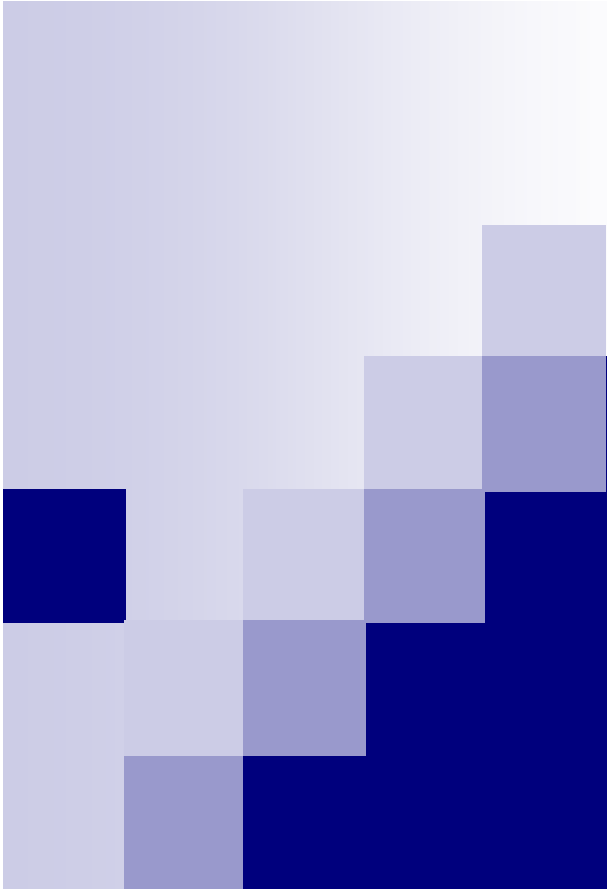




Bioinformática y Biología molecular

Algoritmia para problemas difíciles

24 / 11 / 2014

Elvira Mayordomo

(con comentarios de Jorge Álvarez)



Contenido

- **Introducción y ejemplos**
- Las proteínas
- Los ácidos nucleicos
- El código genético
- Grandes retos algorítmicos:
 - Secuenciación
 - Alineamiento
 - String matching



En los periódicos

- La biotecnología, genética y bioinformática en primera plana
- Todo empezó con el descubrimiento de la estructura del DNA por Watson y Crick en 1953
- En los 90 se inició el proyecto del genoma humano y se clonó a la oveja Dolly
- En el 2000 se anunció la secuenciación completa del genoma humano (proyecto del genoma humano)
- En el 2008 comenzó el proyecto de los 1000 genomas



Ejemplos de bioinformática

- Bioinformática forense
- Detección de enfermedades y tratamientos personalizados
- Estudios epidemiológicos
- Sorpresas

Forenses ...



- Familia del zar Nicolás II
- Supuestamente asesinados en 1918, junto con 4 allegados
- En 1991 se encontraron 9 cuerpos (¿y Anastasia?)
- En 2007 se encontraron 2 cuerpos más



¿Cómo identificar los restos?

- 1) Recuperar material genético
- 2) Compararlo y establecer genealogía

1) Recuperar material genético

```
ATCCATATCTAATCTTACTTATATGTTGTGGAAATGTAAAGAGCCCCATTATCTTAGCCTAAAAAACCTTCTCTTTGGAACCTTTG
AATACGCTTAACTGCTCATTGCTATATTGAAGTACGGATTAGAAGCCGCCGAGCGGGCGACAGCCCTCCGACGGAAGACTCTCCTC
SCGTCTCTCGTCTTCACCGGTCGCGTTCTTGAAACGCAGATGTGCCTCGCGCCGCACGTCTCCGAACAATAAAGATTCACAAATAC
TTTTATGGTTATGAAGAGGAAAAATTGGCAGTAACCTGGCCCCACAAACCTTCAAATTAACGAATCAAATTAACAACCATAGGATC
ATGCGATTAGTTTTTTTAGCCTTATTTCTGGGGTAATTAATCAGCGAAGCGATGATTTTTGATCTATTAAACAGATATATAAATGGAA
CTGCATAACCACTTTAAC TAATAC TTTCAACATTTTCAGTTTGTATTACTTCTTATTCAAATGTCATAAAAGTATCAACAAAAAA
TAATATACCTCTATACCTTTAACGTCAAGGAGAAAAAACTATAATGACTAAATCTCATTTCAGAAGAAGTGATTGTACCTGAGTTCAA
TAGCGCAAAGGAATTACCAAGACCATTGGCCGAAAAAGTGCCCGAGCATAATTAAGAAATTTATAAGCGCTTATGATGCTAAACCGG
TTGTTGCTAGATCGCCTGGTAGAGTCAATCTAATTGGTGAACATATTGATTATTGTGACTTCTCGGTTTTACCTTTAGCTATTGA
SATATGCTTTGCGCCGTCAAAGTTTTGAACGAGAAAAATCCATCCATTACCTTAATAAATGCTGATCCCAAATTTGCTCAAAGGAA
CGATTTGCCGTTGGACGGTTCTTATGTCACAATTGATCCTTCTGTGTCGGACTGGTCTAATTACTTTAAATGTGGTCTCCATGTTC
ACTCTTTTCTAAAGAAACTTGCACCGGAAAGGTTTGGCAGTGCTCCTCTGGCCGGGCTGCAAGTCTTCTGTGAGGGTGATGTACCA
GGCAGTGGATTGTCTTCTTCGGCCGCATTTCATTGTGCCGTTGCTTTAGCTGTTGTTAAAGCGAATATGGGCCCTGGTTATCATAT
CAAGCAAAATTTAATGCGTATTACGGTCGTTGCAGAACATTATGTGGTGTTAAACAATGGCGGTATGGATCAGGC TGCCTCTGT
GTGAGGAAGATCATGCTCTATACGTTGAGTTCAAACCGCAGTTGAAGGCTACTCCGTTTAAATTTCCGCAATTAATAAACCATGAA
AGCTTTTGTTATTGCGAACACCCTTGTTGTAATCTAAACAAGTTTGAAACCGCCCCAACCAATATAATTAAAGAGTGGTAGAAGTCA
AGCTGCAAAATGTTTTAGCTGCCACGTACGGTGTGTTTTACTTTCTTGAAAAAGAAGGATCGAGCACGAATAAAGGTAATCTAAGAC
TCATGAACGTTTTATTATGCCAGATATCACAAACATTTCCACACCCCTGGAACGGCGATATTGAATCCGGCATCGAACGGTTAACAA
CTAGTACTAGTTGAAGAGTCTCTCGCCAATAAGAAACAGGGCTTTAGTGTGACGATGTCGCACAATCCTTGAATTGTTCTCGCGA
ATTACAAAGAGACTACTTAACAACATCTCCAGTGAGATTTCAAGTCTTAAAGCTATATCAGAGGGCTAAGCATGTGTATTCTGAAT
TAAGAGTCTTGAAGGCTGTGAAATTAATGACTACAGCGAGCTTTACTGCCGACGAAGACTTTTTCAAGCAATTTGGTGCCCTTGATC
SAGTCTCAAGCTTCTTGCAGATAAACTTTACGAATGTTCTTGTGCCAGAGATTGACAAAAATTTGTTCCATTGCTTTGTCAAATGGATC
TGGTTCCCGTTTGACCGGAGCTGGCTGGGGTGGTTGTACTGTTCACTTGGTTCCAGGGGGCCCCAAATGGCAACATAGAAAAGGTAA
AAGCCCTTGCCAATGAGTCTTACAAGGTCAAGTACCCTAAGATCACTGATGCTGAGCTAGAAAATGCTATCATCGTCTCTAAACCA
TTGGGCGAGCTGCTATATGAATATAAGTATACCTTCTTTTTTTTACTTTGTTTGCAGAACAACTTCTCATTTTTTTTCTACTCATAAC
GCATCACAAAAATACGCAATAATAACGAGTAGTAACACTTTTATAGTTTATACATGCTTCAACTACTTAATAAATGATTGTATGATA
TTTTCAATGTAAGAGATTTTCGATTATCCACAACTTTTAAACACAGGGACAAAAATTTCTTGATATGCTTTCAACCGCTGCGTTTTG
CCTATTCCTTGACATGATATGACTACCATTTTGTATTGTACGTGGGGCAGTTGACGCTTTATCATATGTCAAAGTCATTTGCGAAC
TTGGCAAGTTGCCAACTGACGAGATGCAGTAAAAAGAGATTGCCGTCTTGAAACTTTTTGTCTTTTTTTTTTCCGGGGACTCTAC
AACCCTTTGTCTCTACTGATTAAATTTGTACTGAATTTGGACAATTGAGATTTTAGTAGACAAGCGCGAGGAGGAAAAAGAAATGAC
AAATTCGGATGGACAAGAAGATAGGAAAAAAGCTTTACACGATTTCC TAGACCGGAAAAAAGTCGTATGACATCAGAATGA
ATTTTCAAGTTAGACAAGGACAAAAATCAGGACAAATTTGTAAGATATAATAAAC TATTTGATTTCAGCGCCAATTTGCCCTTTTCCA
TCCATTAATATCTCTGTTCTCTCTTACTTATATGATGATTAGGTATCATCTGTATAAAACTCCTTTCTTAATTTCACTCTAAAGCA
CCATAGAGAAGATCTTTTCGGTTTCGAAGACATTCCTACGCATAATAAGAATAGGAGGGGAATAATGCCAGACAATCTATCATTACAT
SCGGCTCTTCAAAAAGATTGAACCTCTCGCCAACTTATGGAATCTTCCAATGAGACCTTTGCGCCAAATAATGTGGATTGGA
TATAAGTCATCTCAGAGTAATAAC TACCAAGT TTAGAGGCACTCGAGCTTTGAAGAAAAAGTAAGCTCAGAAAAACCTCAATA
CTCATTCGGAAGAAAAATCTATATGAATATGTGGTTCGTTGACAAATCAATCTTGGGTGTTTTCTATTCTGGATTTCATTTATGTACA
AGGACTTGAAGCCCGTCGAAAAAGAAAGGCGGGTTTGGTCTTGGTACAAT TATTGTTACTTCTGGCTTGCTGAATGTTTCAATATC
ACTTGGCAAAATTGCAGCTACAGGTCTACAAC TGGGTCTAAATTTGGTGGCAGTGTGGATAACAATTTGGATTGGGTACGGTTTCG
```

- Unas 10 pantallas como esta por persona (hay 3000 caracteres por pantalla)

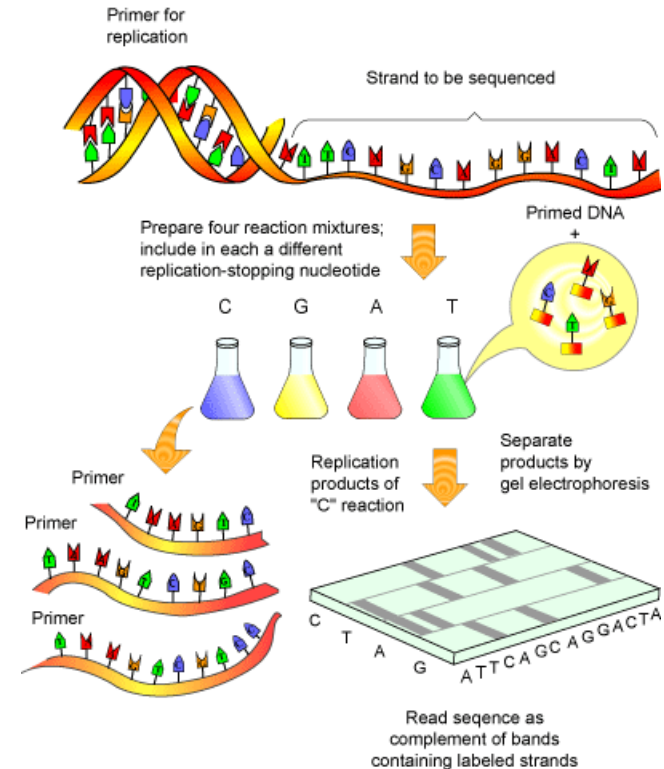
1) Recuperar material genético

¿trabajo de laboratorio?

En el laboratorio se generan muchas copias de la cadena que nos interesa, se trocean (trozos de unos 300 caracteres) y se leen

Imposible leer trozos más grandes ...

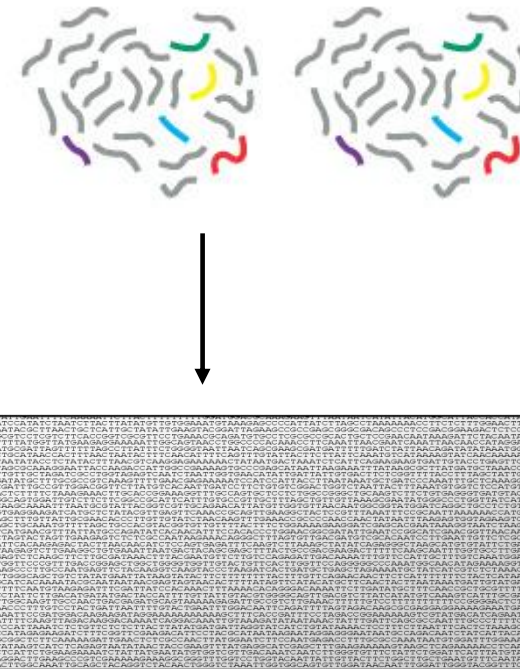
La informática tiene que hacer el resto ...





- A partir de muchos trozos pequeños (con repeticiones y solapes)
- Recuperar el material original

Secuenciación





¿Cómo identificar los restos?

- 1) Recuperar material genético
- 2) Compararlo y establecer genealogía



2) Compararlo y establecer genealogía

Todo es bioinformática:

- Comparar largas secuencias en las que puede haber pequeños errores y omisiones: primero hay que “alinearlas”
- Asegurarse de que en las posiciones significativas hay parecidos no casuales (estudios evolutivos)



2) Compararlo y establecer genealogía

En el caso de los Romanov el material obtenido permitió establecer la línea materna de todos:

- 6 cuerpos (la zarina y sus 5 hijos) estaban emparentados entre sí y con el marido de la reina Isabel
- El 7º (el zar) estaba emparentado con parientes maternos del zar

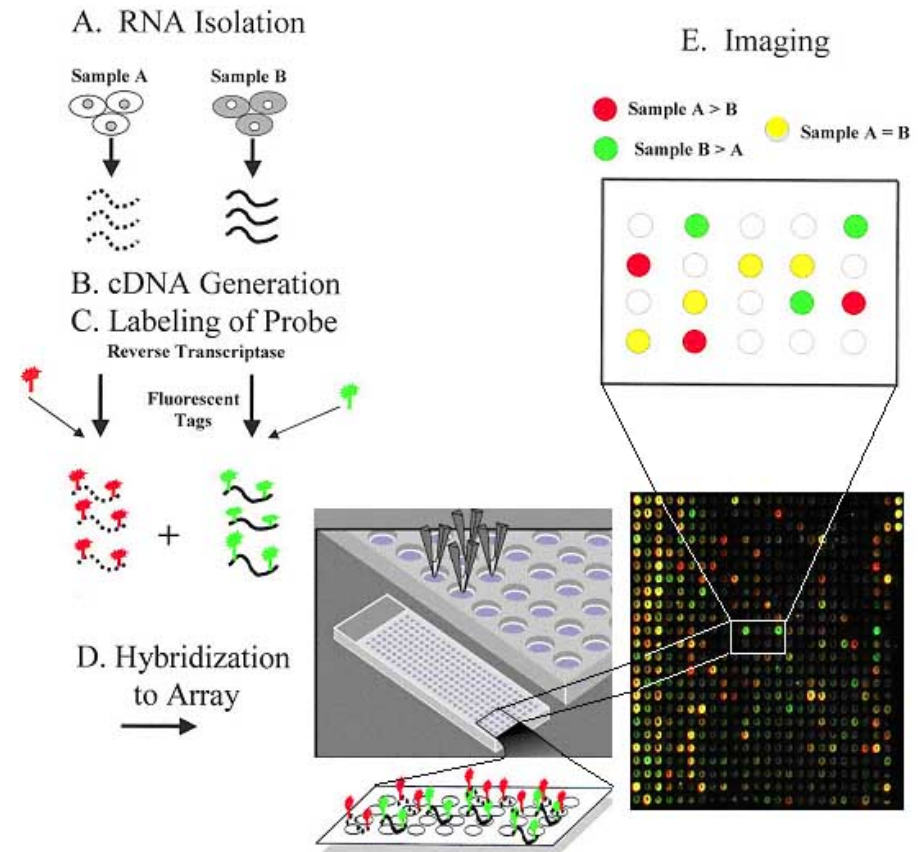


¿Tendré cáncer de ...?

- Para algunos tipos de cáncer existe una tendencia genética a padecerlo
- No todos los miembros de una misma familia tienen esta tendencia
- Además nada es blanco o negro, el tono de gris es importante

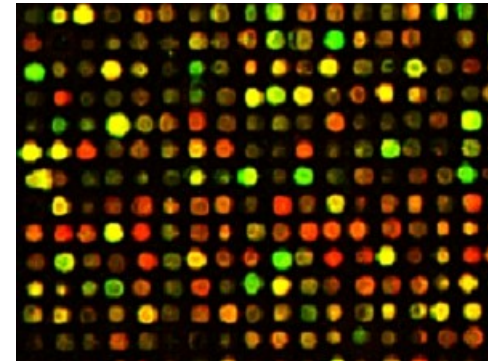
¿Tendré cáncer de ...?

- Se diseña un microarray que recoge diferentes genes que tienen influencia en ese cáncer
- Se extrae el material genético del paciente (A) y otro del control (B)
- Se buscan las coincidencias entre cada uno y el microarray
- Más bioinformática: interpretar resultados



Tratamiento personalizado

- A partir de la “huella” de un microarray la bioinformática permite averiguar cómo de activos están una serie de genes



- Con esta información se diseñan los medicamentos (también usando bioinformática)

Tratamiento personalizado

Ancestry Composition

LOG OUT

Standard E

Sh

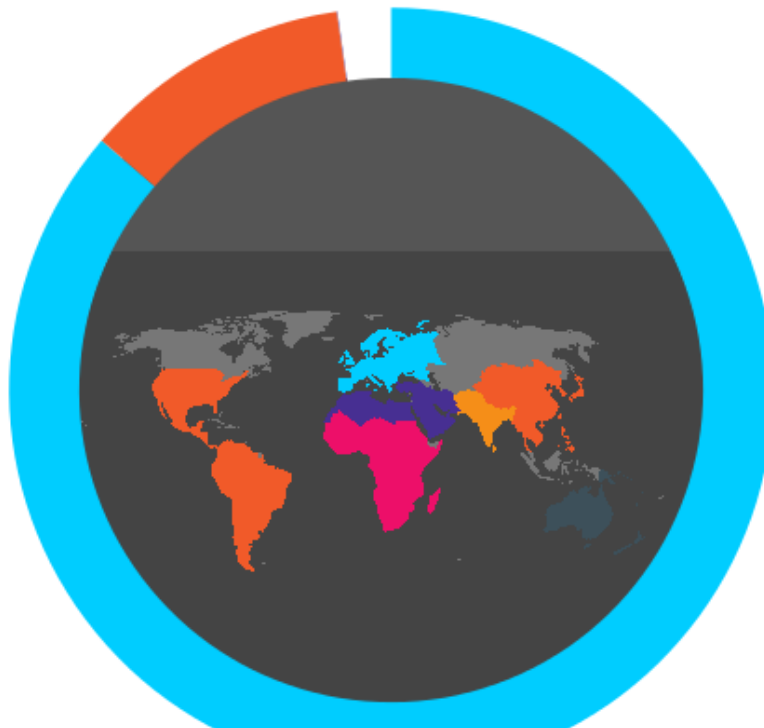
Map View



Global Resolution



Ancestry Composition tells you what percent of your DNA comes from each of 22 populations worldwide. The analysis includes DNA you received from all of your ancestors, on both sides of your family. The results reflect where your ancestors lived 500 years ago, before ocean-crossing ships and airplanes came on the scene.



86.3%	European
11.4%	East Asian & Native American
< 0.1%	Middle Eastern & North African
0.0%	Sub-Saharan African
0.0%	South Asian
0.0%	Oceanian
2.3%	Unassigned

100%

show only [redacted] populations

Tratamiento personalizado

Health Risks (120) ?

↑ ELEVATED RISKS

	YOUR RISK	AVERAGE RISK
Atrial Fibrillation	33.9%	27.2%
Age-related Macular Degeneration	13.8%	6.5%
Rheumatoid Arthritis	6.1%	2.4%
Chronic Kidney Disease	4.2%	3.4%
Melanoma	4.0%	2.9%

[See all 120 risk reports...](#)

Inherited Conditions (50, 2 locked reports) ?

REPORT	RESULT
Alpha-1 Antitrypsin Deficiency	Variant Absent
Agenesis of the Corpus Callosum with Peripheral Neuropathy (ACCPN)	Variant Absent
Autosomal Recessive Polycystic Kidney Disease	Variant Absent
ARSACS	Variant Absent
Beta Thalassemia update	Variant Absent
Bloom's Syndrome	Variant Absent
Canavan Disease	Variant Absent
Congenital Disorder of Glycosylation Type 1a (PMM2-CDG)	Variant Absent

[See all 50 carrier status...](#)

Traits (57) ?

REPORT	RESULT
Alcohol Flush Reaction	Does Not Flush
Bitter Taste Perception	Unlikely to Taste
Earwax Type	Wet

Drug Response (21) ?

REPORT	RESULT
Pseudocholinesterase Deficiency	Increased
Abacavir Hypersensitivity	Typical
Alcohol Consumption, Smoking and Risk of Esophageal	Typical



Estudio epidemiológico

- Comparación entre los ancianos de una residencia de Jaca y otra de Tudela
- Los de Jaca tienen una longevidad mayor
- La bioinformática es capaz de almacenar el material genético de todos ellos, compararlo y encontrar la diferencia
- Los médicos y biólogos explican esa diferencia, se trata de un gen que regula el consumo de oxígeno



Descubrimientos sorpresa

- La empresa de bioinformática de Islandia DeCode Genetics descubre una mutación que provoca un aumento de la fertilidad humana
- Se trata de una inversión de 900.000 bases (caracteres) que se da sobre todo en europeos
- Este tipo de descubrimientos no es posible sin técnicas sofisticadas de tratamiento de información



¿Qué es bioinformática?

■ Bioinformática es:

- La aplicación de la informática y tecnologías cercanas para resolver preguntas de Biología sobre los misterios de la vida
- Principalmente se ocupa de los problemas que tienen que ver con datos extraídos de células de seres vivos, especialmente secuencias (DNA, RNA, proteínas: biología molecular)



Contenido

- Introducción y ejemplos
- **Las proteínas**
- Los ácidos nucleicos
- El código genético
- Grandes retos algorítmicos:
 - Secuenciación
 - Alineamiento
 - String matching

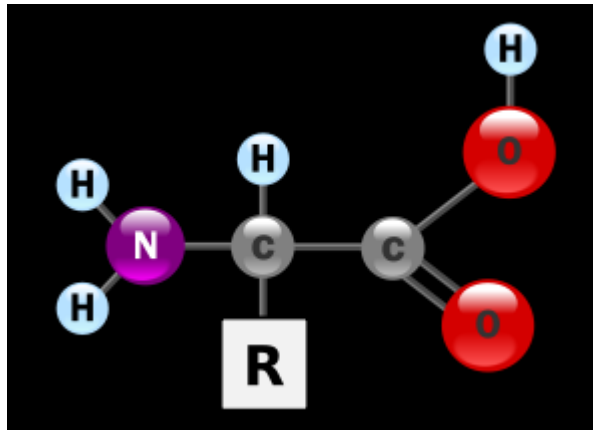


Proteínas

- La clase de moléculas más importantes de los seres vivos
- Funciones: como enzimas (catálisis de procesos metabólicos), en transmisión de señal, mecanismos de defensa, transporte de moléculas, material de construcción

Los aminoácidos forman las proteínas

- Una proteína es una cadena de aminoácidos
- Un aminoácido:





20 aminoácidos

- Según R aparecen 20 aminoácidos distintos que forman parte de las proteínas

Ala A (H)	Val V (H)	Leu L (H)	Ile I (H)	Phe F (H)
Pro P (H)	Met M (H)	Ser S (P)	Thr T (P)	Cys C (P)
Trp W (H)	Tyr Y (P)	Asn N (P)	Gln Q (P)	Asp D (P)
Glu E (P)	Lys K (P)	Arg R (P)	His H (P)	Gly G (P)



Proteínas = cadenas de aminoácidos

- Los aminoácidos se unen mediante enlaces peptídicos
- Podemos representar una proteína como una cadena leída de (H^2N a COOH)

VHLTPEEK ...

- Esto es mucho simplificar, ignorando la estructura espacial ...



Contenido

- Introducción y ejemplos
- Las proteínas
- **Los ácidos nucleicos**
- El código genético
- Grandes retos algorítmicos:
 - Secuenciación
 - Alineamiento
 - String matching

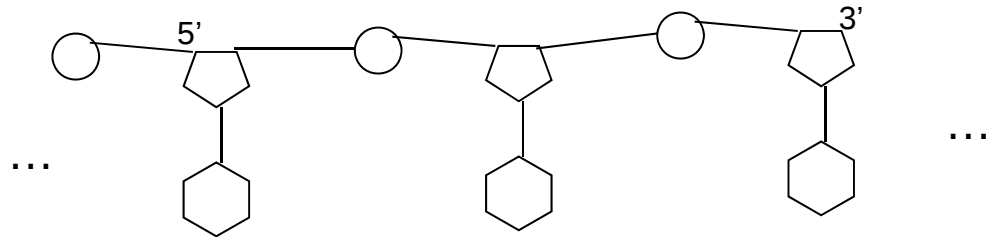
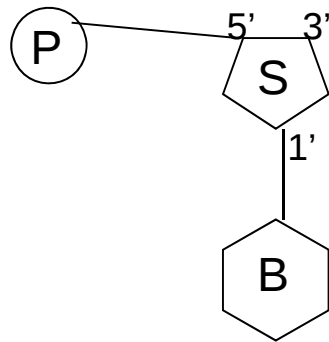


Ácidos nucleicos

- Las moléculas más importantes después de las proteínas
- En todos los seres vivos, son las responsables de codificar y almacenar la información genética
- Permiten la transmisión de información genética de una generación a otra
- Tienen una estrecha conexión con las proteínas: los ácidos nucleicos sirven como mapas para la construcción de las proteínas

Ácido nucleico= cadena de nucleótidos

- Cada ácido nucleico está formado por nucleótidos encadenados





Cadenas ...

- La B (base) caracteriza al nucleótido
- Podemos escribir un ácido nucleico como una secuencia de nucleótidos
- Gran simplificación



DNA y RNA

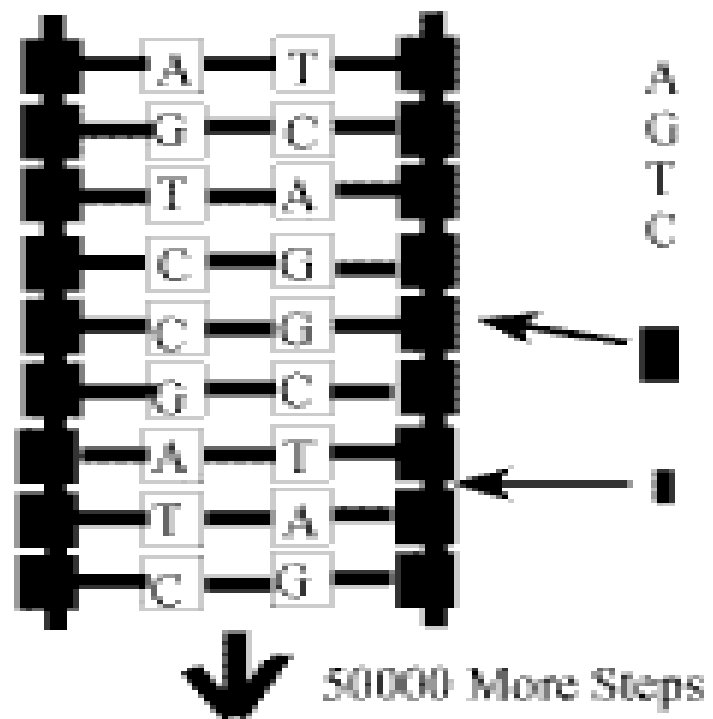
- Los dos tipos de ácidos nucleicos son DNA y RNA
- Se diferencian por los azúcares (S)
- DNA usa cuatro bases A, C, G, T
- RNA usa cuatro bases A, C, G, U



DNA

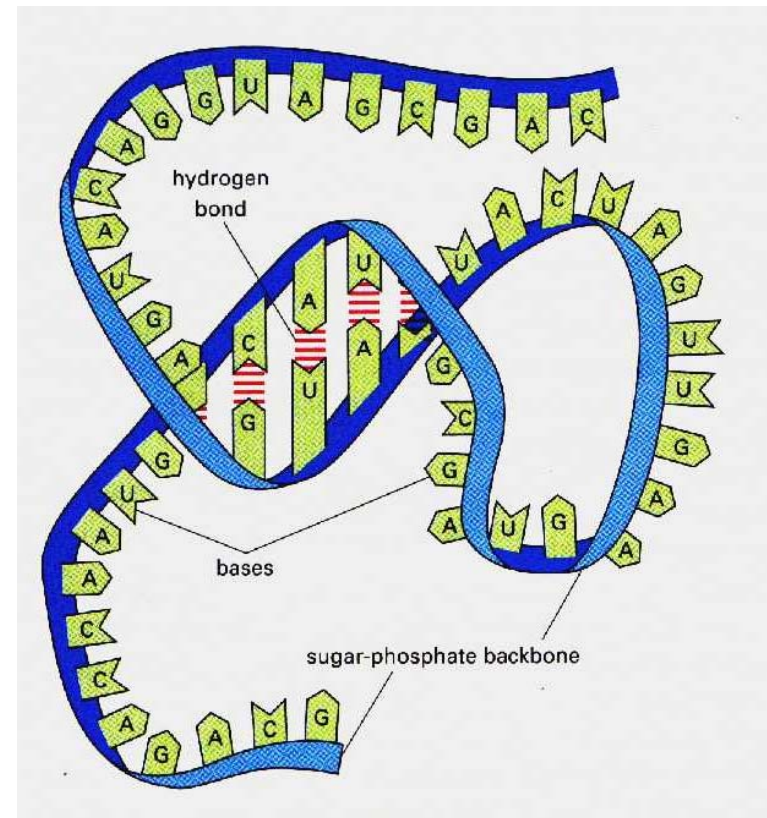
- Está formado por dos cadenas de nucleótidos
 - Complementarias (A-T y C-G)
 - Se leen en direcciones opuestas
 - En forma de hélice

DNA



RNA

- Normalmente una sola cadena
- Trozos de la misma molécula se unen con otros complementarios lo que da formas diversas





Triple estructura

- Primaria: cadena de nucleótidos
- Secundaria: describe los trozos complementarios que están unidos
- Terciaria: cómo está doblada en el espacio



Cómo escribimos DNA

- Usamos la dirección de lectura de la cadena de “arriba” (de 5’ a 3’)

Ejemplo: s=AGACGT es:

s: 5' ...AGACGT ... 3'

$\bar{s}$: 3' ...TCTGCA ... 5'



Longitudes

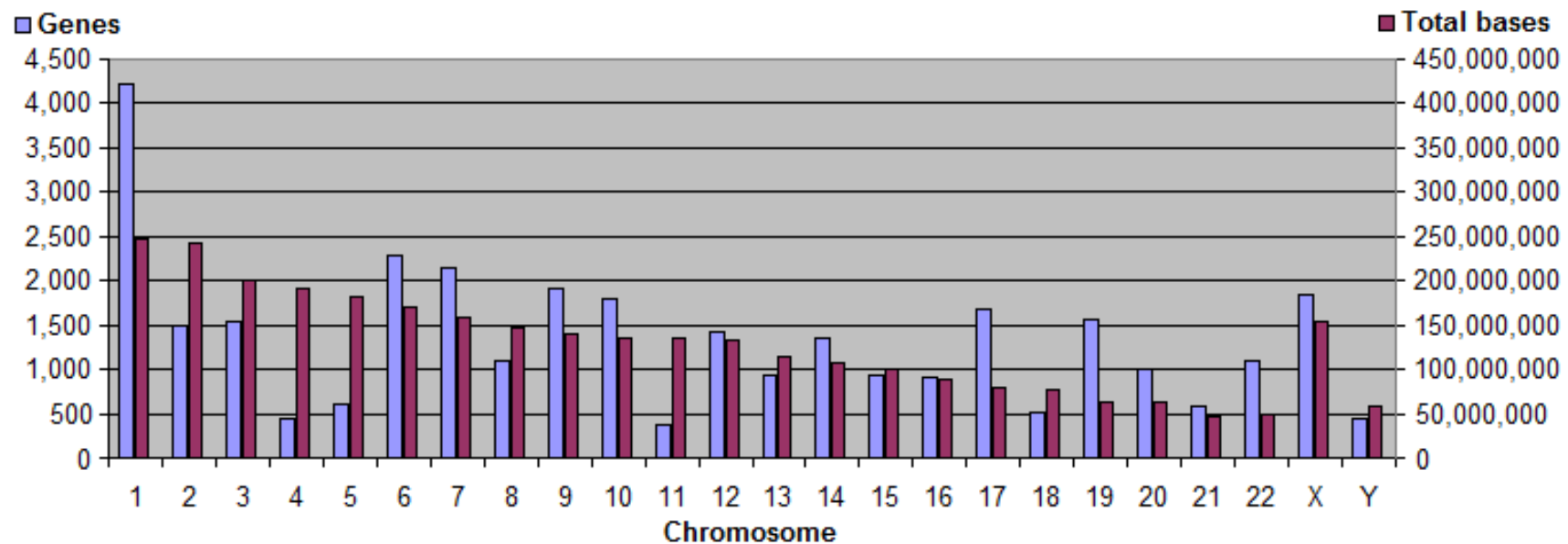
- Se mide la longitud en pares de bases (bp)
- Se usan las unidades kbp (1000 bp) y mbp (1000 kbp)
- AGACGT tiene 6 bp
- También se usan para RNA



Información hereditaria

- Una región de DNA que codifica una proteína se llama **gen**
- Una molécula de DNA que tiene varios genes se llama **cromosoma**
- Los cromosomas suelen aparecer en pares: **cromosomas homólogos** (uno materno y uno paterno)
- Las células humanas tienen 46 cromosomas
- Toda la información hereditaria de una célula se llama **genoma**

23 pares de cromosomas humanos





Contenido

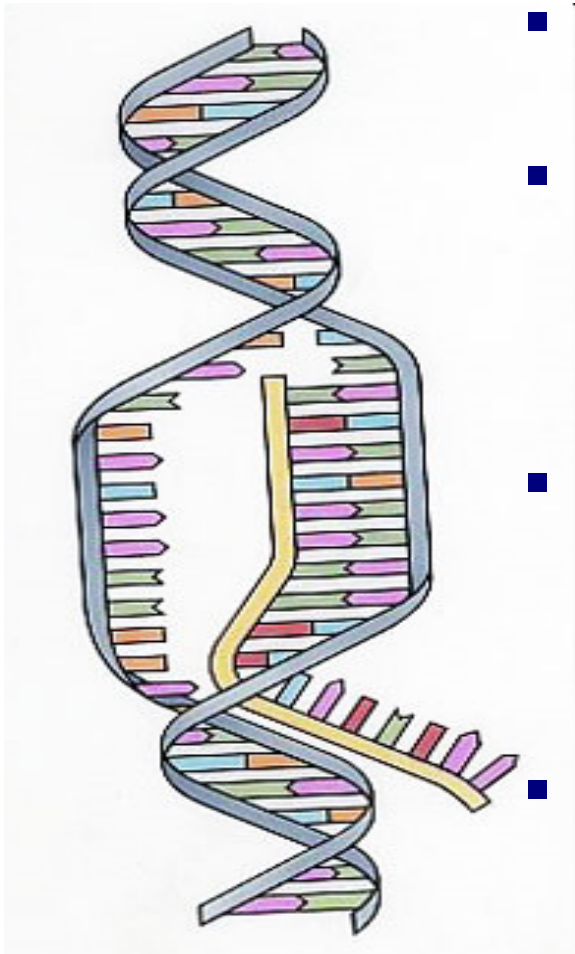
- Introducción y ejemplos
- Las proteínas
- Los ácidos nucleicos
- **El código genético**
- Grandes retos algorítmicos:
 - Secuenciación
 - Alineamiento
 - String matching



Síntesis de proteínas

- El DNA está en el núcleo y las proteínas se sintetizan fuera (en los ribosomas)
- Hay dos pasos: copia y traducción

Transcripción o copia



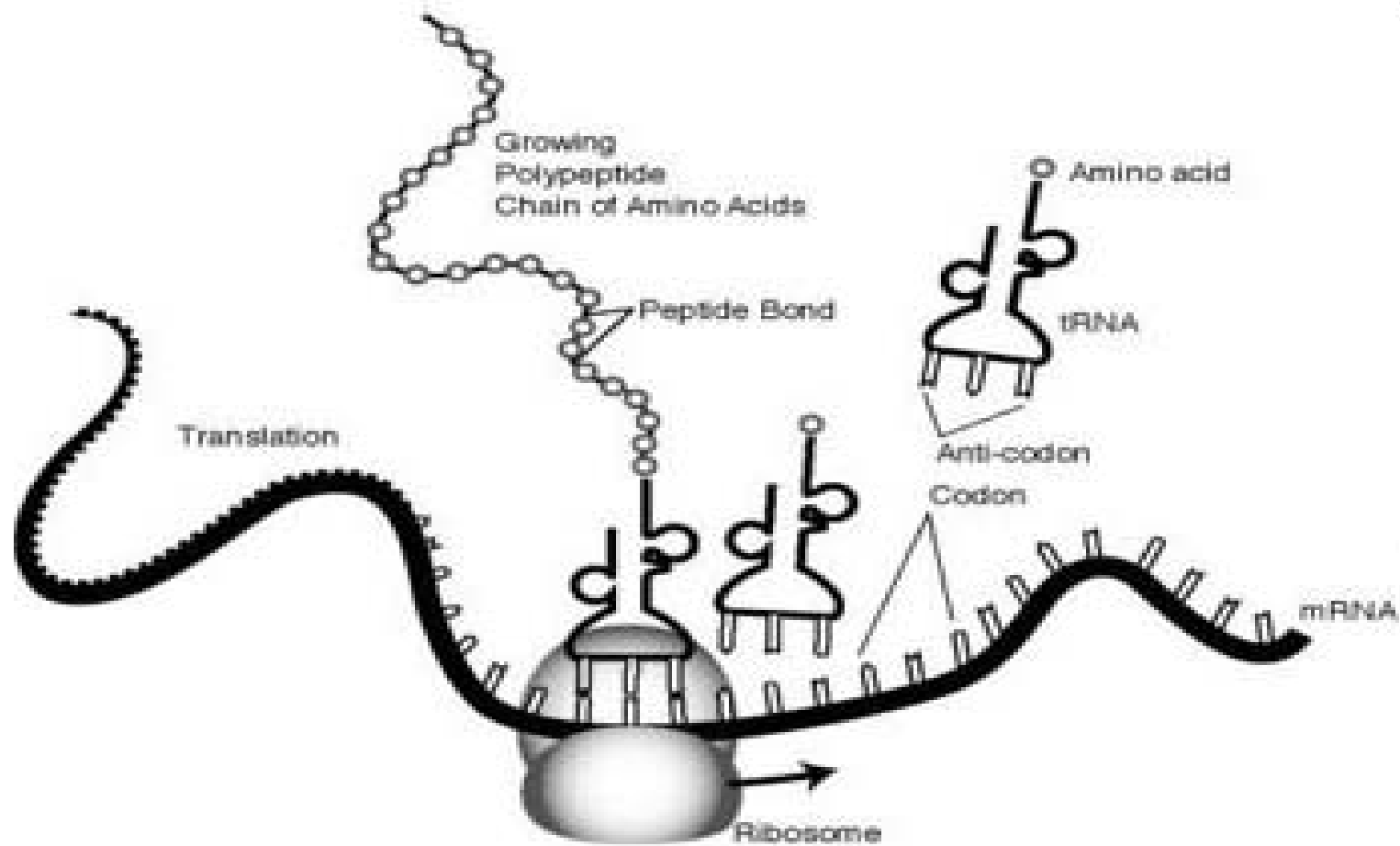
- Se separan las dos copias de DNA
- Se sintetiza una copia complementaria de RNA (amarillo) sustituyendo T por U
- Sólo se copian algunas zonas (los exones) y no los intrones (¿no relevantes?)
- El resultado es el mRNA (RNA mensajero)



Traducción

- La información del mRNA se convierte en una secuencia de aminoácidos
- Un codón son 3 bases. Cada codón codifica un aminoácido
- También hay codones que codifican STOP

Traducción



Codón-aminoácido

		Second Position									
		U		C		A		G			
		code	Amino Acid	code	Amino Acid	code	Amino Acid	code	Amino Acid		
First Position	U	UUU	phe	UCU	ser	UAU	tyr	UGU	cys	U	Third Position
		UUC		UCC		UAC		UGC		C	
		UUA	leu	UCA		UAA	STOP	UGA	STOP	A	
		UUG		UCG		UAG	STOP	UGG	trp	G	
	C	CUU	leu	CCU	pro	CAU	his	CGU	arg	U	
		CUC		CCC		CAC		CGC		C	
		CUA		CCA		CAA	gln	CGA		A	
		CUG		CCG		CAG		CGG		G	
	A	AUU	ile	ACU	thr	AAU	asn	AGU	ser	U	
		AUC		ACC		AAC		AGC		C	
		AUA		ACA		AAA	lys	AGA	arg	A	
		AUG	met	ACG		AAG		AGG		G	
	G	GUU	val	GCU	ala	GAU	asp	GGU	gly	U	
		GUC		GCC		GAC		GGC		C	
		GUA		GCA		GAA	glu	GGA		A	
		GUG		GCG		GAG		GGG		G	



Aminoácido-codón

Inverse table

Ala/A	GCU, GCC, GCA, GCG	Leu/L	UUA, UUG, CUU, CUC, CUA, CUG
Arg/R	CGU, CGC, CGA, CGG, AGA, AGG	Lys/K	AAA, AAG
Asn/N	AAU, AAC	Met/M	AUG
Asp/D	GAU, GAC	Phe/F	UUU, UUC
Cys/C	UGU, UGC	Pro/P	CCU, CCC, CCA, CCG
Gln/Q	CAA, CAG	Ser/S	UCU, UCC, UCA, UCG, AGU, AGC
Glu/E	GAA, GAG	Thr/T	ACU, ACC, ACA, ACG
Gly/G	GGU, GGC, GGA, GGG	Trp/W	UGG
His/H	CAU, CAC	Tyr/Y	UAU, UAC
Ile/I	AUU, AUC, AUA	Val/V	GUU, GUC, GUA, GUG
START	AUG	STOP	UAG, UGA, UAA



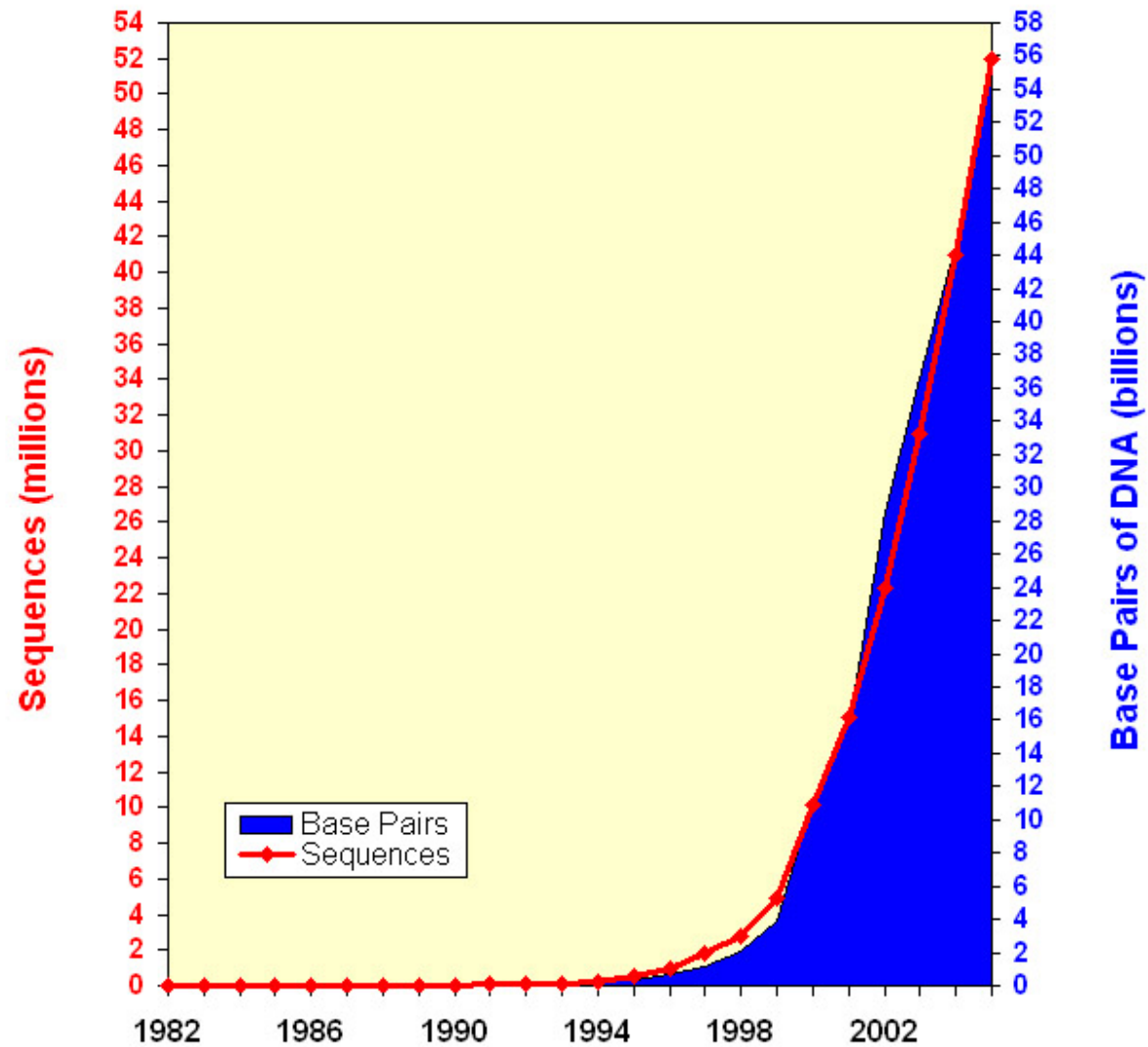
Algunos datos ...

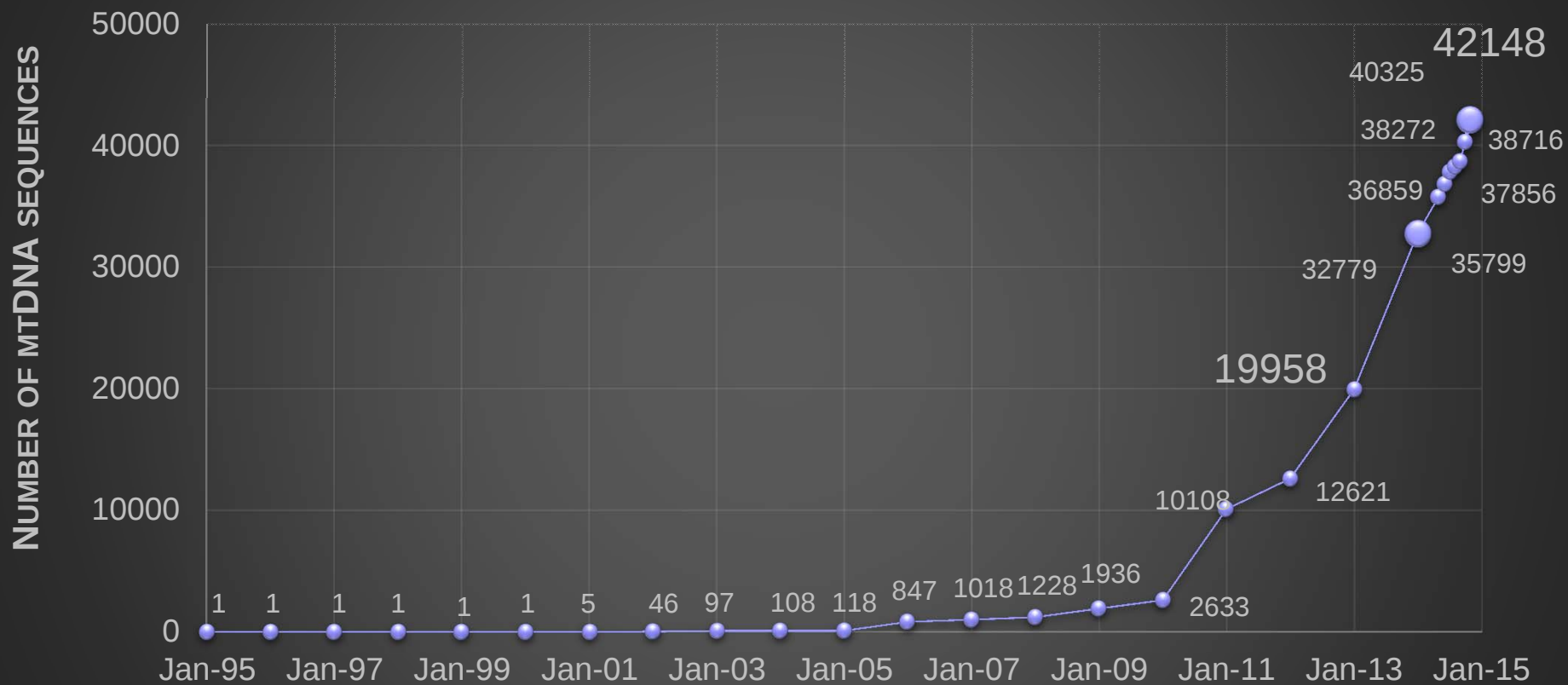
- En los humanos el DNA cromosómico es de 3.000 millones de bp (pares de bases)
- Contiene relativamente poca información (10-20%)
- Se dice que se ha secuenciado el DNA de un individuo cuando se conocen las zonas que se consideran relevantes (genes y otras)
- También hay DNA mitocondrial ...



Growth of GenBank

(1982 - 2005)







Técnicas experimentales

- Necesario conocerlas lo suficiente para entender la explosión de datos actual



Técnicas experimentales

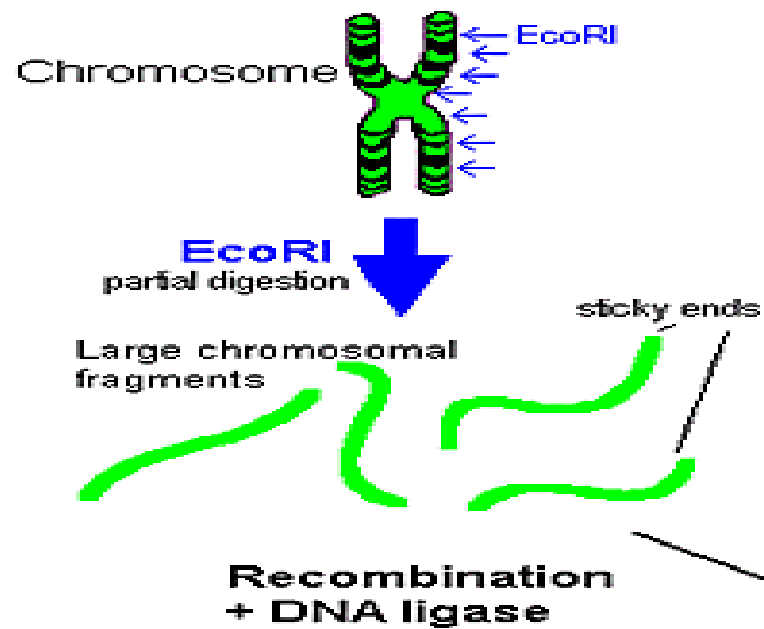
- “Denaturation”: separar las dos hebras del DNA por calor
- Hibridización: asociar bases complementarias o hebras complementarias
- Se puede cortar una doble hebra por un sitio concreto (enzimas de restricción)
- Se pueden reunificar después los trozos (ligasa)



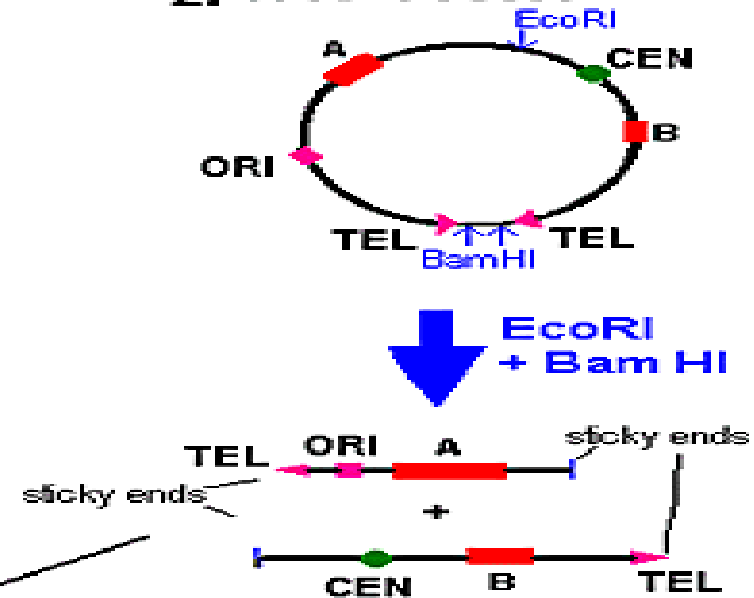
Duplicación de DNA: cloning

- Para los experimentos una sola molécula de DNA no es suficiente, es necesario un gran número de copias idénticas
- Cloning: se inserta el fragmento a copiar en un organismo “host” (anfitrión) , se replica con la reproducción natural y luego se vuelve a extraer
- Los “host” son muy variados, pueden copiar desde 15-50 kbp (bacteria) a varios millones (inserción de cromosomas artificiales)

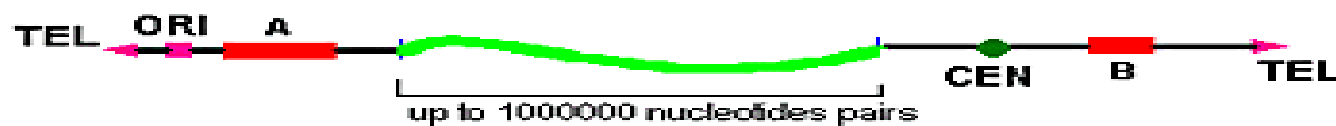
1. Human DNA



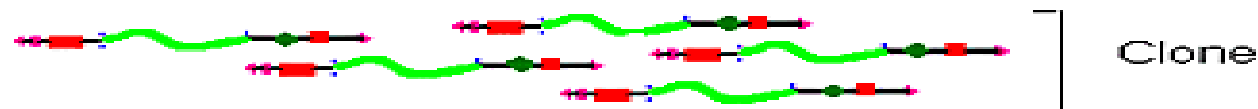
2. YAC vector



3. Yeast artificial chromosome with inserted human DNA



yeast cell transformation



Cloning into a Yeast Artificial Chromosome (YAC)



Problemas del cloning

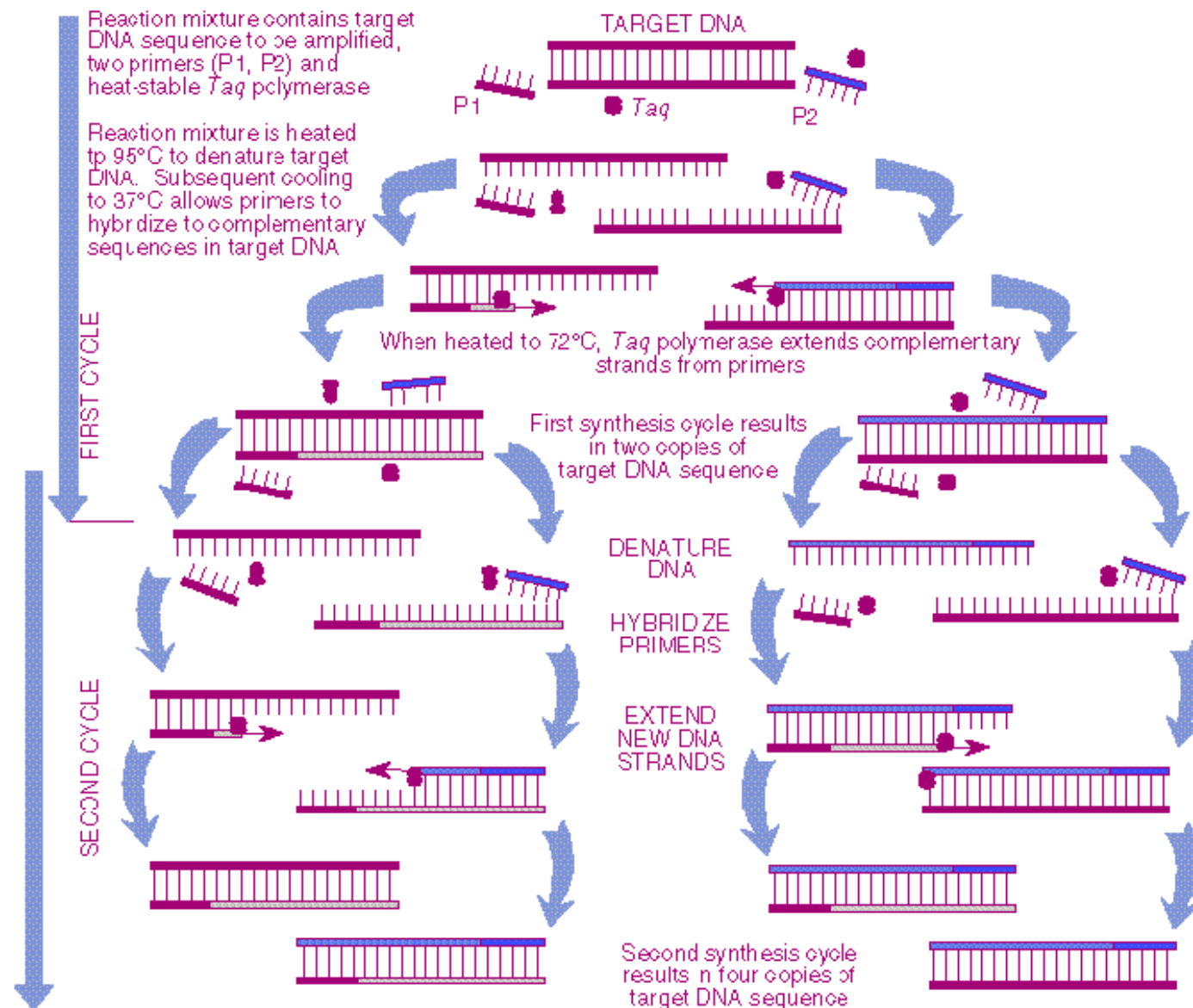
- Contaminación con el DNA del host
- Pérdida de fragmentos completos, cuando la inserción tiene efectos letales en el host
- Dos fragmentos no consecutivos pueden unirse en la clonación (clon quimérico)



Duplicación de DNA: PCR

- Polymerase chain reaction
- Necesitamos conocer un fragmento inicial y otro final
- En cada paso duplica el número de copias
- Los errores al principio son muy peligrosos ...

DNA Amplification Using Polymerase Chain Reaction



Source: *DNA Science*, see Fig. 13.



Gel electrophoresis

- Se trata de separar los fragmentos por tamaño
- Se meten en gel y se aplica un campo eléctrico, la velocidad es inversamente proporcional al tamaño
- Se separan así por longitudes
- Usando trozos de referencia se puede usar para medir la longitud

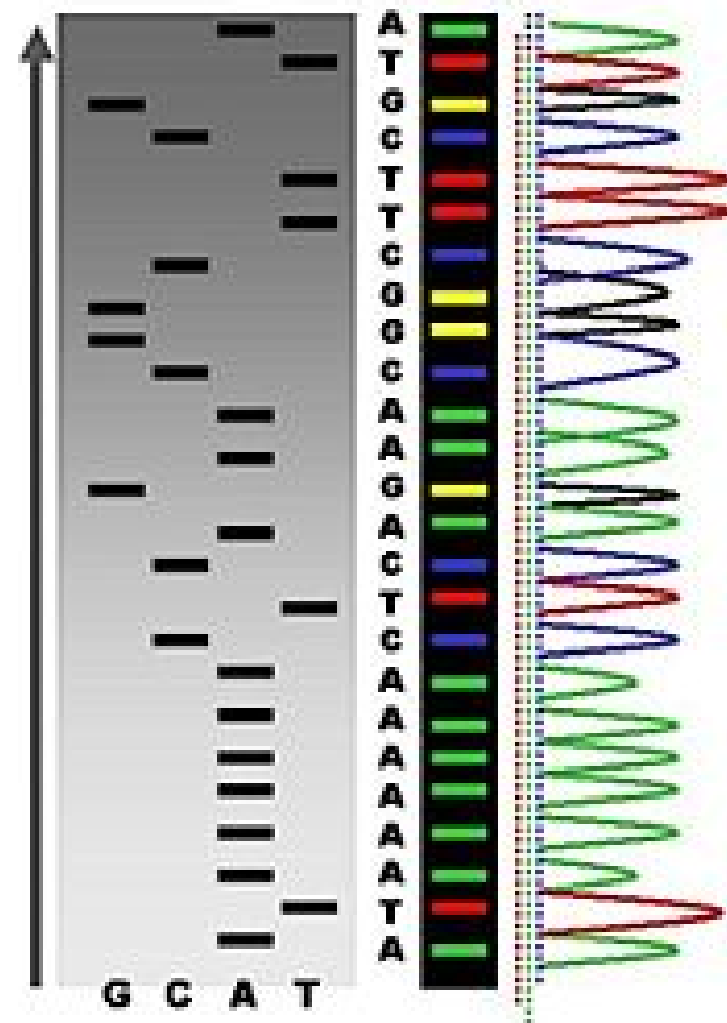


Secuenciar DNA: chain termination method

- El “chain termination method” se basa en el anterior (gel electrophoresis)
- Tenemos un fragmento de DNA desconocido s, hacemos muchas copias
- Paso 1: conseguir que haya 4 tubos de ensayo A, C, G, T cada uno conteniendo los prefijos de s que terminan en A (C,G,T)

chain termination method (2)

- Paso 2: Colocamos los cuatro tubos de ensayo en paralelo y ordenamos por longitud como antes ...





Chain termination method (3)

- Sólo sirve para fragmentos de hasta 1000 bp (más da demasiados errores)
- Puede dar errores de lectura del resultado (llamados errores de secuenciación):
 - ☐ Inserción
 - ☐ Borrado
 - ☐ Sustitución



Experimentos de hibridización

- Para averiguar si un fragmento desconocido s contiene una secuencia t
- Sintetizamos t' la complementaria de t
- Testeamos si s y t' se unen (hibridizan)



DNA chips

- Para hacer varios experimentos de hibridización en paralelo
- Si queremos saber si s contiene $t_1, \dots, t_n$:
 - Colocamos $t'_1, \dots, t'_n$ en sitios fijos (DNA chip)
 - Hacemos copias etiquetadas de s
 - Dejamos que se unan al chip
 - Lavamos las copias de s sueltas y averiguamos las posiciones de hibridización con las etiquetas



DNA chips: errores

- Falsos positivos
- Falsos negativos
- Se pueden usar los DNA chips para RNA
- ...



Mucho por hacer ...

- Hay enormes bases de datos de DNA (en parte gracias a la informática)
- Las preguntas que se quieren contestar son multitud y muy importantes
 - Enfermedades hereditarias
 - Medicina forense y criminología
- ...
- Muchas de estas preguntas necesitan una buena dosis de informática ...



Contenido

- Introducción y ejemplos
- Las proteínas
- Los ácidos nucleicos
- El código genético
- **Grandes retos algorítmicos:**
 - Secuenciación
 - Alineamiento
 - String matching



Secuenciación

- El objetivo es determinar una secuencia de “nucleótidos” que son las piezas que forman el DNA humano, es decir, la molécula que guarda nuestra información genética
- Desde el punto de vista informático buscamos un string hecho con las letras que representan los nucleótidos
- Conocemos métodos para leer estas secuencias desde los 80, pero con longitudes muy restringidas (hoy unos 1000 nucleótidos) → métodos “next generation”
- Nos interesan moléculas de DNA con cientos de miles



Secuenciando DNA ... ¿Cómo?

- Se generan muchas copias de la molécula de DNA que nos interesa
- Rompemos aleatoriamente esas copias en trozos, idealmente pequeños
- Con alta probabilidad esos trozos se solapan entre sí
- Leemos (“secuenciamos”) los trozos
- Nos quedan muchos (miles de) trozos que son subsecuencias de la que buscamos, con solapamientos
- No tenemos idea de cómo combinarlos, el orden se ha perdido
- Aquí entra la informática ...



Secuenciando DNA : 2 métodos

Mapa físico

- Varias copias se cortan en trozos grandes con solape (50-300 kbp). Se pierde el orden de los trozos
- Un mapa físico es un conjunto de trozos pequeños (marcadores) junto con su posición exacta en la secuencia
- Se usa el mapa físico para reconstruir el orden, después los trozos se cortan y se secuencian
- Ejemplo: proyecto del genoma humano

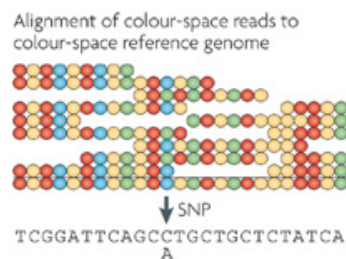
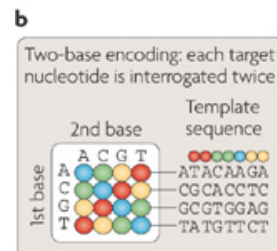
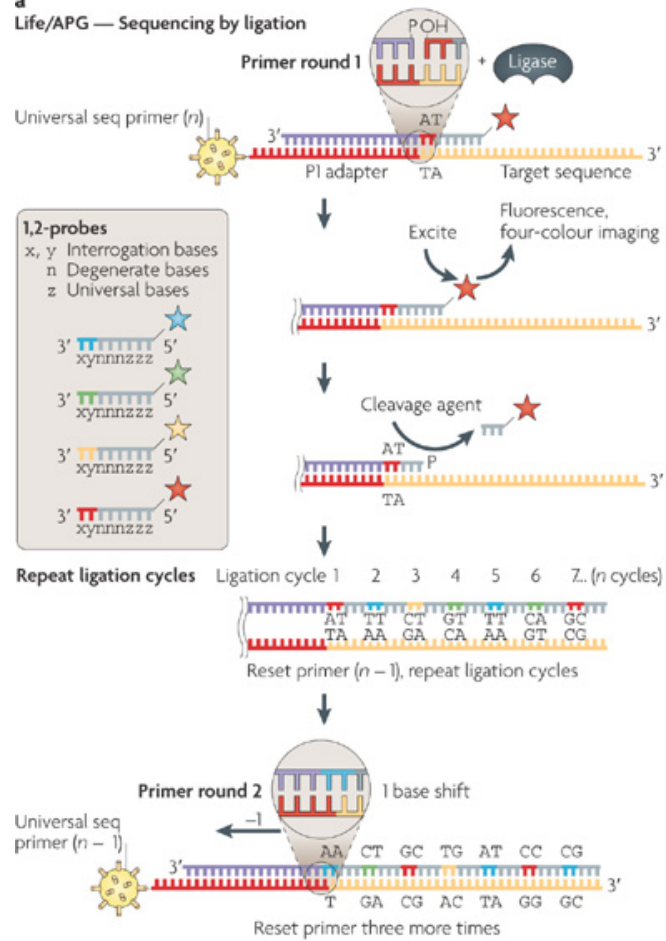


Secuenciando DNA : 2 métodos

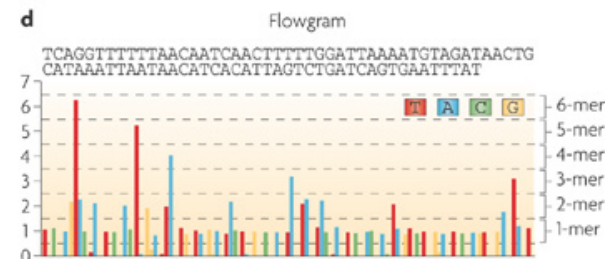
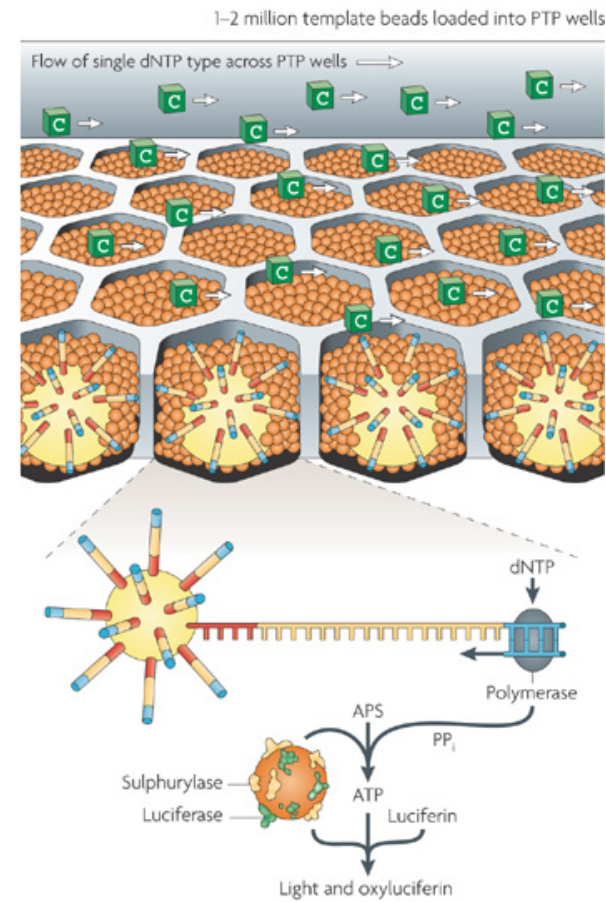
Método “shotgun”

- Varias copias se cortan en trozos con solape.
Se pierde el orden de los trozos
- Reconstrucción directa del orden, si los trozos son largos sólo se secuencian prefijos y sufijos
- Usado por Celera genomics en el proyecto del genoma humano

a
Life/APG — Sequencing by ligation



c
Roche/454 — Pyrosequencing





40421551 40421561 40421571 40421581 40421591 40421601 40421611 40421621 40421631 40421641 40421651 40421661 40421671 40421681 40421691 40421701 40421711 40421721

721tttgagcagacctatataagatggttatgaagattcacacagcggtcatgctgtgatcccagcactttgggaggctgaggcaagtgaggacacctgagatcatgagttcaagaccagcctggccaacatggtgaaacccatctctactaaagatacaaaaattatccagggtggtg

A.....T.....

A tgaacagacctatataagatggtt tgaagattcacacagtggtcatgctgtgatcccagc tgggaggctgagtcagtgaggacacctgagatcatgagtt ACCAGCCTGGCCAACATGGTGAACCCCATCTCTACTAAA ATACAAAAATATCCAGGTGTGGTG
aca cagacctatataagatggtt aagatacacacagtggtcatgctgtgatcccagcactt GGGAGGCTGAGGCAAGTGGAGCACCTGAGATCATGAGTTC cagcctggccaacatggtgaaacccatctctactaaaga AAAAAATATCCAGGTGTGGTG
acatt GACCTATATAAGATGGTTATGAAGATTACACAGTGGCTC CTGTGATCCAGCAGCTTTGGGAGGCTGAGGCAAGTGGAG ACCTGAGATCATGAGTTCAAGACCAGCTGGCAACATGG AACCCCATCTCTACTAAAGATACAAAAATATCCAGGTGT G
ACATTTGAACAG ATATAAGATGGTTATGAAGATTACACAGTGGCTCATGCC tgaatcccagcactttgggagg TGAGGCAAGTGGAGCACCTGAGATCATGAGTTCAAGACCA GCCAACATGGTGAACCCCATCTCTACTAAAGATACAAAA ATCCAGGTGTGGTG
ACATTTGAACAGAC TCAGATGGTTATGAAGATTACACAGTGGCTCATGCCGT ATCCCAGCACCTTTGGGAGGCTGAGGCAAGGGGAGCACCTG ATGAGTTCAAGACCAGCTGGCCAACATGGTGAACCCCA CTCTACTAAAGATACAAAAATTA aggtgtgggtg
acatttgaacagccctata aagatggttatgaagattcacacagtggtcatgctgtg TCCCAGCACCTTTGGGAGGCTGAGGCAAGTGGAGCACCTGA ATGAGTTCAAGACCAGCTGGCCAACATGGTGAACCCCA TATACTAAAGATACAAAAATATCCAGGTGTGGTG
acatttgaacagacctatata gatggttatgaagattcacacagtagctcatgctgtgat AGCACCTTTGGGAGGCTGAGGCAAGGGGAGCACGTGA GAGTTCAAGACCAGCTGGCCAACATGGTGAACCCCATC CTACTAAAGATACAAAAATATCCAGGTGTGGTG
ACAT gacctatataagatggttatgaagattcacacagtggtc CTGTAATCCCATCCTTTGGGAGGCTGAGGCAAGTGGAGC CTGAGATCATGAGTTCAAGA AGCCTGGCCAACATCTGTAACCCCATCTCTACTAAAGAT caaaattatccagggtggtg
ACATTTGAACAGACCTATATA TGGTTATGAAGATTACACAGTGGCTCATGCCGTGTATCC cactttgggagtgctgaggcaagtgaggacacctgagatcat CAAGACCAGCTGGCCAACATGGTGAACCCCATCTCTAC AGAAATACAAAAATATCCATGTGTGGTG
ACATTTG ACCTATATAAGATGGTTATGAAGATTACACAGTGGCTCA TGTGATCCAGCACCTTTGGGAGGCTGAGGCAAGTGGAGCA CTGAGATCATGAGTTCAAGACCAGCTGGCCAACATGGTGAACCCCATCTCTACTAAAGATACAAAAATATCCAGGTGT GG
ACATTTGAACAGACCTATATAAGA GGTATCAAGATTACACAGTGGCTCATGCCGTGTATCCC cactttgggaggctgaggcaagtgaggacacctgagatcat AAGACCAGCTGGCCAACATGGTGAACCCCATCTCTACT AAGATACAAAAATATCCAGGTGTGGTG
acatttgaacagacctatataagat ttaagattcacacagtggtcatgctgtgatcccag CTTTGGGAGGCTGAGGCAAGTGGAGCACCTGAGATCATGA agcctggccaacatggtgaaacccatctctactaaagat AAAATATCCAGGTGTGGTG
acatttgaacagacctatataagatggtt aagattcacacagtggtcatgctgtgatcccagcactt GGGAGGCTGAGGCAAGTGGAGCACCTGAGATCATGAGTTC GCCTGGCCAACATGGTGAAC CCATCTCTACTAAAGATACAAAAATATCCAGGTGTGGTG
ACATTTGAACAGACCTATATAAGATGGTTA agattcacacagtggtcatgctgtgatcccagcactt AGGCTGAGGCAAGTGGAGCACCTGAGATCATGAGTTCAAG CTTGGCCAACATGGTGAACCCCATCTCTACTAAAGATAC AAAATATCCAGGTGTGGTG
acatttgaacagacctatataagatggtta TTCACAGTGGCTCATGCCGTGTATCCCAGCACCTTTGGG GCTGAGGCAAGTGGAGCACCTGAGATCATGAGTTCAAGAC CCAACATGGTGAACCCCATCTCTACTAAAGATACAAAA atccagggtggtg
ACATTTGAACAGACCTATATAAGATGGTTAT CAGTGGCTCATGCCGTGTAT ACTTTGGGAGGCTGAGGCAAGTGGAGCACCTGAGATCATG CAACATGGTGAACCCCATCTCTACTAAAGATACAAAAAT TCCAGGTGTGGTG
ACATTTGAACAGACCTATATAAGATGGTTATGAAG CAGTGGCTCATGCCGTGTATC ACTTCGGGAGGCTGAGGCAAGTGGAGCACCTGAGATCATG AACATGGTGAACCCCATCTCTACTAAAGATACAGAAAT aggtgtgggtg
ACATTTGAACAGACCTATATAAGATGGTTATGAAGAT CAGTGGCTCATGCCGTGTATCC CTTTGGGAGGCTGAGGCAAGTGGAGCACCTGAGATCATG ACATGGTGAACCCCATCTCTACTAAAGATACAAAAATTA GTGTGGTG
ACATTTGAACAGACCTATCTAAGATGGTTATGAAGATT GCGGCTCATGCCGTGTATC CTTTGGGAGGCTGAGGCAAGTGGAGCACCTGAGATCATG ACATGGTGAACCCCATCTCTACTAAAGATACAAAAATTA TGTGGTG
ACATTTGAACAGACCTATATAAGATGGTTATGAAGATT CTCTTGCTGTGATCCCAGCACCTTTGGGAGGCTGAGGCAA TGGAGCACCTGAGATCATGAGTTCAAGACCAGCTGGCCA TGGTGAACCCCATCTCTACTAAAGATACAAAAATATCC ggtg
ACATTTGAACAGACCTATATAAGATGGTTATGAAGATT CTCTTGCTGTGATCCCAGCACCTTTGGGAGGCTGAGGCAA TGGAGCACCTGAGATCATGAGTTCAAGACCAGCTGGCCA GGTGAACCCCATCTCTACTAAAGATACAAAAATATCCCA GG
ACATTTGAACAGACCTATATAAGATGGTTATGAAGATTCA GTGATCCCAGCACCTTTGGGAGGCTGAGGCAAGTGGAGCAC GATCATGAGTTCAAGACCAGCTGGCCAACATGGTGAAC ccatctctactaaagatacaaaaattatccagggtggtg
AGATGGTTATGAAGATTACACAGTGGCTCATGCCGTGTGA CCAGCACCTTTGGGAGGCTGAGGCAAGTGGAGTACCTGAGA GAGTTCAAGACCAGCTGGCCAACATGGTGAACCCCATC TACTAAAGATACAAAAATATCCAGGTGTGGTG
ACATGGTTATGAAGATTACACAGTGGCTCATGCCGTGTGA CTTTGGGAGGCTGAGGCAAGTGGAGCACCTGAGATCATGA CATGGTGAACCCCATCTCTACTAAAGATACAAAAATAT g
GGTTATGAAGATTACACAGTGGCTCATGCCGTGTATCCC CTTTGGGAGGCTGAGGCAAGTGA cccactgagatcatgagttcaagaccagctg**caacat tgaaccccatctctactaaagatacaaaaattatccagg g
TATGAAGATTACACAGTGGCTCA gatcccagcactttgggaggctgaggcaagtgaggacactt agttcaagaccagctggccaacatggtgaaacccatct TACTAAAGATACAAAAATATCCAGGTGTGGTG
ATGAAGATTACACAGTGGCTCATGCCGTGTATCCCAGCA TCTGGGAGGCTGAGGCAAGTGGAGCACCTGAGATCATGAG CATGGTGAACCCCATCTCTACTAAAGATACAAAAATATCCAGGTGTGGTG
gatcccagcactttgggaggctgaggcaagtgaggacactt CATGGTGAACCCCATCTCT CTAAGATACAAAAATATCCAGGTGTGGTG
atcccagcactttgggaggctgaggcaagtgaggacactg CATGGTGAACCCCATCTCTACTAAAGATACAAAAATATGAT C
TCTGAGAGGCTGAGGCAAGTGGAGCACCTGAGATCATGAG GTGAACCCCATCTCTACTAAAGATACAAAAATATCCAG g
GGGATGCTTATGTCATTTGAGCACCTGAGATCATGAGTTC GTGAACCCCATCTCTACTAAAGATACAAAAATATCCAG g
aggctgaggcaagtgaggacactgagatcatgagttcaag tgaaccccatctctactaaagatacaaaaattatccagg g
tggggcaagtgaggacactgagatcatgagttcaagacca gtgaaccccatctctactaaagatacaaaaattatccagg g
tgaagcaagtgaggacactgagatcatgagttcaagacca GAAATCCCATCTCTACTAAAGATACAAAAATATCCAGGT g
GAGGCAAGTGGAGCACCTGAGATCATGAGTTCAAGACCAG GAAACCCCATCTCTACTAAAGATACAAAAATATCCAGGT g
AGGCAAGTGGAGCACCTGAGATCATGAGTTCAAGACCAGC GAAACCCCATCTCTACTAAATAAACA atccagggtggtg
aggcaatttgaatctctgagatcatgagttcaagaccagc gaaccccatctctgtagagtgcaaaaattatccagg g
GCAAGTGGAGCACCTGAGATCA CAAGTGGAGCACCTGAGATCATGAGTTCAAGACCAGCTG A
caagtgaggacactgagatcatgagttcaagaccagctg aaccccatctctactaaagatacaaaaattatccagggt g
AAGTGGAGCACCTGAGATCATGAGTTCAAGACCAGCTGG AACCCCATCTCTACTAAAGATACAAAAATATCCAGGTGT g
AGTGGAGCACCTGAGATCATGAGTTCAAGACCAGCTGGC ACCCCGTTCTACTAAAGATACAAAAATATCCAGGTGT g
AGTGGAGCACCTGAGATCATGAGTTCAAGACCAGCTGGC accccatctctactaaagatacaaaaattatccagggtg g
GTGGAGCACCTGAGATCATGAGTTCAAGACCAGCTGGCC CCCCATCTCTACTAAAGATAC atccagggtggtg
GGAGCACCTGAGATCATGAGTTCAAGACCAGCTGGCCAA CATCTCTACTAAAGATACAAAAATATCCAGGTGTGGTG
ggagcactgagatgagttcaagaccaggtggccaa CATCTCTACTAAAGATACAAAAATATCCAGGTGTGTG
ggagcactgagatcatgagttcaagaccagctggccaa CGTCTCTACTAAAGATACAAAAATATCCAGGTGTGGTG
GAGCACCTGAGATCATGAGTTCAAGACCAGCTGGCCAAC CATCTCTACTAAAGATACAAAAATATCCAGGTGTGGTG



Alineamiento

- Se trata de considerar los problemas de los errores y omisiones en los datos biológicos, es decir, cambios puntuales y huecos en la secuencia
- Por ejemplo podemos comparar dos versiones del mismo gen hechas por distintos laboratorios, buscar un string como subsecuencia de todos los de una BdD, o cálculo aproximado de overlaps
- También se utiliza alineamiento para comparar organismos o especies y para secuenciación



Alineamiento de 2 strings

- **Definición:** Sean $s=s_1\dots s_m$ y $t=t_1\dots t_n$ dos strings sobre Σ . Sea $h: (\Sigma \cup \{-\})^* \rightarrow \Sigma^*$ la función “quita huecos”

$(h(-)=\lambda, h(a)=a \text{ para } a \in \Sigma)$

Un alineamiento de s y t es un par de strings (s', t') que cumplen

1. $h(s')=s \quad h(t')=t \quad |s'|=|t'|$
2. No hay posiciones que contengan – tanto en s' como en t'



Ejemplo de alineamiento

- **s = GACGGATTATG**

- **t = GATCGGAATAG**

- **Alineamiento**

- **s' = GA-CGGATTATG**

- **t' = GATCGGAATA-G**



Columnas del alineamiento

- $s' = \text{GA-CGGATTATG}$
- $t' = \text{GATCGGAATA-G}$
- Inserción
- Borrado
- Coincidencia
- Sustitución/desacuerdo



Puntuación de un alineamiento

- Distancia de edición o de Levenshtein

$p(a,b)=1$ si $a \neq b$, $p(a,a)=0$,

$p(a,-) = p(-,a) = 1$

- A menor puntuación mejor alineamiento



Ejemplo de puntuación

- **s = GACGGATTATG**

- **t = GATCGGAATAG**

- **Alineamiento**

- **s' = GA-CGGATTATG**

- **t' = GATCGGAATA-G**



Ejemplo de puntuación

- **s = GACGGATTATG**

- **t = GATCGGAATAG**

- Alineamiento

- **s' = GA-CGGATTATG**

- **t' = GATCGGAATA-G**



Ejemplo de puntuación

- $s = \text{GACGGATTATG}$

- $t = \text{GATCGGAATAG}$

- Alineamiento

- $s' = \text{GA} \boxed{-} \text{CGGATTATG}$

- $t' = \text{GAT} \boxed{C} \text{GGA} \boxed{A} \text{TA} \boxed{-} \text{G}$

- puntuación = 3



Objetivo de optimización

- El mejor alineamiento es el de puntuación mínima



Alineamientos múltiples

- Ahora se trata de alinear varias secuencias a la vez
- El problema es mucho más difícil (por ejemplo la programación dinámica es muy lenta)



Alineamiento múltiple

Definición: Sean $s_1 = s_{11} \dots s_{1m_1}, \dots, s_k = s_{k1} \dots s_{km_k}$ k strings sobre Σ . Sea $h: (\Sigma \cup \{-\})^* \rightarrow \Sigma^*$ la función “quita huecos”

$(h(-) = \lambda, h(a) = a \text{ para } a \in \Sigma)$

Un alineamiento múltiple de $s_1, \dots, s_k$ es una k -tupla de strings $(s'_1, \dots, s'_k)$ que cumplen

1. $h(s'_i) = s_i$
2. $|s'_i| = |s'_j|$ para cada i, j
3. No hay posiciones que contengan – alineados en todos los s'_i



Ejemplo de alineamiento múltiple

■ $s_1 = \text{AATGCT}$, $s_2 = \text{ATTC}$, $s_3 = \text{TCC}$

$s'_1 = \text{AATGCT}$

$s'_2 = \text{A-TTC-}$

$s'_3 = \text{---TCC}$



Puntuando alineamientos múltiples

- Calculamos la puntuación del multialineamiento como la suma de las puntuaciones de los alineamientos entre parejas (SP)



Ejemplo de multilineamiento

■ $s_1 = \text{AATGCT}$, $s_2 = \text{ATTC}$, $s_3 = \text{TCC}$

$s'_1 = \text{AATGCT}$

$s'_2 = \text{A-TTC-}$

$s'_3 = \text{---TCC}$

puntuación = $2+2+2+2+0+3 = 11$



El problema del string matching

- Consiste en encontrar un string (corto), el *patrón*, como substring de un string (largo), el *texto*
- En bioinformática lo más frecuente es buscar un fragmento nuevo de DNA (un gen) en una colección de secuencias
- En este caso permitimos un cierto error, pero el string matching exacto es una subrutina



Enunciado del problema ...

- Entrada: Dos strings $t = t_1 \dots t_n$, $p = p_1 \dots p_m$ sobre Σ
- Salida: El conjunto de posiciones de t donde aparece p , es decir, $I \subseteq \{1, \dots, n-m+1\}$ tales que $i \in I$ sii $t_i \dots t_{i+m-1} = p$



Variantes

- Encontrar todas las ocurrencias de ...
un patrón de tamaño m en N textos de tamaño n
- Encontrar todas las ocurrencias de ...
M patrones de tamaño m en un texto de tamaño n



El problema del substring

- Entrada: Un patrón $p = p_1 \dots p_m$ y N textos $t_1, \dots, t_N$
- Salida: El conjunto de textos en los que aparece el patrón (es decir, $I \subseteq \{1, \dots, N\}$ tal que $i \in I$ si y solo si p es un substring de t_i)
- Utilidad: para búsqueda en una base de datos de DNA



Substring común más largo

- Si tenemos una base de datos de DNA de organismos parecidos, encontrar strings comunes a todos ellos puede ser importante
- Si un string aparece en todas quiere decir que ha mutado muy poco y por tanto es importante para la supervivencia del organismo



Substring común más largo

Problema de optimización

- Entrada: Un conjunto de textos $T = \{t_1, \dots, t_N\}$
- Soluciones aceptables: Todos los strings x que son substrings de t_i para todo i
- Coste de una solución x : $|x|$
- Objetivo de optimización: Maximizar



Generalizaciones

- Calcular el substring común más largo de al menos k de los N textos
- Calcular la posición del substring común más largo
- Todos los substrings comunes de una longitud dada l
 - Muy útil para testear la contaminación de DNA ($N=2$)



Cálculo de Overlaps

- El *solape* de s y t es el string y más largo que cumple $s=xy$, $t=yz$
- Se trata de calcular los solapes de cada par de textos de entre $t_1, \dots, t_N$
- Se usa mucho para secuenciación de ADN



Repeticiones

- Muy usadas para modelizar y tratar una secuencia de ADN
- Se usan repeticiones exactas y aproximadas, con solape o sin él



Definición

- Sean t y p dos strings, $t=t_1 \dots t_n$, $p=p_1 \dots p_m$
- Añadimos principio y final a t , $t_0=\$, t_{n+1}=\$$
- p es una **repetición exacta en t** si existen i, j distintos que cumplen
$$p=t_{i+1} \dots t_{i+m}=t_{j+1} \dots t_{j+m}$$
- p es una repetición exacta **maximal** en t si además $t_i \neq t_j$ y $t_{i+m+1} \neq t_{j+m+1}$



Problema

- Entrada: Un string $t=t_1 \dots t_n$
- Salida: El conjunto de todas las repeticiones exactas maximales en t