionki [**task 34**]