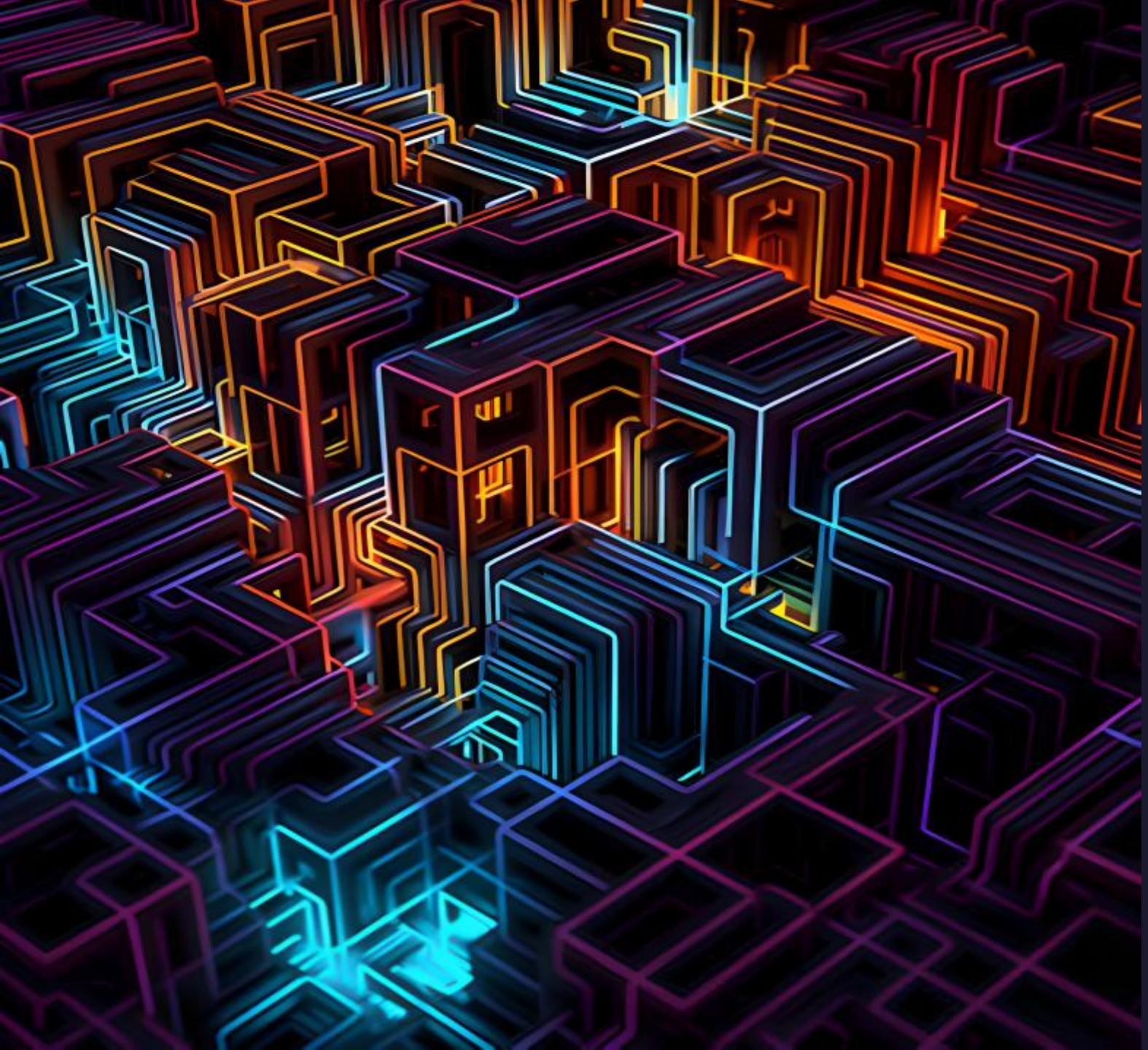


Situación problema: secuencias de ADN

Análisis y diseño de algoritmos avanzados

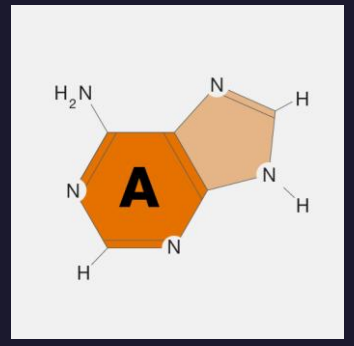
Dra. Valentina Narváez Terán

Tecnológico de Monterrey

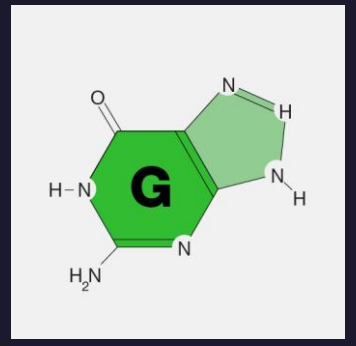
Moléculas del ADN

El ADN esta formado por cuatro moléculas, llamadas bases de nucleótidos

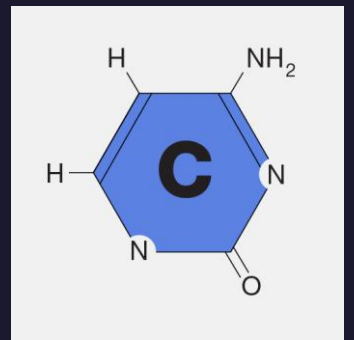
Adenina(A)



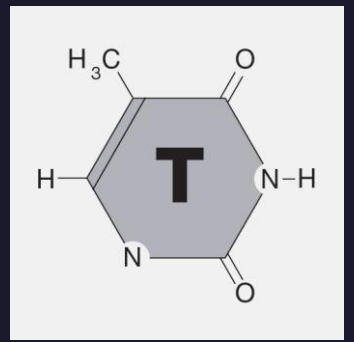
Guanina (G)



Citosina (C)



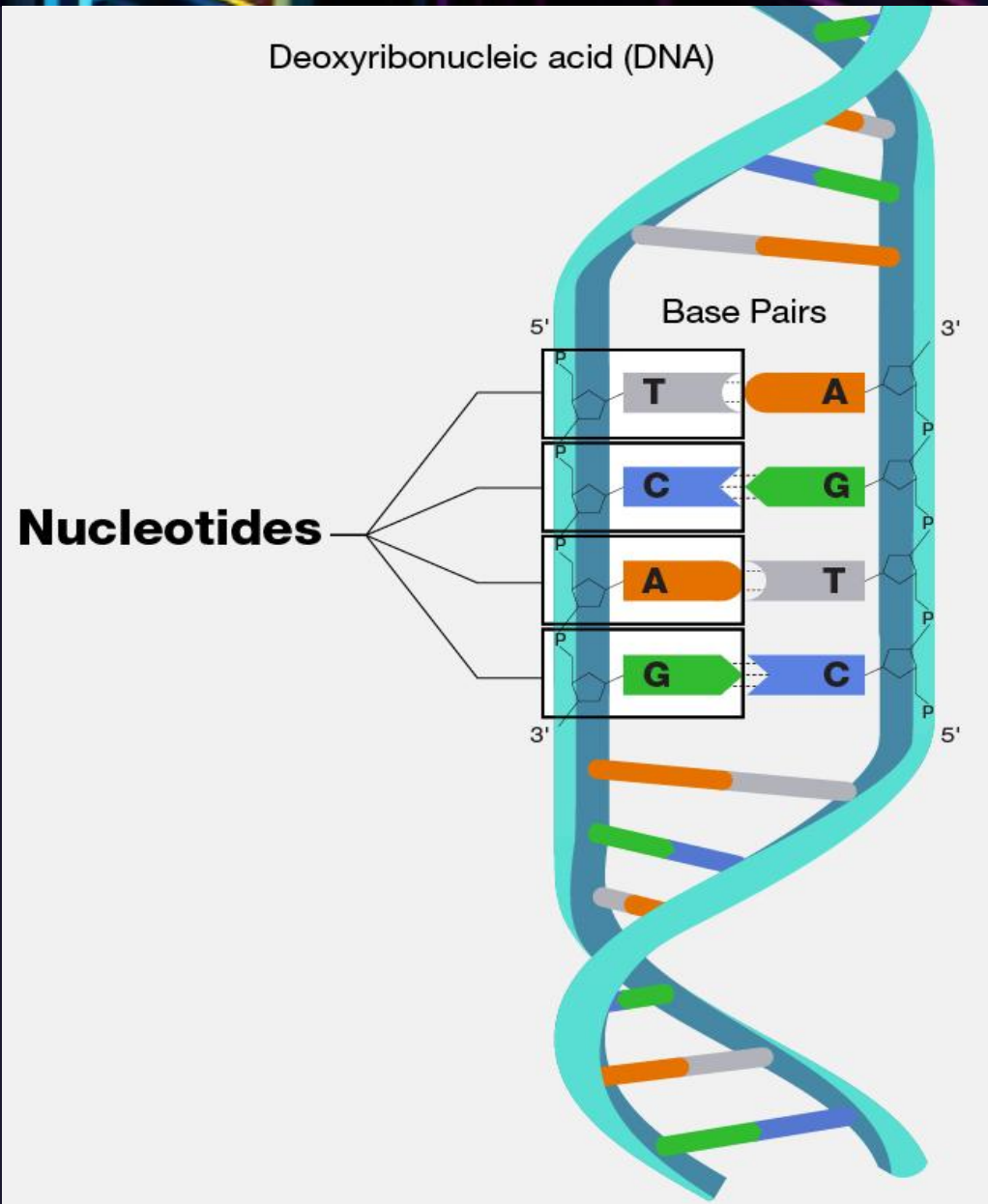
Timina (T)



Las bases siempre forman los mismos pares:

adenina con timina

citosina con guanina



Moléculas del ADN

Las secuencias de ADN se representan como largas cadenas formadas por estas cuatro letras

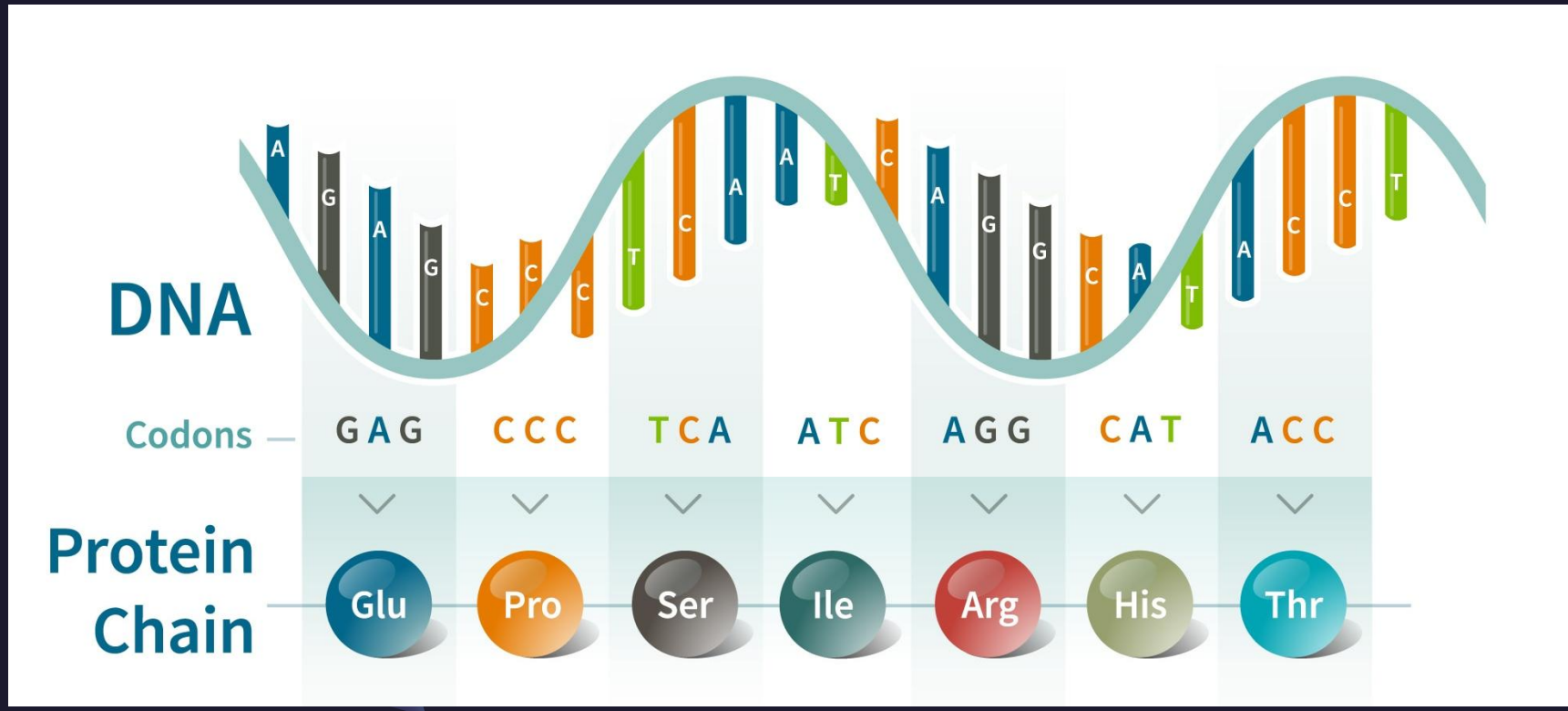
Dado que las bases siempre se emparejan igual, no es necesario representar todos los pares, basta con una letra por par

De nucleótidos a aminoácidos

Cada grupo de 3 letras forman un "codon"

Puedes imaginarlos como códigos, o instrucciones que hacen que los ribosomas de las células "impriman" aminoácidos

Los aminoácidos se encadenan, formando proteínas



De nucleótidos a aminoácidos

Tabla de aminoácidos

Sirve para traducir cuales tripletas forman cada aminoácido

Amino acid	DNA codons	Amino acid	DNA codons
Ala, A	GCT, GCC, GCA, GCG	Ile, I	ATT, ATC, ATA
Arg, R	CGT, CGC, CGA, CGG; AGA, AGG	Leu, L	CTT, CTC, CTA, CTG; TTA, TTG
Asn, N	AAT, AAC	Lys, K	AAA, AAG
Asp, D	GAT, GAC	Met, M	ATG
Asn or Asp, B	AAT, AAC; GAT, GAC	Phe, F	TTT, TTC
Cys, C	TGT, TGC	Pro, P	CCT, CCC, CCA, CCG
Gln, Q	CAA, CAG	Ser, S	TCT, TCC, TCA, TCG; AGT, AGC
Glu, E	GAA, GAG	Thr, T	ACT, ACC, ACA, ACG
Gln or Glu, Z	CAA, CAG; GAA, GAG	Trp, W	TGG
Gly, G	GGT, GGC, GGA, GGG	Tyr, Y	TAT, TAC
His, H	CAT, CAC	Val, V	GTT, GTC, GTA, GTG
START	ATG	STOP	TAA, TGA, TAG

Nota que varias tripletas pueden resultar en el mismo aminoácido

Por ejemplo, CAT y CAC son Histidina (H)

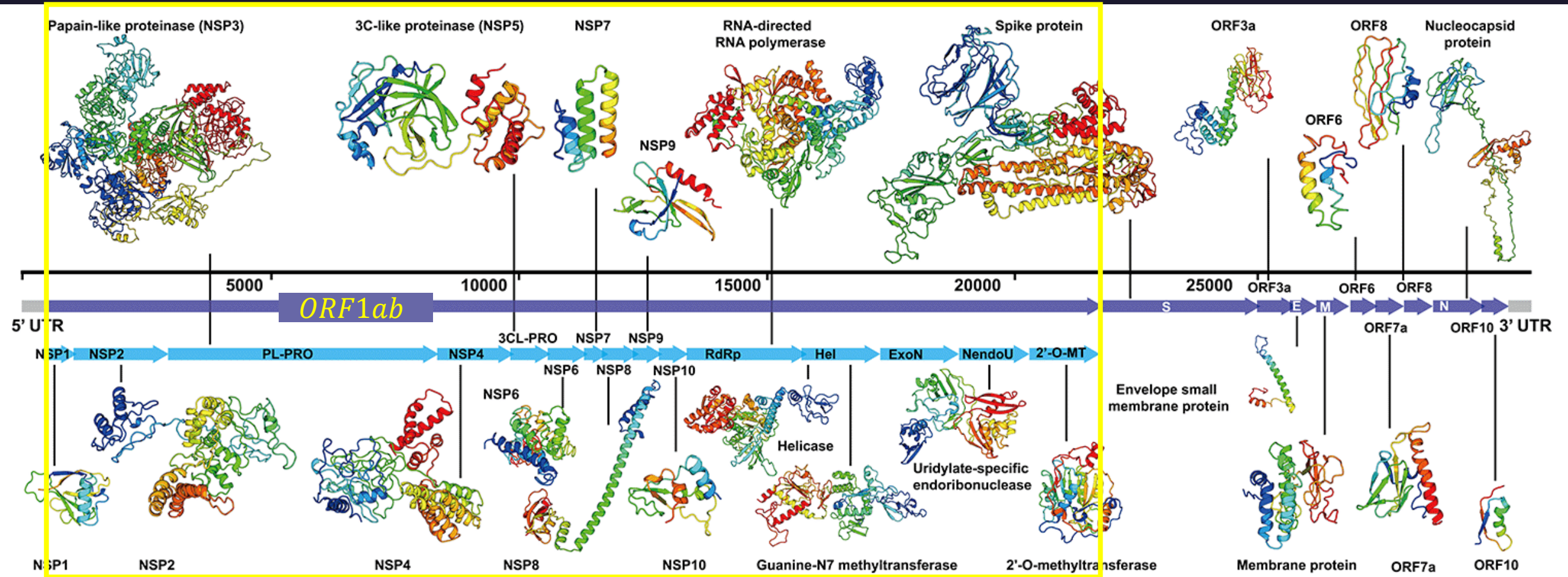
Amino acid code

A – Alanine	G – Glycine	M – Methionine	S – Serine
C – Cysteine	H – Histidine	N – Asparagine	T – Threonine
D – Aspartic acid	I – Isoleucine	P – Proline	V – Valine
E – Glutamic acid	K – Lysine	Q – Glutamine	W – Tryptophan
F – Phenylalanine	L – Leucine	R – Arginine	Y – Tyrosine

El código genético de SARS-COV-2

Cada gen induce la creación de varias proteínas

Por ejemplo, el gen *ORF1ab*, abarca la sección resaltada en rojo, y contiene instrucciones para varias proteínas



Situación problema

Un pobre virólogo perdió parte de su investigación. Ayúdale a recuperarla.

- El archivo SARS-COV-2-MN908947.3.txt contiene la secuencia genómica de SARS-COV-2, de Wuhan, 2019
- Los siguientes archivos contienen las secuencias de tres genes
  - Gen M: gen-M.txt
  - Gen ORF1ab: gen-ORF1AB.txt
  - Gen S: gen-S.txt
- El archivo seq-proteins.txt contiene las secuencias de aminoácidos de las proteínas del virus
- El archivo SARS-COV-2-MT106054.1.txt se obtuvo de muestras de Texas, en 2020

1. ¿Cómo encontrar genes en la secuencia del virus?
2. Dentro de un gen, los palíndromos son importantes, porque son regiones propensas a mutaciones. Dado un gen, determina la sección propensa a mutaciones mas larga bajo este criterio
3. ¿Cómo encontrar las secciones del virus donde se produce una proteína?
4. Compara las versiones del genoma del virus de Wuhan, 2019 vs Texas, 2020.
   ¿Dónde son iguales?
   ¿Donde que difieren?
   ¿Las diferencias resultan en aminoácidos diferentes?

Situación problema: entregables

Código Python o C/C++ (80pts, divididos en 20pts por cada funcionalidad)

- Las funciones de comparación de strings y búsqueda de palíndromos deben ser programadas por ustedes

  NO pueden usar str.find o similares

Video: (5pts)

- Aprox. 10 min, donde participe todos, expliquen que hicieron y muestren resultados

Reporte: (15pts)

- Explica el problema, que algoritmos y estructuras de datos usaste. Incluye breves reflexiones individuales al final sobre como fue tu experiencia resolviéndola

¿Cuándo?

4 de octubre (semana 3 de periodo 2)

En parejas

Revisa Canvas para saber que debe dar como salidas.

Situación problema

El archivo del virus contiene la secuencia del virus, en formato fasta

La primera línea describe a la secuencia

El resto son las letras que representan las bases de nucleótidos: Adenina, Citosina, Guanina, Timina

SARS-COV-2-MN908947.3.txt

```
>MN908947.3 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAA
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACTCACGCAGTATAATTAATAAC
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTG
TTGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTC
CCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTAC
GTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGG
```

Situación problema

Los archivos de cada gen tienen el mismo formato

¿Qué algoritmos puedes utilizar para saber donde ocurre un gen especifico en la secuencia del virus?

gen-ORF1AB.txt

```
>MN908947.3:266-21555 Severe acute respiratory syndrome coronavirus 2 isolate
Wuhan-Hu-1, complete genome
ATGGAGAGCCTTGTCCCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTC
GCGACGTGCTCGTACGTGGCTTTGGAGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAA
AGATGGCACTTGTGGCTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTC
ATCAAACGTTCGGATGCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAG
GCATTCAGTACGGTCGTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGGCGAAATACCAGTGGC
TTACCGCAAGGTTCTTCTTCGTAAGAACGGTAATAAAGGAGCTGGTGGCCATAGTTACGGCGCCGATCTA
AAGTCATTTGACTTAGGCGACGAGCTTGGCACTGATCCTTATGAAGATTTTCAAGAAAACTGGAACACTA
AACATAGCAGTGGTGTTACCCGTGAACTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGA
```

Situación problema

seq-proteins.txt

```
>QHD43415_1
MESLVPGFNEKTHVQLSLPVLQVRDVLVRGFGDSVEEVLSEARQHLKDGTCGLVEVEKGVLPQLEQPYVFIKRSDA
VAELEGIQYGRSGETLGVLVPHVGEIPVAYRKVLLRKNGNKGAGGHSYGADLKSFDLGDELGTDPYEDFQENWNTK
LNGG
>QHD43415_2
AYTRYVDNNFCGPDGYPLECIKDLLARAGKASCTLSEQLDFIDTKRGVYCCREHEHEIAWYTERSEKSYELQTPFE
ECPNFVFPLNSIIKTIQPRVEKKKLDGFMGRIRSVYPVASPNECNQMCLSTLMKCDHCGETSWQTGDFVKATCEFC
CGYLPQNAVVKIYCPACHNSEVGPEHSLAEYHNESGLKTILRKGGRTIAFGGCVFSYVGCHNKCAYWVPRASANIG
GLNDNLLEILQKEKVNINIVGDFKLNEEIAIILASFSASTSAFVETVKGLDYKAFKQIVESCGNFKVTKGKAKKGA
LYAFASEAARVVRSIFSRTLETAQNSVRVLQKAAITILDGISQYSLRLIDAMMFTSDLATNNLVVMAYITGGVVQL
YEKLKPVLDWLEEKFKEGVEFLRDGWEIVKFISTCACEIVGGQIVTCAKEIKESVQTFFKLVNKFLALCADSIIIG
FVTHSKGLYRKCVKSREETGLLMPLKAPKEIIFLEGETLPTEVLTEEVVLKTGDLQPLEQPTSEAVEAPLVGTPVC
INGLMLLEIRDTEKYCALAPNMMVTNNTFTLKGG
>QHD43415_3
APTKVTFGDDTVIEVQGYKSVNITFELDERIDKVLNEKCSAYTVELGTEVNEFACVVADAVIKTLQPVSELLTPLG
IDELMQMNSCSFYPPDEDEEEGDCEEEEFEPSTQYEYGTEDDYQGKPLEFGATSAALQPEEEQEEDWLDDDSQQA
LFDESGEFKLASHMYCSFYPPDEDEEEGDCEEEEFEPSTQYEYGTEDDYQGKPLEFGATSAALQPEEEQEEDWLDD
EDNQTTTIQTIVEVQPQLEMELTPVVQTIEVNSFSGYLKLTDNVYIKNADIVEEAKKVKPTVVVNAANVYLKHGGG
MQVESDDYIATNGPLKVGGSCVLSGHNLAKHCLHVVGPNVNKGEDIQLLKSAYENFNQHEVLLAPLLSAGIFGADP
TNVYLAVFDKNLYDKLVSSFLEMKSEKQVEQKIAEIPKEEVKPFITESKPSVEQRKQDDKKIKACVEEVTTTLEET
```

El archivo seq-proteínas.txt contiene las cadenas de aminoácidos de las proteínas. Por cada proteína se lista su nombre, seguido de los aminoácidos que la forman

La primera es:
>QHD43415_1
MESLVPGFNEKTHVQLS…

La segunda es:
>>QHD43415_2
AYTRYVDNNFCGPDGYPL…

Y la tercera:
>QHD43415_3
APTKVTFGDDTV…

Situación problema

Amino acid	DNA codons	Amino acid	DNA codons
Ala, A	GCT, GCC, GCA, GCG	Ile, I	ATT, ATC, ATA
Arg, R	CGT, CGC, CGA, CGG; AGA, AGG	Leu, L	CTT, CTC, CTA, CTG; TTA, TTG
Asn, N	AAT, AAC	Lys, K	AAA, AAG
Asp, D	GAT, GAC	Met, M	ATG
Asn or Asp, B	AAT, AAC; GAT, GAC	Phe, F	TTT, TTC
Cys, C	TGT, TGC	Pro, P	CCT, CCC, CCA, CCG
Gln, Q	CAA, CAG	Ser, S	TCT, TCC, TCA, TCG; AGT, AGC
Glu, E	GAA, GAG	Thr, T	ACT, ACC, ACA, ACG
Gln or Glu, Z	CAA, CAG; GAA, GAG	Trp, W	TGG
Gly, G	GGT, GGC, GGA, GGG	Tyr, Y	TAT, TAC
His, H	CAT, CAC	Val, V	GTT, GTC, GTA, GTG
START	ATG	STOP	TAA, TGA, TAG

Amino acid code

A – Alanine	G – Glycine	M – Methionine	S – Serine
C – Cysteine	H – Histidine	N – Asparagine	T – Threonine
D – Aspartic acid	I – Isoleucine	P – Proline	V – Valine
E – Glutamic acid	K – Lysine	Q – Glutamine	W – Tryptophan
F – Phenylalanine	L – Leucine	R – Arginine	Y – Tyrosine

Para encontrar cual sección del virus produce a una proteína, debes decodificarla

Ejemplo:
La primera proteína es QHD43415_1, y sus primeros tres aminoácidos son

M E S

M es ATG
E es GAG… y también GAA
S podría ser TCT, TCC, TCA, AGT, o AGC

¿Qué estructura permitiría encontrar rápido los diferentes códigos posibles de un aminoácido?

Fuentes y recursos

Secuencias de Sars-Cov-2 y genes (tomados de MN908947.3)

https://www.ncbi.nlm.nih.gov/nuccore/MN908947.3
https://www.ncbi.nlm.nih.gov/nuccore/MT106054.1

Secuencias de las proteínas (corresponden a MN908947.3)
https://zhanggroup.org/COVID-19/index.html#download

Investiga:
¿Cómo se decide a partir de cual posición se comienzan a agrupar tripletas en codones?
Pistas: conceptos de marco de lectura (frame), UTR-5, UTR-3

¿Cómo calcular la subcadena común mas larga de dos cadenas?
Pista: programación dinámica y tabulación

¿En que podría ser útil en la situación problema?