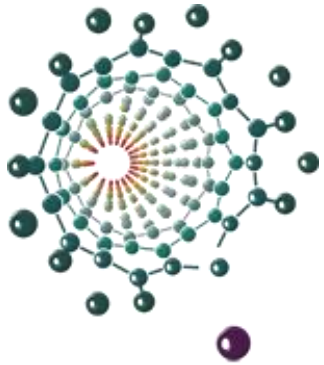


b101nf0rmat1cs

ACCATGGATTACATA 1101101100011010 10
GATTCCATTATAAGGA 1100111000000100
TGCCGGCAATAGGCAA 1110101000110101
CAATAAGCATTCCAC 0101010110101 1011

Računarska vizualizacija i manipulacija DNK sekvencijom

Miloš Mojović, v. prof.

Osnovni pojmovi u genetici

- Bazične metode manipulacije DNK sekvencama danas se u potpunosti zasnivaju na upotrebi savremenih računarskih metoda.
- Vizualizacija i statistička obrada oblasti DNK zaduženih za kodiranje proteinskih sekvenci predstavlja veoma koristan alat za proučavanja u oblasti genetičkih istraživanja.

Osnovni pojmovi u genetici

- Genetski kod predstavlja skup pravila pomoću kojih se informacija enkodirana u genetskom materijalu (nukleotidna sekvenca u DNK ili iRKN) prenosi na proteine (sekvencu aminokiselina).
- Kodon predstavlja niz od tri nukleotida (triplet) na iRNK koji predstavlja šifru za jednu aminokiselinu. Niz kodona šifruje polipepdidni lanac.
- Antikodon predstavlja sekvencu od tri susedna nukleotida na tRNK (koji se vezuju za odgovarajuće kodone na iRNK) i određuju specifičnu aminokliselinu u procesu sinteze proteina.
- Geni su funkcionalni delovi DNK koji sadrže informacije o sekvenci aminokiselina u nekom proteinu.
- Genom nekog organizma predstavlja skup svih naslenih podataka nekog organizma. Genom u sebi sadrži skup svih gena i nekodirajućih sekvenci u DNK.

Osnovni pojmovi u genetici

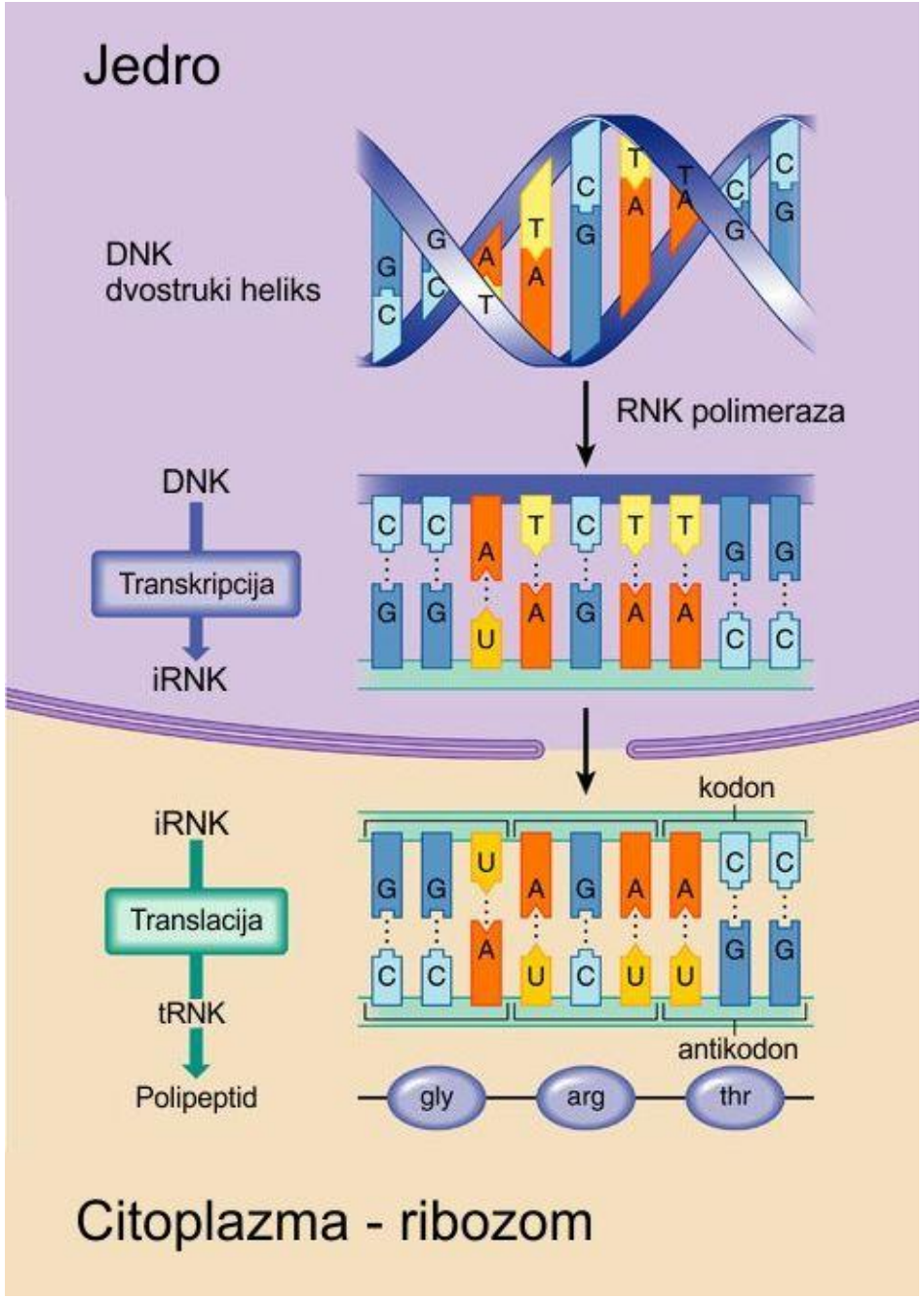
- Transkripcija predstavlja proces sinteze molekula iRNK kao kopije jednog dela lanca DNK (gena). Ovaj proces katalizuje enzim RNK polimeraza i odvija se jedru eukariota. U procesu transkripcije naspram adenina (A) u DNK postavlja se uracil (U) u iRNK, naspram guanina (G) DNK postavlja se citozin (C) u iRNK, naspram citozina (C) u DNK postavlja se guanin (G) u iRNK i naspram timina (T) u DNK postavlja se adenin (A) u iRNK.

DNK		iRNK	Primer:
A	→	U	DNK molekul: 3' AAATTCCCG 5'
G	→	C	iRNK molekul: 5' UUUAAGGGC 3'
C	→	G	
T	→	A	

Slika 1. Princip transkripcije molekula DNK u molekul iRNK.

Osnovni pojmovi u genetici

- Translacija predstavlja proces dekodiranja nukleotidne sekvence i biološku sinteza proteina.
- Odvija se u ribozomima u kojima se aminokiseline povezuju po redosledu koji je zapisan u iRNK korišćenjem molekula tRNK koji su zaduženi za prenos aminokiselina i njihovu implementaciju prema kodnom rasporedu.



Slika 2. Shematski prikaz transkripcije i translacije.

Kodiranje aminokiselinske sekvencije

- Kako je svaki kodon sastavljen od po tri nukleotida, a pošto postoje četiri različite baze, moguće je napraviti $4^3 = 64$ različitih kombinacija baza u kodonu. Na slici 3. nalazi se svih 64 kodona i odgovarajuće amonokiseline koje oni kodiraju.

Aminokiselina	iRNK kodoni	Aminokiselina	iRNK kodoni
Ala/A	GCU, GCC, GCA, GCG	Leu/L	UUA, UUG, CUU, CUC, CUA, CUG
Arg/R	CGU, CGC, CGA, CGG, AGA, AGG	Lys/K	AAA, AAG
Asn/N	AAU, AAC	Met/M	AUG
Asp/D	GAU, GAC	Phe/F	UUU, UUC
Cys/C	UGU, UGC	Pro/P	CCU, CCC, CCA, CCG
Gln/Q	CAA, CAG	Ser/S	UCU, UCC, UCA, UCG, AGU, AGC
Glu/E	GAA, GAG	Thr/T	ACU, ACC, ACA, ACG
Gly/G	GGU, GGC, GGA, GGG	Trp/W	UGG
His/H	CAU, CAC	Tyr/Y	UAU, UAC
Ile/I	AUU, AUC, AUA	Val/V	GUU, GUC, GUA, GUG
START	AUG	STOP	UAG, UGA, UAA

Slika 3. Tabela kodona na iRNK i aminokiselina koje oni kodiraju.

Kodiranje aminokiselinske sekvencije

- Treba primetiti da se kodon AUG ponavlja pošto kodira i aminokiselinu Metionin a istovremeno kodira i region na iRNK molekulu gde počinje proces translacije u protein. Pored kodona koji inicira proces translacije (START kodon) postoje i kodoni koji je zaustavljaju (STOP kodoni).
- Za tok sinteze proteina, veoma je važno imati u potpunosti definisanu početnu poziciju čitanja.
- Na primer, ukoliko imamo sledeću nukleotidnu sekvencu na iRNK 5'- GGGAAACCC -3', ako počnemo čitanje sa prvog G na levoj strani imaćemo sekvencu koja se sastoji od kodona: GGG, AAA, CCC. Međutim, ako počinjemo čitanje sa trećeg G sa leve strane imaćemo sekvencu: GAA, ACC (poslednji nukleotid se ignoriše jer ne čini potpun kodon).
- Svaka od pomenutih sekvenci će dati drugačiji raspored aminokiselina, tj. u prvom slučaju to će biti Gly-Lys-Pro a u drugom slučaju Glu-Thr.

Kodiranje aminokiselinske sekvencije

- Sama činjenica da se proteini sastoje iz mnoštva aminokiselina, kao i da nekoliko različtihih kodona može kodirati istu aminokiselinu, čini proces identifikacije nukleotidne sekvence veoma komplikovanim bez upotrebe računara.
- Danas na javno dostupnim bazama postoji veliki broj nukleotidnih sekvenci koje kodiraju različite vrste proteina kao i kompletne genome ćelijskih organela i živih bića.

Kodiranje aminokiselinske sekvencije

- Sama činjenica da se proteini sastoje iz mnoštva aminokiselina, kao i da nekoliko različtihih kodona može kodirati istu aminokiselinu, čini proces identifikacije nukleotidne sekvence veoma komplikovanim bez upotrebe računara.
- Danas na javno dostupnim bazama postoji veliki broj nukleotidnih sekvenci koje kodiraju različite vrste proteina kao i kompletne genome ćelijskih organela i živih bića.

http://www.ncbi.nlm.nih.gov/

Osnovnim računarskim principima istraživanja DNK sekvenci (genoma)

- Upoznaćemo se sa osnovnim računarskim principima istraživanja DNK sekvenci (genoma). Kao primer poslužiće nam sekvence humanog i goveđeg proteina albumina (HSA i BSA).
- Informacije o genomima mogu se pronaći na javno dostupnim servisima na Internetu kao što je: "Genome repository at the National Center for Biotechnology Information (NCBI)".
- Za izradu ove vežbe koristićemo program MATLAB®.
- Za ovaj korak potrebno je da računar ima pristup Internetu. Proveriti da li je u MATLAB programu setovan pristup Internetu. Za pretragu takođe možemo koristiti i bilo koji raspoloživi Internet pretraživač.

Računarska vizualizacija i manipulacija DNK

```
% Ucitavanje nukleotidne sekvence genoma. Formiramo nove promenljive 'HSA'
% i 'BSA'

load 'BSA.mat'
load 'HSA.mat'

% Alternativno, mozemo se povezati na Internet i sa adekvatnog servera
% preuzeti nukleotidne sekvence genoma. Te sekvence se mogu naci na sajtu
% http://www.ncbi.nlm.nih.gov/
% Kada pronadjemo adekvatne sifre za sekvenci genoma, uvozimo ih upotrebom
% komandi:

% HSA = getgenbank('AF542069','SequenceOnly',true);
% BSA = getgenbank('M73993.1','SequenceOnly',true);

% Sada mozemo prikazati informacije o velicini preuzetih sekvanci upotrebom
% 'whos' komande:

whos HSA
whos BSA
```

Integralni prikaz kompozicije nukleotidne sekvencije

```
% Za integralni prikaz kompozicije nukleotidne sekvence koristimo komandu
% "ntdensity" cija je sintaksa:

figure(1)
ntdensity(HSA)
title('HSA')

figure(2)
ntdensity(BSA)
title('BSA')
```

```
%% Zadatak:

% Probajte da napravite nasumicnu nukleotidnu sekvenciju i prikazite informacije o toj sekvenci:
% s = randseq(1000, 'alphabet', 'dna');
% ntdensity(s)
```

Računanje statistike nukleotidnih sekvenci

```
% Izracunajmo statistiku nukleotidnih sekvenci za HSA i BSA. Za tu svrhu
% koristiæemo komandu "basecount" cija je sintaksa:

HSAbases=basecount(HSA);
BSAbases=basecount(BSA);

% Dobili smo dve strukturne promenljive 'HSAbases' i 'BSAbases' koje
% predstavljaju broj pojedinacnih baza u sekvencama. Njihove vrednosti
% mozemo videti dvoklikom u 'Workspace' prozoru.

% Broj komplementnih baza mozemo dobiti upotrebom komande "seqrcomplement"
% cija je sintaksa:

compHSA=basecount(seqrcomplement(HSA));
compBSA=basecount(seqrcomplement(BSA));
```

Grafički prikaz distribucije baza po genomima

```
% Graficki prikaz distribucije baza po genomima mozemo dobiti upotrebom
% sledece sintakse:

figure(3)
basecount(HSA,'chart','pie');
title('Distribucija nukleotidnih baza za HSA');

figure(4)
basecount(BSA,'chart','pie');
title('Distribucija nukleotidnih baza za BSA');
```

Uvid u kodone

```
% Uvid u kodone dobija se upotrebom komande "codoncount".
% Ovom komandom brojimo uvig u zatupljenost svih 64 mogucih kodona
% pocevsi od prvog rapolozivog nukleotida (prvi frejm).
% Medjutim, kodoni ce se razlikovati ukoliko pocinjemo citanje sa drugog
% ili treceg nukleotida (drugi i treci frejm).
% Za svrhu prikaza zatupljenosti svih kodona (u zavisnosti od nukeotida od
% koga pocinjemo sekvencioniranje) napravicemo sledece petlje:

for frame = 1:3
figure(5)
subplot(3,1,frame);
codons{1,frame}=codoncount(HSA,'frame',frame,'figure',true);
title(sprintf('Zastupljenost mogucih kodona od frejma %d za HSA',frame));
end

for frame = 1:3
figure(6)
subplot(3,1,frame);
codons{1,frame}=codoncount(BSA,'frame',frame,'figure',true);
title(sprintf('Zastupljenopst mogucih kodona od frejma %d za BSA',frame));
end
```

Ispitivanje genoma humane mitohondrije

```
% Ispitajmo genom humane mitohondrije. Izdvojicemo gene koji mogu kodirati
% razlicite proteine, odredimo njihovu aminokiselinsku sekvencu i predvidimo
% neke osobine tih proteina.

% Kompletan genom humane mitohondrije mozemo preuzeti sa javno dostupne
% Internet baze ciji deo se nalazi u MATLAB-u:

% load mitochondria (tu se formira promenljiva 'mitochondria_gbk')

% ... ili ga, ukoliko postoji u nasem folderu, ucitavamo komandom:

load mitochondria_gbk

% Ukoliko zelimo da ispitamo koje oblasti u ovoj nukleotidnoj sekvenciji
% mogu biti translirane u protein (pronalazenjem START i STOP kodona tj.
% trazenjem ORF - Open Reading Frames), koristicemo komandu 'seqshoworfs':

figure(7)
seqshoworfs(mitochondria_gbk);
```

Izdvajanje samo odredjene proteinske sekvencije

```
% Uzmimo, na primer, da nas zanima protein ciji ORF pocinje na poziciji
% 73 a zavrsava se na poziciji 198. Da bi izdvojili sekvenciju za ovaj
% protein upotrebicemo sintaksu:

proteinSeq = mitochondria_gbk.Sequence(73:198);
```

Konverzija nukleotidnih sekvencija u aminokiseline

```
% Koverziju nukleotidnih sekvencija u AK radimo komandom 'nt2aa':

protein = nt2aa(proteinSeq);

% AK su ovde prikazane samo kodnim slovom. Ukoliko zelimo da vidimo koje
% kodno slovo odgovara kojoj AK, upotrebicemo komandu 'amonolookup':

aminolookup

% Na primer, ako zelimo da prevedemo dobijenu AK sekvenciju iz kodnih slova
% u skraceno ime, uradicemo sledece:

proteinfull=aminolookup(protein);

% Sada mozemo prikazati zastupljenost odredjanih AK komandom 'aacount':

figure(7)
proteinaaCount = aacount(protein,'chart','bar');
title('Histogram AK za protein u okviru mitohondrijalnog genoma');
```

Izračunavanje zastupljenost odredjenih atoma i molekulske mase proteina

```
% Da bi izracunali zastupljenost odredjenih atoma u proteinu koristicemo
% komandu 'atomiccomp'

proteinAtomicComp = atomiccomp(protein) % Izostavimo ";" da bi dobili ispis
                                        % u Command Window

% Da bi izracunali molekulsku masu proteina koristicemo komandu 'molweight'

proteinMolWeight = molweight(protein) % Izostavimo ";" da bi dobili ispis
                                      % u Command Window
```

Predviđanje strukture proteina na osnovu AK sekvencije

```
% Da bi predvideli strukturu i ostale osobine proteina koji ima odredjenu
% AK sekvanciju koristimo MATLB potprogram koji pozivamo komandom
% 'proteinplot'. Sintaksa je sledeca:

proteinplot(protein)

% Ukoliko zelimo, mozemo (u okviru ovog potprograma) da uvezemo AK
% sekvencije razlicitih proteina pomocu "Import Sequence" opcije.
% Ove sekvencije mozemo naci na: "http://www.ncbi.nlm.nih.gov/protein".
```

```
%% Zadatak:

% Pronadjite u proteinskoj bazi podataka AK sekvenciju za nekoliko razlicitih proteina: keratin,
% humani interferon, citohrom c, itd.
% Za uvoz sekvence AK (npr. za interferon) koristiti komandu 'getgenepept':
INTERF = getgenpept('AAA36123.1','SequenceOnly',true);
% Uporedite strukturne osobine odabranih proteina (zastupljenost alfa-heliksa, beta-ravni, itd.)
```