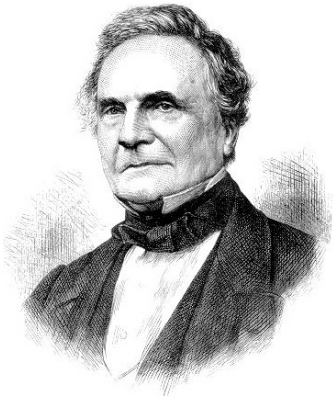
Bioinformática

(presentación nueva optativa)

Elvira Mayordomo
Univ. de Zaragoza 29-4-15

¿Qué tiene que ver la informática con la biología?: la bioinformática

Alrededor de 1800 …



- Charles Babbage: diseñó el considerado primer computador

- Louis Pasteur: primera vacuna de bacterias debilitadas

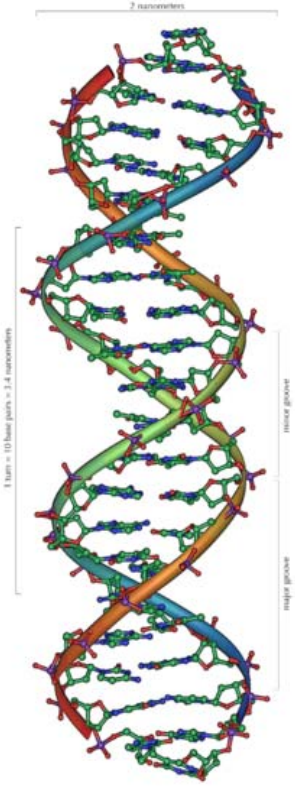
En el siglo XX

- Descubrimiento de la estructura del DNA por Watson y Crick en 1953

- Primer computador digital, Atanasoff-Berry Computer (ABC) desde 1937, ENIAC desde 1943 ...

En los periódicos

- La biotecnología, genética y bioinformática en primera plana
- 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
- En el 2008 comenzó el proyecto de los 1000 genomas



Esta charla

- **Ejemplos de bioinformática**
  - Bioinformática forense
  - Detección de enfermedades y tratamientos personalizados
  - Estudios epidemiológicos
  - Sorpresas
- ¿Qué es la bioinformática?
- Trabajar/investigar en bioinformática
- La nueva optativa
- Grandes retos algorítmicos

Ejemplo 1:
Bioinformática forense

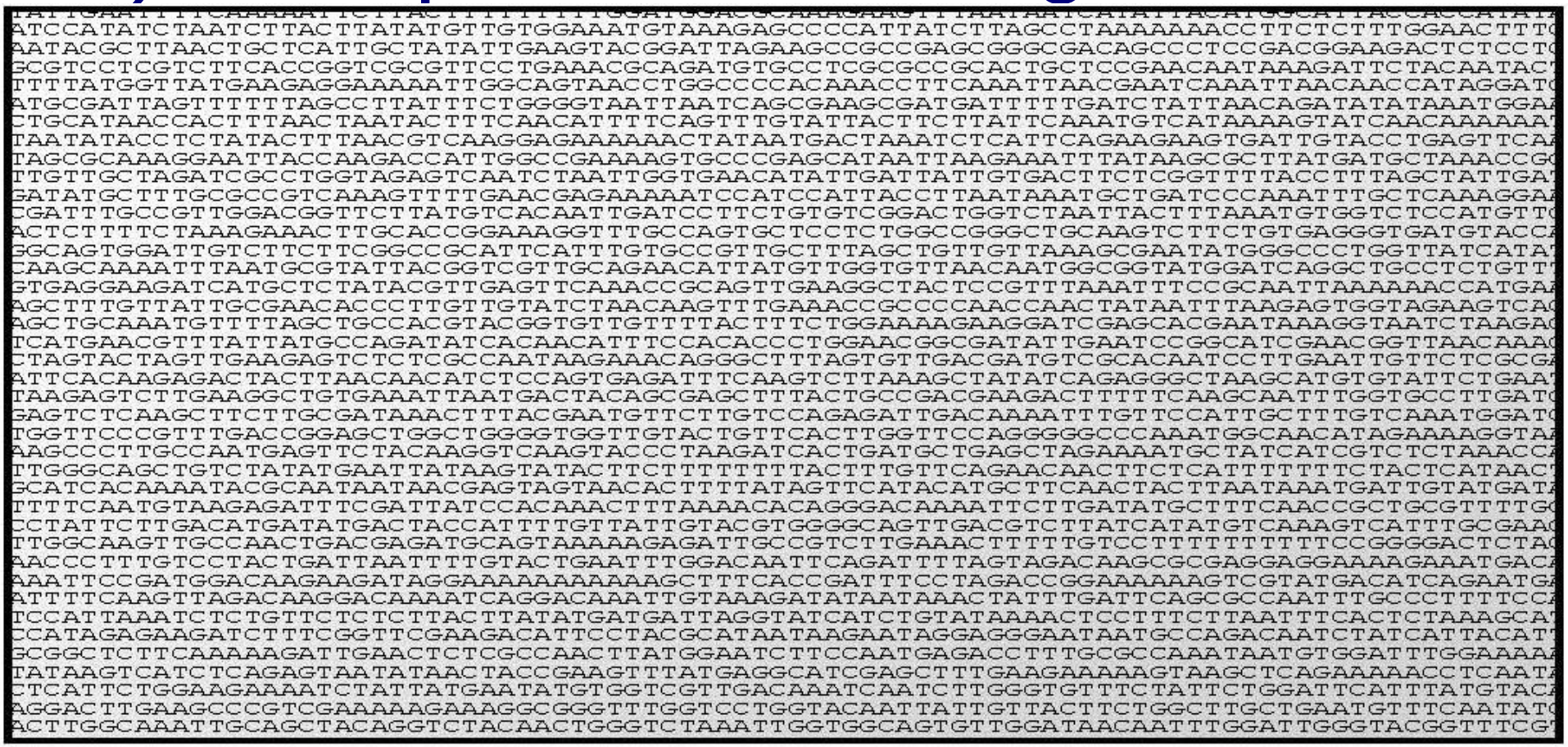
Forense …

- 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

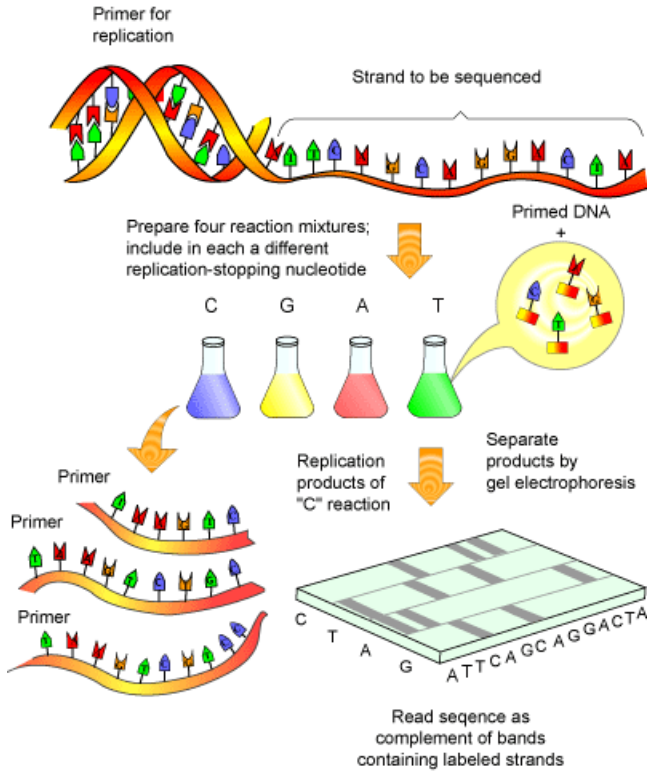


- Unas 10 pantallas como esta por persona (hay 3000 caracteres por pantalla)

1) Recuperar material genético

¿trabajo de laboratorio?

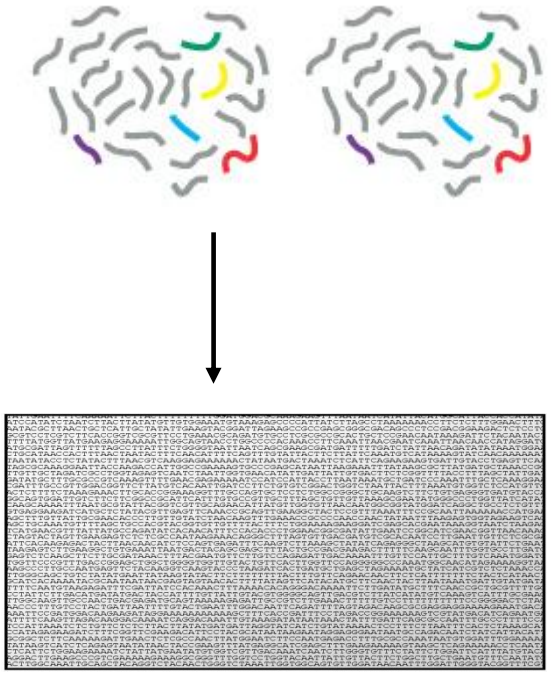
- En el laboratorio se generan muchas copias de la cadena que nos interesa, se trocean (trozos de unos 300 caracteres) y se leen
- Imposible leer trozos más grandes …
- La informática tiene que hacer el resto …



1) Recuperar material genético

Trabajo de bioinformática:

- 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 de Inglaterra
- El 7º (el zar) estaba emparentado con parientes maternos del zar

Para el Hola …

- Todos los hijos de Nicolas II fueron asesinados junto con él y su esposa

Ejemplo 2:

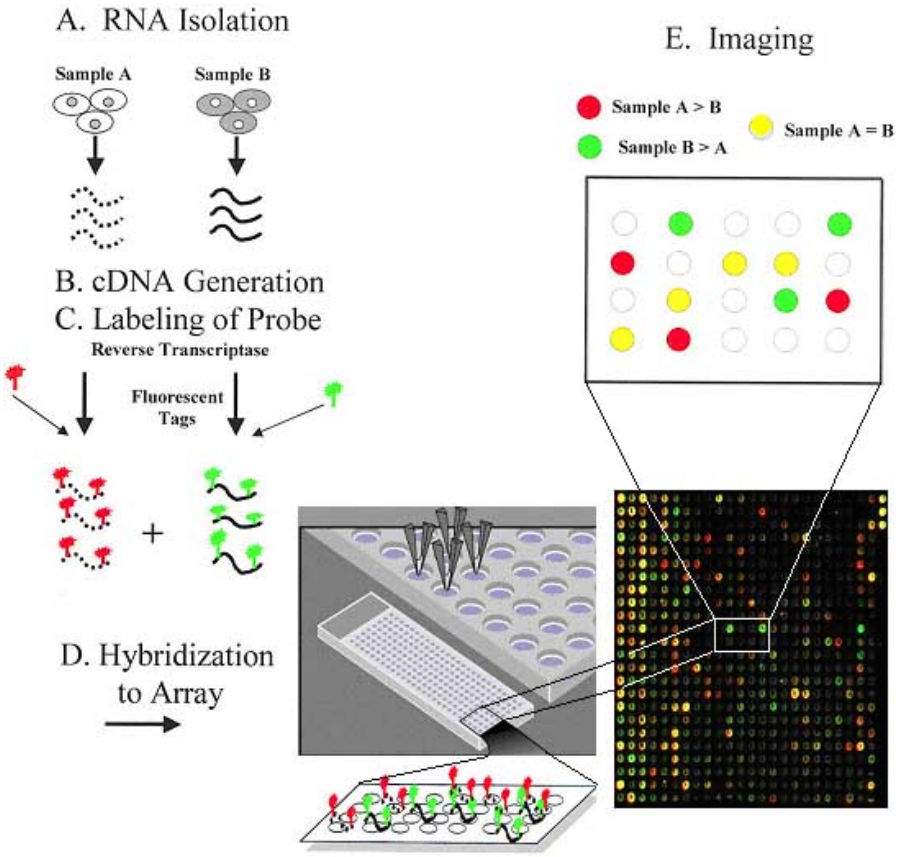
Detección de enfermedades y tratamientos personalizados

¿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 todo de gris es importante

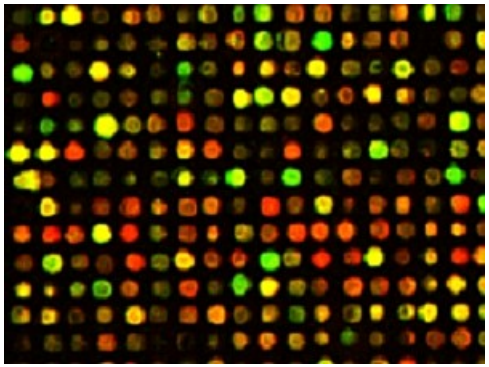
¿Tendré cáncer de …?

- 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 está información se diseñan los mediamentos (también usando bioinformática)

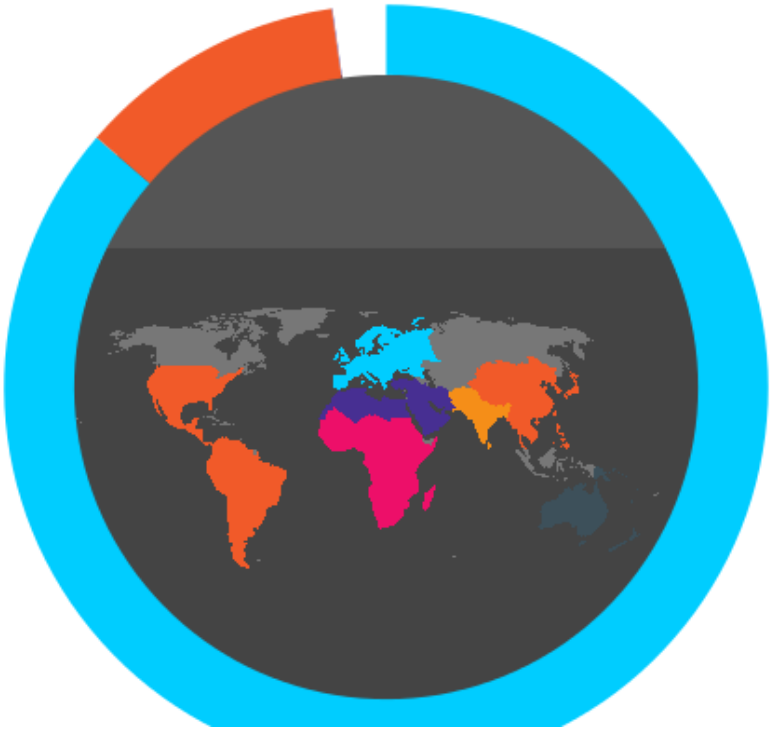
Tratamiento personalizado

Ancestry Composition

[redacted] Standard E: LOG OUT

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%	[redacted]

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

Ejemplo 3:

Estudios epidemiológicos

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

Ejemplo 4:

Sorpresas

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

*** En 2012 AMGEN compra DeCode**

Ejemplo X:

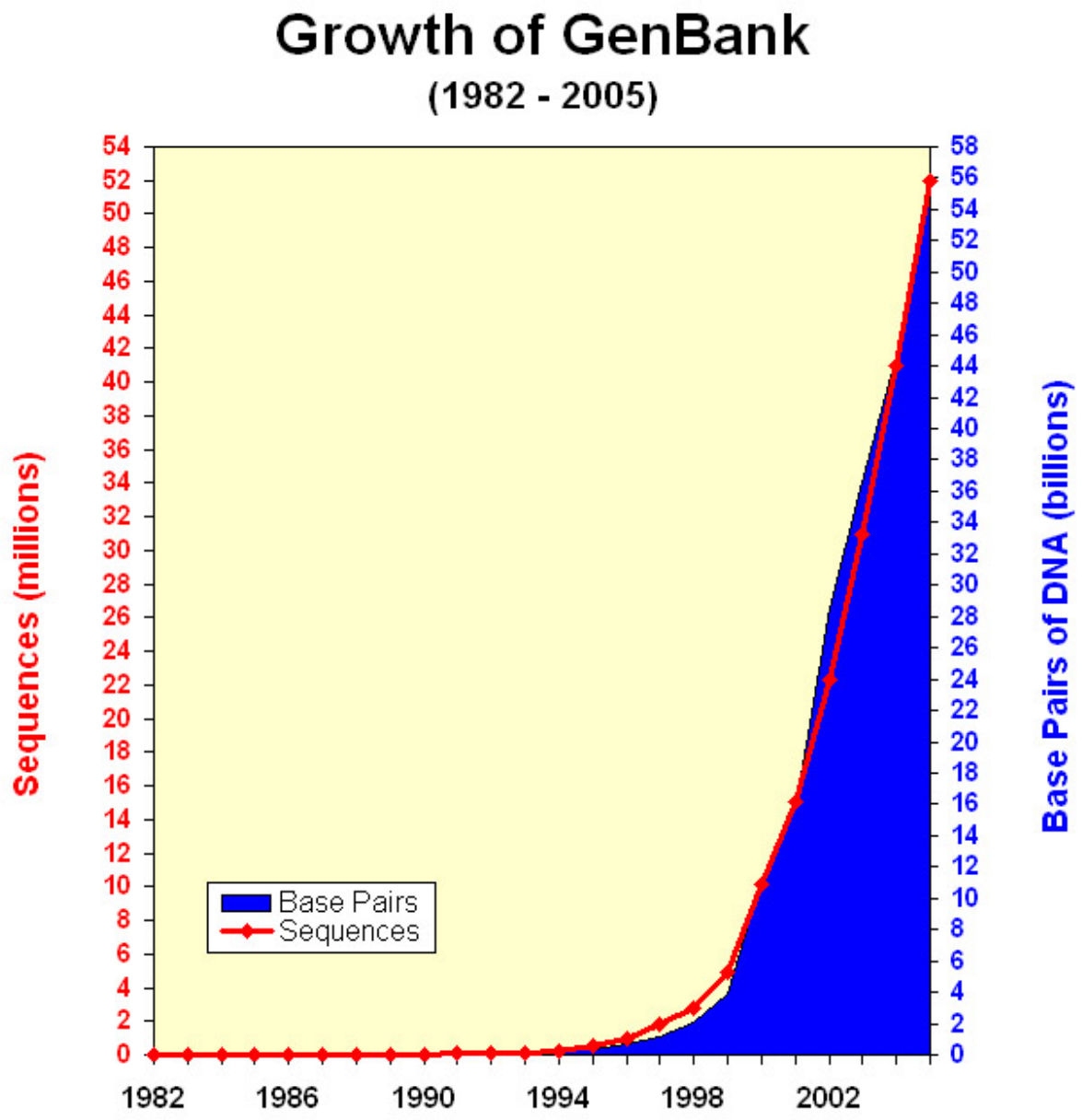
...

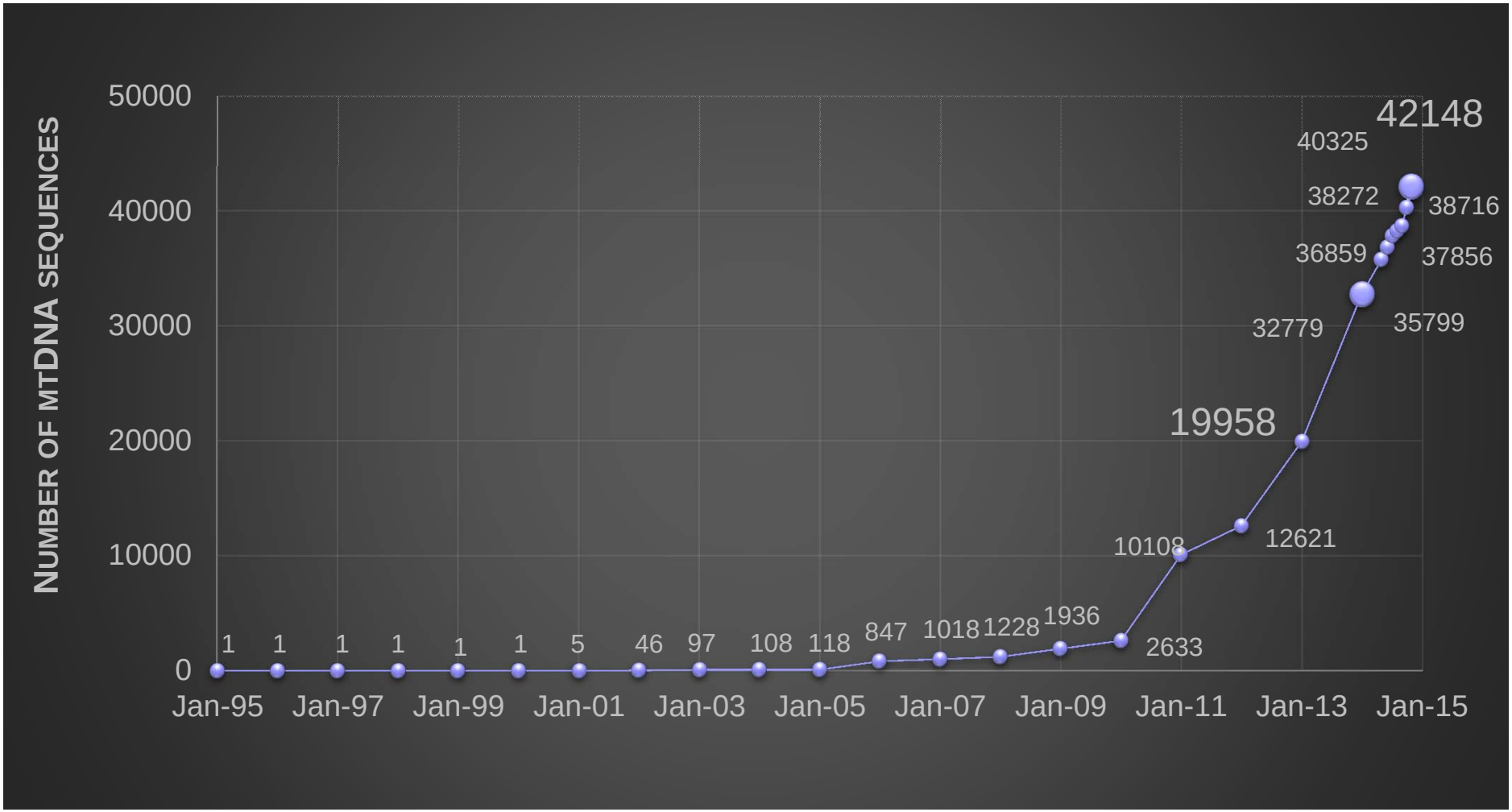
Esta charla

- Ejemplos de bioinformática
  - Bioinformática forense
  - Detección de enfermedades y tratamientos personalizados
  - Estudios epidemiológicos
  - Sorpresas
- **¿Qué es la bioinformática?**
- Trabajar/investigar en bioinformática
- La nueva optativa
- Grandes retos algorítmicos

¿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, especialemente secuencias (DNA, RNA, proteínas: biología molecular)


Growth of GenBank
(1982 - 2005)
Sequences (millions)
Base Pairs of DNA (billions)
Base Pairs
Sequences
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
46
48
50
52
54
56
58
1982
1986
1990
1994
1998
2002


NUMBER OF MTDNA SEQUENCES
50000
40000
30000
20000
10000
0
Jan-95
Jan-97
Jan-99
Jan-01
Jan-03
Jan-05
Jan-07
Jan-09
Jan-11
Jan-13
Jan-15
1
1
1
1
1
1
5
46
97
108
118
847
1018
1228
1936
2633
10108
12621
19958
32779
35799
36859
37856
38272
38716
40325
42148

Métodos computacionales en bioinformática

- Algoritmia
- Inteligencia Artificial (aprendizaje, minería de datos)
- Gestión de bases de datos
- Estadística

Mucho por hacer …

- Hay enormes bases de datos de DNA, RNA y proteínas (en parte gracias a la informática)
- Las preguntas que se quieren contestar son multitud y muy importantes
  - Enfermedades hereditarias
  - Medicina forense y criminología
  - Evolución de los seres vivos

  …
- Muchas de estas preguntas necesitan una buena dosis de informática …

Esta charla

- Ejemplos de bioinformática
  - Bioinformática forense
  - Detección de enfermedades y tratamientos personalizados
  - Estudios epidemiológicos
  - Sorpresas
- ¿Qué es la bioinformática?
- **Trabajar/investigar en bioinformática**
- La nueva optativa
- Grandes retos algorítmicos

Trabajar/investigar en bioinformática

- Informáticos, biólogos, médicos, bioinformáticos
- Trabajar en bioinformática, trabajar para biólogos/médicos, colaborar
- Compañías y laboratorios de bioinformática
- Distribución geográfica
- Ya hablaremos …

Esta charla

- Ejemplos de bioinformática
  - Bioinformática forense
  - Detección de enfermedades y tratamientos personalizados
  - Estudios epidemiológicos
  - Sorpresas
- ¿Qué es la bioinformática?
- Trabajar/investigar en bioinformática
- **La nueva optativa**
- Grandes retos algorítmicos

Asignatura de bioinformática

- Introducción de la mínima biología necesaria
- Los temas más importantes de la bioinformática:
  - ideas asentadas
  - temas abiertos
  - productos disponibles

¿Qué veremos?

- Un poco de biología
- Problemas de biología que resolver con informática: secuenciación, alineamiento, análisis, …
- Énfasis en algoritmia, también otras técnicas (estadística, IA)
- Ejemplos concretos
- Prácticas con herramientas bioinformáticas

Contenido
0. Intro

- Introducción a la bioinformática.
  Conceptos básicos de biología molecular.
  Algoritmos para cadenas.

Contenido

1. Lectura o recuperación

- Métodos de alineamiento. Métodos heurísticos para búsqueda en repositorios biológicos. Alineamientos múltiples.

- Secuenciación

Contenido
2. Análisis

- Predicción de genes y promotores. Búsqueda de señales.
- Filogenética computacional.
- Bioinformática estructural.

Contenido
3. Temas abiertos y novedosos

- Epigenética
- ...

Esta charla

- Ejemplos de bioinformática
  - Bioinformática forense
  - Detección de enfermedades y tratamientos personalizados
  - Estudios epidemiológicos
  - Sorpresas
- ¿Qué es la bioinformática?
- Trabajar/investigar en bioinformática
- La nueva optativa
- **Grandes retos algorítmicos**

Grandes retos algorítmicos

- Secuenciación
- Alineamiento
- String matching
- … (la lista es larga)

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

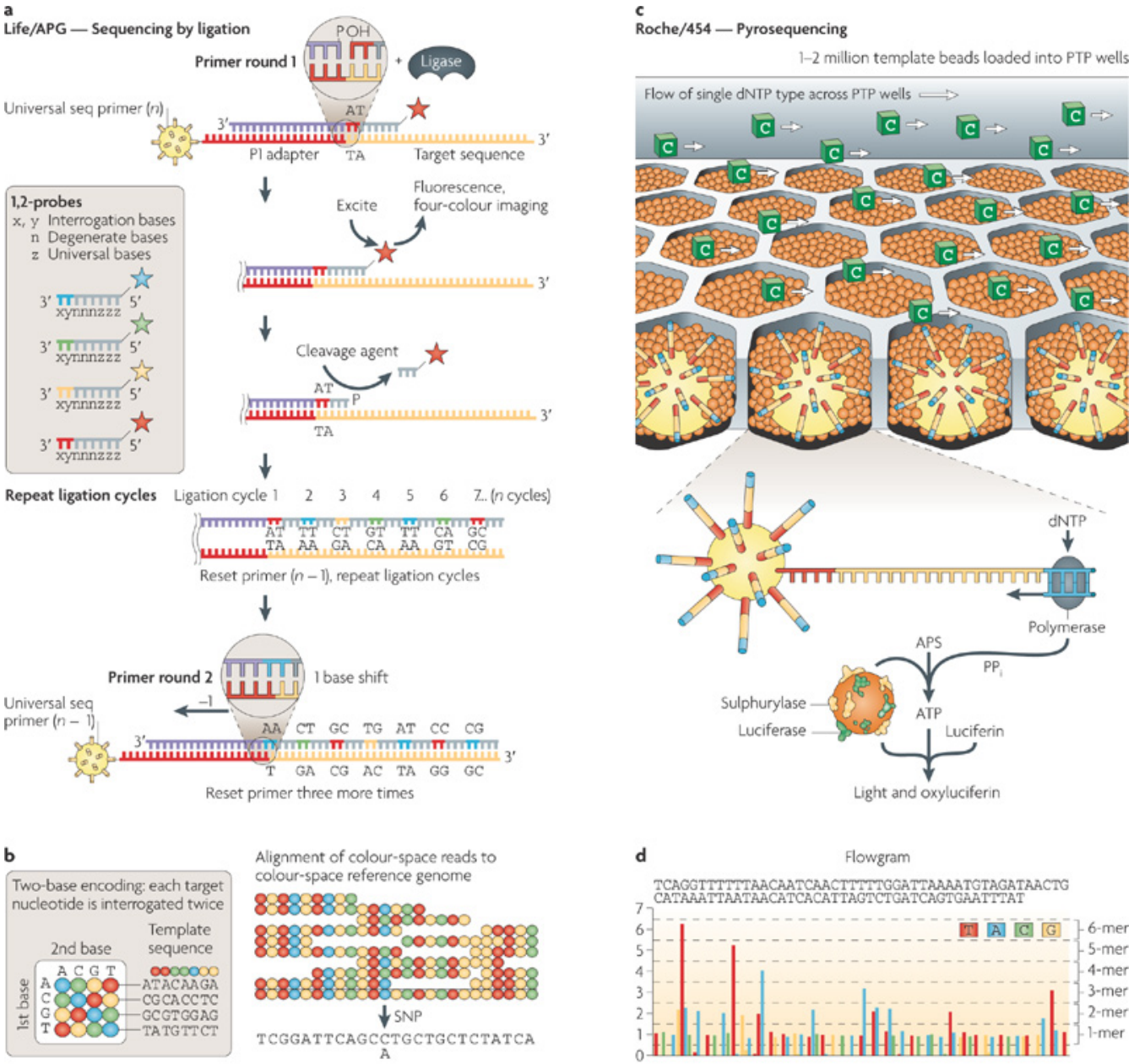
Mapa físico

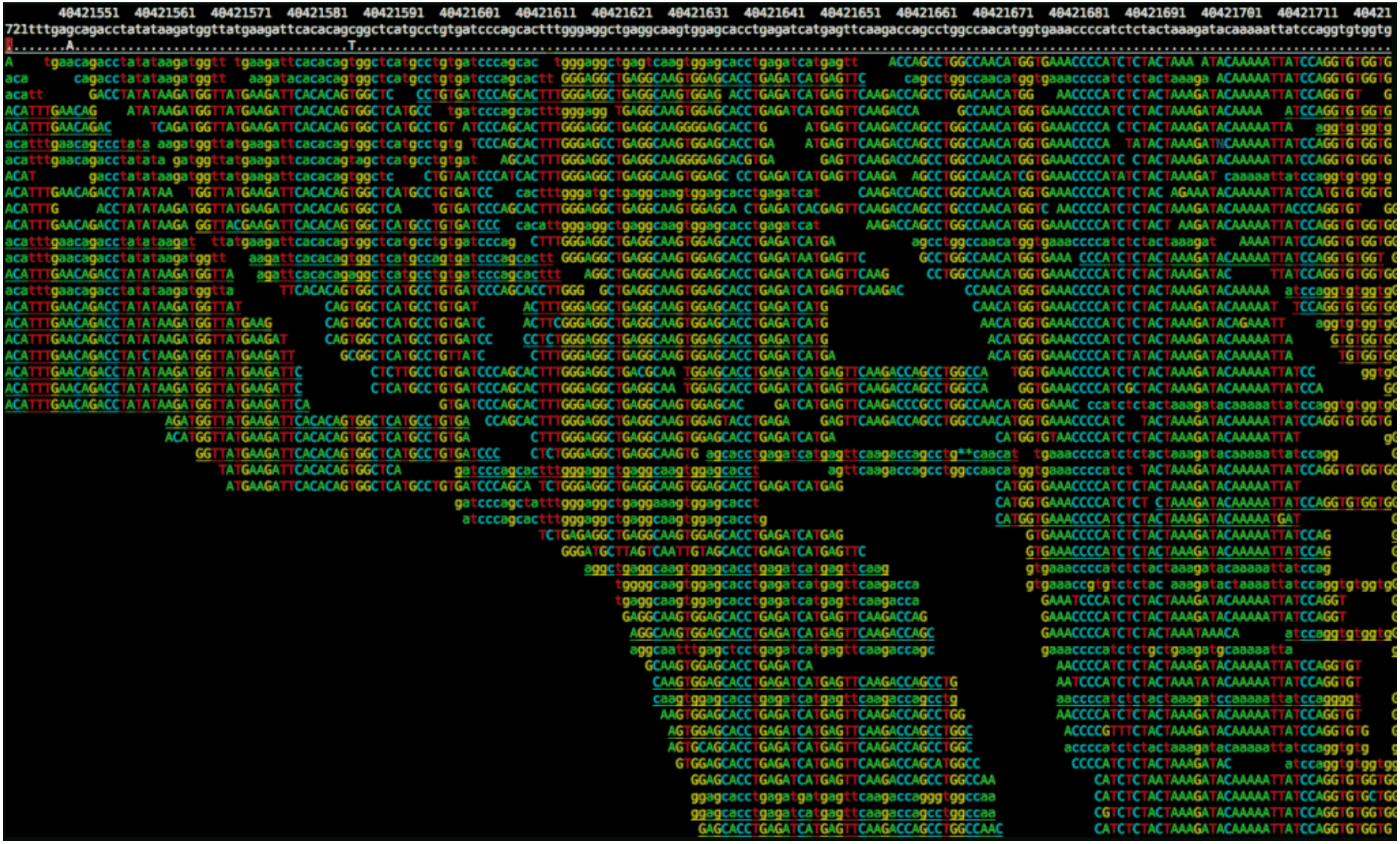
- Varias copias se cortan en trozos grandes con solape (50-300 kbp). Se pierde el orden de los trozos
- Un mapa físico es un conjunto de trozos pequeños (marcadores) junto con su posición exacta en la secuencia
- Se usa el mapa físico para reconstruir el orden, después los trozos se cortan y se secuencian
- 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
Primer round 1
POH
Ligase
Universal seq primer (n)
3′
AT
P1 adapter
TA
Target sequence
3′
1,2-probes
x, y Interrogation bases
n Degenerate bases
z Universal bases
3′ xynnnzzz 5′
Excite
Fluorescence, four-colour imaging
Cleavage agent
AT
P
TA
Repeat ligation cycles
Ligation cycle 1 2 3 4 5 6 7.. (n cycles)
AT TT CT GT TT CA GC
TA AA GA CA AA GT CG
Reset primer (n – 1), repeat ligation cycles
Primer round 2
1 base shift
–1
Universal seq primer (n – 1)
AA CT GC TG AT CC CG
T GA CG AC TA GG GC
Reset primer three more times
b
Two-base encoding: each target nucleotide is interrogated twice
2nd base
Template sequence
A C G T
1st base
A
C
G
T
ATACAAGA
CGCACCTC
GCGTGGAG
TATGTTCT
Alignment of colour-space reads to colour-space reference genome
SNP
TCGGATTCAGCCTGCTGCTCTATCA
A
c
Roche/454 — Pyrosequencing
1–2 million template beads loaded into PTP wells
Flow of single dNTP type across PTP wells
dNTP
Polymerase
APS
PPi
Sulphurylase
Luciferase
ATP
Luciferin
Light and oxyluciferin
d
Flowgram
TCAGGTTTTTTAACAATCAACTTTTTGGATTAAAATGTAGATAACTG
CATAAATTAATAACATCACATTAGTCTGATCAGTGAATTTAT
T A C G
6-mer
5-mer
4-mer
3-mer
2-mer
1-mer
0 1 2 3 4 5 6 7


40421551 40421561 40421571 40421581 40421591 40421601 40421611 40421621 40421631 40421641 40421651 40421661 40421671 40421681 40421691 40421701 40421711 40421
721tttgagcagacctatataagatggttatgaagattcacacagcggctcatgcctgtgatcccagctttgggaggctgaggcaagtggagcacctgagatcatgagttcaagaccagcctggccaacatggtgaaaccccatctctactaaagatacaaaaattatccaggtgtggtg

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

Ejemplo de alineamiento

- **s= GACGGATTATG**
- **t= GATCGGAATAG**

- Alineamiento
- **s’= GA-CGGATTATG**
- **t’= GATCGGAATA-G**

Columnas del alineamiento

- **s’= GA-CGGATTATG**
- **t’= 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= GACGGATTATG**
- **t= GATCGGAATAG**

- Alineamiento
- **s'= GA-CGGATTATG**
- **t'= GATCGGAATA-G**

- puntuación = 3

Objetivo de optimización

- El mejor alineamiento es el de puntuación mínima

Algoritmos

- Se pueden alinear 2 secuencias con programación dinámica (eficiente)
- Hay algunas pegas cuando se trata de elegir el alineamiento “con mayor significado biológico”

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)

Ejemplo de alineamiento múltiple

- s_1= AATGCT, s_2= ATTC, s_3= TCC

s'_1= AATGCT
s'_2= A-TTC-
s'_3= ---TCC

Puntuando alineamientos múltiples

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

Ejemplo de multialineamiento

- s_1= AATGCT, s_2= ATTC, s_3= TCC

s'_1= AATGCT
s'_2= A-TTC-
s'_3= ---TCC

puntuación = 2+2+2+2+0+3 = 11

Multialineamiento

- Problema intratable
- Heurísticas poco satisfactorias
- Aproximaciones poco satisfactorias
- Se gasta mucho tiempo y dinero haciendo multialineamientos (y criticándolos)

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

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

Hay muchos más retos en bioinformática …

Terminando …

- La asignatura de Bioinformática empieza el **curso 2015-16** (primavera)
- Es **optativa** de la especialidad de Computación
- Si queréis **saber más** de la asignatura contactar conmigo elvira@unizar.es
- Si os interesa un **Trabajo Fin de Grado** en Bioinformática contactar conmigo elvira@unizar.es

Esta charla

- Ejemplos de bioinformática
  - Bioinformática forense
  - Detección de enfermedades y tratamientos personalizados
  - Estudios epidemiológicos
  - Sorpresas
- ¿Qué es la bioinformática?
- Trabajar/investigar en bioinformática
- La nueva optativa
- Grandes retos algorítmicos

Gracias por vuestra atención!

?

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 …