Laborator 2 – Bioinformatică

I. Simularea procesului de transcriptie din ADN in ARN si pregatirea procesului de translatie

a. Să se implementeze o funcţie *ADN(n)* care să genereze o secvenţă aleatoare de ADN ca o succesiune de baze azotate *Adenina*, *Guanina*, *Timina*, *Ciozina* sub forma unei liste de tupluri de câte 3 baze azotate. Lista va avea dimensiunea *n*, iar succesiunea de baze este obţinută utilizând modulul *random*.

Apelul *ADN(5)* întoarce, de exemplu:

```
[(A,T,C), (G,A,A), (C,A,A), (T,C,A), (A,G,C)]
```

Indicii:

- să se utilizeze funcţiile *seed(x)* şi *randint(a,b)* din modulul *random*;
- *seed(x)* trebuie apelată o singură dată pentru a iniţia algoritmul de generare de numere aleatoare (pseudoaleatoare);
- pentru ca la fiecare apel al funcţiei *ADN()* să se genereze o altă secvenţă de numere aleatoare, este recomandat să se seteze ca *seed* timpul sistemului;
- timpul sistemului în secunde este întors de funcţia *time()* care se găsesşte în modulul *time*;
- funcţia *seed(time())* trebuie apelată o singură dată în funcţia *ADN()*;
- numerele aleatoare se generează cu funcţia *randint(a,b)* care întoarce un numar aleator între *a* şi *b*.

b. Să se scrie o funcţie *Afisare(secv)* care primeşte ca parametru o secvenţă precum cea de la punctul **a.** şi o transformă într-un şir de caractere eliminând parantezele şi virgulele.

Pentru exemplul de mai sus, *Afisare(secv)* va întoarce: `ATCGAACAATCAAGC`

c. Să se implementeze o functie *ARN(secADN)* care generează o secvenţă de ARN mesager, care copiaza secvenţa de ADN primită ca parametru. Secventa de ADN este primita in directia 5'3'. Tratati ambele cazuri, considerand ca secventa care se doreste a fi copiata se poate afla fie pe catena primita ca parametru, fie pe cea complementara. Secvenţa de ARN este memorată in directia 5’3’ tot într-o listă de tupluri de baze azotate precum cea de ADN. Un tuplu de 3 baze azotate reprezintă un codon care va fi tradus într-un aminoacid în procesul de translaţie care are loc la nivelul ribozomilor.

Observaţie: Bazele complementare care formează legături sunt A-T si G-C. ARN conţine Adenina, Guanina si Citozina, iar Timina este inlocuită cu Uracil.

II. Numărarea perechilor de baze A-T şi C-G dintr-o secvenţă

1. Să se scrie o funcţie pbADN(n) care primeşte ca parametru o secvenţă de ADN sub forma unui şir de caractere. Funcţia trebuie să numere perechile de baze azotate A=T, respectiv C≡G, conţinute în regiunea de ADN reprezentată prin secvenţa primită ca parametru. Funcţia va întoarce un tuplu (at, cg) în care:

- at reprezintă numărul de perechi de baze azotate A=T,
- gt reprezintă numărul de perechi de baze azotate C≡G.

2. Sa se studieze si sa se testeze urmatorul cod. Este o implementare optima? Sa se cautate in documentatie (http://docs.python.org/library) la Built-in Functions functia *filter*. O modalitate de a defini functii intr-o forma foarte comprimata este utilizarea expresiei lambda (http://docs.python.org/reference/expressions.html#lambda).

```
secv = 'AAATTTCTCTGGTAGA'
def number2(secv):
        n = len(filter(lambda x: x =='A' or x == 'T', secv))
        return ( n, len(secv)-n)
```

III. Simularea procesului de translaţie – traducerea unei secvenţe de ADN într-o secvenţă proteică

1. Nucleotidele sunt codificate standard prin literele: A, T, C, G (în cazul ADN) şi A, U, C, G (în cazul ARN). O nucleotidă conţine o bază azotată, de aceea denumirea nucleotidelor pleacă de la baza azotată pe care o conţine. Acizii nucleici reprezintă lanţuri de nucleotide.
Ştiind că un codon (trei nucleotide) codifică un aminoacid, să se explice de ce codul genetic standard poate fi exprimat atât sub formă de codoni din ARN, cât şi sub formă de codoni din ADN.
Să se descrie pe scurt procesele prin care se face trecerea de la informaţia genetică din ADN la proteine (transcripţia şi translaţia): ADN → ARN (ARNm, ARNt, ARNr) → secvenţă proteică (catena polipeptidică – lanţ de aminoacizi). Codul genetic este ilustrat în figurile urmatoare:

Prima bază	A doua bază: T	A doua bază: C	A doua bază: A	A doua bază: G	A treia bază
T	TTT (Phe/F) Fenilalanină	TCT (Ser/S) Serină	TAT (Tyr/Y) Tirozină	TGT (Cys/C) Cisteină	T
	TTC (Phe/F) Fenilalanină	TCC (Ser/S) Serină	TAC (Tyr/Y) Tirozină	TGC (Cys/C) Cisteină	C
	TTA (Leu/L) Leucină	TCA (Ser/S) Serină	TAA (Stop)	TGA (Stop)	A
	TTG (Leu/L) Leucină	TCG (Ser/S) Serină	TAG (Stop)	TGG (Trp/W) Triptofan	G
C	CTT (Leu/L) Leucină	CCT (Pro/P) Prolină	CAT (His/H) Histidină	CGT (Arg/R) Arginină	T
	CTC (Leu/L) Leucină	CCC (Pro/P) Prolină	CAC (His/H) Histidină	CGC (Arg/R) Arginină	C
	CTA (Leu/L) Leucină	CCA (Pro/P) Prolină	CAA (Gln/Q) Glutamină	CGA (Arg/R) Arginină	A
	CTG (Leu/L) Leucină	CCG (Pro/P) Prolină	CAG (Gln/Q) Glutamină	CGG (Arg/R) Arginină	G
A	ATT (Ile/I) Izoleucină	ACT (Thr/T) Treonină	AAT (Asn/N) Asparagină	AGT (Ser/S) Serină	T
	ATC (Ile/I) Izoleucină	ACC (Thr/T) Treonină	AAC (Asn/N) Asparagină	AGC (Ser/S) Serină	C
	ATA (Ile/I) Izoleucină	ACA (Thr/T) Treonină	AAA (Lys/K) Lizină	AGA (Arg/R) Arginină	A
	ATG (Met/M) Metionină (Start)	ACG (Thr/T) Treonină	AAG (Lys/K) Lizină	AGG (Arg/R) Arginină	G
G	GTT (Val/V) Valină	GCT (Ala/A) Alanină	GAT (Asp/D) Acid aspartic	GGT (Gly/G) Glicină	T
	GTC (Val/V) Valină	GCC (Ala/A) Alanină	GAC (Asp/D Acid aspartic	GGC (Gly/G) Glicină	C
	GTA (Val/V) Valină	GCA (Ala/A) Alanină	GAA (Glu/E) Acid glutamic	GGA (Gly/G) Glicină	A
	GTG (Val/V) Valină	GCG (Ala/A) Alanină	GAG (Glu/E) Acid glutamic	GGG (Gly/G) Glicină	G

Codonii ADN care codifică cei 20 de aminoacizi

Prima bază	A doua bază: U	A doua bază: C	A doua bază: A	A doua bază: G	A treia bază
U	UUU (Phe/F) Fenilalanină	UCU (Ser/S) Serină	UAU (Tyr/Y) Tirozină	UGU (Cys/C) Cisteină	U
	UUC (Phe/F) Fenilalanină	UCC (Ser/S) Serină	UAC (Tyr/Y) Tirozină	UGC (Cys/C) Cisteină	C
	UUA (Leu/L) Leucină	UCA (Ser/S) Serină	UAA (Stop)	UGA (Stop)	A
	UUG (Leu/L) Leucină	UCG (Ser/S) Serină	UAG (Stop)	UGG (Trp/W) Triptofan	G
C	CUU (Leu/L) Leucină	CCU (Pro/P) Prolină	CAU (His/H) Histidină	CGU (Arg/R) Arginină	U
	CUC (Leu/L) Leucină	CCC (Pro/P) Prolină	CAC (His/H) Histidină	CGC (Arg/R) Arginină	C
	CUA (Leu/L) Leucină	CCA (Pro/P) Prolină	CAA (Gln/Q) Glutamină	CGA (Arg/R) Arginină	A
	CUG (Leu/L) Leucină	CCG (Pro/P) Prolină	CAG (Gln/Q) Glutamină	CGG (Arg/R) Arginină	G
A	AUU (Ile/I) Izoleucină	ACU (Thr/T) Treonină	AAU (Asn/N) Asparagină	AGU (Ser/S) Serină	U
	AUC (Ile/I) Izoleucină	ACC (Thr/T) Treonină	AAC (Asn/N) Asparagină	AGC (Ser/S) Serină	C
	AUA (Ile/I) Izoleucină	ACA (Thr/T) Treonină	AAA (Lys/K) Lizină	AGA (Arg/R) Arginină	A
	AUG (Met/M) Metionină (Start)	ACG (Thr/T) Treonină	AAG (Lys/K) Lizină	AGG (Arg/R) Arginină	G
G	GUU (Val/V) Valină	GCU (Ala/A) Alanină	GAU (Asp/D) Acid aspartic	GGU (Gly/G) Glicină	U
	GUC (Val/V) Valină	GCC (Ala/A) Alanină	GAC (Asp/D Acid aspartic	GGC (Gly/G) Glicină	C
	GUA (Val/V) Valină	GCA (Ala/A) Alanină	GAA (Glu/E) Acid glutamic	GGA (Gly/G) Glicină	A
	GUG (Val/V) Valină	GCG (Ala/A) Alanină	GAG (Glu/E) Acid glutamic	GGG (Gly/G) Glicină	G

Codonii ARN care codifică cei 20 de aminoacizi

2. Să se genereze o secvenţă de ADN aleatoare sub forma unui şir de caractere sau a unei liste de tuppluri (ca la 1.), având n=3*k nucleotide. Ştiind că un codon codifică un aminoacid, să se construiască secvenţa proteică corespunzatoare secvenţei de ADN generate, sub forma unui şir de caractere in care aminoacizii sunt separati printr-un spatiu. Să se afişeze înşiruirea de aminoacizi în două moduri:

a. utilizând denumirile lor prescurtate din codul genetic standard:

```
stdcode = {
'TTT':'Phe', 'TTC':'Phe', 'TTA':'Leu', 'TTG':'Leu',
'TCT':'Ser', 'TCC':'Ser', 'TCA':'Ser', 'TCG':'Ser',
'TAT':'Tyr', 'TAC':'Tyr', 'TAA':'Stop', 'TAG':'Stop',
'TGT':'Cys', 'TGC':'Cys', 'TGA':'Stop', 'TGG':'Trp',

'CTT':'Leu', 'CTC':'Leu', 'CTA':'Leu', 'CTG':'Leu',
'CCT':'Pro', 'CCC':'Pro', 'CCA':'Pro', 'CCG':'Pro',
'CAT':'His', 'CAC':'His', 'CAA':'Gln', 'CAG':'Gln',
'CGT':'Arg', 'CGC':'Arg', 'CGA':'Arg', 'CGG':'Arg',

'ATT':'Ile', 'ATC':'Ile', 'ATA':'Ile', 'ATG':'Met',
'ACT':'Thr', 'ACC':'Thr', 'ACA':'Thr', 'ACG':'Thr',
'AAT':'Asn', 'AAC':'Asn', 'AAA':'Lys', 'AAG':'Lys',
'AGT':'Ser', 'AGC':'Ser', 'AGA':'Arg', 'AGG':'Arg',

'GTT':'Val', 'GTC':'Val', 'GTA':'Val', 'GTG':'Val',
'GCT':'Ala', 'GCC':'Ala', 'GCA':'Ala', 'GCG':'Ala',
'GAT':'Asp', 'GAC':'Asp', 'GAA':'Glu', 'GAG':'Glu',
'GGT':'Gly', 'GGC':'Gly', 'GGA':'Gly', 'GGG':'Gly'
}
```

b. utilizând codificarea standard de o literă pentru aminoacizi – în acest caz, dacă în secvenţa de ADN se generează codoni de stop care nu codifică aminoacizi, se va afişa secvenţa proteică până la codonul de Stop.

```
aminocode = {
'Gly' : 'G', 'Ala' : 'A', 'Pro' : 'P', 'Val' : 'V',
'Ile' : 'I', 'Leu' : 'L', 'Phe' : 'F', 'Met' : 'M',
'Ser' : 'S', 'Cys' : 'C', 'Thr' : 'T', 'Asn' : 'N',
'Gln' : 'Q', 'His' : 'H', 'Tyr' : 'Y', 'Trp' : 'W',
'Asp' : 'D', 'Glu' : 'E', 'Lys' : 'K', 'Arg' : 'R'
}
```

Observaţie. Se poate utiliza funcţia sir.split(' ') care întoarce o listă cu subşirurile din şir separate de un separator dat ca parametru (în acest caz, separatorul este un spaţiu).

3. Se da o secvenţă de ARN sub forma unei liste de tupluri generata aleator. Să se construiască secvenţa proteică corespunzatoare. Un tuplu va contine 3 simboluri, fiind corespunzator unui codon ARN alcatuit din 3 nucleotide.

```
[(A,U,C), (G,A,A), (C,A,A), (U,C,A), (A,G,C)]
```

IV. Formate de fişiere pentru secvenţe şi baze de date biologice

1. Noţiuni introductive

Bioinformatica se ocupă cu prelucrarea cantităţii enorme de date biologice. În urma secvenţierii unui genom sau a unei proteine, secvenţa respectivă, alcatuită din cele patru nucleotide, în cazul unui genom, sau din aminoacizi, în cazul unei proteine, este stocată într-o bază de date publică pentru a putea fi studiată de numeroşi cercetători.

Există diferite programe de analiză şi prelucrare a secvenţelor biologice. Astfel, pentru ca datele referitoare la o secvenţă să poată circula între diferite aplicaţii informatice, au fost definite mai multe formate de fişiere utilizate în bioinformatică, precum:

FASTA
ASN. 1
GenBank
PDB
XML

Pentru a putea publica descoperirea unei noi gene sau a unui fragment genomic, trebuie trimisă secvenţa către bazele de date biologice publice pentru a putea obţine un număr de acces („accession number"). Numărul de acces este unic pentru fiecare secvenţă trimisă, pentru a putea fi identificată printre numeroasele secvenţe descoperite zilnic de cercetători din toată lumea.

2. Baza de date biologice

NCBI (National Center for Biotechnology Information) pune la dispoziţie baze de date biologice, împreună cu un SGBD (care include motorul de căutare *Entrez* şi diferite programe şi utilitare pentru gestiunea şi prelucrarea înregistrărilor): http://www.ncbi.nlm.nih.gov

Secvenţele de nucleotide (ADN) sau cele de aminoacizi (proteine) au un număr de acces unic într-o bază de date. Motorul de căutare *Entrez* caută informaţii disponibile în mai multe baze de date, astfel se poate obţine aceeaşi secvenţă de mai multe ori cu numere de acces diferite. NCBI pune la dispoziţie o bază de date *RefSeq* în care sunt incluse informaţii non-redundante.

Secvenţele din *RefSeq* au numere de acces de forma:
NT_123456 – regiuni de cromozomi;
NC_123456 – cromozom;
NP_123456 – proteina;
NG_123456 – gena;
Genomurile secvenţiate până la ora actuală pot fi găsite pe NCBI:
http://www.ncbi.nlm.nih.gov/sites/genome

a. Să se acceseze următoarea adresă de la NCBI (National Center for Biotechnology Information): http://www.ncbi.nlm.nih.gov/guide/genomes-maps/

b. Să se acceseze baza de date *Genome*.

c. Să se acceseze un grup de organisme şi, apoi, un număr de acces (*accession number)* pentru un genom cromozomial al unui organism, de exemplu: NC_005562 .

d. Să se acceseze baza de date *RefSeq*: NC_005562.

e. Să se salveze informaţia găsită, în formatele: FASTA, ASN. 1, GenBank.

f. Să se salveze doua secvente genomice diferite (unul procariot si unul eucariot) în fiecare din cele trei formate.