



**fcfm**

FACULTAD DE CIENCIAS  
FÍSICAS Y MATEMÁTICAS  
UNIVERSIDAD DE CHILE

## ALMACENAMIENTO Y MANEJO DE INFORMACIÓN BIOLÓGICA:

# ALCANCES DEL CÓDIGO GENÉTICO: DATOS, CÓDIGO Y SECUENCIAS

PARTE 1

BT<sub>1211</sub> - APLICACIONES DE LA BIOLOGÍA A LA INGENIERÍA Y CIENCIAS



# Contenidos

- Información:
  - ¿Qué es la información?
  - Almacenamiento de la información
  - Entropía de la información
  - Ruido e información
  - Compresión de la información
- Ejemplos

# Contenidos

- Procesamiento de información biológica:
  - Información biológica
  - Bases de datos
  - Bioinformática
- Ejemplos de aplicación

# Materiales biológicos y no biológicos

- La vida está compuesta por diferentes moléculas que forman los materiales permiten su **complejidad**.
- La ciencia de los **biomateriales** estudia estos materiales y sus aplicaciones.
- La **ingeniería y la biología** se relacionan de diferentes formas.
- Existe un enorme potencial para utilizar organismos vivos en la solución problemas ambientales, la medicina y la industria.

# Materiales biológicos y no biológicos

- La **biomimética** es la aplicación del conocimiento de los seres vivos para solucionar problemas de ingeniería.
- La **biónica** Ciencia que estudia la creación y desarrollo de aparatos y procedimientos tecnológicos que sustituyen o sirven de ayuda a las funciones naturales de los seres vivos.
- La **bionanotecnología** crea o modifica componentes del tamaño de las macromoléculas de la vida. La nanoescala posee ventajas.
- Las **elecciones de materiales y diseño** necesitan criterios en **múltiples ámbitos**, además del material-estructural.

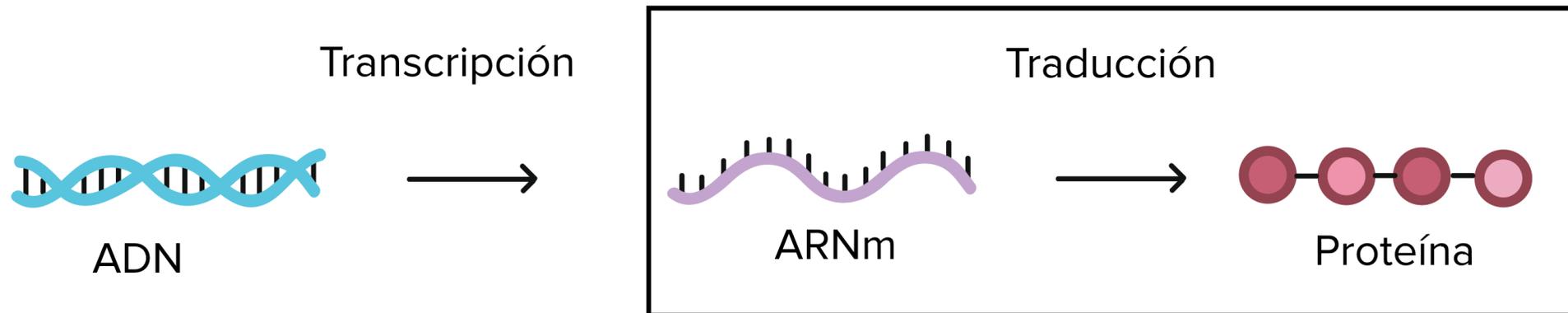
# INFORMACIÓN

---

# Información

Conjunto organizado de datos que constituyen un mensaje que cambia el “estado de conocimiento” o nivel de información de un sistema

Para un sistema biológico, existe información almacenada internamente que se interpreta para cumplir funciones:



# Información

Y también existe información externa que se procesa para adaptar su funcionamiento al entorno :

## **Estímulo sensorial**

Se recibe información a través de un receptor

## **Procesamiento**

La información se interpreta y se convierte en una respuesta

## **Reacción**

La información genera una acción que responde a ella

# Almacenamiento de la información

La información almacenada sigue ciertos patrones de codificación, es decir, existe un código para interpretar los datos guardados

- En el caso de información digital, se utiliza el sistema binario (ceros y unos)
- En el caso biológico, el código de almacenamiento químico se compone de 4 nucleótidos (A, T, G y C) que codifican la información genética



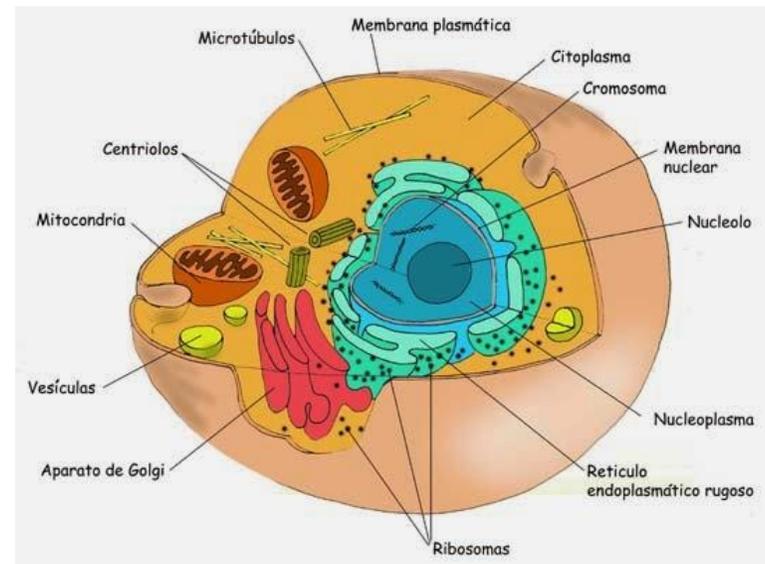
# Almacenamiento de la información

La información también puede almacenarse en forma de orden e interacciones

En otras palabras, la ubicación de un objeto confiere información al sistema, que se expresa en propiedades emergentes.

Por ejemplo:

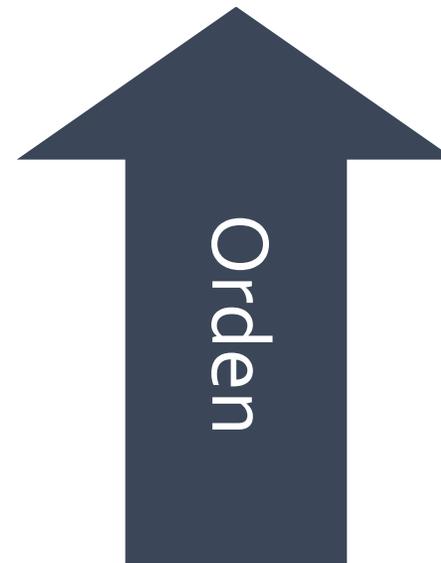
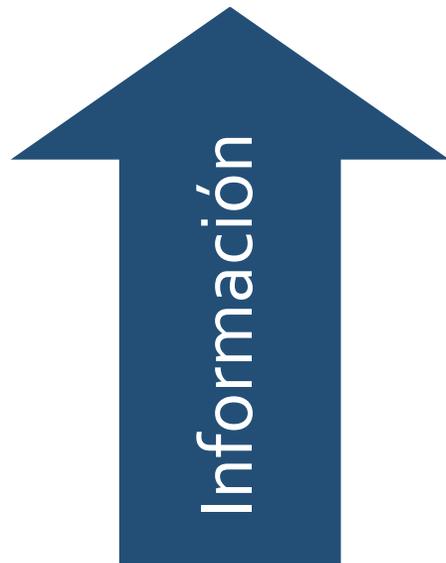
- Organelos en células
- Memoria



# Entropía de la información

Intuitivamente parece ser que la cantidad de información y la cantidad de orden son, de alguna manera, la misma cosa.

En otras palabras, la cantidad de información de un sistema es equivalente a la cantidad de orden que se puede generar a partir de él.



# Entropía de la información

Así, si tomamos la **entropía** como una medida del desorden de un sistema, entonces a mayor información, menor entropía podría tener un sistema, y viceversa.



# Recordatorio

Como habíamos visto anteriormente:

- El mantenimiento del orden de un sistema de cuerpos, requiere ejercer trabajo sobre el sistema
- El trabajo requiere energía, la cual es extraída del medio (no es energía interna del sistema)
- El ordenamiento de un sistema cerrado se obtiene a costa del desordenamiento de la materia fuera del sistema
- Por lo tanto, el mantenimiento de la vida (orden), requiere un gasto permanente de energía

# Contenido de información de Shannon

La definición del contenido de información de un sistema que tiene  $i$  posibles estados es:

$$I = - \sum_i p_i \log_n p_i$$

donde:

$I$ : contenido de información de un sistema

$p_i$ : probabilidad de ocurrencia de un estado  $i$  de ordenamiento

$\log_n$ : logaritmo en base  $n$  ( $I$  tiene escala logarítmica)

Si  $n = 2$ ,  $I$  está expresado en unidades de **bits**

## Características de $I$

$$I = - \sum_i p_i \log_n p_i$$

Como  $p_i$  es una probabilidad,  $p_i \in [0,1]$ . Si  $p_i = 0$ , el estado  $i$  nunca ocurre. En este caso, por definición (y continuidad) se dice que  $I = 0$ .

Luego, siempre se tiene que:  $\log_n p_i < 0 \Rightarrow I > 0$ .

# Características de $I$

$$I = - \sum_i p_i \log_n p_i$$

Supongamos que en un sistema puede presentarse (ocurrir) el evento  $i$  o puede no ocurrir el evento  $i$ . Es decir, **el sistema tiene 2 estados**.

Si  $i$  **siempre** se presenta en el sistema, tiene probabilidad  $p_i = 1$  de ser observado. Luego, al observar el sistema, no se obtiene información desconocida acerca de su estado ( $I = 0$ ).

Igualmente, si  $i$  **nunca** se presenta en el sistema, tiene probabilidad  $p_i = 0$  de ser observado. Como dicho evento nunca se observa, conocer el estado del sistema tampoco entrega información ( $I = 0$ ).

# Características de $I$

Por lo tanto, cuando  $I = 0$ , el estado observado del sistema no contiene información para el observador respecto al estado.

En otras palabras, el sistema siempre está en un estado conocido: siempre está en el estado  $i$  o nunca está en el estado  $i$ .

De esto se deriva que siempre se puede predecir el estado de un sistema que no contiene información.

Preguntas propuestas:

- 1) ¿Para qué valor de la probabilidad  $p_i$  de que ocurra un evento la información  $I$  obtenida al observar el estado del sistema es máxima?
- 2) ¿Cuál es el valor máximo de  $I$  para un evento que tiene dos estados (puede ocurrir o no)?

# “Sorpresa” contenida en la información

La cantidad de información depende de la **sorpresa** que el el estado del sistema contenga.

Cuanto menos **predecible** es un sistema, más sorprendente es constatar la observación, y más información puede obtenerse de una observación.



# Relación entre entropía e información

Se puede demostrar (Shannon) que la entropía termodinámica de un sistema ordenado se relaciona con la información necesaria para ordenarlo como:

$$\Delta S = -k \cdot \ln 2 \cdot I$$

donde:

$$k = 1,3802 \times 10^{-23} \text{ N} \cdot \text{m}/\text{K} \text{ (constante de Boltzmann)}$$

$$\ln 2 = 0,693$$

# Ruido e información

El ruido de una información corresponde a la señal no deseada que se mezcla con la señal útil que lleva el mensaje, resultando en una corrupción del mensaje original.

Los sistemas biológicos sensan las señales que provienen del ambiente y las procesan para eliminar el ruido que impide su correcto entendimiento.

# Ejemplos de procesamiento biológico de señales ruidosas

Procesamiento de señales sensoriales:

- Oído y sonidos de fondo

- Olores de fondo

- Ruido retinal

- Imágenes residuales en la retina

# Compresión de la información

La información puede ser comprimida para ocupar menos espacio y de esta manera ser más fácil de almacenar o transportar. Hay dos conceptos claves que se deben tener presentes:

- **Entropía:** cantidad de información contenida en el mensaje en función de la sorpresa
- **Redundancia:** cuanto mayor redundancia (repetición) haya, menor información hay, ya que conociendo una parte de los datos es posible conocer otra igual repetida

¿Cómo comprimirías la información contenida en esta foto?



# Compresión de la información biológica

Se ha encontrado en los virus capacidad de comprimir información de sus genes, en forma de ADN, para minimizar el espacio que ocupan. Esto lo llevan a cabo traslapando secuencias de los genes.

Ejemplo análogo: Buscar las palabras que están en la siguiente secuencia de letras.

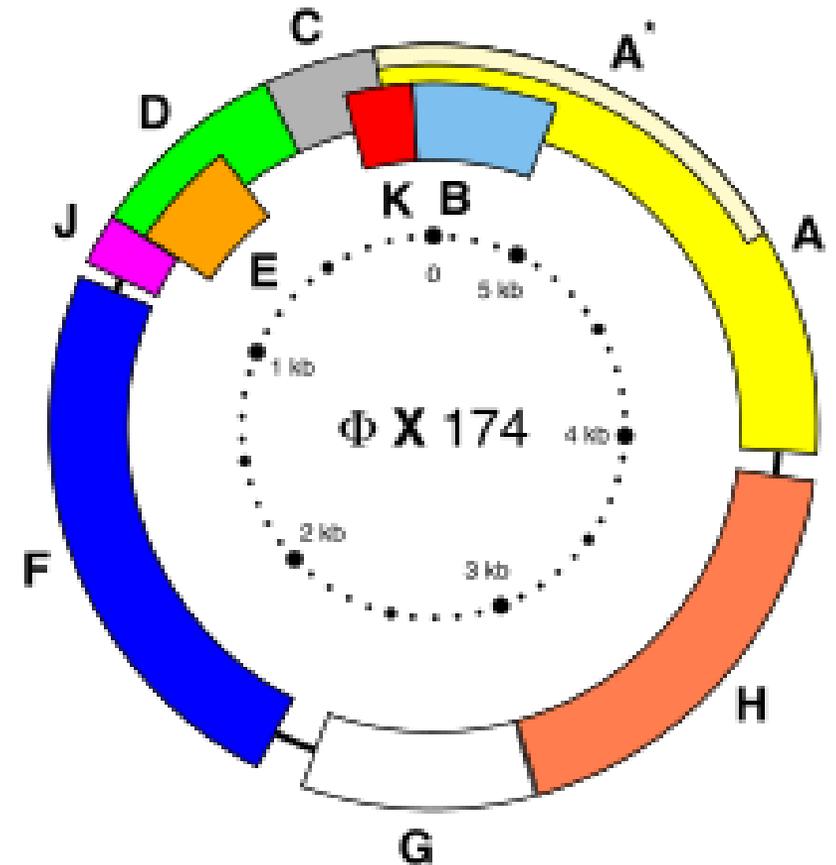
**PRIMAVERA**

# Compresión de la información biológica

Se ha encontrado en los virus capacidad de comprimir información de sus genes, en forma de ADN, para minimizar el espacio que ocupan. Esto lo llevan a cabo sobrelapando secuencias de los genes.

Por ejemplo: encontrar las palabras contenidas en la siguiente secuencia de letras.

**PRIMAVERA**



# EJEMPLOS

---

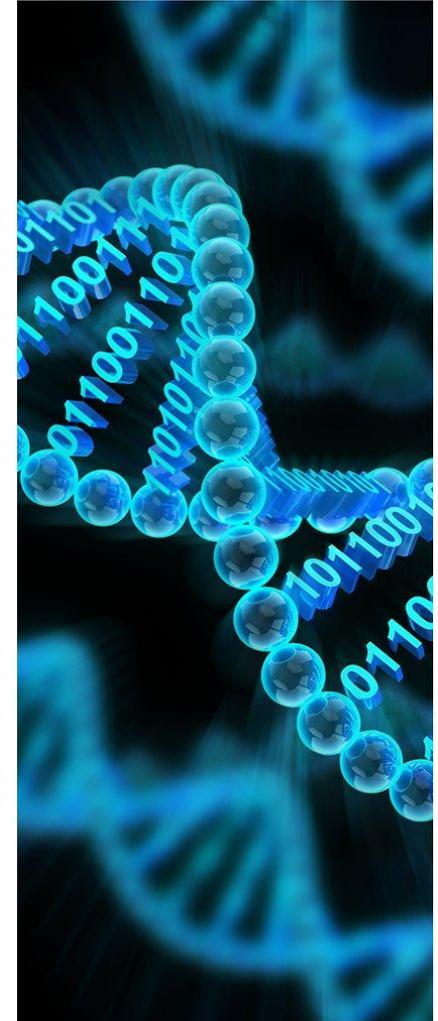
# Sistema de encriptación en ADN

Existe una diversidad de técnicas y códigos utilizados para codificar el lenguaje e información a secuencias de ADN que dependen de la naturaleza de lo que se quiera realizar.

Por ejemplo:

*Next Generation Digital Information Storage:* utiliza el código binario.

*Genesis Project:* utiliza el código Morse para armar el código.



# Ejemplo: codificación de imagen GIF

[https://youtu.be/to\\_hZ7V7Yzs](https://youtu.be/to_hZ7V7Yzs)

Mediante la codificación de un GIF a binario y luego a código genético fue posible transferir la información de una bacteria a otra y luego leer el archivo.

¿Cuál crees que podrían ser las ventajas de guardar información en ADN? ¿Y las desventajas?

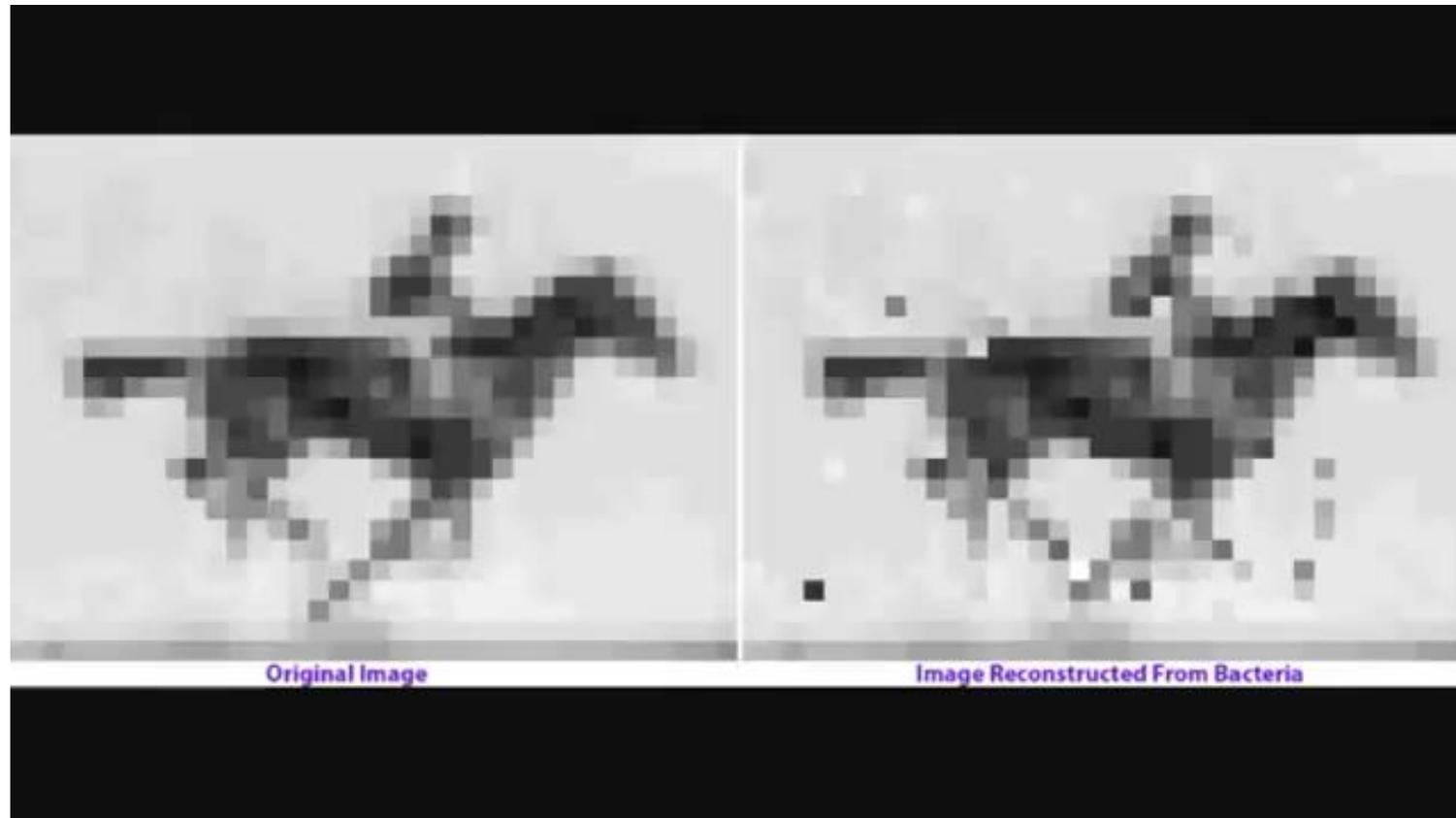


Original Image



Image Reconstructed From Bacteria

# Codificación de GIF

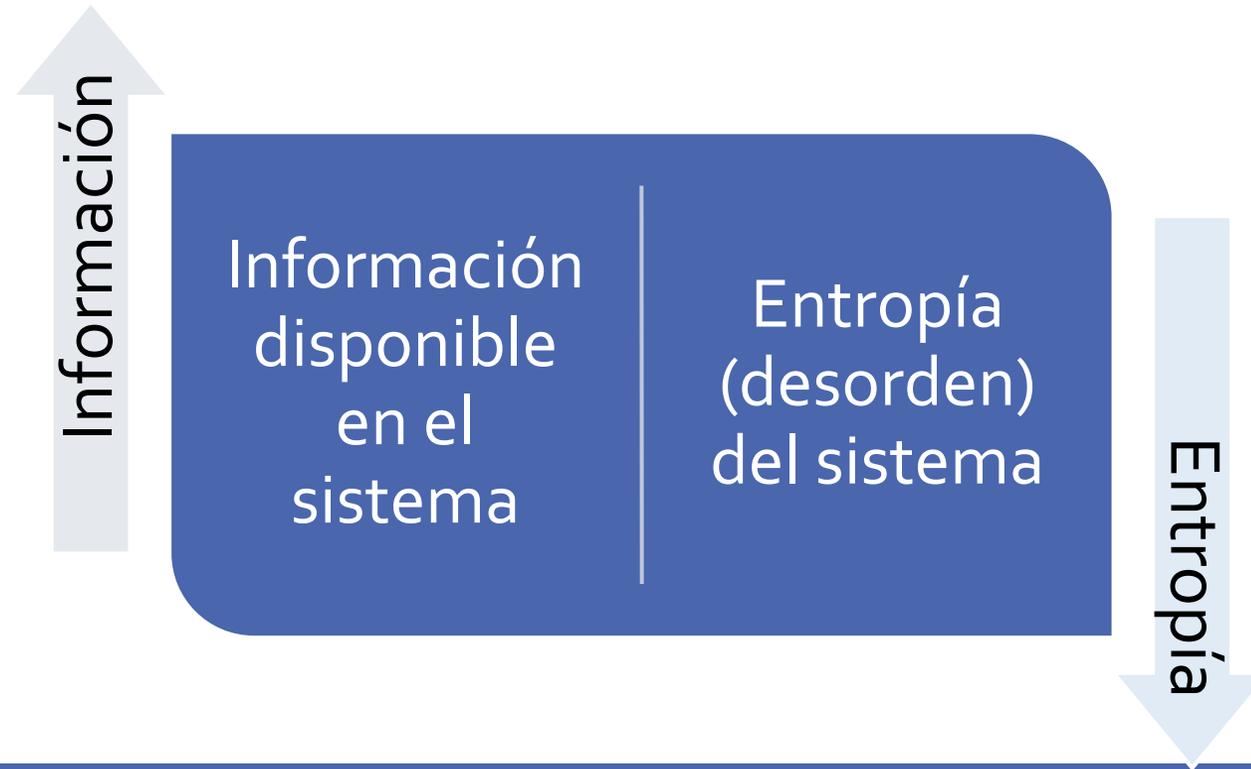


GIF codificado original

GIF obtenido de bacterias

# Resumen de primera parte

Información: conjunto organizado de datos que constituyen un mensaje que cambia el “estado de conocimiento” de un sistema



# PROCESAMIENTO DE INFORMACIÓN BIOLÓGICA

---

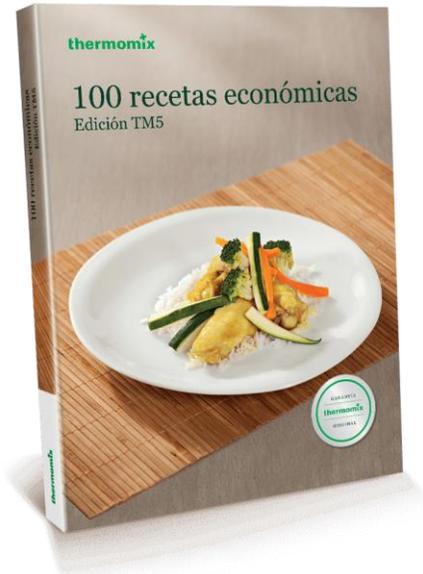
# Información biológica

Actualmente y en un futuro cercano, se genera información de secuencias biológicas del orden de 1 zetabases ( $10^{21}$  bytes) por año, sin considerar otro tipo de información

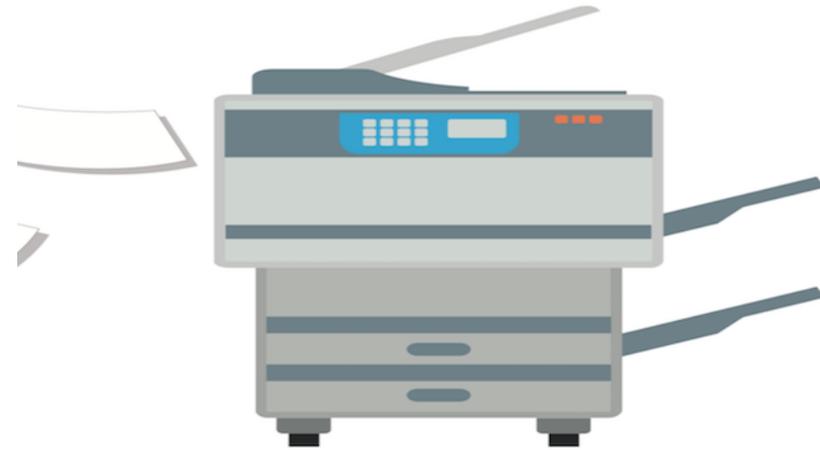
Contempla además toda la información que puede ser recogida de los organismos vivos, en particular de sus funcionamiento interno con grandes volúmenes de datos y en 4 grandes disciplinas.

- Genómica
- Transcriptómica
- Proteómica
- Metabolómica

# Ejemplo de recetas



Colección de recetas en un libro (ADN)



Selección y copia de las recetas (ARN)

# Ejemplo de recetas



Materiales utilizados  
(Enzimas y Proteínas)



Platos preparados  
(Metabolitos)

# Información biológica

## Genómica

- Estudio de la estructura, función, evolución y mapeo de genes

## Transcriptómica

- Estudio de la presencia de los distintos tipos de ARN en una célula

## Proteómica

- Estudio a gran escala de las proteínas, en su función y estructura

## Metabolómica

- Estudio de metabolitos en una célula u organismo

# Bases de datos

Los volúmenes gigantescos de información han planteado desafíos de cómo organizarlos y tener disponible la información de manera rápida. Además, es necesario buscar patrones y establecer relaciones entre estructuras o secuencias similares.

```

                                signal peptide
CsnV26 MEISMQK-PGFWKGAIISSFMSPCFPLPMMSDIVFAAGLNKDKQKRRAEQLTISIFENGITTE 59
csn168 MKISMQK-ADFWKGAII SLLVFIMFFILMSEIVFAAGLNKDKQKRRAEQLTISIFENGITTE 59
DAU101 MKISLKGKAGFWKKTAVSSLIFFIMFFILMMSGIVFAAGLNKDKQKRRAEQLTISIFENGKTE 60
***** ..*****:* * :: * : **** *****:*****:***

CsnV26 IQYGYVEPLGDGRGYTCGRAGFTTATGDALEVVEVYTKAVPNNKLRKYLPELRLRAEEES 119
csn168 IQYGYVERLDDGRGYTCGRAGFTTATGDALEVVEVYTKAVPNNKLRKYLPELRLRAKEES 119
DAU101 IQYGYVEALDDGRGYTCGRAGFTTATGDALEVVEVYTKAVPNNKLRKYLPELRLRAKDES 120
***** * .*****:*****:*****:***

CsnV26 DDISNLKGFASAWRSLGNDKDFRAAQDKVNDRLYYQPAMKRSENAGLKTALAKAVMYDTV 179
csn168 DDTISNLKGFASAWKSLANDKEFRAAQDKVNDHLYYQPAMKRSDNAGLKTALARAVMYDTV 179
DAU101 DDISNLKGFASAWRSLGNDKAFRAAQDKVNDRLYYQPAMKRSENAGLKTALAKAVMYDTV 180
** *****:*** ** ***** *****:*****:*****

CsnV26 IQHGDGDDPDSFYALIKRTINKKAGGSPKDGIDEKKWLNKFLDVRYYDDLMPADPDRDEW 239
csn168 IQHGDGDDPDSFYALIKRTINKKAGGSPKDGIDEKKWLNKFLDVRYYDDLMPANHDTREW 239
DAU101 IQHGDGDDPDSFYALIKRTINKKAGGSPKDGIDEKKWLNKFLDVRYYDDLMPSEDTQDEW 240
***** ***** *****:*****:***

CsnV26 RESVVRVDVLRSLAKSNYYNLNGPINIHSAYGDFVIK 277
csn168 RESVARVDVLRSLAKENYYNLNGPIHVR SNEYGNFVIK 277
DAU101 RESVARVDVFRDIVKEKNYYNLNGPIHVR SSEYGNFTIQ 278
****.*****:*.*.*****:*** ***:*.**

```

# Aplicación de las Bases de Datos

Muffin Conocido



Pastel desconocido



# Aplicación de las Bases de Datos

Muffin Conocido



Pastel desconocido



# Aplicación de las Bases de Datos

Muffin Conocido

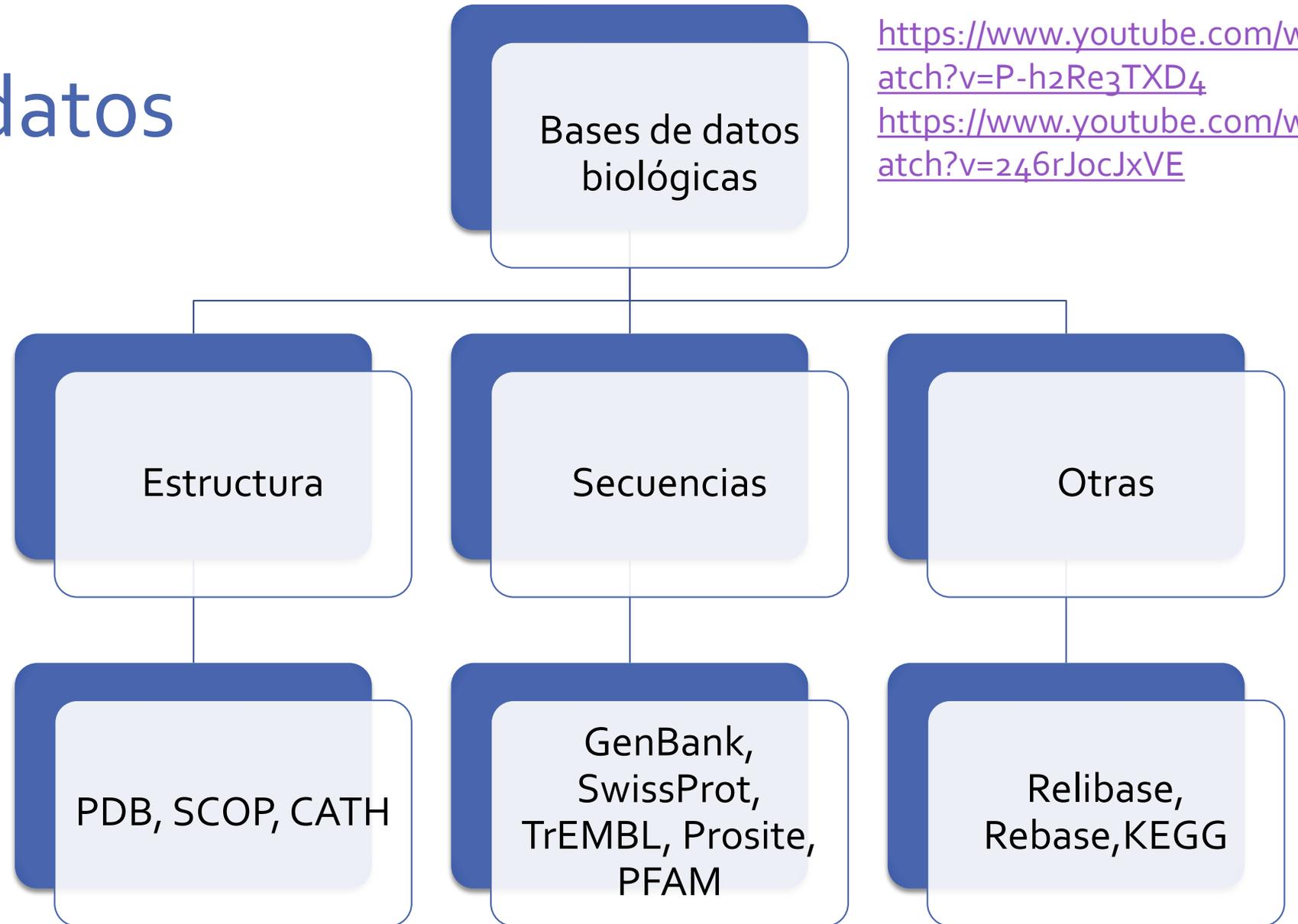


Pastel desconocido



# Bases de datos

Para los distintos tipos de información biológica (ómicas) es posible encontrar bases de datos especializadas.





## GenBank

GenBank

- Secuencias genéticas



## PDB

- Estructuras de proteínas

UniProt

## UniProt

- Funcionalidad de proteína
- Secuencia de proteínas

# Bioinformática

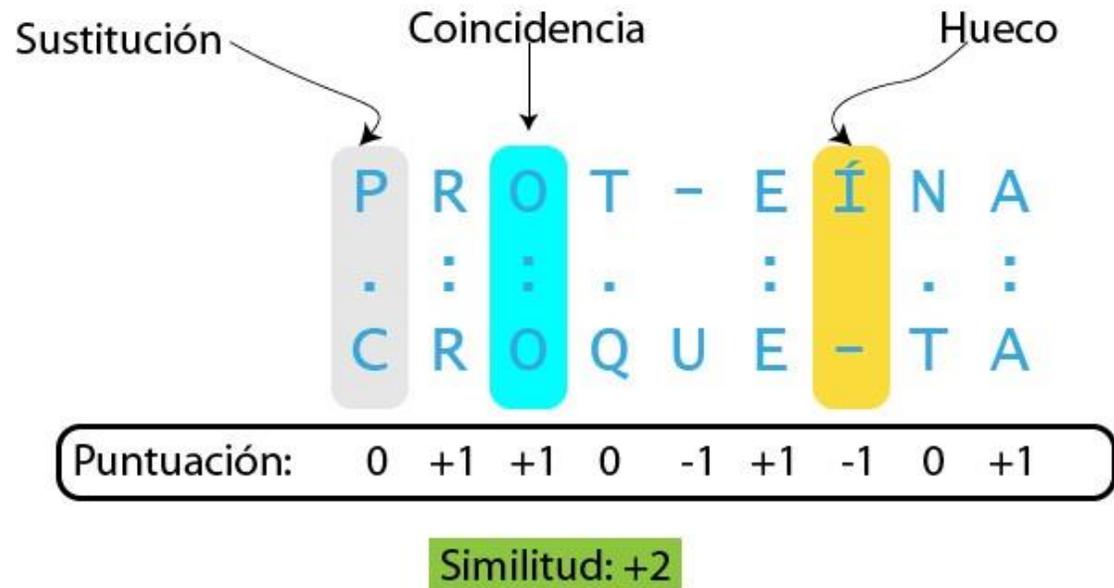
Disciplina que se preocupa de la recolección y análisis de datos biológicos complejos, mediante el desarrollo de métodos computacionales y herramientas de software avanzadas.



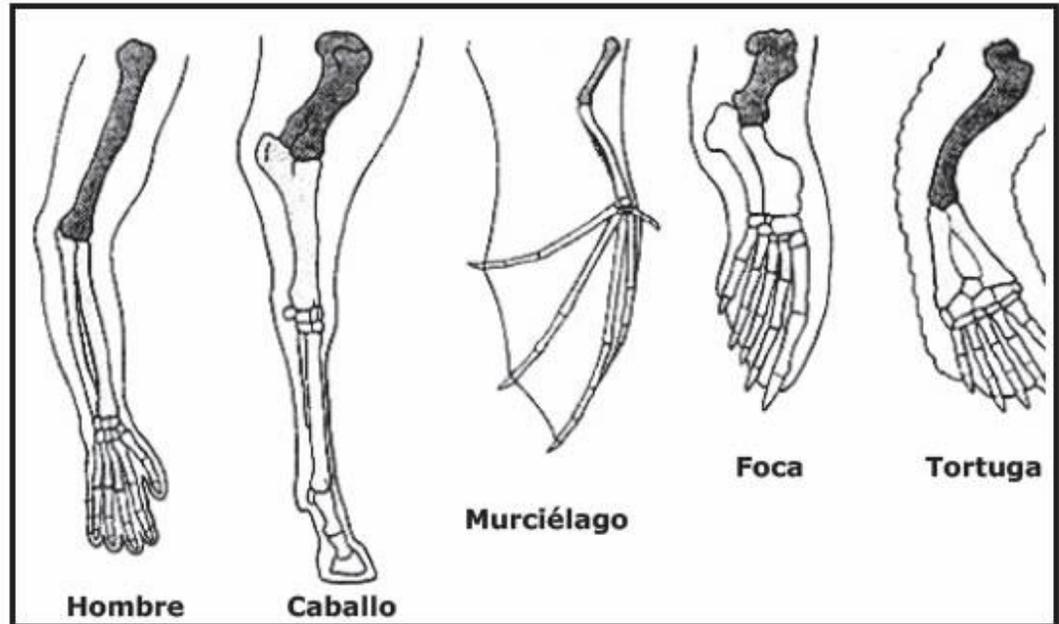
# EJEMPLOS DE APLICACIÓN

---

# Comparación de secuencias y estructuras



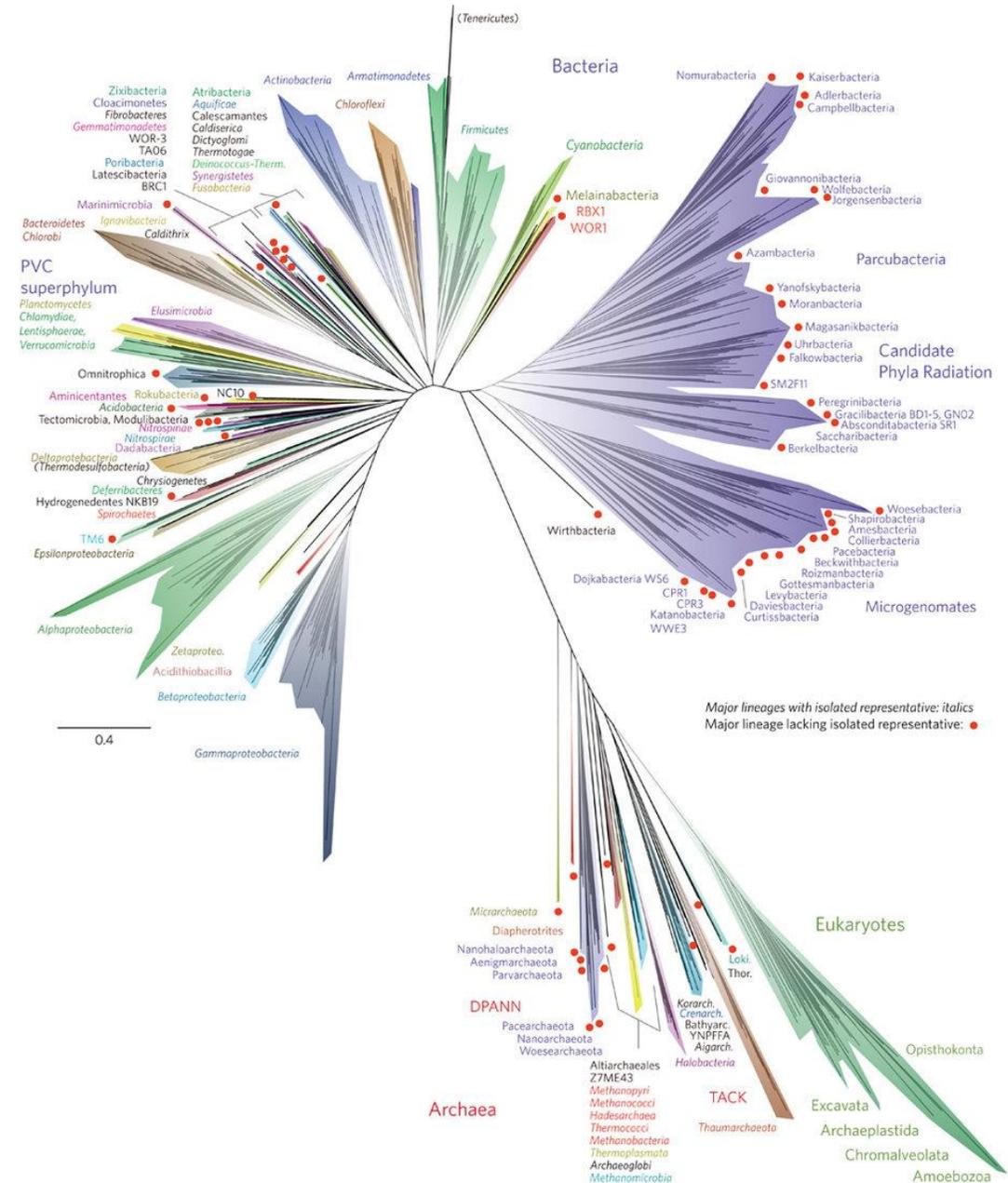
Alineamiento de  
secuencias



Predicción de  
comportamiento de  
estructuras

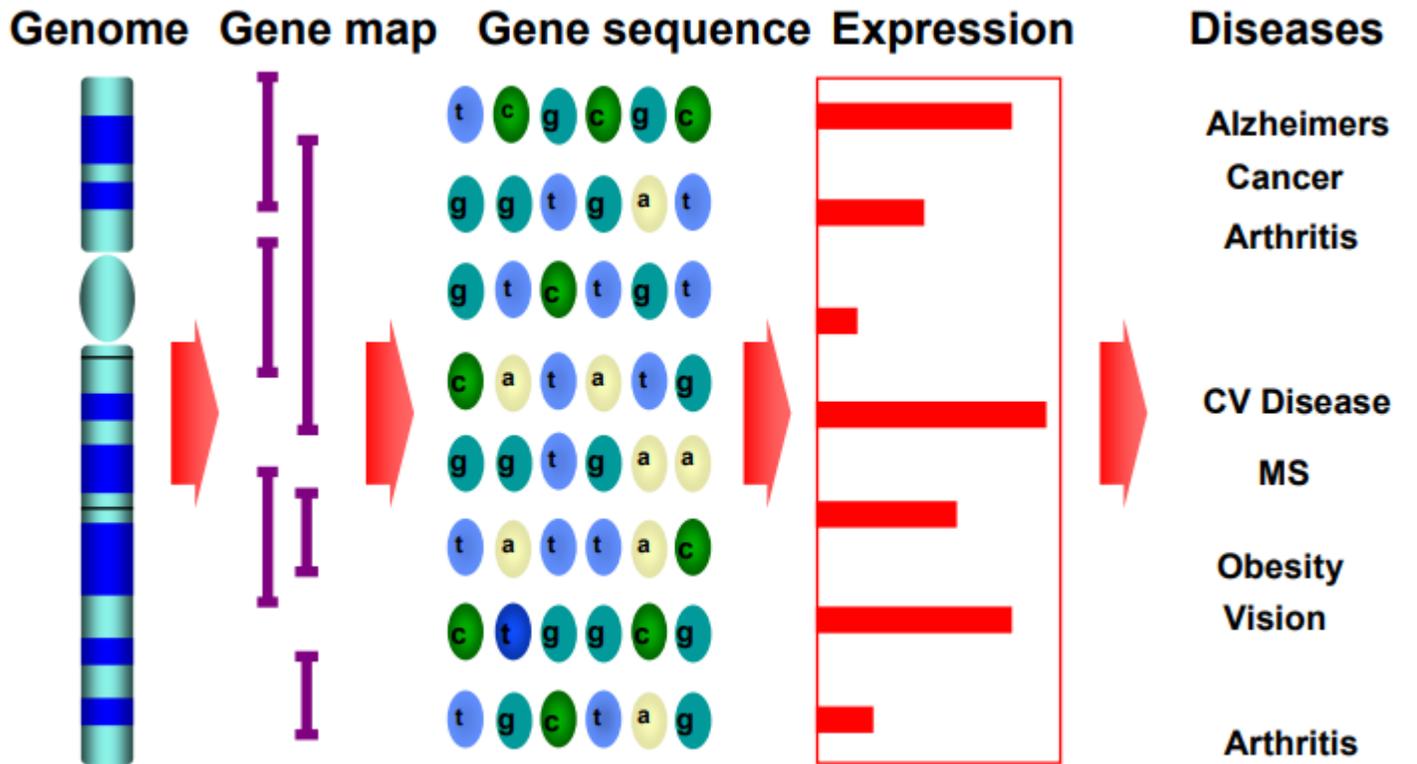
# Búsqueda de filogenia

Con súpercomputadores se ha podido encontrar relaciones entre los genomas de las especies, en particular se armó el árbol de la vida con más de 3000 genomas.



# Detección de enfermedades

Con el estudio del genoma y genes expresados es detectar enfermedades como el cáncer y analizar predisposiciones a los genes.



# BIOALGORÍTMICA

---

# Bioalgorítmica

Un algoritmo corresponde a una serie de instrucciones y/o operaciones que permiten llegar a un resultado esperado. Y dentro de la naturaleza existen varios que han sido estudiados como inspiración.



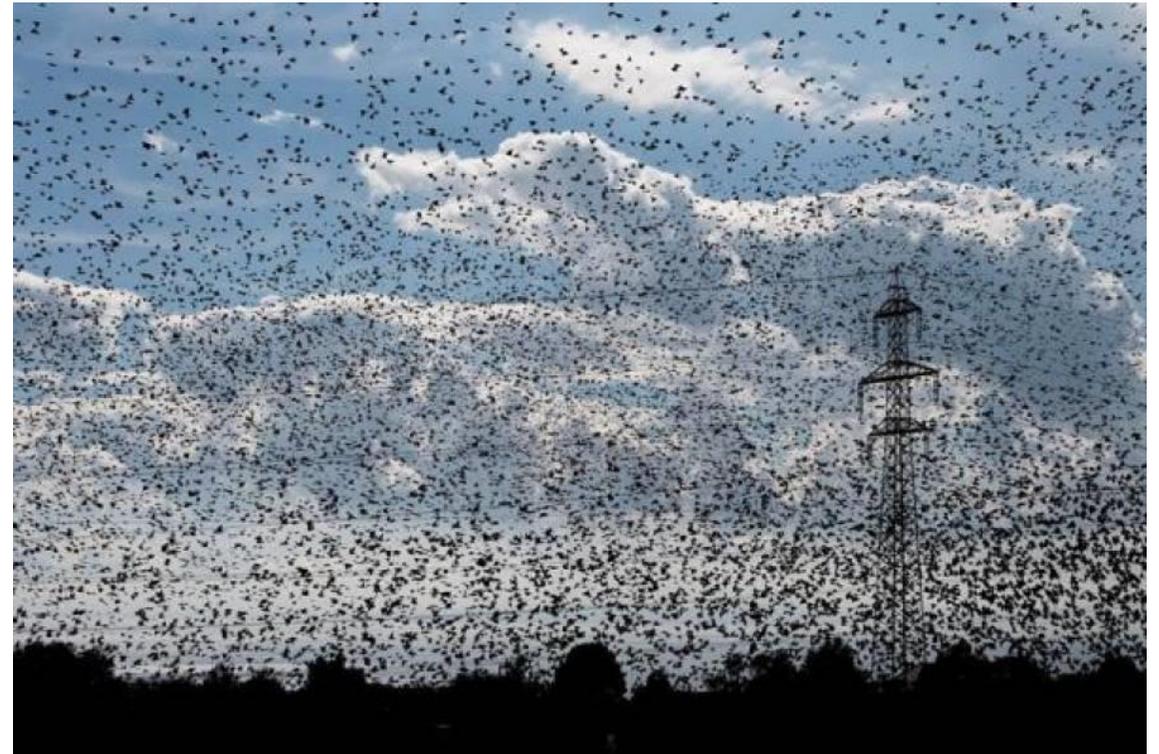
# Algoritmos de bandadas y enjambres

## Algoritmo Biológico

En las ciudades las aves buscan zonas con una densidad arbórea alta para reposar. Esto lo hacen dividiéndose y comunicándose hasta encontrar el mejor lugar.

## Principio del Algoritmo

Con puntos repartidos aleatoriamente en un espacio, éstos se mueven hacia un punto óptimo que se ha encontrado dentro de su conjunto, hasta que finalmente los puntos convergen.



# Algoritmos de colonia de hormigas

## Algoritmo Biológico

Las hormigas marcan el camino que recorren liberando feromonas. Luego, si éstas buscan comida, el camino más corto tendrá más presencia de este odorante, por lo que las hormigas usarán ese trayecto.

## Principio del Algoritmo

Con una solución inicial (camino) se buscan soluciones con ligeras variaciones de éste, hasta encontrar una que minimice el largo y el conjunto de soluciones converja a ésta.



Explicación en video:

<https://www.youtube.com/watch?v=qfeymoF8pb4>

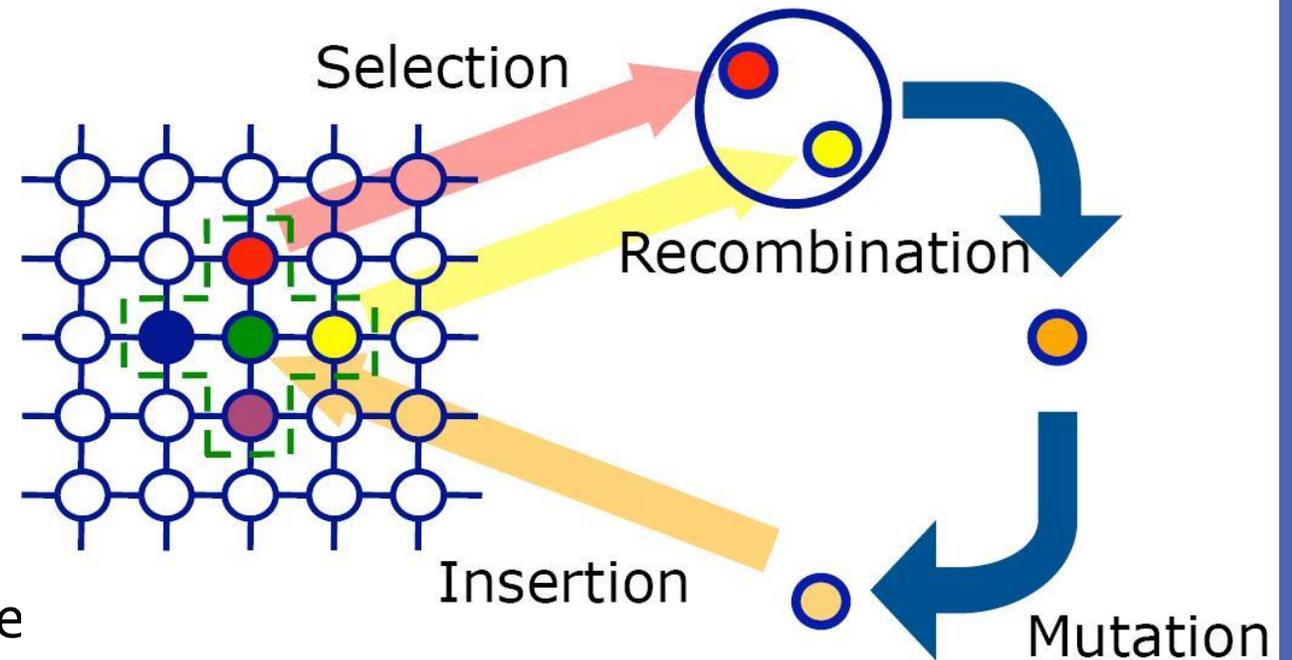
# Algoritmos genéticos

## Algoritmo Biológico

Los organismos de distintas especies tienen ligeras variaciones que les permiten ser más aptos para sobrevivir al ambiente y de esta manera evolucionar, mediante selección natural.

## Principio del Algoritmo

Basado en la selección natural, se tiene un conjunto de soluciones que se entrecruzan entre sí, además de introducir mutaciones, para crear nuevas generaciones de soluciones mejoradas.



Explicación en video:

<https://www.youtube.com/watch?v=Bhmezi8jHpU>

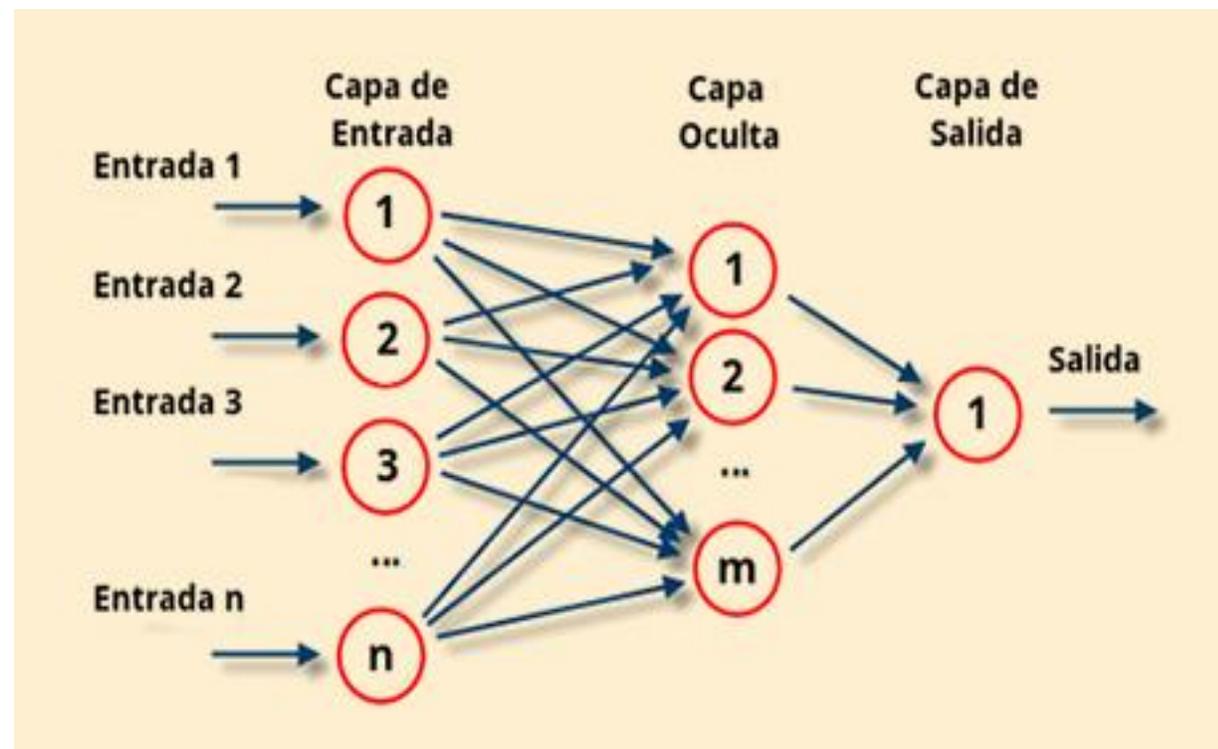
# Redes neuronales

## Algoritmo Biológico

El cerebro y el sistema nervioso están compuestos de miles de neuronas con comportamientos simples que recogen estímulos y, al interaccionar entre ellas, los procesan como información y conocimiento que se almacena.

## Principio del Algoritmo

La interacción entre neuronas artificiales bajo reglas simples permite el procesamiento de la información. De esta manera, al utilizar una serie de neuronas para analizar un problema, el sistema puede adaptar su respuesta para conformarse a una salida conocida, aprendiendo la solución.





**fcfm**

FACULTAD DE CIENCIAS  
FÍSICAS Y MATEMÁTICAS  
UNIVERSIDAD DE CHILE

## ALMACENAMIENTO Y MANEJO DE INFORMACIÓN BIOLÓGICA:

# ALCANCES DEL CÓDIGO GENÉTICO: DATOS, CÓDIGO Y SECUENCIAS

PARTE 2

BT<sub>1211</sub> - APLICACIONES DE LA BIOLOGÍA A LA INGENIERÍA Y CIENCIAS

