Making sense of Sequences, in R

Taller de Bioinformática nivel intermedio

CodeMyGen by GenoBit

Sofia Acuña, Eduardo Cepeda, Angel Peña.

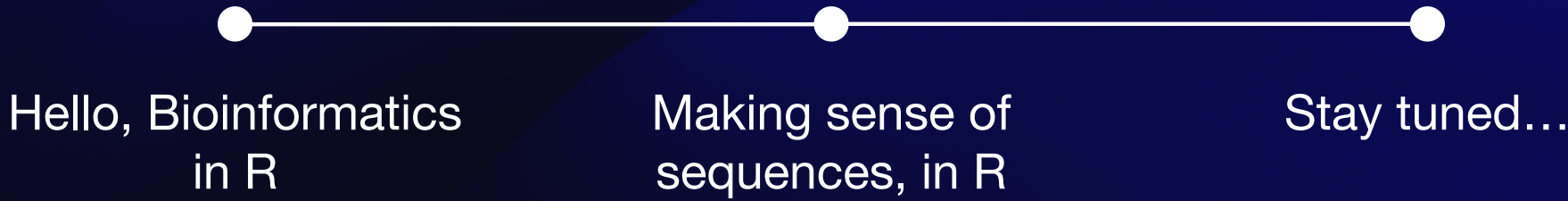
CodeMyGen

Talleres ofrecidos por **GenoBit: Genomics and Bioinformatics**
Para aquellos que buscan desarrollar habilidades de programación, aplicada a diferentes ámbitos.

Registro

Regístrate para obtener tu certificado.
Solo si no te registraste antes

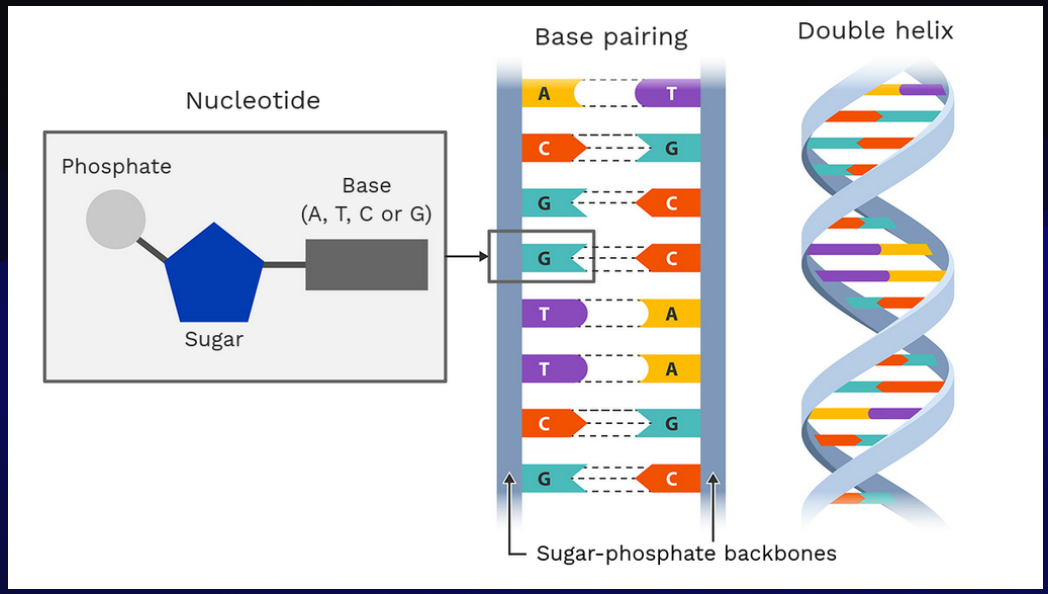
CodeMyGen
Ruta de aprendizaje



Overview

El ADN: El lenguaje de la vida

> Contiene las instrucciones para construir y mantener a los seres vivos.
> Se forma por una cadena de 4 nucleótidos: A (adenina), T (timina), C (citosina), G (guanina)
> El orden de estas bases define la información genética.
> Un gen es un fragmento que codifica una proteína.
> 3 bases ADN = 1 codon → 1 amino ácido

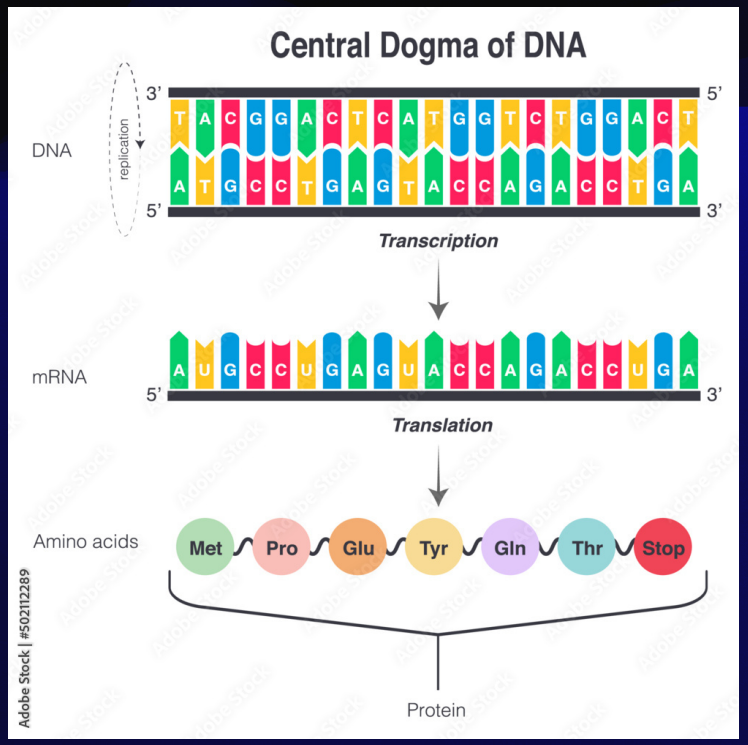


Overview

Del ADN a la proteína: cómo se expresa la información genética

> La información del ADN se expresa en dos pasos:
> 1. Transcripción: El ADN se copia en una molécula de ARN mensajero (mRNA).
> 2. Traducción: El mRNA se lee en grupos de tres bases (codones) y cada codón codifica un aminoácido.

> Traducción *in silico*:
Proceso computacional para convertir una secuencia de ADN o ARN en su correspondiente secuencia de aminoácidos (proteína).



Overview

¿Qué es un archivo FASTA?

> El formato FASTA es la manera estándar de guardar secuencias biológicas (ADN, ARN o proteínas).

> Es un archivo de texto muy simple que tiene:

- Línea de encabezado, que empieza con > y contiene el nombre o descripción de la secuencia.
- La secuencia (las letras A, T, G, C, o los aminoácidos).

Con R, podremos leer estos archivos FASTA, extraer información y visualizarla.

```
>NG_008679.1:5001-38170 Homo sapiens paired box 6 (PAX6)
ACCCTCTTTTCTTATCATTGACATTTAAACTCTGGGGCAGGTCCTCGCGTAGAACGCGGCTGTCAGATCT
GCCACTTCCCCTGCCGAGCGGCGGTGAGAAGTGTGGGAACCGGCGCTGCCAGGCTCACCTGCCTCCCCGC
CCTCCGCTCCCAGGTAACCGCCCGGGCTCCGGCCCCGGCCCGGCTCGGGGCCCGCGGGGCCTCTCCGCTG
CCAGCGACTGCTGTCCCCAAATCAAAGCCCGCCCCAAGTGGCCCCGGGGCTTGATTTTTGCTTTTAAAAG
GAGGCATACAAAGATGGAAGCGAGTTACTGAGGGAGGGATAGGAAGGGGGGTGGAGGAGGGACTTGTCTT
TGCCGAGTGTGCTCTTCTGCAAAAGTAGCAAAATGTTCCACTCCTAAGAGTGGACTTCCAGTCCGGCCCT
GAGCTGGGAGTAGGGGGCGGGAGTCTGCTGCTGCTGTCTGCTAAAGCCACTCGCGACCGCGAAAAATGCA
GGAGGTGGGGACGCACTTTGCATCCAGACCTCCTCTGCATCGCAGTTCACGACATCCACGCTTGGGAAAG
TCCGTACCCGCGCCTGGAGCGCTTAAAGACACCCTGCCGCGGGTCGGGCGAGGTGCAGCAGAAGTTTCCC
GCGGTTGCAAAGTGCAGATGGCTGGACCGCAACAAAGTCTAGAGATGGGGTTCGTTTCTCAGAAAGACGC
```

Overview

De una secuencia a la biología: qué podemos analizar con FASTA y R

> Con un archivo FASTA, podemos usar R para analizar distintos aspectos de una secuencia genética:

GC content:	**Porcentaje de bases Guanina (G) y Citosina (C).**
Motifs:	Pequeños patrones de secuencia que se repiten y suelen tener funciones biológicas importantes (por ejemplo, el codón de inicio ATG o regiones promotoras).
Traducción:	Conversión de la secuencia de ADN o ARN a aminoácidos, que forman una proteína.
Frecuencia de aminoácidos:	Permite conocer la composición y posibles propiedades de la proteína (hidrofobicidad, carga, estructura).

Descarga R y RStudio

> Descarga R

> Descarga RStudio para Windows o Mac

Carpeta de Drive con cuadernillos

> Descarga el archivo en el que trabajaremos
> CodeMyGen2.rmd

Case Analysis

Contexto:

Eres parte de un pequeño laboratorio de biología molecular que está analizando la secuencia de un gen candidato relacionado con la resistencia a antibióticos. Te entregan una secuencia FASTA y te piden hacer un análisis exploratorio básico

Objetivo: Comprender qué tan rica en GC es la secuencia.

Feedback