BIOINFORMATYKA - STUDIA PODYPLOMOWE

Adrian Kania

NCBI

Baza danych PubMed zawiera informacje dotyczące artykułów z zakresu nauk biologicznych i medycznych. Wyszukiwanie może odbywać się poprzez przeglądarkę internetową.

Zadanie1 Odszukaj pracę o identyfikatorze *14697198* w bazie danych *PubMed (*PubMed (nih.gov)*,* a następnie odpowiedz na pytania:

- jaki jest tytuł tej pracy?
- w jakim wydawnictwie została wydana ta praca?
- ilu autorów odpowiada za tę prace?
- podaj przykładowe MeSH termy z tej pracy.

Zadanie2 Zaproponuj hasło do wyszukania prac posiadających w tytule *mRNA*, opublikowanych w *2017* roku w czasopiśmie *BMC Genomics*. Ile jest takich prac?

Dopasowanie sekwencji

Zadanie3 Z bazy danych *Nucleotide* pobierz sekwencje o następujących identyfikatorach:

- HM370969.1

- CY138562.1

- HQ185381.1

- JX081142.1

- HQ185383.1

- Jakie białko kodują te sekwencje?
- Skopiuj sekwencje HM370969.1 oraz HQ185383.1 do programu dotmatcher (EMBOSS: dotmatcher (bioinformatics.nl)) aby utworzyć macierz kropkową. Skomentuj otrzymany wynik.

- Zapisz je w jednym pliku w formacie FASTA, a następnie dokonaj ich porównania przez Clustal Omega (Clustal Omega < Multiple Sequence Alignment < EMBL-EBI).
- Obejrzyj jak wygląda otrzymane dopasowanie oraz skopiuj proponowane drzewo filogenetyczne charakteryzujące te sekwencje.
- Czy jest ono kompatybilne z typami białek podanych w nawiasach?

BLAST

Najważniejszym zastosowaniem przyrównywania sekwencji parami jest wyszukiwanie sekwencji biologicznych w bazach danych na podstawie podobieństwa. Wymaga to wysłania badanej sekwencji jako zapytania i przyrównania jej parami z wszystkimi sekwencjami w bazie.

Zadanie4 Poniżej zamieszczono sekwencje aminokwasową:

MKLKNTLGVVIGSLVAASAMNAFAQGQNSVEIEAFGKRYFTDSVRNMKNADLYGGSIGYFLTD
DVELALSYGEYHDVRGTYETGNKKVHGNLTSLDAIYHFGTPGVGLRPYVSAGLAHQNITNINSD
SQGRQQMTMANIGAGLKYYFTENFFAKASLDGQYGLEKRDNGHQGEWMAGLGVGFNFGGSK
AAPAPEPVADVCSDSDNDGVCDNVDKCPDTPANVTVDANGCPAVAEVVRVQLDVKFDFDKSK
VKENSYADIKNLADFMKQYPSTSTTVEGHTDSVGTDAYNQKLSERRANAVRDVLVNEYGVEGG
RVNAVGYGESRPVADNATAEGRAINRRVEAEVEAEAK

Korzystając z *BLASTP* (BLAST: Basic Local Alignment Search Tool (nih.gov)) odpowiedz na poniższe zapytania:

- Co to za białko? (na podstawie nazwy (adnotacji)).
- Z jakiego organizmu pochodzi to białko?

Dodatkowo, z użyciem bazy danych *UniProt* (UniProt) podaj funkcje molekularne tego białka.

Zadanie5 Dana jest sekwencja nukleotydowa:

`GTCCTTCATAGCCTAACCTGTTACCACTAGATTACCCACCGGCCGTTCCTACCCGCTCTCACCAGCCC`
`TC`

Korzystając z *BLASTN* (BLAST: Basic Local Alignment Search Tool (nih.gov)):

- Zidentyfikuj z jakiego organizmu pochodzi ta sekwencja i częścią jakiego genu jest.
- Ile wyników zwrócono?
- Ile wynosi E-value najlepszego wyniku?

Następnie wykonaj analogiczne wyszukiwanie dla `GTCCTTCATAGCCTAACCTGTTA` (czyli początkowego fragmentu poprzedniej sekwencji). Ile wynosi E-value dla wyniku z poprzedniego etapu?

Dane mikromacierzowe

GeneMania

Serwis ten pozwala na zbiorczą analizę grupy genów. Tworzona jest sieć połączeń ze względu na takie cechy jak: genetyczne interakcje, fizyczne interakcje, współdzielone domeny, kolokalizację czy koekspresję.

Zadanie 6 Wejdź na http://genemania.org/ a następnie przeanalizuj zestaw potencjalnych genów markerowych dla prognozy raka piersi.

EFNA1

EGFR

ERBB2

GATA3

GZMB

MST1

MYB

MYBL2

MYC

PLAT

SOX4

SOX9

SRF

XBP1

- Który gen jest połączony z *GATA3* jeżeli chodzi o kolokalizację?
- Który gen jest połączony z *XBP1* jeżeli chodzi o genetyczną interakcję?
- Jaka występuje najbardziej znacząca funkcja w tej grupie genów (tzn. taka która ma najmniejszy FDR).

NCBI (GEO DataSets) udostępnia dane z eksperymentów mikromacierzowych a także pozwala na ich analizę online. W tym kroku poddamy analizie wybrany taki eksperyment.

Zadanie7 Przypadek do analizy z NCBI GEO DataSets

- Wejdź w link poniżej: https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS810
- Czego dotyczyło badanie?
- Gdzie badano ekspresje genów? (jaki materiał/tkanka)
- Ile było wszystkich próbek i na ile grup były one podzielone ***(Experiment design and value distribution oraz Sample Subsets)***
- Według jakich kryteriów podzielono próbki na grupy? Opisz wykorzystane parametry. ***(Experiment design and value distribution oraz Sample Subsets)***
- Jak wygląda przebieg ekspresji dla genów *SPARC, VSNL1* oraz *COL5A2* w kolejnych grupach? ***(Expression Profiles)*** Za co odpowiadają te geny? Czy obserwujesz jakieś tendencje zmiany poziomu ich ekspresji w kolejnych grupach? Poszukaj w źródłach zewnętrznych informacji na temat ich związku z chorobą Alzheimera.
- Czym są *housekeeping genes?* Jaką pełnią rolę w eksperymencie mikromacierzowym? Wybierz trzy przykładowe geny tej kategorii i sprawdź ich ekspresje w kolejnych próbkach.