DNA

Aquesta tasca se centra en les **llistes** i el **processament de fitxers/textos**. Crearàs un fitxer anomenat dna.py. També necessitaràs els dos fitxers d'entrada dna.txt i ecoli.txt, disponibles al campus virtual de l'assignatura.. Desa aquests fitxers a la mateixa carpeta que el teu programa.

La tasca consisteix a processar dades de fitxers de genomes. El teu programa hauria de funcionar amb els dos fitxers d'entrada proporcionats. Si tens curiositat (no és necessari), el National Center for Biotechnology Information publica molts altres fitxers de genomes de bacteris. A la darrera pàgina s'explica com fer servir el teu programa per processar altres fitxers de genomes publicats.

Informació sobre l'ADN:

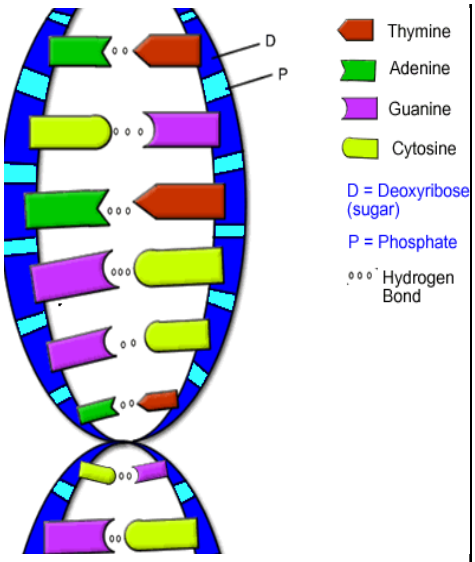
Nota: Aquesta secció explica informació de l'àmbit de la biologia relacionada amb aquesta tasca.
És només per a la teva informació; no cal entendre-la completament per completar la tasca.

L'àcid desoxiribonucleic (ADN) és una macromolècula bioquímica complexa que transporta informació genètica per a formes de vida cel·lulars i alguns virus. L'ADN és també el mecanisme mitjançant el qual es transmet la informació genètica dels progenitors durant la reproducció. L'ADN està format per llargues cadenes de compostos químics anomenats nucleòtids. Hi ha quatre nucleòtids presents en l'ADN: Adenina (A), Citosina (C), Guanina (G) i Timina (T). L'ADN té una estructura de doble hèlix (vegeu el diagrama a continuació) que conté cadenes complementàries d'aquests quatre nucleòtids connectades per enllaços d'hidrogen.

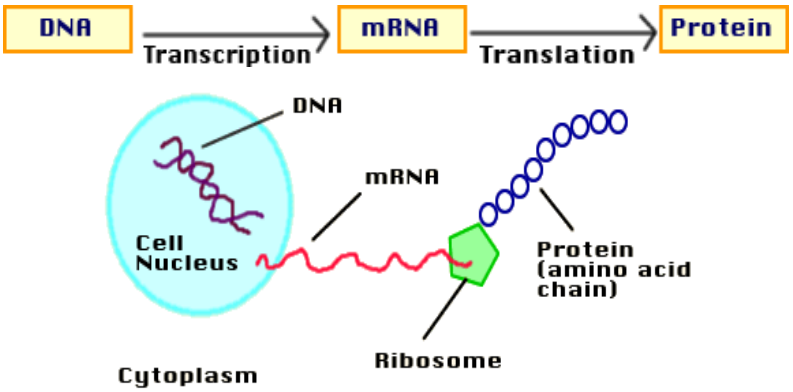
Algunes regions de l'ADN s'anomenen gens. La majoria dels gens codifiquen instruccions per a la construcció de proteïnes (són anomenats "gens codificadors de proteïnes"). Aquestes proteïnes són responsables de portar a terme la major part dels processos vitals de l'organisme.

Els nucleòtids d'un gen s'organitzen en *codons*. Els *codons* són grups de tres nucleòtids i s'escriuen amb les primeres lletres dels seus nucleòtids (p. ex., TAC o GGA). Cada *codon* codifica de manera única un sol aminoàcid, un bloc de construcció de les proteïnes.

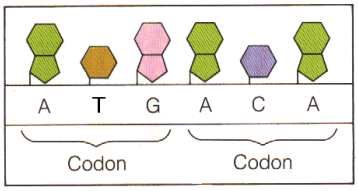
El procés de construcció de proteïnes a partir de l'ADN té dues fases principals anomenades transcripció i traducció. Durant aquestes fases, un gen es replica en una forma intermèdia anomenada ARN missatger (ARNm), que després és processada per una estructura anomenada ribosoma per construir la cadena d'aminoàcids codificada pels *codons* del gen.



The chemical structure of DNA.



DNA translation.



Les seqüències d'ADN que codifiquen proteïnes es troben entre un *codon* d'inici (que assumirem que és **ATG**) i un *codon* de parada (que pot ser qualsevol dels següents: **TAA**, **TAG** o **TGA**). No totes les regions de l'ADN són gens; hi ha grans porcions que no es troben entre un *codon* d'inici i un *codon* de parada vàlid. Aquestes regions es coneixen com a ADN intergènic i tenen altres funcions (possiblement desconegudes). Els biòlegs computacionals analitzen grans fitxers de dades d'ADN per trobar patrons i informació rellevant, com ara quines regions són gens. De vegades, també els interessa calcular el percentatge de massa que correspon a cadascun dels quatre tipus de nucleòtids. Sovint, percentatges elevats de Citosina (**C**) i Guanina (**G**) són indicadors de dades genètiques importants.

Per obtenir més informació, visita la pàgina de Wikipedia sobre l'ADN:
http://en.wikipedia.org/wiki/DNA

En aquesta tasca, llegiràs un fitxer d'entrada que conté seqüències de nucleòtids identificades amb un nom i produiràs informació sobre elles. Per a cada seqüència de nucleòtids, el teu programa haurà de:

1. Comptar les ocurrències de cadascun dels quatre nucleòtids (**A**, **C**, **G** i **T**).
2. Calcular el percentatge de massa ocupada per cada tipus de nucleòtid, arrodonit a una xifra decimal.
3. Informar sobre els *codons* (grups de tres nucleòtids) presents a cada seqüència.
4. Predir si la seqüència és un gen codificador de proteïnes.

Per a nosaltres, un gen codificador de proteïnes és una cadena que compleix totes les condicions següents:*

- Comença amb un ***codon* d'inici vàlid** (**ATG**).
- Acaba amb un ***codon* de parada vàlid** (un dels següents: **TAA**, **TAG** o **TGA**).
- Conté **almenys 5 *codons* en total** (incloent-hi el *codon* d'inici i el *codon* de parada final).
- La **Citosina (C)** i la **Guanina (G)** combinades representen **almenys el 30% de la seva massa total**.

(*Aquests són aproximacions per a la nostra tasca, no condicions exactes utilitzades en biologia computacional per identificar proteïnes.)

Les dades d'entrada de l'ADN consisteixen en parells de línies. La primera línia conté el **nom de la seqüència de nucleòtids**, i la segona és la **seqüència de nucleòtids** pròpiament dita. Cada caràcter en una seqüència de nucleòtids pot ser **A**, **C**, **G**, **T** o un guió **"-"**. Els nucleòtids de l'entrada poden estar en majúscules o minúscules.

Input file `dna.txt` (partial):

```
cure for cancer protein
ATGCCACTATGGTAG
captain picard hair growth protein
ATgCCAACATGgATGCCcGATAtGGATTgA
bogus protein
CCATt-AATgATCa-CAGTt
```

...

Els caràcters guió **"-"** representen regions "residus" en la seqüència. En la major part del programa, aquests caràcters s'han d'ignorar en els teus càlculs, tot i que sí que contribueixen a la **massa total de la seqüència**, tal com s'explicarà més endavant.

Comportament del programa:

El teu programa comença amb una introducció i demana els noms dels fitxers d'entrada i sortida. Pots assumir que l'usuari introduirà el nom d'un fitxer d'entrada existent i en el format correcte. El programa llegeix el fitxer d'entrada per processar les seves seqüències de nucleòtids i escriu els resultats al fitxer de sortida indicat.

Tingues en compte que:

La **seqüència de nucleòtids** es mostra en **majúscules** a la sortida.
Els **comptatges de nucleòtids** i els **percentatges de massa** es presenten en l'ordre **A, C, G, T**.
Un mateix ***codon***, com ara **GAT**, pot aparèixer més d'una vegada en la mateixa seqüència.

Log d'execució (l'entrada de l'usuari està subratllada):

```
This program reports information about DNA
nucleotide sequences that may encode proteins.
Input file name? dna.txt
Output file name? output.txt
```

Output file output.txt després de l'execució anterior (parcial):

```
Region Name: cure for cancer protein
Nucleotides: ATGCCACTATGGTAG
Nuc. Counts: [4, 3, 4, 4]
Total Mass%: [27.3, 16.8, 30.6, 25.3] of 1978.8
Codons List: ['ATG', 'CCA', 'CTA', 'TGG', 'TAG']
Is Protein?: YES

Region Name: captain picard hair growth protein
Nucleotides: ATGCCAACATGGATGCCCGATATGGATTGA
```

```
Nuc. Counts: [9, 6, 8, 7]
Total Mass%: [30.7, 16.8, 30.5, 22.1] of 3967.5
Codons List: ['ATG', 'CCA', 'ACA', 'TGG', 'ATG', 'CCC', 'GAT', 'ATG', 'GAT', 'TGA']
Is Protein?: YES

Region Name: bogus protein
Nucleotides: CCATT-AATGATCA-CAGTT
Nuc. Counts: [6, 4, 2, 6]
Total Mass%: [32.3, 17.7, 12.1, 29.9] of 2508.1
Codons List: ['CCA', 'TTA', 'ATG', 'ATC', 'ACA', 'GTT']
Is Protein?: NO
```

Directrius d'implementació, consells i estratègia de desenvolupament:

L'objectiu principal d'aquesta tasca és demostrar la teva comprensió de les llistes i dels recòrrer-les amb bucles for. Per tant, has d'utilitzar llistes per emmagatzemar les dades de cada seqüència. En particular, els teus comptatges de nucleòtids, percentatges de massa i *codons* haurien de ser emmagatzemats utilitzant llistes. A més, has d'utilitzar llistes i bucles for per transformar les dades d'una forma a una altra, com segueix:

1. De la cadena original de la seqüència de nucleòtids als comptatges de nucleòtids.
2. De comptatges de nucleòtids a percentatges de massa.
3. De la cadena original de la seqüència de nucleòtids a triplets de *codons*.

Aquestes transformacions es resumeixen en el següent diagrama utilitzant les dades de la proteïna "*cure for cance*r":

Nucleotides:	`"ATGCCACTATGGTAG"`	
↓	What is computed	Output to file
Counts:	`{4, 3, 4, 4}`	`Nuc. Counts: [4, 3, 4, 4]`
↓		
Mass %:	`{27.3, 16.8, 30.6, 25.3}`	`Total Mass%: [27.3, 16.8, 30.6, 25.3] of 1978.8`
↓		
Codons:	`{ATG, CCA, CTA, TGG, TAG}`	`Codons List: [ATG, CCA, CTA, TGG, TAG]` `Is protein?: YES`

Recorda que pots imprimir qualsevol llista utilitzant **str()**. Per exemple:

```
numbers = [10, 20, 30, 40]
print("my data is " + str(numbers))      # my data is [10, 20, 30, 40]
```

Per calcular els percentatges de massa, utilitza les següents masses de cada nucleòtid (grams/mol). Els guions que representen regions "escombraries" s'exclouen de molts dels teus càlculs, però sí que contribueixen a la massa total.

- **Adenina** (A): 135.128
- **Citosina** (C): 111.103
- **Guanina** (G): 151.128
- **Timina** (T): 125.107
- **Residu** (-): 100.000

Per exemple, la massa de la seqüència **ATGG-AC** és (135.128 + 125.107 + 151.128 + 151.128 + 100.000 + 135.128 + 111.103) o **908.722**. D'aquesta massa, 270.256 (29.7%) provenen de les dues **Adenines**; 111.103 (12.2%) provenen de la **Citosina**; 302.256 (33.3%) provenen de les dues **Guanines**; 125.107 (13.8%) provenen de la **Timina**; i 100.000 (11.0%) provenen del "residu" guió.

Et suggerim que comencis aquest programa escrivint el codi per llegir el fitxer d'entrada. Prova d'escriure codi per llegir simplement el nom de cada proteïna i la seva seqüència de nucleòtids i imprimir-los.

A continuació, escriu codi per recórrer una seqüència de nucleòtids i comptar el nombre d'**A**, **C**, **G** i **T**. Desa els teus comptatges en una llista de mida 4. Per mapar entre els nucleòtids i els índexs de la llista, potser voldràs escriure una funció que converteixi un caràcter individual (és a dir, **A**, **C**, **T**, **G**) en índexs (és a dir, de 0 a 3).

Un cop tinguis els comptatges funcionant correctament, pots convertir aquests comptatges en una nova llista de **percentatges de massa** per a cada nucleòtid utilitzant les masses dels nucleòtids que hem donat prèviament. Si has escrit codi per mapar entre les lletres dels nucleòtids i els índexs de la llista, també et pot ajudar a consultar les masses en una llista com la següent:

```
masses = [135.128, 111.103, 151.128, 125.107]
```

Pots desar els percentatges de massa ja arrodonits a una xifra decimal o arrodonir-los al moment de mostrar la llista de percentatges de massa

Recorda que els guions ("junk") sí que contribueixen a la massa total. Per a altres parts del programa, potser voldràs eliminar els guions de l'entrada.

Després de calcular els percentatges de massa, hauràs de separar la seqüència en ***codon*s** i examinar cada *codon*. Potser et serà útil revisar funcions de string com **upper()**, i **lower()**.

També et suggerim que primer facis funcionar el teu programa correctament imprimint la sortida a la consola abans de desar-la al fitxer de sortida.

Pots assumir que el fitxer d'entrada existeix, és llegible i conté dades vàlides. (En altres paraules, no has de tornar a demanar els noms dels fitxers d'entrada o sortida). També pots assumir que el nombre de nucleòtids de cada seqüència (sense els guions) serà un múltiple de 3, tot i que els nucleòtids en una línia poden estar en majúscules, minúscules o una combinació d'ambdues. El teu programa hauria d'*overwrite* qualsevol dada existent en el fitxer de sortida.

Directrius d'estil:

Per aquesta tasca, es requereix que tinguis les següents quatre constants:

1. **Una per al nombre mínim de *codons* que una proteïna vàlida ha de tenir**, com a enter (per defecte, 5).

2. **Una segona per al percentatge de massa de C i G necessari perquè una proteïna sigui vàlida**, com a enter (per defecte, 30).
3. **Una tercera per al nombre de nucleòtids únics** (4, representant A, C, G i T).
4. **Una quarta per al nombre de nucleòtids per *codon*** (3).

Hauria de ser possible canviar els dos primers valors constants (mínim de *codons* i percentatge mínim de massa) i fer que el teu programa canviï el seu comportament per avaluar la validesa de les proteïnes. Les altres dues constants no canviaran mai, però són útils per fer que el teu codi sigui més llegible. Fes referència a aquestes constants en el teu codi i no facis servir números directament com el 4 o el 3. Pots utilitzar constants addicionals si això fa que el teu codi sigui més clar.

Estructura de funcions:
Per aquesta tasca, es requereix que utilitzis almenys **quatre funcions no trivials**, a més de la funció main. Aquestes funcions han d'utilitzar paràmetres i retorns, incloent llistes, quan sigui necessari. Les funcions han d'estar ben estructurades i evitar la redundància. Cap funció ha de fer una part massa gran de la tasca global. No es permet anidar funcions dins d'altres ni dins de la funció main.

En particular, es requereix que tinguis la següent funció al teu programa:

- **Una funció per imprimir tota la sortida al fitxer per a una proteïna potencial** (nucleòtids, comptatges, percentatges, si és una proteïna, etc.).

Dit d'una altra manera, tota la sortida al fitxer ha de ser feta mitjançant una única funció cridada per cada seqüència de nucleòtids de l'entrada. La resta de funcions han de fer els càlculs per obtenir la informació que serà passada a aquesta funció d'impressió.

La funció main ha de ser un resum concís del programa global. Està bé que main contingui algun codi com per exemple instruccions d'impressió. Però main no ha de fer una part massa gran de la tasca en si mateixa, com per exemple examinar cada caràcter d'una línia d'entrada. A més, evita la **"cadena"** de funcions, és a dir, quan moltes funcions es criden entre si sense mai tornar a main.

Si tens un tros de codi molt similar repetit diverses vegades al teu programa, elimina la redundància, com per exemple creant una funció, utilitzant bucles **for** sobre els elements de les llistes, i/o separant els blocs **if/else**.

Segueix les directrius d'estil passades com la **indentació**, els **noms de les variables**, les **llargades de les línies**, i els **comentaris** (al principi del teu programa, a cada funció i a seccions complexes del codi). No es permet l'ús de **variables globals**.

Fitxers d'entrada addicionals (opcionals):
Si vols generar fitxers d'entrada addicionals per provar el teu programa, pots crear-los a partir de dades genètiques reals de l'NCBI. El següent lloc web conté molts fitxers de dades que inclouen genomes complets d'organismes bacterians:

ftp://ftp.ncbi.nih.gov/genomes/Bacteria/

Aquest lloc conté moltes subcarpetes amb noms d'organismes. Després d'entrar a una subcarpeta, pots trobar i desar un fitxer de genoma (un fitxer que acaba amb l'extensió `.fna`) i una taula de proteïnes (un fitxer que acaba amb l'extensió `.ptt`).