**DotPlot, BLAST**

**Zad. 1**

Korzystając z techniki dot-plot porównaj sześć par sekwencji DNA w pliku <http://www.combio.pl/files/dotplot.xlsx>. Przyporządkuj poniższe obserwacje do sześciu otrzymanych wykresów dot-plot (a-f):

1. Zgodność sekwencji,
2. Niezgodność sekwencji
3. Insercja/delecja
4. Sekwencja palindromowa
5. Tandemowe powtórzenia
6. Sekwencje powtórzone

Wypełniony arkusz dołącz do sprawozdania.

**Zad. 2**

Poniżej znajduje się fragment sekwencji mRNA genu insuliny gryzonia koszatniczki pospolitej (*Octodon degus*).

>insulin Octodon degus insulin (Ins), mRNA

TGAGGCATTCTCTAACAGGTTCTCGACCCTCCGCCATGGCCCCGTGGATGCATCTCCTCACCGTGCTGGC

CCTGCTGGCCCTCTGGGGACCCAACTCTGTTCAGGCCTATTCCAGCCAGCACCTGTGCGGCTCCAACCTA

TCTGCAGAAGCGCGGCATTGTGGATCAGTGCTGTAATAACATTTGCACATTTAACCAGCTGCAGAACTAC

TGCAATGTCCCTTAGACACCTGCCTTGGGCCTGGCCTGCTGCTCTGCCCTGGCAACCAATAAACCCCTTG

AATGAG

Ze strony serwisu [NCBI](https://www.ncbi.nlm.nih.gov/) otwórz stronę programu BLAST. Wybierz program `Nucleotide BLAST`. W formularzu, w polu `Enter Query Sequence` umieść powyższą sekwencję w formacie FASTA. Użyj następujących ustawień:

- W polu `Database` wybierz bazę `Reference RNA sequences (refseq\_rna)`.

- W panelu `Program Selection` wybierz `Somewhat similar sequences (blastn)`.

Z listy otrzymanych trafień (panel `Descriptions`) zwróć uwagę na sekwencję, która uzyskała najwyższą wartość punktacji (`Max score`).

1. Podaj numer dostępu sekwencji najbardziej podobnej do sekwencji zapytania.
2. Czy znaleziona sekwencja jest identyczna do sekwencji zapytania?
3. Ile lokalnych przyrównań sekwencji wyznaczył BLAST między sekwencją zapytania a ` XM\_004627084.1`?
4. O czym mówią parametry `Max score` i `Total score`? Wskazówka: [NCBI](https://www.ncbi.nlm.nih.gov/Web/Newsltr/V15N2/BLView.html).
5. Ile wynosi procent identyczności między sekwencją zapytania a `XM\_004627084.1`?
6. le wynosi wartość `Query cover`?
7. O czym informuje parametr `Query cover`. Wskazówka: [NCBI](https://www.ncbi.nlm.nih.gov/Web/Newsltr/V15N2/BLView.html).
8. Ile wynosi wartość `E-value`?
9. O czym informuje E-value?
10. Ile przerw wprowadzono w tym przyrównaniu?
11. W zakładce `Search Summary` znajdują się informacje na temat parametrów i bazy danych użytych w tym przeszukiwaniu BLAST. Ile sekwencji znajduje się w bazie danych, która została przeszukana?

**Zad. 3**

Poniżej znajduje się sekwencja białkowa genu FOXP2 człowieka.

>NP\_055306.1 forkhead box protein P2 isoform I [Homo sapiens]

MMQESATETISNSSMNQNGMSTLSSQLDAGSRDGRSSGDTSSEVSTVELLHLQQQQALQAARQLLLQQQT

SGLKSPKSSDKQRPLQVPVSVAMMTPQVITPQQMQQILQQQVLSPQQLQALLQQQQAVMLQQQQLQEFYK

KQQEQLHLQLLQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQHPGKQAKEQQQQQQQQQQL

AAQQLVFQQQLLQMQQLQQQQHLLSLQRQGLISIPPGQAALPVQSLPQAGLSPAEIQQLWKEVTGVHSME

DNGIKHGGLDLTTNNSSSTTSSNTSKASPPITHHSIVNGQSSVLSARRDSSSHEETGASHTLYGHGVCKW

PGCESICEDFGQFLKHLNNEHALDDRSTAQCRVQMQVVQQLEIQLSKERERLQAMMTHLHMRPSEPKPSP

KPLNLVSSVTMSKNMLETSPQSLPQTPTTPTAPVTPITQGPSVITPASVPNVGAIRRRHSDKYNIPMSSE

IAPNYEFYKNADVRPPFTYATLIRQAIMESSDRQLTLNEIYSWFTRTFAYFRRNAATWKNAVRHNLSLHK

CFVRVENVKGAVWTVDEVEYQKRRSQKITGSPTLVKNIPTSLGYGAALNASLQAALAESSLPLLSNPGLI

NNASSGLLQAVHEDLNGSLDHIDSNGNSSPGCSPQPHIHSIHVKEEPVIAEDEDCPMSLVTTANHSPELE

DDREIEEEPLSEDLE

Użyj serwisu NCBI BLAST (`Protein BLAST`) w celu przeszukania bazy danych RefSeq, ograniczając wyszukiwanie do sekwencji zwierząt (*Metazoa*) i wykluczając z nich sekwencje pochodzące z naczelnych (*Primates*).

Z listy otrzymanych trafień wybierz jedną sekwencję, która najbardziej odpowiada sekwencji FOXP2.

1. Z jakiego organizmu pochodzi ta sekwencja?
2. Ile wynosi `E-value` tego dopasowania?
3. Podaj procent identyczności i podobieństwa tego dopasowania.

**Lokalny program NCBI BLAST**

Program NCBI BLAST można zainstalować na Windows, Linux i MacOS ([pomoc](https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download)).

**Zad. 4 - blastn**

W pliku <http://www.combio.pl/files/mito_genes.fasta> znajdują się sekwencje trzech genów mitochondrialnego DNA człowieka (COX1, ND6, tRNA-Pro). W pliku <http://www.combio.pl/files/mito_genomes.fasta> znajdują się sekwencje całych genomów mitochondrialnych pochodzące z różnych organizmów (np.: mysz, szympans). Zapisz oba pliki na dysku i umieść je w jednym katalogu. Twoim zadaniem jest użycie lokalnej wersji programu BLAST w celu zidentyfikowania lokalizacji trzech genów w sekwencjach genomowych.

Przygotowanie bazy sekwencji nukleotydowych:

makeblastdb -in mito\_genomes.fasta -dbtype nucl

Uruchomienie programu blastn:

blastn -query mito\_genes.fasta -db mito\_genomes.fasta

blastn -query mito\_genes.fasta -db mito\_genomes.fasta -out results.txt

1. W których genomach mitochondrialnych występuje gen tRNA-Pro?
2. Czy sekwencja tego genu jest identyczna we wszystkich genomach?
3. Podaj lokalizację tego genu w genomie mitochondrialnym człowieka.
4. Sprawdź jakie parametry może przyjmować program blastn (`blastn –help`). Wykonaj ponowne przeszukiwanie, tym razem wyświetlając wyniki w formie tabeli (format tabularny z komentarzami). Podaj numery kolumn, w których znajdują się `E-value` i `score`.
5. Podaj pozycję startu i końca genu tRNA-Pro w sekwencji szympansa.
6. Zmodyfikuj poprzednie polecenie, aby wyświetlić wyniki w formacie tabularnym bez komentarzy.
7. Zmodyfikuj poprzednie polecenie zmieniając wartość parametru `task` z `megablast` na `blastn`. Czy w wyniku otrzymano mniej, czy więcej wyników?
8. Do polecenia z pkt. 7 dodaj odpowiednią opcję, aby wyświetlić przyrównania o wartość E-value <= 1e-05.
9. Użyj odpowiedniego polecenie Linuxa, aby posortować otrzymany wynik, aby w obrębie każdego organizmu (genomu) trafienia były uszeregowane zgodnie z ich lokalizacją w genomie.
10. Użyj odpowiedniego polecenia Linuxa, aby odpowiedzieć na pytanie ile trafień znalazł program BLAST w obrębie każdego organizmu.
11. Do polecenia z pkt. 10 dodaj kolejny potok, aby uszeregować wynik ze względu na malejącą liczbę trafień i nazwę organizmu.

**Zad. 5 - blastp**

W pliku <http://www.combio.pl/files/yeast_query.fasta> znajduje się 10 sekwencji białkowych pochodzących z drożdży piekarniczych (*Saccharomyces cerevisiae*), z kolei w pliku <http://www.combio.pl/files/spombe.fasta> znajdują się wszystkie sekwencje białek drożdży *Schizosaccharomyces pombe*.

Przygotowanie bazy sekwencji białkowych:

makeblastdb -in spombe.fasta -dbtype prot

Uruchomienie programu blastp:

blastp -query yeast\_query.fasta -db spombe.fasta -outfmt 7

1. Czy w wynikach znaleziono sekwencje podobne dla wszystkich 10 sekwencji zapytania?
2. Uruchom ponownie blastp ograniczając przyrównania do evalue <= 1e-05. Dla ilu sekwencji zapytania znaleziono sekwencje podobne?
3. Podaj numer dostępu sekwencji *S. pombe*, która uzyskała najwyższą wartość punktacji w przyrównaniu z sekwencją zapytania `sp|P11484|SSB1\_YEAST`.
4. Ile wynosi E-value tego przyrównania?
5. Ile wynosi procent identyczności?
6. Ile wynosi długość przyrównania?
7. Na ilu pozycjach w przyrównaniu aminokwasy są niezgodne (*mismatch*)?
8. Ile przerw występuje w przyrównaniu?

**Zad. 6**

Utwórz skrypt, który wczyta dwie sekwencje DNA w formacie FASTA (każda w osobnym pliku) i wykona prostą analizę typu dotplot (wielkość okna = 1, wartość graniczna = 1).

Format pliku wynikowego:

 A C G G A T A

A X . . . X . X

C . X . . . . .

G . . X X . . .

C . X . . . . .

A X . . . X . X

G . . X X . . .

T . . . . . X .

A X . . . X . X

**Zad. 7\* (Python dla chętnych)**

W pliku <http://www.combio.pl/files/ReAV.fasta> znajduje się białkowa sekwencja zapytania, a pod adresem <http://www.combio.pl/files/genomes.tar.gz> znajduje się 5 plików, w których każdy zawiera zestaw wszystkich białek danego gatunku roślin. Napisz skrypt, który uruchomi wyszukiwanie BLAST pomiędzy sekwencją zapytania a każdym z pięciu plików. Skrypt powinien również wywołać komendę makeblastdb dla każdego z 5 plików. W wyniku skrypt powinien zwrócić - dla każdego z pięciu plików - identyfikator sekwencji (wraz z wartością score i E-value), która wykazuje najwyższą punktację do sekwencji zapytania.