
Búsqueda de Expresiones Regulares

Con un enfoque en biología computacional

Disclaimer

**Esta presentación no pretende ser
precisa en términos biológicos,
químicos ni de ningún otro tipo.**



Objetivos

- **Expresiones regulares:**
¿Qué son y para qué sirven?.
- **Secuencias de nucleótidos:**
Un breve contexto.
- **Búsqueda:**
¿Por qué es útil?.

Texto:

Caracteres en donde se hace una búsqueda.

Patrón:

lo que estamos buscando.

Ocurrencia:

Cada vez que el patrón se encuentra en el texto.

Ejemplo:

Texto:

“Pepe Pecas pica papas con un pico, con un pico pica papas Pepe Pecas. Si Pepe Pecas pica papas con un pico, ¿dónde está el pico con que Pepe Pecas pica papas?”

Patrón:

“pic”

Ocurrencias: 8

“Pepe Pecas pica papas con un pico, con un pico pica papas Pepe Pecas. Si Pepe Pecas pica papas con un pico, ¿dónde está el pico con que Pepe Pecas pica papas?”

Expresión regular

Una expresión regular es una secuencia de caracteres que conforma un patrón de búsqueda.

Algo así como una forma reducida de escribir múltiples patrones en uno solo.

Ejemplo:

Patrones:

("pic" o "pec") o ("pep" o "pap")

Ocurrencias: 8,4,4,4.

"Pepe Pecas pica papas con un pico, con un pico pica papas Pepe Pecas. Si Pepe Pecas pica papas con un pico, ¿dónde está el pico con que Pepe Pecas pica papas?"

Operadores

Unión “ | ” (o)

Concatenación “ . ” (y)

Clausura de Kleene “ * ” (repetido)

Ejemplo:

Patrones:

("pic" o "pec") o ("pep" o "pap")

Expresión Regular:

p . (((i | e) . c) | ((e | a) . p))

Porque:

1 un solo recorrido del texto

soporta infinitos patrones



ER: A.(E*).P

Patrones:

AP

AEP

AEEP

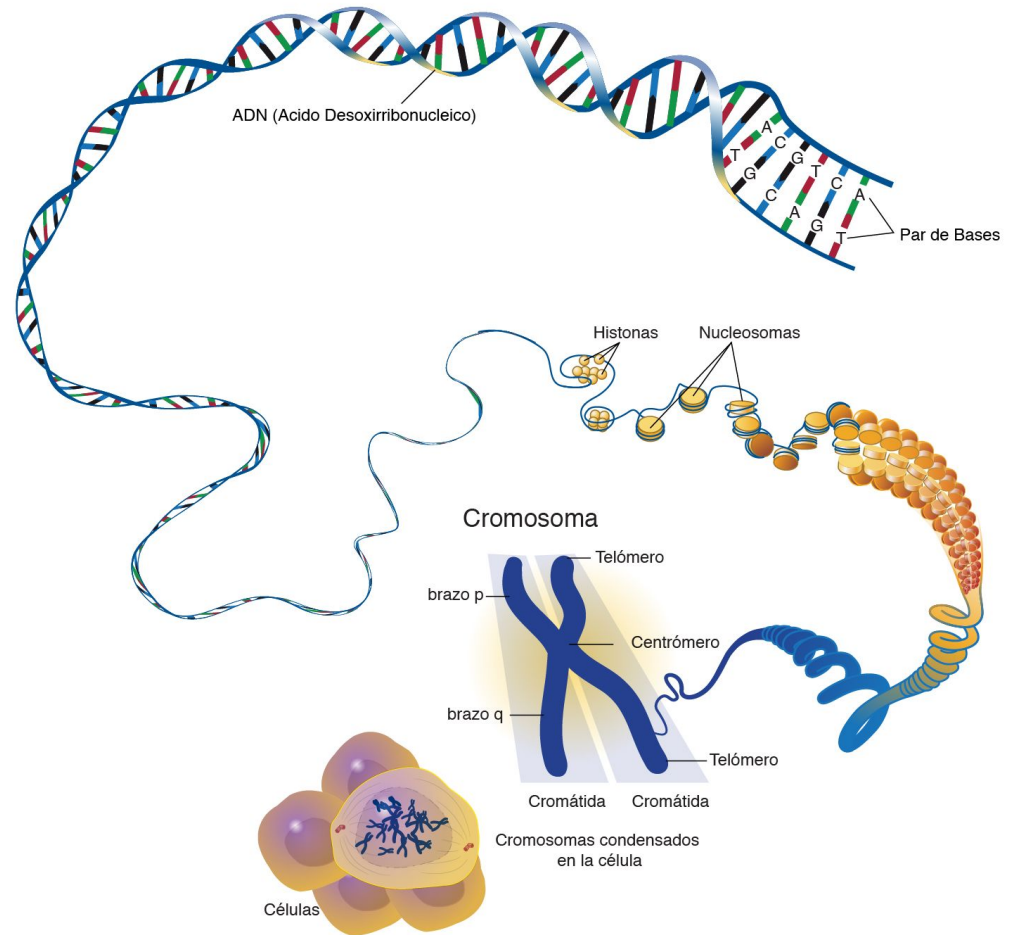
AEEEP

AEEEEEP

A.....P

Contexto.

Aminoácidos, proteínas,
genes, nucleótidos,
desoxirribosa,
ribonucleótidos,
procariotas, eucariotas



—

Contexto.

Un gen será una pequeña porción de ADN que contiene información en forma codificada necesaria para producir una determinada molécula, la cual cumplirá con una función definida en el ser vivo. Al conjunto del ADN codificante (los genes) y el ADN no codificante de un determinado organismo se le conoce como GENOMA.

El genoma humano
contiene aproximadamente
3 200 millones de bases y
está dividido en 23
cromosomas.

si cada b.n. fuera un
hombre, el núcleo de una
célula podría contener a
todos los hombres **del**
mundo

Bases nitrogenadas

Adenina

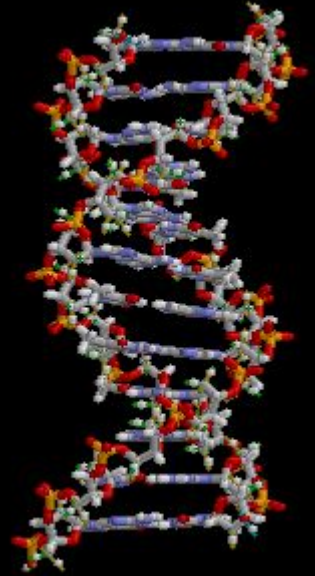
Timina

Guanina

Citosina

La Biblia contiene
3.5666.480 letras

Un solo genoma
podría imprimirse
en 90 biblias



The pathological severity of HD correlates with the number of (CAG)_n repeats in exon-1 of the gene *htt* which encodes the protein huntington. In Huntington's disease, a higher number of repeats means an earlier onset of disease and a more rapid disease progression. The CAG codon specifies glutamine, and HD belongs to a broad class of polyglutamine diseases. Healthy (wild-type) variants of this gene feature between 6–35 tandem repeats, whereas more than 35 repeats virtually assure the disease.

The Codon CAA also encodes glutamine.

(CAA|GAG)*

—

TAGNGG, se vuelve **TAG(A|C|T|G)**GG

Q = Glutamina (aminoácido)

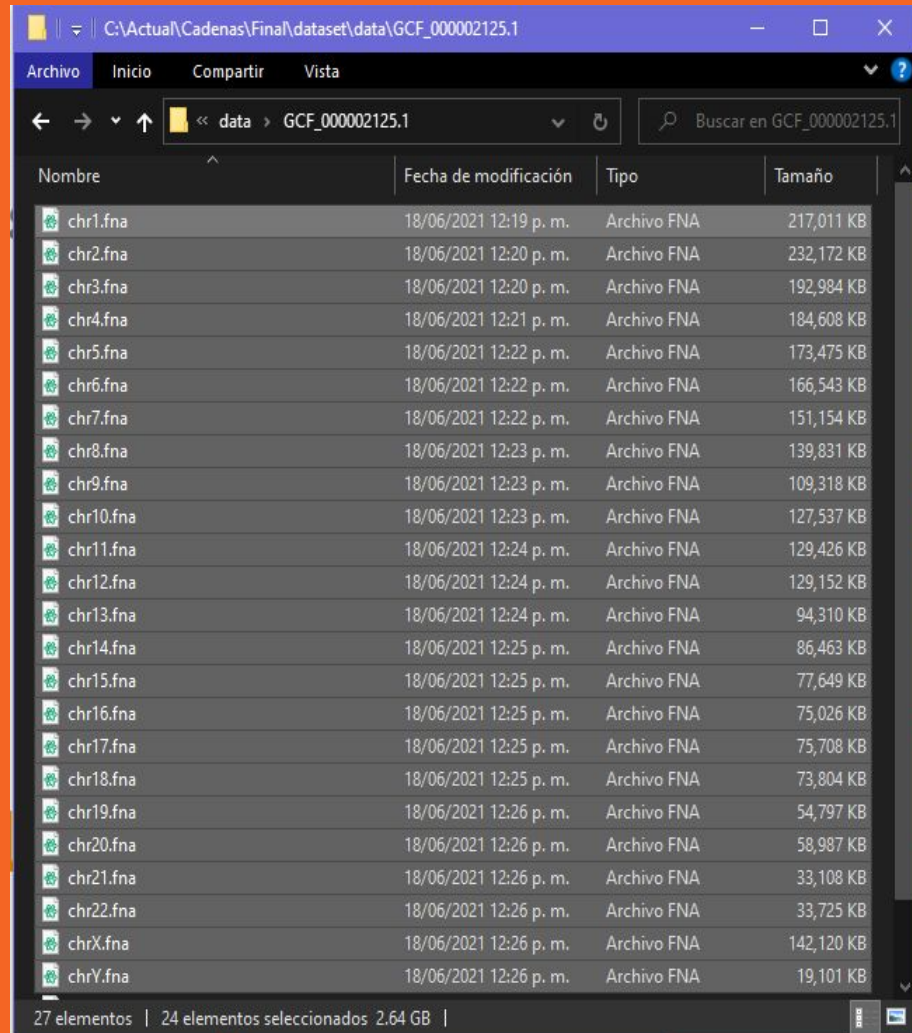
QWD se vuelve

(CAA|CAG)(TGG)(GAT|GAC)

Nucleic acid notation

Symbol ^[2]	Description	Bases represented				
A	Adenine	A				1
C	Cytosine		C			
G	Guanine			G		
T	Thymine				T	
U	Uracil				U	
W	Weak	A			T	2
S	Strong		C	G		
M	aMino	A	C			
K	Keto			G	T	
R	puRine	A		G		
Y	pYrimidine		C		T	3
B	not A (B comes after A)		C	G	T	
D	not C (D comes after C)	A		G	T	
H	not G (H comes after G)	A	C		T	
V	not T (V comes after T and U)	A	C	G		4
N or -	any Nucleotide (not a gap)	A	C	G	T	

en esta búsqueda se toma
un fragmento (1/1k) del
cromosoma 4 de un homo
sapiens con 170k caracteres



The screenshot shows a Windows File Explorer window with the address bar set to `C:\Actual\Cadenas\Final\dataset\data\GCF_000002125.1`. The window displays a list of files in a table format. The table has four columns: 'Nombre', 'Fecha de modificación', 'Tipo', and 'Tamaño'. All files are of type 'Archivo FNA' and are named 'chr1.fna' through 'chr22.fna', 'chrX.fna', and 'chrY.fna'. The files are all dated 18/06/2021. The status bar at the bottom indicates '27 elementos | 24 elementos seleccionados 2.64 GB |'.

Nombre	Fecha de modificación	Tipo	Tamaño
chr1.fna	18/06/2021 12:19 p. m.	Archivo FNA	217,011 KB
chr2.fna	18/06/2021 12:20 p. m.	Archivo FNA	232,172 KB
chr3.fna	18/06/2021 12:20 p. m.	Archivo FNA	192,984 KB
chr4.fna	18/06/2021 12:21 p. m.	Archivo FNA	184,608 KB
chr5.fna	18/06/2021 12:22 p. m.	Archivo FNA	173,475 KB
chr6.fna	18/06/2021 12:22 p. m.	Archivo FNA	166,543 KB
chr7.fna	18/06/2021 12:22 p. m.	Archivo FNA	151,154 KB
chr8.fna	18/06/2021 12:23 p. m.	Archivo FNA	139,831 KB
chr9.fna	18/06/2021 12:23 p. m.	Archivo FNA	109,318 KB
chr10.fna	18/06/2021 12:23 p. m.	Archivo FNA	127,537 KB
chr11.fna	18/06/2021 12:24 p. m.	Archivo FNA	129,426 KB
chr12.fna	18/06/2021 12:24 p. m.	Archivo FNA	129,152 KB
chr13.fna	18/06/2021 12:24 p. m.	Archivo FNA	94,310 KB
chr14.fna	18/06/2021 12:25 p. m.	Archivo FNA	86,463 KB
chr15.fna	18/06/2021 12:25 p. m.	Archivo FNA	77,649 KB
chr16.fna	18/06/2021 12:25 p. m.	Archivo FNA	75,026 KB
chr17.fna	18/06/2021 12:25 p. m.	Archivo FNA	75,708 KB
chr18.fna	18/06/2021 12:25 p. m.	Archivo FNA	73,804 KB
chr19.fna	18/06/2021 12:26 p. m.	Archivo FNA	54,797 KB
chr20.fna	18/06/2021 12:26 p. m.	Archivo FNA	58,987 KB
chr21.fna	18/06/2021 12:26 p. m.	Archivo FNA	33,108 KB
chr22.fna	18/06/2021 12:26 p. m.	Archivo FNA	33,725 KB
chrX.fna	18/06/2021 12:26 p. m.	Archivo FNA	142,120 KB
chrY.fna	18/06/2021 12:26 p. m.	Archivo FNA	19,101 KB

—

Con la Expresión Regular (1)

A.((AA|AT)*.(GA|TT)*).G



Algunas posibles
ocurrencias

AAAG

AATG

AG

AGAG

ATTG

AAAGAG

ATTTTG ...

—

Con la Expresión Regular (2)

(GA|AT).((AG|AAA)*)



**Algunas posibles
ocurrencias**

GAAG

GA

GAAAA

ATAG

ATAAAAG

GAAGAAA

GAATAG ...

—

Con la Expresión Regular (3)
(A.A*).(CG|T).((GG|AAG)*)



Algunas posibles
ocurrencias

AT

AACGGG

ACGAAG

AAAAAAT

ATAAGAAGCG

AAAAAAG

ACGAAGCGCG ...