Bioinformatyka – Laboratorium nr 3
Podstawowe bioinformatyczne formaty plików.

Format **FASTA** – format zapisu sekwencji kwasów nukleinowych oraz białek używanym w bioinformatyce. Nukleotydy (dla DNA i RNA) oraz aminokwasy (dla białek) oznaczone są jednoliterowymi skrótami. Format FASTA uwzględnia również możliwość dodawania opisów i komentarzy do sekwencji.

Np.

```
>Keratyna 5, ekson 2, Homo sapiens
GTGCGGTTCCTGGAGCAGCAGAACAAGGTTCTGGACACCAAGTGGACCCTGCTGCAGGAG
CAGGGCACCAAGACTGTGAGGCAGAACCTGGAGCCGTTGTTCGAGCAGTACATCAACAAC
CTCAGGAGGCAGCTGGACAGCATCGTGGGGGAACGGGGCCGCCTGGACTCAGAGCTGAGA
AACATGCAGGACCTGGTGGAAGACTTCAAGAACAA

>Keratyna 5, ekson 6, Homo sapiens
CTTCTGCAGCCCTACGCAGAAGCGGCGGAAAACCTCCTTCATGTTGCCCCCCTTCTCCATGGAGATGAC
CCGGAGGTGATCCTCCTCGTTCACCCACACCAGGAAGCTCTTGTTGTCATTGTGC
```

GenBank format (GenBank Flat File Format)

Format ten oprócz sekwencji (zaczynającej się od słowa ORIGIN a kończącej znakami //) zawiera także sekcje opisującą dane fragmenty sekwencji (sekcja LOCUS) oraz informacje o genach, regionach o zidentyfikowanym znaczeniu biologicznym, sekwencje kodujące białka i wiele innych (sekcja FEATURES)
Przykład:

Example

```
LOCUS       AF068625                 200 bp    mRNA    linear
            ROD 06- DEC-1999
DEFINITION  Mus musculus DNA cytosine-5 methyltransferase 3A
(Dnmt3a) mRNA,
            complete cds.
ACCESSION   AF068625 REGION:
1..200 VERSION  AF068625.2
GI:6449467 KEYWORDS   .
SOURCE      Mus musculus (house
  mouse) ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Mus.
REFERENCE   1  (bases 1 to 200)
  AUTHORS   Okano,M., Xie,S. and
  Li,E.
  TITLE     Cloning and characterization of a family of novel
mammalian DNA(cytosine-5) methyltransferases JOURNAL Nat. Genet. 19
(3), 219-220 (1998)
   PUBMED   9662389
REFERENCE   2  (bases 1 to 200)
  AUTHORS   Xie,S., Okano,M. and
  Li,E. TITLE   Direct Submission
  JOURNAL   Submitted (28-MAY-1998) CVRC, Mass. Gen. Hospital, 149
13th Street,
            Charlestown, MA
02129, USA REFERENCE  3  (bases 1
to 200)
  AUTHORS   Okano,M., Chijiwa,T., Sasaki,H. and
  Li,E. TITLE   Direct Submission
```

```
  JOURNAL   Submitted (04-NOV-1999) CVRC, Mass. Gen. Hospital, 149
13th Street,
            Charlestown, MA 02129, USA
  REMARK    Sequence update by
  submitter
COMMENT     On Nov 18, 1999 this sequence version replaced gi:3327977.
FEATURES
                     Location/Qualifier
     s source        1..200
                     /organism="Mus musculus"
                     /mol_type="mRNA"
                     /db_xref="taxon:10090"
                     /chromosome="12"
                     /map="4.0 cM"
     gene            1..>200
                     /gene="Dnmt3a"
ORIGIN
        1 gaattccggc ctgctgccgg gccgcccgac ccgccgggcc acacggcaga
gccgcctgaa
       61 gcccagcgct gaggctgcac ttttccgagg gcttgacatc agggtctatg
tttaagtctt
      121 agctcttgct tacaaagacc acggcaattc cttctctgaa gccctcgcag
ccccacagcg
      181 ccctcgcagc cccagcctgc
//
```

FASTQ

Format do przechowywania danych z sekwencjonowania - 4 linie na fragment

- Linia zaczyna się od znaku @, następnie identyfikator sekwencji i opcjonalnie opis
- Linia zawiera wyłącznie sekwencję
- Linia zaczyna się od znaku + i może opcjonalnie zawierać identyfikator i opis
- Linia zawiera symbolicznie zapisaną ocenę jakości dla każdego nukleotydu. Musi zawierać dokładnie tyle samo znaków co sekwencja.

Np.

@M01939:25:000000000-AC3BN:1:1101:12876:1706 2:N:0:9

CAAAAACAATAAAAAAGTCCATACCCCAAGGATACAGCTTTCCTTAGTTTTTTATAGAGGGTTTTGCT
AGACCCTCTCCCGCAGTTTCAAGCGGATTTTATTTCCATTTTCCTTCTCCATCTTATCACAACTCGCACT
TTTCGCC

+

To jest 2 linia (tylko się nie zmieściła)

6,8<FFDGGA@99+8,=C,C9FC@C,,,,,,;,C,,;CCEAFFF9,<@CCFC,C9,,,,,8BCD,9@,,,,:64B6BEB7:++8
A?@,,,54++8@E<,CAFF9,@FFFDAFCF,9=,@@DE,@
>,=,9+9=+6+6+4@===@?

To jest 4 linia

@M01939:25:000000000-AC3BN:1:1101:8626:1712 2:N:0:9

CTAATTTTTGTTGATTGTTTATTTCCATATTTGCTTGTTTATGGAGTGCAGTAGCGTTATGATACATCAC
AAATTGATTACAATAACAACATACAAACCTTTGTTGTAATTTGAAACAATGGCTAACCCTATTCGTAAT

+

- 8,@CFDGGGGGG,CECEE,,CEFEF,C,CCF9@EC,CEE,C,,,,;,;,,C@,;+BF,B,,C,B,C,,;,,,,;,,C,,,,,:,,:,,:<F,=,
9,,9FEC,AE8@E,44C9,,>AE,C,49>,4>=BE,=;,,94,@

SAM/BAM

• SequenceAlignment Map – format tekstowy oryginalnie zaprojektowany do przechowywania danych dopasowania sekwencji do sekwencji referencyjne. Szeroko używany w sekwencjonowaniu nowej generacji. Może zawierać wywołane zasady, jakość dopasowania, ilość niedopasowanych nukleotydów oraz sekwencje unikalne.

BAM jest wersją binarną SAM

MOL/MOL2

Najbardziej typowy format przechowywania informacji o cząsteczkach chemicznych i reakcjach, zawierający informacje o atomach, wiązaniach, współrzędnych i wielu innych. Jest wspierany przez większość programów bioinformatycznych, informatyki chemicznej i innych. Oparty o otwarty standard CT File (chemical table file).

PDB

Format pdb to oryginalny format banku (wyjaśnienie poniżej). Przewodnik po tym formacie był kilkakrotnie poprawiany aktualnie wersja 5.1 z września 2022 (http://www.wwpdb.org/documentation/biocuration).

Baza **danych białek** lub **BDP** the *Research Collaboratory for Structural Bioinformatics* , bardziej znana jako ***Protein Data Bank*** lub **PDB.** Każdy model jest oznaczony w banku unikalnym identyfikatorem składającym się z 4 znaków, pierwszy to zawsze znak numeryczny, a kolejne trzy to znaki alfanumeryczne. Ten identyfikator nazywa się „kodem pdb".

Plik zawierają współrzędne kartezjańskie atomów, bibliografię, informacje strukturalne, czynniki struktury krystalograficznej oraz eksperymentalne dane NMR. Pierwotnie format pdb był podyktowany użyciem i szerokością dziurkowanych kart komputerowych. W efekcie każdy wiersz zawiera dokładnie 80 znaków.

Plik w formacie pdb to plik tekstowy, w którym każda kolumna ma swoje znaczenie: każdy parametr jest niezmienny. Zatem pierwsze 6 kolumn, czyli pierwsze 6 znaków dla danego wiersza, określa pole pliku. Znajdujemy na przykład pola „TITLE_" (czyli tytuł badanej makrocząsteczki), „KEYWDS" (słowa kluczowe hasła), „EXPDTA", które dostarczają informacji o zastosowanej metodzie eksperymentalnej, „SEQRES" (słowo kluczowe hasła). sekwencja badanego białka), „ATOM__" lub „HETATM", pola zawierające wszystkie informacje

związane z danym atomem. Ostatni przykład, w tych ostatnich polach nazwa atomu jest opisana w kolumnach od 13 do 16 (tj. od trzynastego do szesnastego znaku wiersza).

Linie „ATOM__” dotyczą aminokwasów lub kwasów nukleinowych, a linie „HETATM” dedykowane są innym cząsteczkom (rozpuszczalnik, substrat, jon, detergent itp.). Linii „ATOM__” i „HETATM” jest tyle, ile atomów obserwowanych przez eksperymentatora dla danej makrocząsteczki lub kompleksu.

Ograniczenia formatu pdb

Format 80-kolumnowy plików pdb jest stosunkowo restrykcyjny. Maksymalna liczba atomów w pliku pdb to 99 999, ponieważ na numery atomów jest przydzielonych tylko 5 kolumn. Podobnie liczba reszt w ciągu wynosi maksymalnie 9999: dla tej liczby dozwolone są tylko 4 kolumny. Liczba ciągów jest ograniczona do 62: dostępna jest tylko jedna kolumna, a możliwymi wartościami są jedna z 26 liter alfabetu, pisana małymi lub dużymi literami lub jedna z cyfr od 0 do 9. Ilość tego formatu została zdefiniowana, ograniczenia te nie wydawały się restrykcyjne, ale były wielokrotnie przezwyciężane przy deponowaniu bardzo dużych struktur, takich jak wirusy, rybosomy lub kompleksy multienzymatyczne.

Filtrowanie danych i formuły tekstowe.

Przydatne skróty w programie MS Excel lub LibreOffice Calc.

Ctrl + ↓- przejście do ostatniej wypełnionej komórki w tabeli.
Shift + Ctrl + ↓ - zaznaczenie wszystkich komórek od bieżącej do ostatniej wypełnionej w danej kolumnie.
Ctrl + Home – przejście do pierwszej komórki w arkuszu (A1).
Ctrl + End – przejście do ostatniej komórki w arkuszu.

Przydatne formuły (funkcje) tekstowe i logiczne:

DŁ (tekst) – oblicza długość ciągu tekstowego;
LEWY (tekst;liczba) – zwraca początkowe (z lewej strony) znaki (liczba) ciągu tekstowego;
FRAGMENT.TEKSTU (tekst;początek;liczba) - zwraca fragment tekstu od wskazanej pozycji (początek;
ZŁĄCZ.TEKST (tekst1; tekst2;…..) - łączy ciągi tekstowe;
JEŻELI (test logiczny;wartość dla prawdy;wartość dla fałszu) – wyświetla odpowiednią wartość w zależności od wyniku testu logicznego;
LICZ.JEŻĘLI (zakres;kryteria) – liczy argumenty, które spełniają podane warunki (kryteria)

Ćwiczenia

Plik Bacillus.fasta.

1. Znaleźć fragment i podać jego pozycję względem początku sekwencji CACGACTTCAAACACTAACGAAACATTGCGTTTTTTCACA
2. Policzyć ilość wystąpień kodonu TAC
3. Zaznaczyć fragment sekwencji od początku linii 48254 do końca i zapisać w pliku Bacillus_part2.fa
4. Rozdzielić plik tak aby po terminowej sekwencji TTGAAGCCATC następował nowa sekwencja (znacznik >).

Plik 5eaa.pdb

5. Wyświetlić cząsteczkę białka 5eaa przy użyciu programu RasMol lub PMV
6. Obliczyć ilość atomów w tej cząsteczce
7. Policzyć ilość aminokwasów TYR (tyrozyna) z których zbudowana jest ta cząsteczka (sekcja SEQRES w pliku pdb).

Plik wheatM1.fastq

8. Usunąć z pliku wszystkie sekwencje krótsze niż 41 nukleotydów
9. Utworzyć histogram długości sekwencji
10. Odszukać, zapisać do osobnego pliku wszystkie sekwencje rozpoczynające się od CGGACC. Policzyć ilość sekwencji rozpoczynających się od takiego adaptera (podzielić liczbę wystąpień CGGACC przez 4).

Plik 5eaa.pdb

11. Policzyć ilość wystąpień atomów C alfa (CA).
12. Utworzyć tabele liczności wszystkich atomów wchodzących w skład tej cząsteczki.

Plik wheatB1_to_ref.sam

13. Przekonwertować plik SAM na fasta (czyli każda sekwencja ma mieć unikalny identyfikator)
14. Otworzyć plik *B1_1_u.sort.bam*.

Zadania z gwiazdką albo dwoma

15. ** Napisać program który automatycznie wykonana ćwiczenie 6. Język programowania dowolny. Umiejętność programowania niewymagana. (10 pkt)
16. ** Napisać program który automatycznie wykonana ćwiczenie 11.

Język programowania dowolny. Umiejętność programowania niewymagana.
(10 pkt)

17 *. Odszukać i zapisać do osobnego pliku wszystkie sekwencje kończące się następującymi sekwencjami TCC lub TTCT. Sekwencje wynikowe powinny być posortowane według długości. (5 pkt.)

Biologiczne bazy danych

- **EMBL**
- **NCBI**
- **DDBJ**

Genomy eukariotyczne

Genom jądrowy jest podzielony na dwie lub więcej liniowych cząsteczek DNA – chromosomów (u człowieka 22 autosomy i dwa chromosomy płci X i Y)

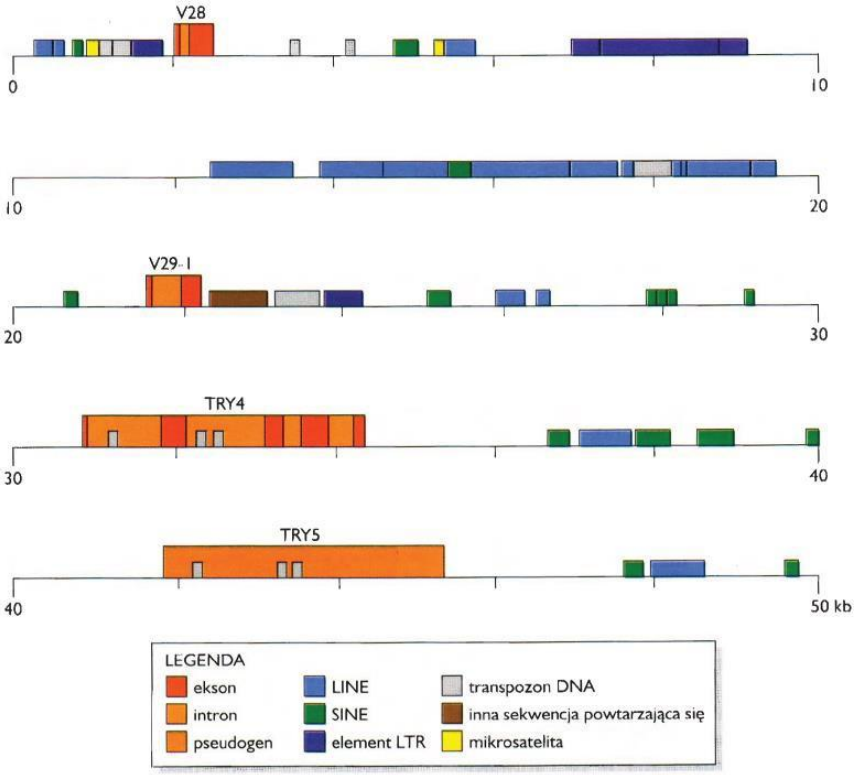
Genom mitochondrialny

Genom chloroplastowy tylko w organizmach fotosyntezujących

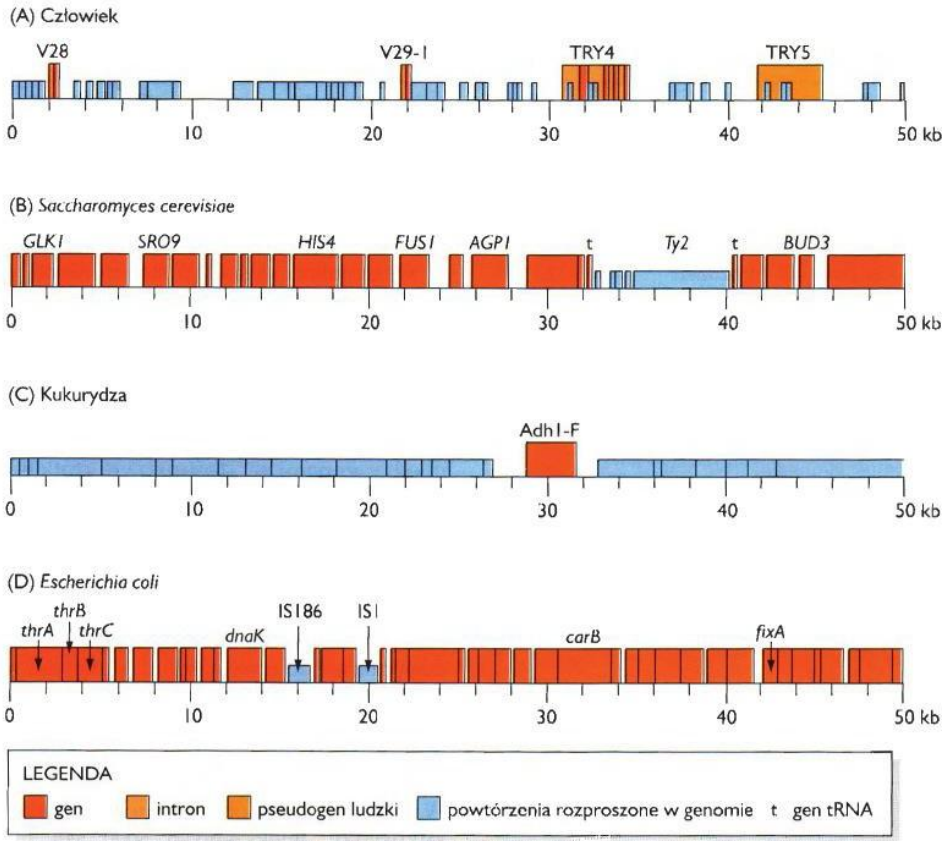
Organizm	Wielkość genomu (Mb)
Prokariota	
Mycoplasma genitalium	0,58
Escherichia coli	4,64
Bacillus megaterium	30
Eukariota	
Grzyby	
Saccharomyces cerevisiae (drożdże)	12,1
Aspergillus nidulans	25,4
Pierwotniaki	
Tetrahymena pyriformis	190
Bezkręgowce	
Caenorhabditis elegans (nicień)	100
Drosophila melanogaster (muszka owocowa)	140
Bombyx mori (jedwabnik)	490
Strongylocentrus purpuratus (jeżowiec)	845
Locusta migratoria (szarańcza)	5000
Kręgowce	
Fugu rubripes (ryba najeżka)	400
Homo sapiens (człowiek)	3000
Mus musculus (mysz)	3300
Rośliny	
Arabidopsis thaliana (rzodkiewnik)	100
Oryza sativa (ryż)	565
Pisum sativum (groch)	4800
Zea mays (kukurydza)	5000
Triticum aestivum (pszenica)	17000
Fritillaria assyriaca (szachownica)	120 000

Dane za: Brown (1998).

Zawartość genetyczna genomu człowieka



Porównanie genomów



Ćwiczenia

15. Zapisać na dysku referencyjną sekwencje chromosomu 1 Kropidlaka popielatego (Aspergillus fumigatus)

16. Odszukać w bazie danych sekwencję genu FMR1 człowieka i zapisać ją w formacie FASTA.

17. Zapisać sekwencje aminokwasów kodowanych przez drugi ekson ludzkiego genu CREB5

BLAST (ang. *Basic LocalAlignmentSearchTool*) – narzędzie bioinformatyczne (algorytm) służące do lokalnego przyrównywania sekwencji aminokwasówbiałek lub nukleotydówDNA. BLAST umożliwia naukowcom porównywanie zadanej sekwencji z sekwencjami zawartymi w biologicznych bazach danych i statystyczną ocenę ich podobieństwa.

Różne typy BLAST-a służą do porównywania różnych rodzajów sekwencji. Dla przykładu: po odkryciu nieznanego genu u myszy, przeszukują bazę z ludzkim genomem pod kątem obecności podobnych genów. BLAST znajdzie sekwencje podobne w bazie o ustalonych z góry parametrach (takich jak stopień podobieństwa). Dzięki prostocie obsługi jest bardzo użytecznym narzędziem.

Projektantami algorytmu byli: Stephen Altschul, Warren Gish, Webb Miller, Eugene Myers, i David J. Lipman z National Institutes of Health.

Ćwiczenia

18. Dopasować sekwencję:
    AAAAAACAAAAAACAAGATTGATGTTCGGATGATGGCTGCTAGCTATCCTTTTATGAAGG
    TGGGATATGTGGTCTCACTGCAGGGTGGCCTAGGAAGGGGAATCTGCTCATAGACATTAC
    CCAGAGATCTGTTCTCCTGCATTCTGACTGGAGGGGAGCTAAGACCCTAGAGGAGTCGGC
    CTGTCACCCATAGACCAATCAATCACCTGTCATCCATAGACCAGTCACCTGTCACCTAAA
    CCCATGGCTGCTCTTTTTTCATCTTCATCTTCCAAGACCTAGTGACCCAGGTTCATGC

19. Dopasować sekwencję:
    CCTCCTCTCTCTTTTCTTCTTTCTTCTCTTTTTCTTTTCTCTTCTTCTTCTTCTCCTTCTTTCCTCTTT
    CACTCCTCTTTCTTTTCTTTTTCTCTTCTTTTTTTTTCTTCCCTCCTTCCTCTTTCTCCCCCCTTTTC
    TATTTTCCCTCCCTCCTCTTTCCCCTCTTCTCTCCCCCCCCCCCTTCCCCCCCCCCTCCTCTCTCCC
    TTTTCTCCTTCTCCTCCCCCCTCCCCCCTTTCTCCCCCTCTCCCTTCTTCCTCTCCCCCCCCCCCCC
    CCCCCTCCCCCCCCCTTCCCTCTCCCTCCT

20. Dopasować sekwencję:

```
caggaggctt ataagttttg ccaagagtat aggtatctga attaactgta atcgacttaa
 tggttttcac taaatcctcc cgttaatacc attttaccat aataccatta ccatcattaa
 aaaaaggtaa caaacagacc taacataaag aaaatacctc ccgccaccaa aattagatat
 atgaatccaa tcaagactct aaacacaaaa tatttactgc tcatgacaat aaaaa
```

21. Dopasować sekwencję:

    MVGGGTKRRRRRLQLSKLYTLTCAQACFKQDHSQIGGPGFSRVVYCNEPDSPEADSRNYSDNY
    VRTTKYTLATFLPKSLFEQFRRVANFYFLVTGVLAF

22. Dopasować sekwencję:

    CTGCATGGCCGTCCTTTGAAGGTTCAAAACGAACCTTTTTACAATTTGCACTCACTGTTG