Computación y Ciencia

La pareja perfecta

http://vis.usal.es/rodrigo/documentos/bie/charla.pdf

Rodrigo Santamaría
2014

Digan lo que digan…

EL TAMAÑO IMPORTA

Un tema de escalas

10^{3}	Kilo	10^{-3}	Mili
10^{6}	Mega	10^{-6}	Micro
10^{9}	Giga	10^{-9}	Nano
10^{12}	Tera	10^{-12}	Pico
10^{15}	Peta	10^{-15}	Femto
10^{18}	Exa	10^{-18}	Atto
10^{21}	Zetta	10^{-21}	Zepto
10^{24}	Yotta	10^{-24}	Yocto

http://htwins.net/scale2/

Escalas (información)

10^3	Kilo	KB	Un informe
10^6	Mega	MB	Un libro
10^9	Giga	GB	Una película
10^{12}	Tera	TB	La biblioteca del congreso de EEUU
10^{15}	Peta	PB	Todas la bibliotecas de EEUU

burikmodeldesign.com/search/How_Many_Bytes.htm

*The collections of the Library of Congress include more than **32 million cataloged books** and other print materials in 470 languages; more than **61 million manuscripts**; […] over 1 million US government publications; **1 million issues of world newspapers** spanning the past three centuries; 33,000 bound newspaper volumes; 500,000 microfilm reels; over 6,000 titles in all, totaling more than 120,000 issues comic book titles; films; 5.3 million maps; 6 million works of sheet music; 3 million sound recordings; more than **14.7 million prints and photographic images***

http://en.wikipedia.org/wiki/Library_of_Congress

LHC

- Large Hadron Collider (CERN, Suiza)
  - 27 TB / día → 10 PB / año
  - Una conexión de 10 Gbps

http://en.wikipedia.org/wiki/Worldwide_LHC_Computing_Grid

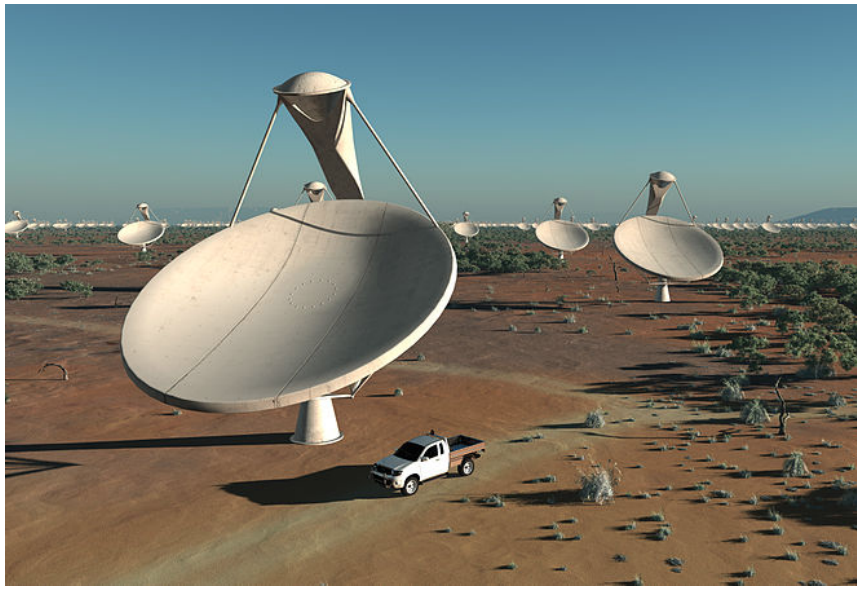
GTC

- Gran Telescopio de Canarias
  - 600 GB en archivo*
  - Uno de los telescopios más grandes del mundo

* http://gtc.sdc.cab.inta-csic.es/gtc/help/overview.jsp

SKA

- Square Kilometre Array
  - Australia – Sudáfrica (2020)
  - Radio Telescopio formado por cientos de antenas
    - Pruebas a la teoría de la relatividad
    - Pruebas sobre materia y energía oscura
    - Primeros momentos del Big-Bang
  - 1000 PB /día → 1 ExaByte / día



http://en.wikipedia.org/wiki/Square_Kilometre_Array

Internet

- Tráfico de datos en Internet en 2012: 31 EB / mes
- Datos indexados en Internet (2013): ~672 EB
- Datos totales en Internet (2013): ~4 Zetta Bytes
  - 1 000 000 000 000 000 000 000 Bytes

Más ejemplos:
http://en.wikipedia.org/wiki/Petabyte
http://en.wikipedia.org/wiki/Exabyte
http://en.wikipedia.org/wiki/Zettabyte

http://en.wikipedia.org/wiki/Internet_traffic

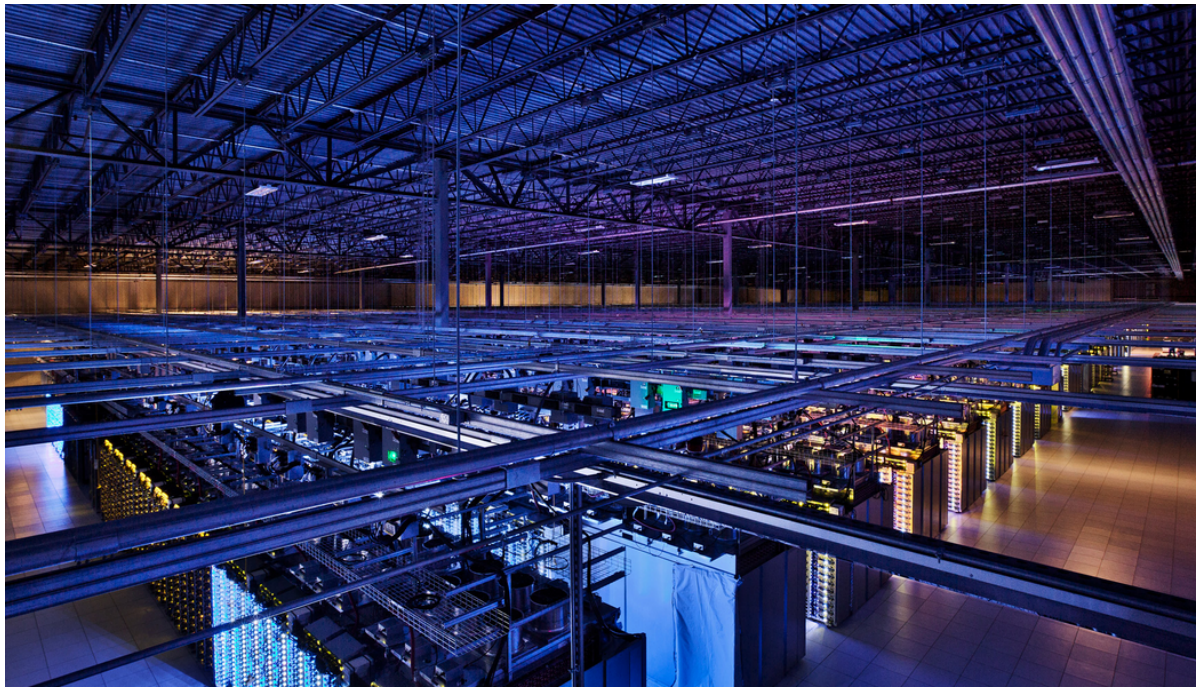
http://www.factshunt.com/2014/01/total-number-of-websites-size-of.html

No preocuparse...

INFORMÁTICA AL RESCATE!

Tecnologías de la Información (IT)

- Gestión de la información
  - **Almacenamiento**: memoria, servidores
  - **Transmisión**: redes, protocolos de transporte
  - **Consulta**: bases de datos, servicios web



http://www.google.com/about/datacenters/gallery/#/

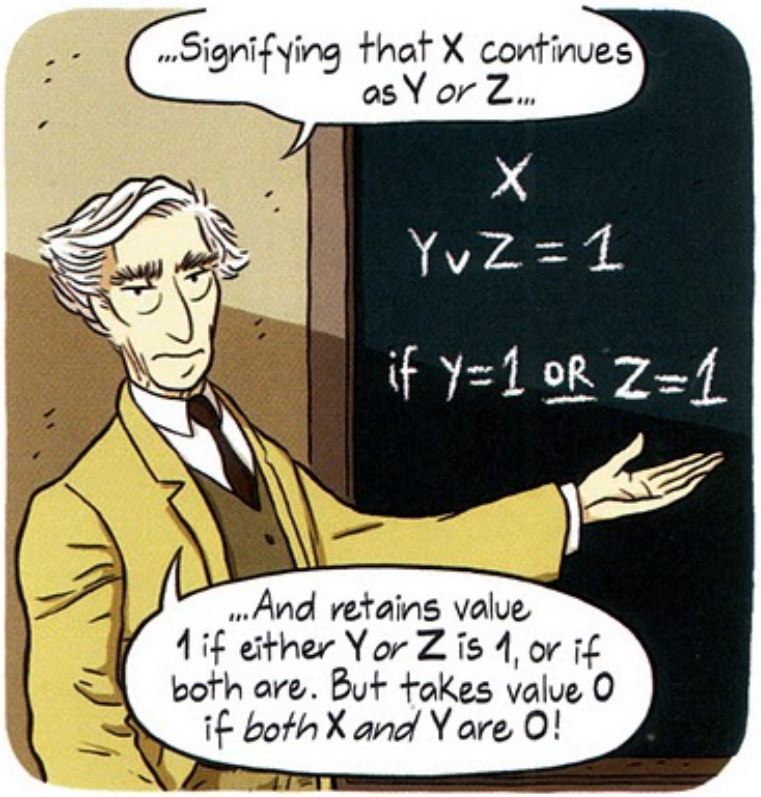
Ciencias de la Computación (CS)

- Análisis de la información
  - **Búsqueda** de patrones
  - **Clasificación** de datos
  - **Predicción** de comportamiento

http://xkcd.com/308/

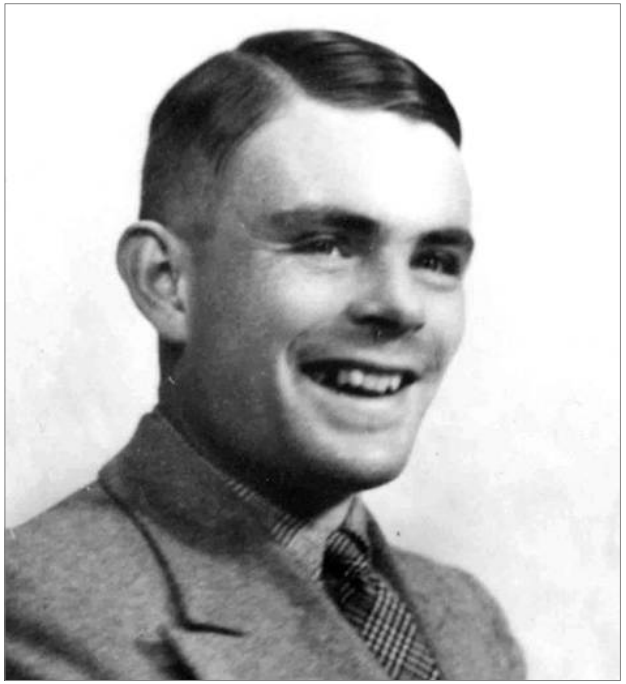
Programación

- Forma de comunicarnos con un computador
- Evolución natural de la lógica



Bertrand Russell en *Logicomix-The Epic Search of Truth*

Ludwig Wittgenstein



Alan Turing

Lenguajes de programación

- Lógica convertida en lenguaje
  - Tiene su propia gramática, sintaxis, vocabulario…
  - Interpretable por un ordenador
  - ‘Comprensible’ por un humano
- Varios lenguajes de programación
  - Java, C, Perl, Fortran, Basic, Cobol, …
  - Usaremos **Python**
    - Muy sencillo y utilizado

Code.org

- Iniciativa para enseñar a programar en colegios e institutos
  - www.codeadademy.com → Python

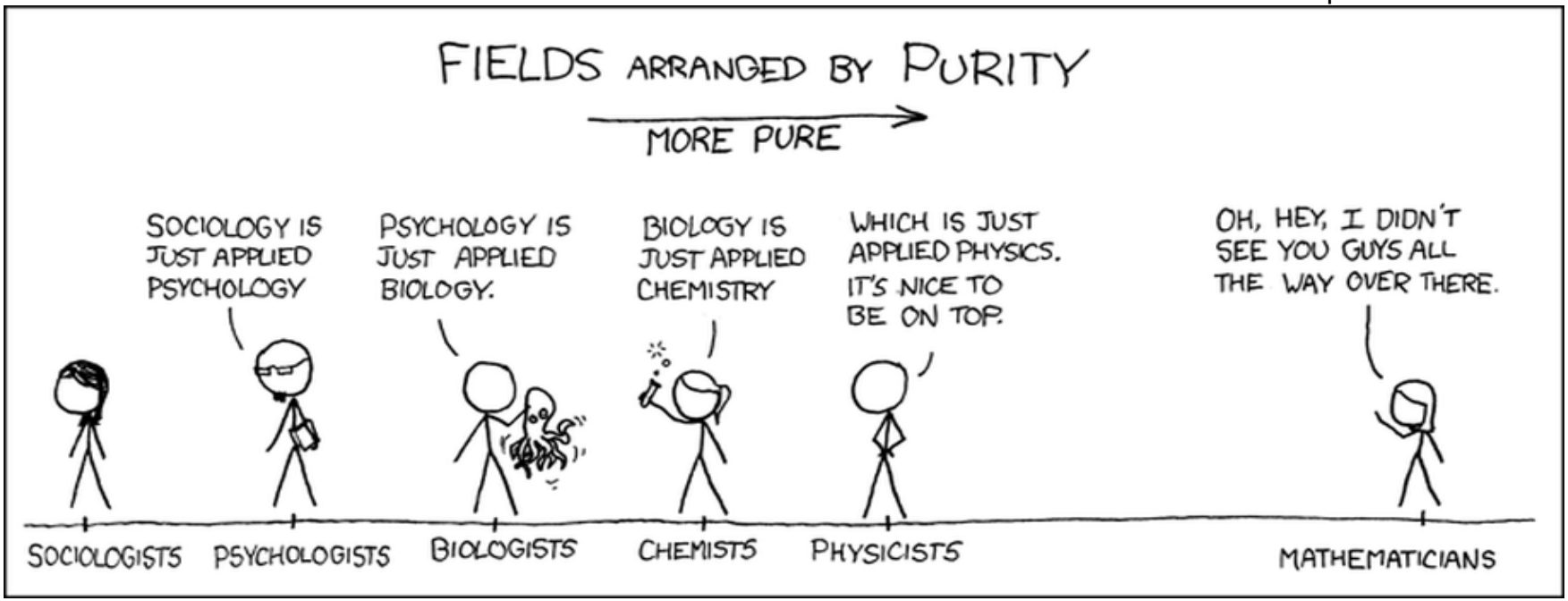
- Hora 2:
  - Sintaxis
  - Cadenas y salida
  - Control de flujo
- Hora 3:
  - Funciones
  - Listas y diccionarios
  - Bucles

Ahora que sabemos programar...

HAGAMOS ALGO DE CIENCIA!

Aplicaciones

http://xkcd.com/435/



psicoinformática

bioinformática

quemoinformática

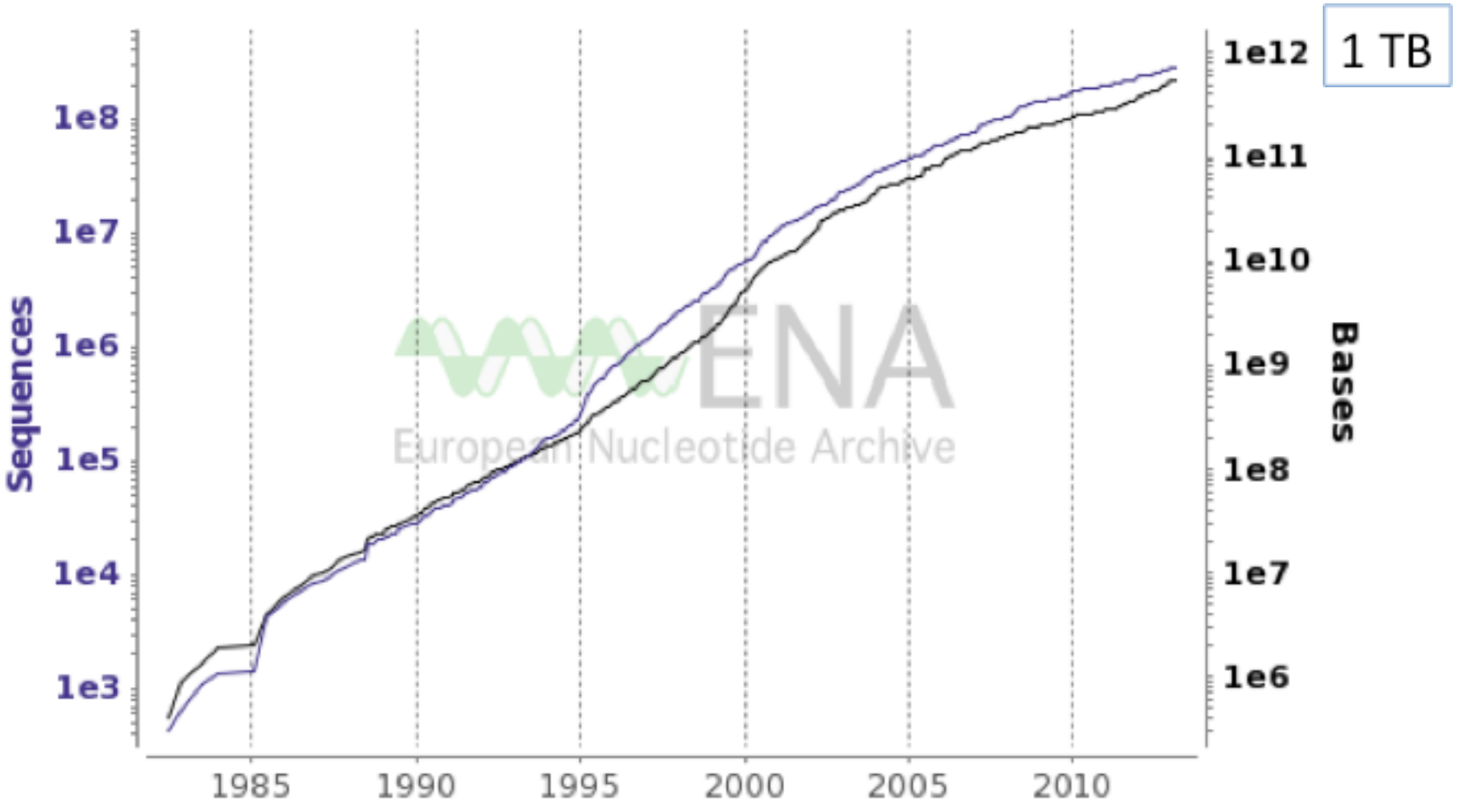
Genomas

Organismo	Nº de pares de bases (aprox)	Tamaño (aprox)
Bacteria	$4 \cdot 10^6$ (4 millones)	2 MB
Levadura	$2 \cdot 10^7$ (20 millones)	10 MB
Gusano	$8 \cdot 10^7$ (80 millones)	40 MB
Mosca	$2 \cdot 10^8$ (200 millones)	200 MB
Ratón	$2.5 \cdot 10^9$ (2500 millones)	1.25 GB
Humano	$3 \cdot 10^9$ (3000 millones)	1.5 GB

1 base (4 letras posibles) → se puede almacenar en 2 bits
1 par de bases (8 posibilidades) → se puede almacenar en 4 bits
1 byte (8 bits) → puede almacenar 2 pares de bases

Bacteria: $4 \cdot 10^6$ pares / 2 pares por byte = $2 \cdot 10^6$ bytes= 2MB

Todos los genomas secuenciados



Búsqueda de información

Guerra y Paz

- 2900 páginas
- 3.5 millones de letras
- Ruso
- 5.8 MB
- Estructura en capítulos, párrafos, frases
- Sabemos de qué va

Genoma Humano

- 20000 genes
- 3000 millones de letras
- ADN (A,C,G,T)
- 1.5 GB
- Estructura en cromosomas, genes, exones?
- Sabemos de qué va?

El genoma de una bacteria

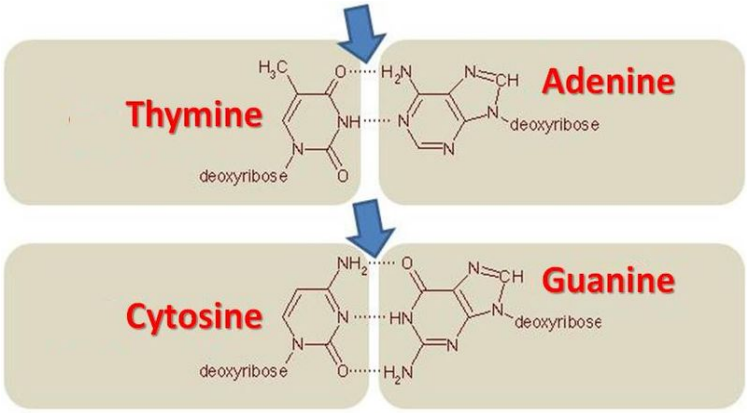
Escherichia coli (E. coli)

- La tenemos en los intestinos
  - *Hasta 2 Kg del peso corporal son microorganismos!*
- Organismo modelo en genética y biología molecular
- Genoma completo:
  - `http://vis.usal.es/rodrigo/documentos/bie/E-coli.txt`

Contenido en GC

En el DNA, los pares de bases pueden ser AT o GC

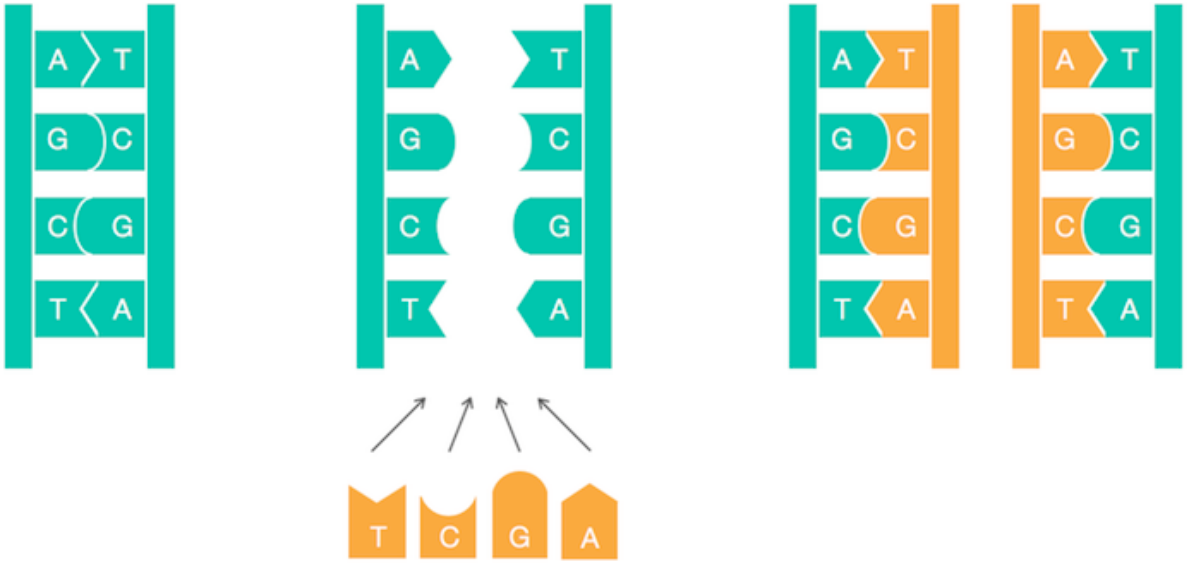
- Los enlaces GC son más estables (aunque a nivel de DNA no es muy significativo)
- **Contenido en GC**: % de bases que son G o C respecto al total
  - Este % varía entre especies!



Calculemos el contenido en GC de E. coli

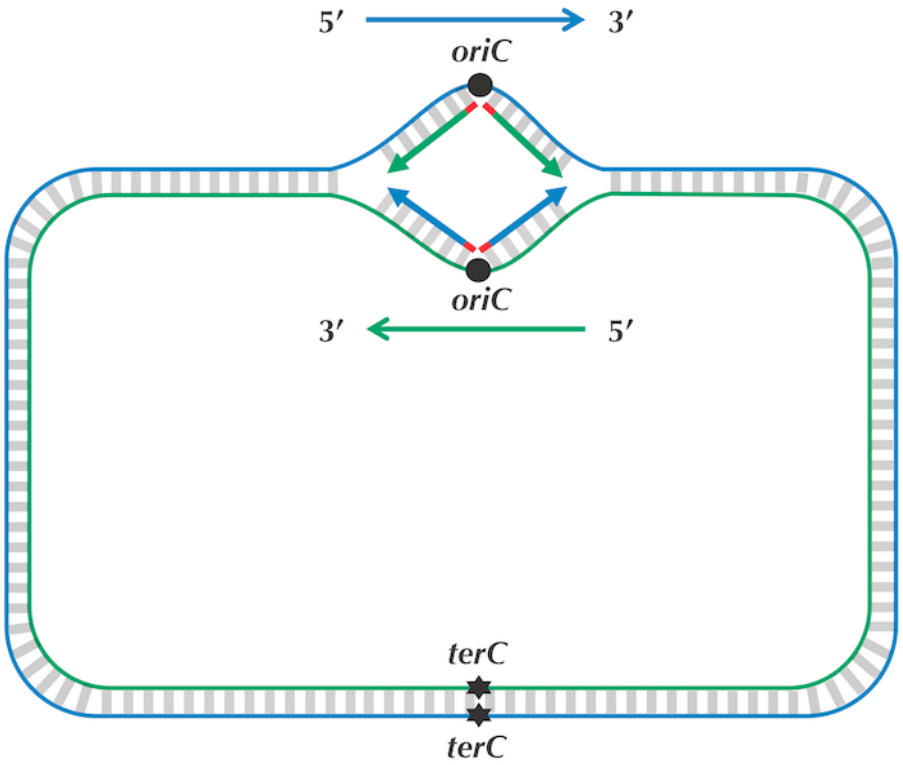
Replicación

- La replicación del genoma es uno de los procesos más importantes de una célula
  - Necesario antes de poder dividirse para dar una copia genética a cada célula hija



Origen de replicación (oriC)

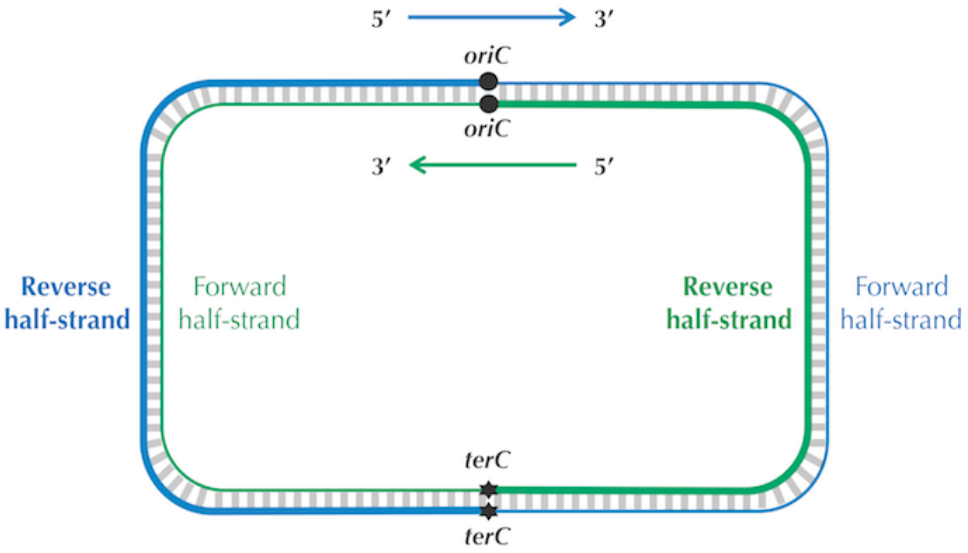
- Región donde comienza la replicación
  - Hasta terminar en otro punto (terC)

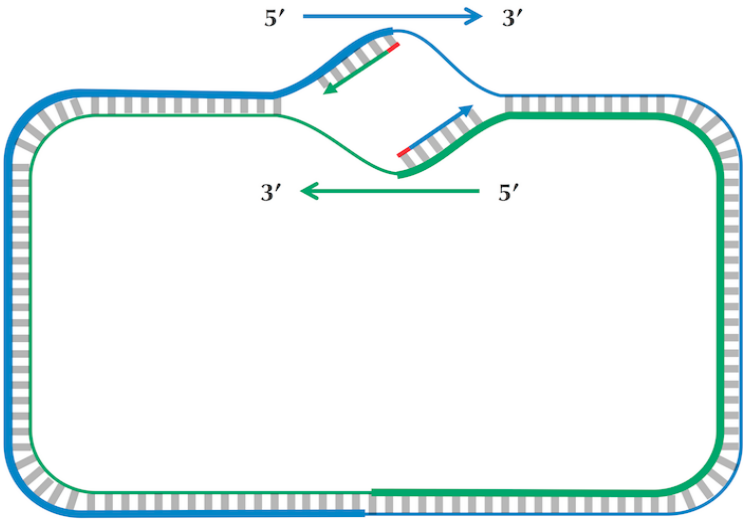


Cuatro DNA polimerasas (rojo) replicando el genoma desde oriC

Asimetría en la replicación

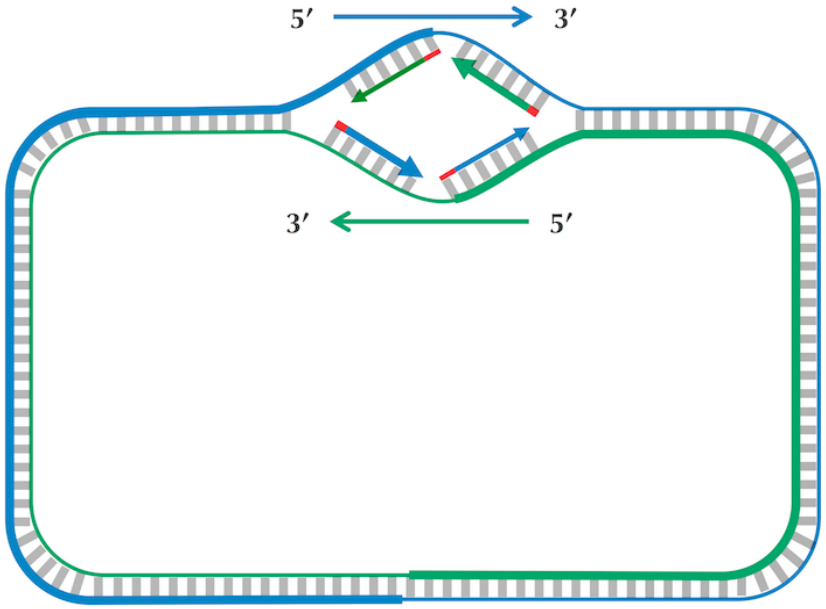
- La DNA polimerasa es una enzima que va 'añadiendo nucleótidos' complementarios a una cadena simple de ADN
  - Sólo lo puede hacer en un sentido: 3' → 5'

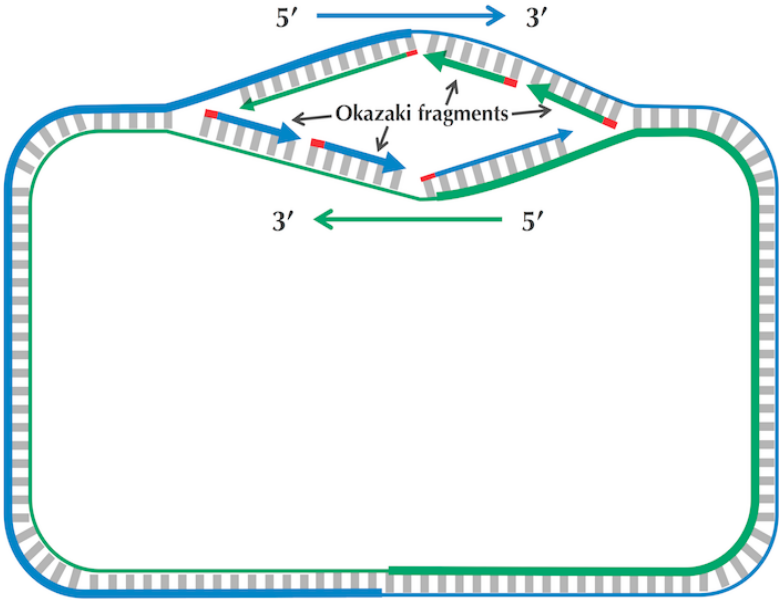




Asimetría en la replicación

- El sentido 5' → 3' se replica 'hacia atrás' cuando la cadena se ha abierto un poco
  - Lo hace trocito a trocito → fragmentos de Oyazaki





Buscando OriC

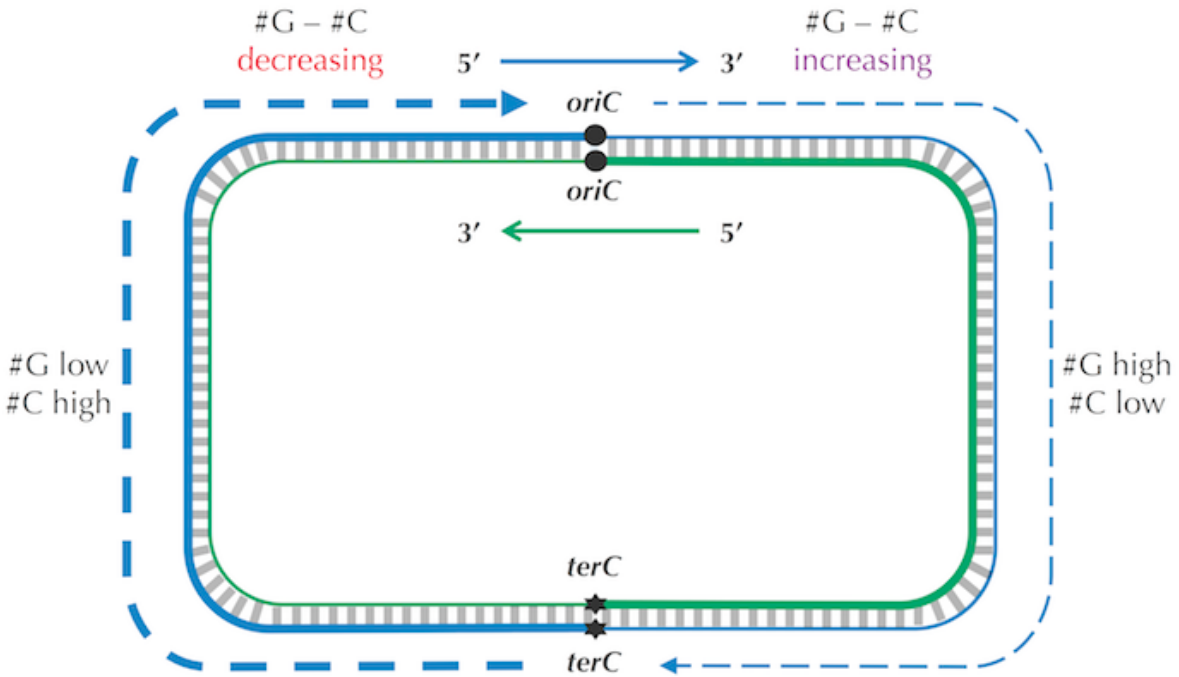
- La cadena 'inversa' tiene que esperar un poco hasta que puede replicarse
- Es decir, pasa más tiempo como cadena 'simple' → es más fácil que sufra mutaciones

	C	G	A	T
Entire strand	427419	413241	491488	491363
Reverse half-strand	219518	201634	243963	246641
Forward half-strand	207901	211607	247525	244722
Difference	**+11617**	**-9973**	-3562	-1919

de nucleótidos en *Thermotoga petrophila*

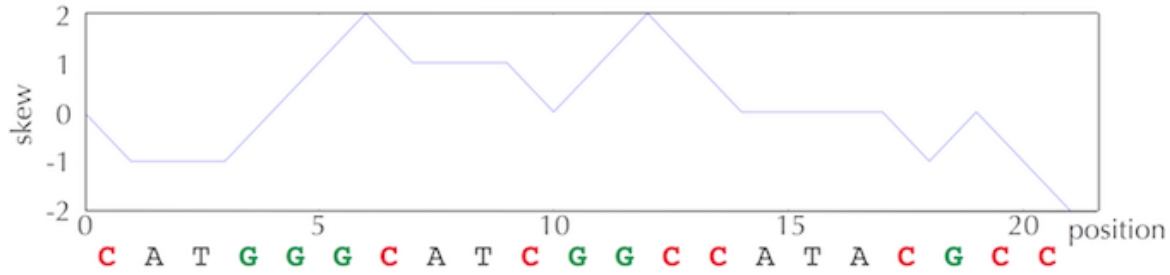
Buscando OriC

- Cómo podemos usar esta estadística para intentar localizar OriC?
  - Observando la diferencia entre G y C a lo largo del ADN!



Buscando OriC

- Calcular el **desvío** *(skew)*: recorremos el ADN desde una posición cualquiera
  - +1 al desvío cuando encontremos una G
  - -1 al desvío cuando encontremos una C



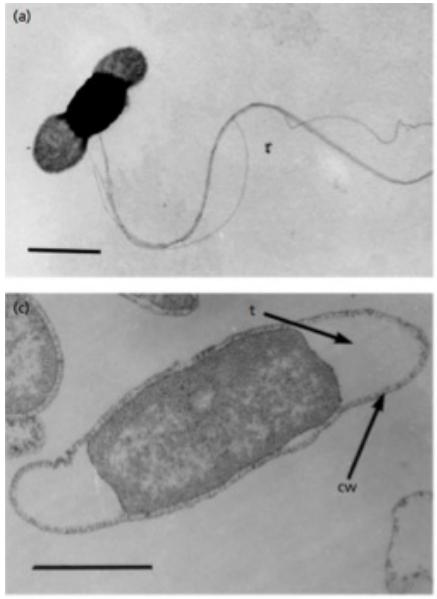
CATGGGCATCGGCCATACGCC → 0 -1 -1 -1 0 1 2 1 1 1 0 1 2 1 0 0 0 0 -1 0 -1 -2

Buscando OriC

- Empecemos con algo sencillo
  - Entrada (cadena de ADN):
    - TAAAGACTGCCGAGAGGCCAACACGAGTGCTAGAACGAGGGGCGTAAACGCGGGTCCGAT
  - Salida (posiciones con un valor mínimo de desvío)
    - 11, 24

Buscando OriC

- Vamos a buscar dónde empieza la replicación en *Thermotoga petrophila*
  - Resiste temperaturas muy altas
  - Descubierta en un depósito de petróleo, a 70º C
  - Entrada: genoma completo:



Takahata et al. 2001

`http://vis.usal.es/rodrigo/documentos/bie/Thermotoga-petrophila.txt`

  - Salida: posiciones de menor desvío (opcional: gráfica)