Ćwiczenie 1 - Biologiczne bazy danych i formaty zapisu rekordów

---

W tym ćwiczeniu:
1) dowiesz się jak uzyskać informacje na temat genów i ich produktów na przykładzie ludzkiego genu kodującego białko prionowe. Poznasz jego:

- kontekst genomowy, kodowane białka
- pełnione funkcje kodowanego białka
- występowanie genu u innych organizmów
- strukturę 3D kodowanego białka

2) zapoznasz się z podstawowymi formatami danych wykorzystywanymi przez najpopularniejsze bazy biologiczne znajdujące się na serwerach NCBI i EMBL. Nauczysz się:

- rozpoznawać poszczególne formaty rekordów
- wyszukiwać potrzebne informacje w rekordach
- krytycznej oceny zawartości danych

Słowa kluczowe: egzon, intron, splicing, ontologia genów, geny homologiczne, baza RefSeq, baza PDB, format EMBL, FASTA, format GenBank, numer dostępu, identyfikator GI, baza UniProt, apolipoproteina E (APOE), choroba Alzheimer'a, izoforma, alternatywny splicing

---

Pytanie 1. Czym są priony?

Przeszukaj serwis NCBI (http://www.ncbi.nlm.nih.gov/) zapytaniem 'prion protein'. Z bazy *Gene* wybierz gen prionowy o nazwie PRNP.
Pytanie 2. W którym miejscu w komórce zlokalizowane jest białko kodowane przez ten gen?
Pytanie 3. Jakie są inne nazwy tego genu?
Pytanie 4. Na którym chromosomie znajduje się ten gen?
Pytanie 5. Jaka jest długość sekwencji tego genu na chromosomie?
Pytanie 6. Jaka jest struktura tego genu: z ilu egzonów i intronów składa się ten gen?
Pytanie 7. Ile wariantów splicingowych odnotowano dla tego genu?
Pytanie 8. Ile artykułów naukowych opisuje ten gen?

Korzystając z panelu pokrewnych informacji (*Related information*) znajdź geny sąsiadujące (*Gene neighbors*) z genem PRNP na chromosomie.
Pytanie 9. Czy w otoczeniu genu PRNP znajdują się inne geny prionowe?

Wróc do rekordu genu PRNP. Korzystając z *Related information* wyszukaj geny homologiczne (*HomoloGene*) białka PRNP u innych organizmów. Następnie wyświetl informacje na temat podobieństwa znalezionych sekwencji (*Display - Alignment score*).
Pytanie 10. Ile wynosi procent identyczności sekwencji między dwoma białkami PRNP: myszy i człowieka?

Pytanie 11. Dlaczego identyczność jest wyższa między sekwencjami białkowymi, a nie DNA?

Wróć do rekordu genu PRNP.
Pytanie 12. W panelu *Phenotypes* odszukaj dwie choroby związane z mutacjami w tym białku.

Wróć do rekordu genu PRNP. W celu znalezienia białek kodowanych przez ten gen, z panelu *Related information* wybierz *Protein.*
Pytanie 13. Ile białkowych sekwencji dostępnych jest dla tego genu?

Wróć do rekordu PRNP. W celu wyświetlenia struktur 3D białka prionowego z panelu *Related Information* wybierz *3D structures.* Następnie wybierz struktury wyznaczone metodą NMR. Spośród nich wybierz strukturę patogennego białka prionowego *Pathogenic Mutant (D178n)* u człowieka. Następnie przejdź do rekordu tej struktury w bazie PDB i zapisz jego numer dostępu. Odszukaj w rekordzie numer dostępu struktury niepatogennej *(Wild type)*. Aby porównać struktury obu białek skorzystaj z menu *Analyze* > *Sequence & Structure Alignment.* Podaj numery dostępu obu struktur I wybierz metodę przyrównania *jFATCAT rigid*, naciśnij *Align*.

Pytanie 14. Jakie są różnice w sekwencji pomiędzy dwoma białkami?
Pytanie 15. Jakie są różnice w strukturze pomiędzy dwoma białkami?

Formaty zapisu rekordów w bazach danych

GenBank

Format GenBank jest charakterystycznym formatem rekordów sekwencji pochodzących z bazy o tej samej nazwie znajdującej się na serwerach NCBI. Pełne objaśnienie formatu znajdziesz http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html

Zapoznaj się z rekordem znajdującym się pod adresem http://www.ncbi.nlm.nih.gov/nuccore/48762938.

Pytanie 16. Jaki jest numer dostępu (Accession number), numer wersji (version) oraz identyfikator GI? Które z powyższych informacji dowodzą, że rekord był modyfikowany i w jaki sposób można to rozpoznać?
Pytanie 17. Z jakiego organizmu pochodzi rekord?
Pytanie 18. Jaka jest pełna nazwa genu znajdującego się pod tym rekordem, jakiej jest długości i na jakim chromosomie jest zlokalizowany? Podaj nazwy pól, w których znalazłeś tą informację.

Poszukiwana informacja	Odpowiedź	Nazwa pola*
Pełna nazwa genu		

Długość genu		
Chromosom		

*główna nazwa napisana w rekordzie dużymi literami

Pytanie 19. Podaj miejsce początku i końca sekwencji kodującej (CDS) w obrębie tego genu?
Pytanie 20. Podaj numer dostępu sekwencji białkowej (protein_id) kodowanej przez ten gen oraz jej identyfikator GI. Czy jest taki sam jak GI sekwencji nukleotydowej?

Za pomocą numeru dostępu znalezionego w pytaniu 20 przejdź do rekordu białkowego.
Pytanie 21. Porównaj numery dostępu sekwencji nukleotydowej i białkowej. Podaj ich wspólne cechy.
Pytanie 22. Z jakiej bazy danych pochodzą oba rekordy?
Pytanie 23. W rekordzie białkowym odszukaj i podaj długość sekwencji białkowej.
Pytanie 24. Co znajduje się w pozycji 104..299 sekwencji białkowej?

ENA

Format ENA jest źródłowym formatem charakterystycznym dla bazy o tej samej nazwie.

Zapoznaj się z analogicznym rekordem dotyczącym APOE z GenBank znajdującym się pod adresem http://www.ebi.ac.uk/ena/data/view/K00396. Po wczytaniu rekordu wyświetl go jako zwykły tekst (View: text).

Pytanie 25. Porównaj datę ostatniej modyfikacji oraz długość sekwencji z informacjami uzyskanymi z poprzedniej bazy. Jak oceniasz współpracę między obiema bazami?

UNIPROT

Przejdź do rekordu białkowego z bazy UniProtKB/Swiss-Prot pod adresem www.uniprot.org/uniprot/P02649.
Pytanie 26. Podaj nazwę rekordu i w bazie danych UniProtKB/Swiss-Prot.
Pytanie 27. Podaj daty utworzenia i ostatniej aktualizacji rekordu. Który z rekordów białkowych – z RefSeq czy UniProtKB/Swiss-Prot jest bardziej aktualny?

FASTA
Pozostając w bazie UniProt, wyświetl rekord jako FASTA.

Pytanie 28. Podaj dwie kluczowe cechy charakteryzujące ten format danych? Jakie informacje znajdują się teraz w rekordzie?

Pytanie 29. Czy format FASTA jest charakterystyczny tylko dla sekwencji białkowych? W oparciu o rekord nukleotydowy z GenBank, znajdź argument, który podtrzyma Twoje zdanie.