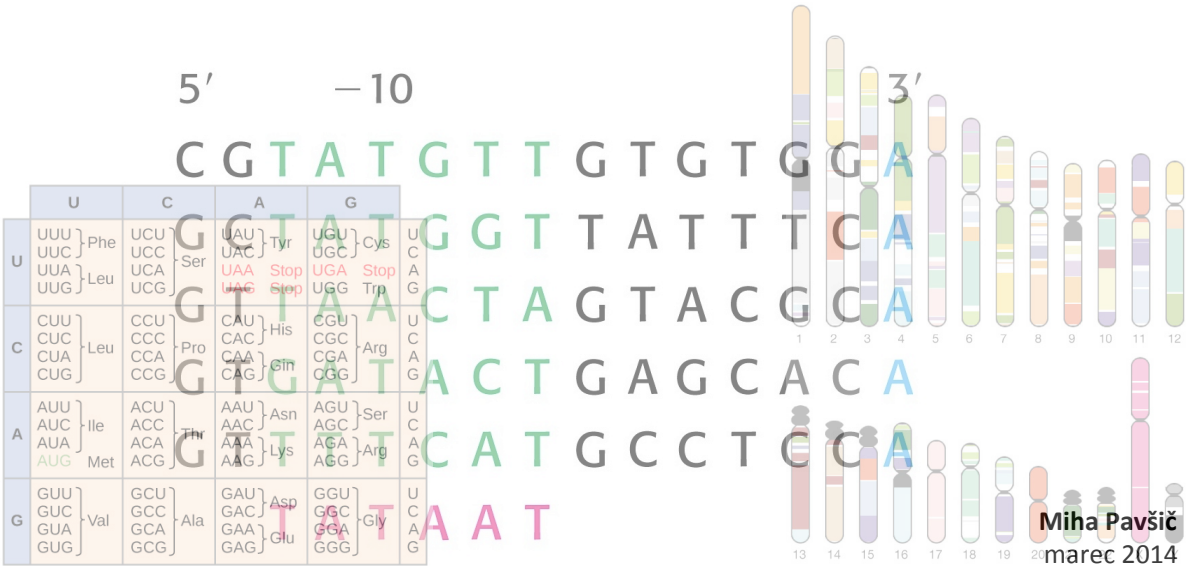
Praktične vaje pri predmetu
Biokemijska informatika

4. vaja

Baze in analiza nukleotidnih zaporedij



Miha Pavšič
marec 2014

Pregled vaje

Baze nukleotidnih zaporedij
- pregled
- osnovni tipi baz nukleotidnih zaporedij
- iskanje po bazah
- genomska zaporedja in zaporedja cDNA
- bazi EST in GSS v okviru GenBank
- OMIM – Online Mendelian Inheritance in Man

v obliki kviza v spletni učilnici

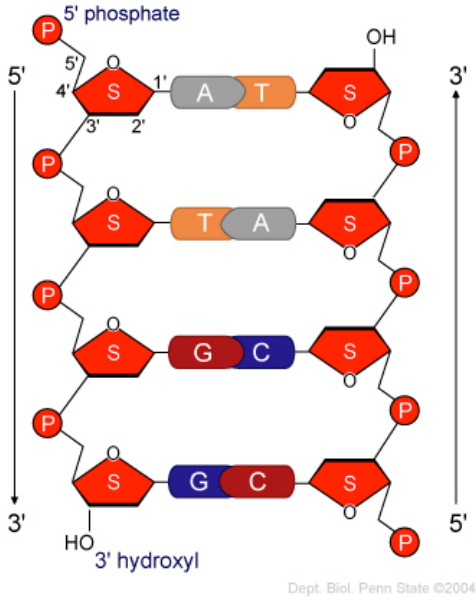
Manipulacija nukleotidnih zaporedij in analiza
- obratno in/ali komplementarno zaporedje
- iskanje odprtih bralnih okvirjev (ORF)
- prevajanje zaporedij nt v ak (+1, +2, +3, -1, -2, -3)

- skupno zaporedje (*consensus sequence*)

v Excel-u

Zapisovanje nukleotidnih zaporedij

Nukleotidna zaporedja **ZMERAJ (!!!)** pišemo od 5'- proti 3'-koncu (razen, če **IZRECNO** navedemo drugače).




```
5'-ATGG-3'
3'-TACC-5'
```

zapišemo kot

ATGG

- zaporedje komplementarne verige je avtomatsko določeno
- pogosto uporaba pisave s fiksno širino črk (npr. `Courier`) za lažje poravnave večih zaporedij oz. pišemo v *Notepad*-u (*plain text*)

Format FASTA

- je navaden **tekstni format**
- prva vrstica v datoteki se začne z znakom **>**, ki mu sledi **ime/opis/... zaporedja**
- od druge vrstice naprej je zapisano **zaporedje (brez presledkov, številk, ...)**
- zaporedje je lahko podano v eni vrstici (uporabno za iskanje!) ali v večih
- v isti datoteki je lahko več zapisov, ki si sledijo eden za drugim
- enak format se uporablja tudi za zaporedja ak

ime/opis/... **1** ← `>GRT87-Homo_sapiens`

zaporedje **1** 5'→3' ←
```
AGTCGCGTAGGCTGATCGGCTAGATTTCGCTAGAGATCGATGGCTGACA
GGCTAGCTGATCGGCTAGGATCGGATCGGCTTAGAGTGGATGCGGCTGA
GGCTTAGGCTAGGGCTGAAT
```

ime/opis/... **2** ← `>His54-Mus_musculus`

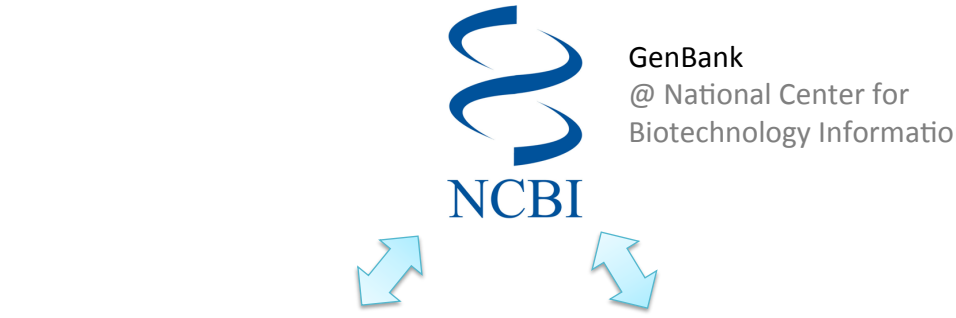
zaporedje **2** 5'→3' ←
```
TGCGCTAGAGAGAAAATATATAAACACTCGTAGGGATCGGATGCGGAGG
ATCGAGATCGGAGCTGAGTCGGAGTCTGGAGGAGAGAGTTCTCC
```

ime/opis/... **3** ← `>T32-Bos_taurus 35..44`

zaporedje **3** 5'→3' ←
```
AGCTGGGATG
```

Osnovne baza nukleotidnih zaporedij

International Nucleotide Sequence Database Collaboration (INSDC):



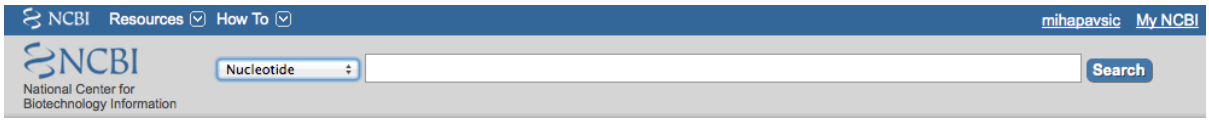
GenBank
@ National Center for Biotechnology Information

European Nucleotide Archive
@ European Molecular Biology Laboratory (EMBL)

DNA DataBank of Japan

Izmenjava podatkov med temi bazami poteka dnevno.

GenBank



- prosto dostopna baza nukleotidnih zaporedij v okviru NCBI
- **158 × 10^9 nt** v 171 milijonov zaporedij (15. februar 2014; izdaja 200.0)
- baza **(Core)Nucleotide** – osnovna zbirka (http://www.ncbi.nlm.nih.gov/nuccore/) in "podbazi":
  - **dbEST** – Expressed Sequence Tags (http://www.ncbi.nlm.nih.gov/nucest/)
  - **gbGSS** – Genome Survey Sequences (http://www.ncbi.nlm.nih.gov/nucgss/)
- iskanje je možno preko **osnovnega vmesnika** ali preko **BLAST** (iskanje podobnih zaporedij – to bomo obravnavali pri eni kasnejših vaj)
- "flat file format": http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
```

GenBank – format zapisa

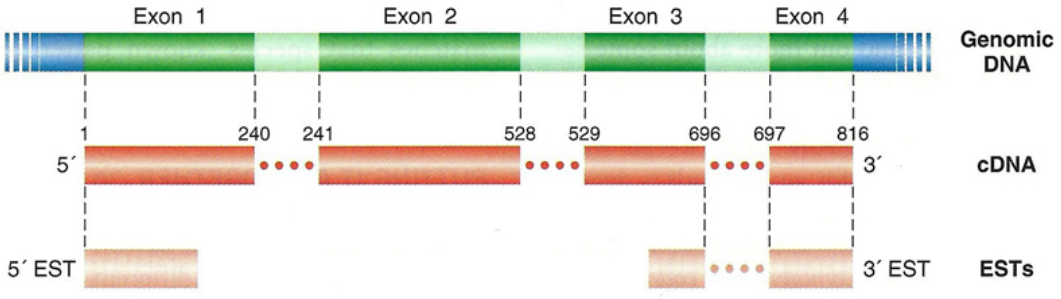
Nekatera **polja**, ki jih vsebuje posamezen zapis v bazo:

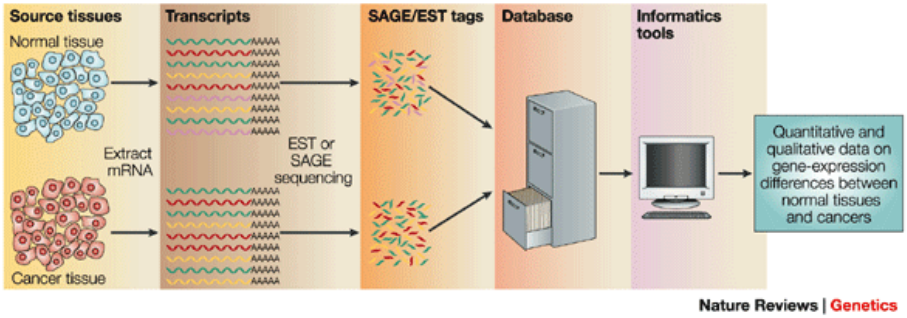
- sequence length – dolžina zaporedja v bp
- molecule type – genomic DNA, genomic RNA, mRNA (cDNA), ...
- division – primati, glodalci, ..., sintetična zaporedja, EST, GSS, ENV, ...
- modification date – datum zadnje spremembe zapisa
- definition – kratkek opis (organizem, ime gena/proteina, ...)
- accession – unikatna koda za posamezen vnos v bazo
- keywords, source (organism), reference (avtorji, naslov članka in revije, PubMed ID)
- features (npr. CDS – *coding sequence*); pozor pri začetku/koncu, določen segment ni nujno, da je popolen (znaka < oz. >), lahko je na komplementarni verigi)
- pri zapisih, ki kodirajo polipeptidno verigo, je dodan prevod (translation)

Expressed Sequence Tags (baza EST @ NCBI)

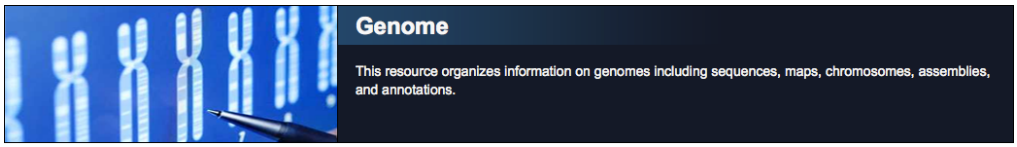
reverzna transkripcija

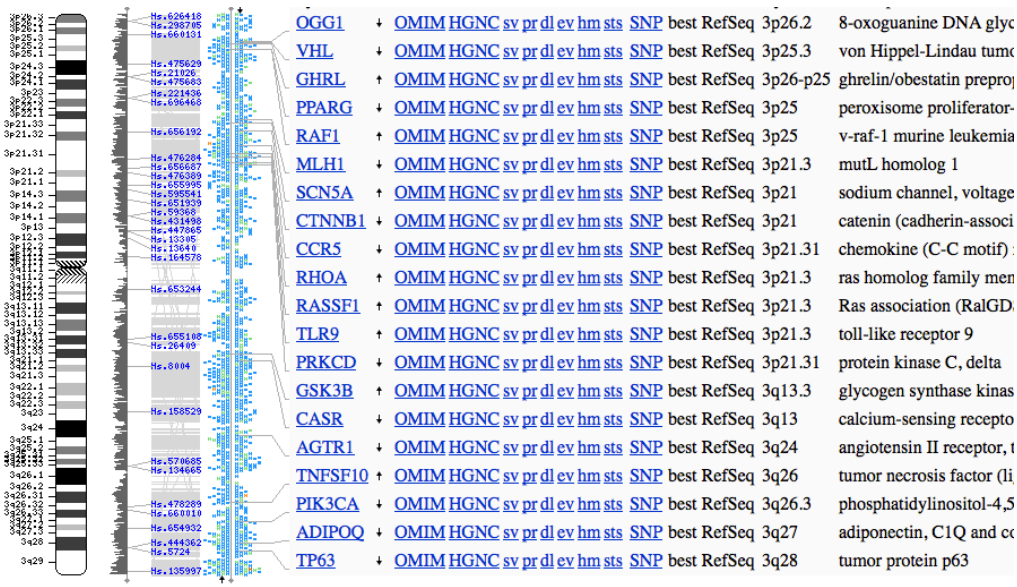
mRNA → cDNA → DNA → sekveniranje s 5'- in 3'-konca





Genome @ NCBI





kromosom gen povezave ime proteina, ...

Obratno / komplementarno zaporedje (*Reverse / Complementary Sequence*)

dano zaporedje

AGTCGAGCTG

obratno zaporedje (REVERSE)
(dano zaporedje preberemo v obratni smeri)

GTCGAGCTGA
AGTCGAGCTG

komplementarno zaporedje (COMPLEMENT)
(nt danega zaporedja zamenjamo s komplementarnimi)

TCAGCTCGAC
AGTCGAGCTG

obratno komplementarno zaporedje
(REVERSE COMPLEMENT)

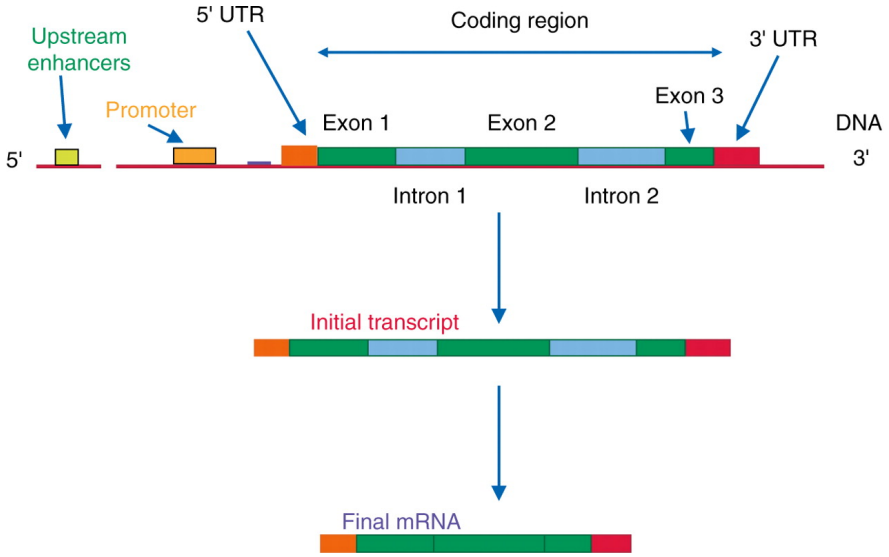
CAGCTCGACT
AGTCGAGCTG

zaporedje + obratno komplementarno zaporedje → dsDNA

5'–CAGCTCGACT–3'
3'–AGTCGAGCTG–5'

Kodirajoče zaporedje (CDS, *Coding Sequence*)

- splošno: del DNA ali RNA, ki kodira za zaporedje
  - RNA *coding sequence* – zaporedje na DNA, ki kodira za RNA
  - *protein coding sequence* – mRNA oz. eksoni na DNA, ki kodirajo za protein



Bralni okvir (*Reading Frame*)

dano zaporedje: GATGGTACTGAGTCGTAGTGGGGCGTAAGC

```
          1                  10                  20                  30
   5' GATGGTACTGAGTCGTAGTGGGGCGTAAGC 3'
+1    D   G   T   E   S   *   W   G   V   S
+2      M   V   L   S   R   S   G   A   *
+3        W   Y   *   V   V   V   G   R   K
        I   T   S   L   R   L   P   A   Y   A   -1
          P   V   S   D   Y   H   P   T   L     -2
        H   Y   Q   T   T   T   P   R   L       -3
   3' CTACCATGACTCAGCATCACCCCGCATTCG 5'
```

* = STOP kodon

Odprt bralni okvir (ORF, *Open Reading Frame*)

- definicija: del bralnega okvirja, ki ne vsebuje STOP kodonov
- ponavadi obravnavamo kot odsek DNA oz. RNA od START do STOP kodona, pri čemer je število nukleotidov med njima deljivo s 3:

ATG — $(3nt)_N$ — TAA ochre
TAG amber
TGA opal

START STOP

- v zaporedju cDNA je najdaljši bralni okvir načeloma tisti, ki kodira protein

Odprt bralni okvir (ORF, *Open Reading Frame*)

dano zaporedje: GATGGTACTGAGTCGTAGTGGGGCGTAAGC

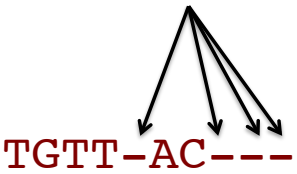
```
          1                   10                  20                  30
5' GATGGTACTGAGTCGTAGTGGGGCGTAAGC 3'
+1   D  G  T  E  S  *  W  G  V  S
+2   M  V  L  S  R  S  G  A  *
+3    W  Y  *  V  V  V  G  R  K
      I  T  S  L  R  L  P  A  Y  A   -1
       P  V  S  D  Y  H  P  T  L     -2
        H  Y  Q  T  T  T  P  R  L     -3
3' CTACCATGACTCAGCATCACCCCGCATTCG 5'
```

* = STOP kodon

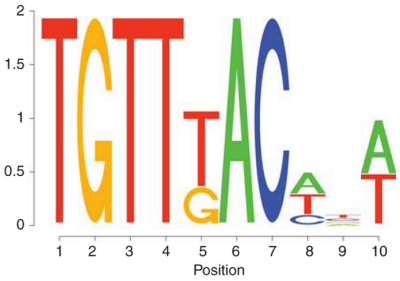
Skupno zaporedje (*Consensus Sequence*)

- **izračunamo** ga na podlagi večih zaporedij, ki so med seboj **PORAVNANA**
- lahko gre za nt ali ak zaporedje
- v skupnem zaporedju je na določenem mestu zapisan tisti nt/ak, ki se najpogosteje pojavlja v izhodnih zaporedjih na tem istem mestu
- lahko vključimo **prag**: na posameznem mestu izpišemo nt/ak le, če je frekvenca najpogostejšega večja od neke vrednosti
- na primer, prag 0,5 pomeni, da bo na posameznem mestu skupnega zaporedja izpisan nt/ak le v primeru, da se najpogostejši pojavlja v več polovici analiziranih zaporedij

ti nt niso prikazani ker je frekvenca < prag



klasična oblika
(izguba informacije!)



oblika “sequence logo”