Biología computacional para informáticos

Francesc Rosselló

UIB

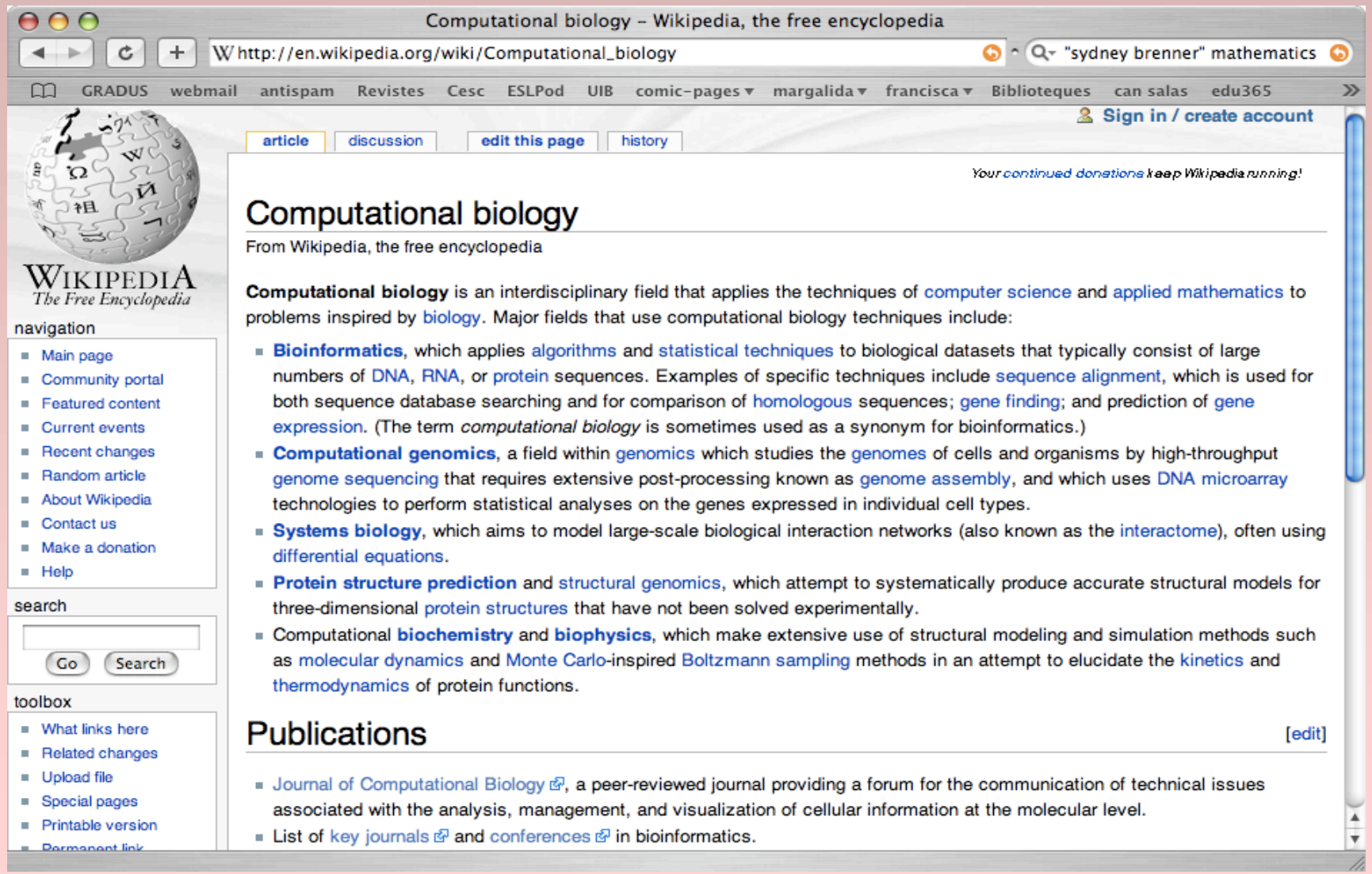
http://bioinfo.uib.es/~cesc

Índice

- Por qué los biólogos de hoy en día no son nada sin los informáticos
- Algunos problemas de biología computacional
  - Alineamiento de secuencias
  - Secuenciado de ADN
  - Búsqueda de genes
  - Estructura 3D de las proteínas
  - Metabolómica

La biología tiene más de 500 años de problemas en los que podemos trabajar (D. Knuth)

¿Qué es la biología computacional?



¿Qué es la biología computacional?

La biología molecular es todo aquello que interesa a los biólogos moleculares

F. Crick, "Molecular biology in 2000", Nature (1970)

¿Qué es la biología computacional?

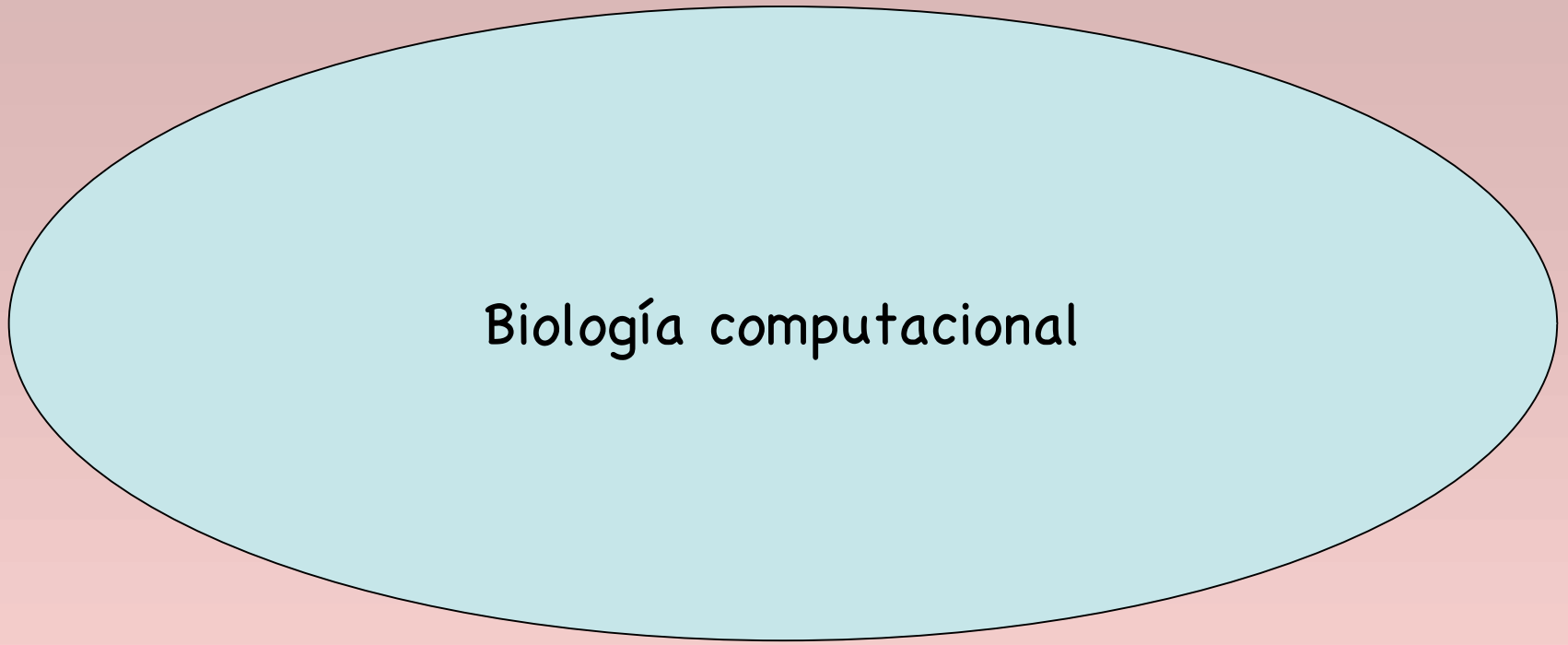
La biología molecular es todo aquello que interesa a los biólogos moleculares

F. Crick, "Molecular biology in 2000", *Nature* (1970)

La biología computacional es lo que hacen los biólogos computacionales

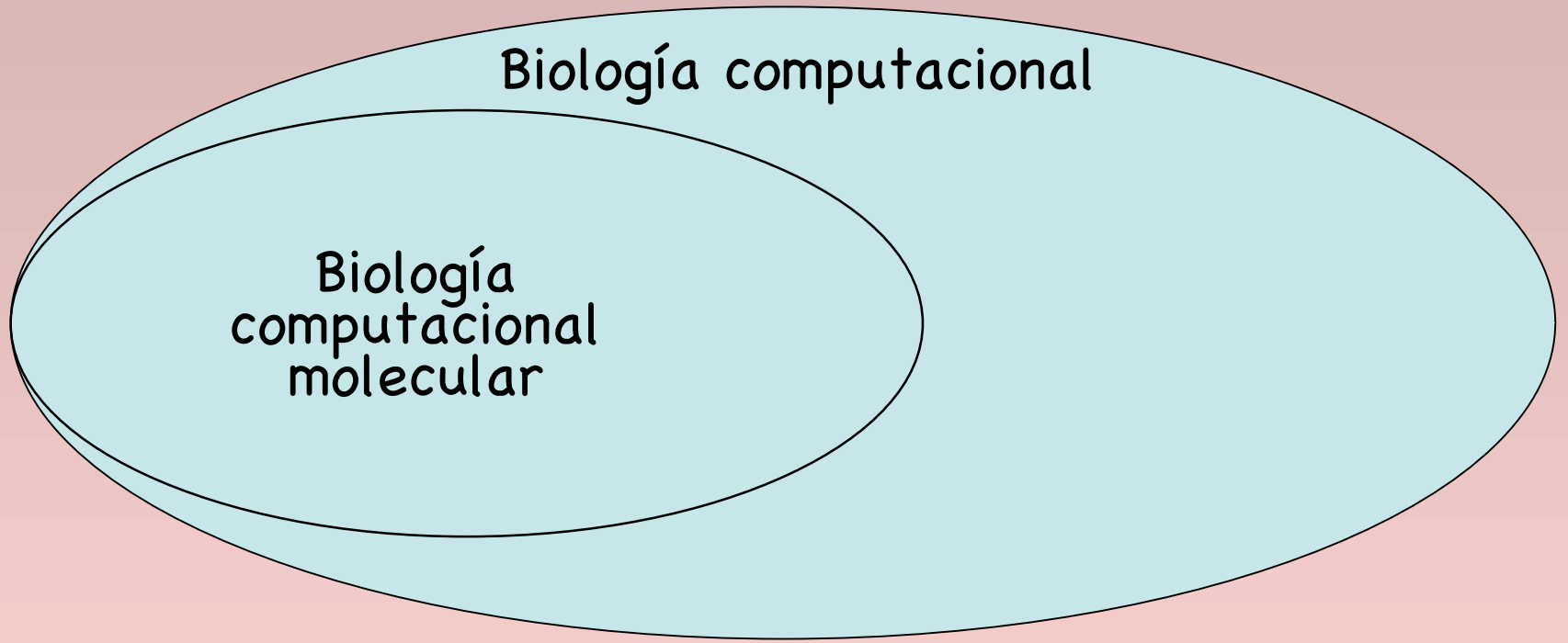
F. Rosselló, *Imaginática07* (2007)

¿Qué es la biología computacional?



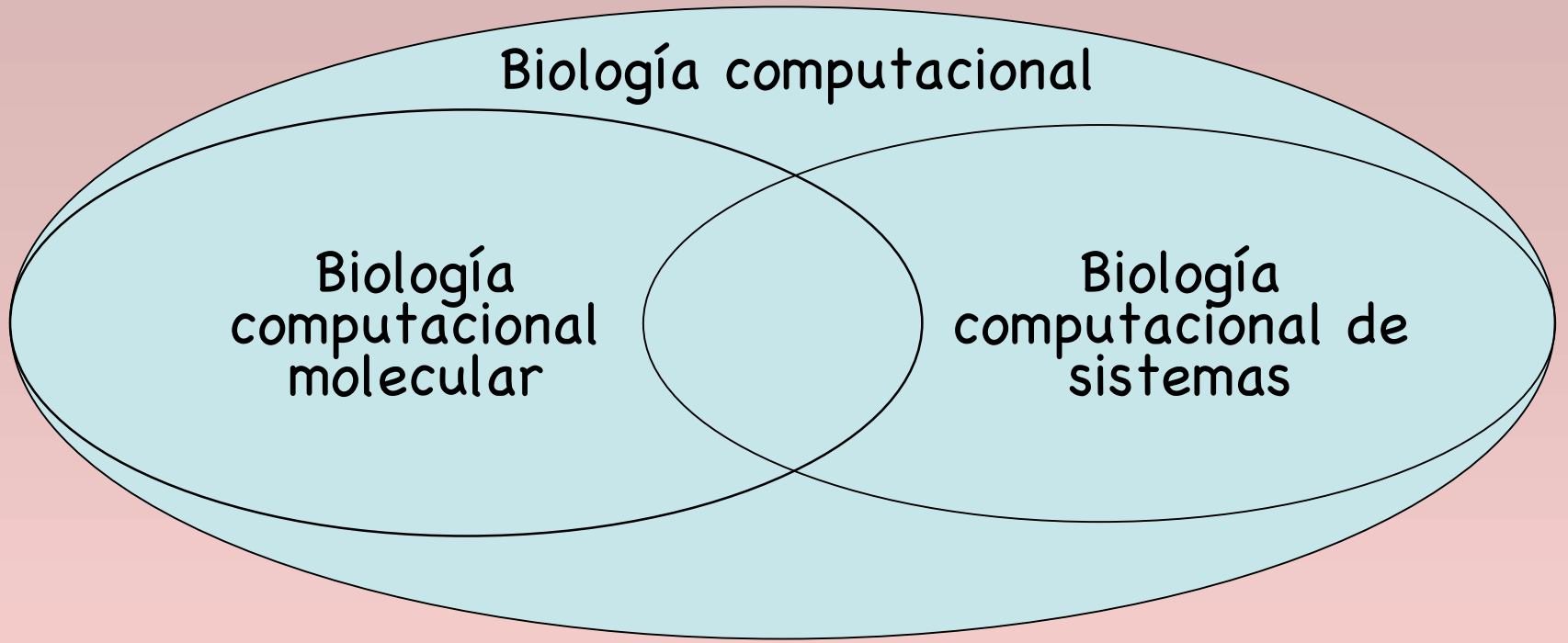
Aplicación de las `matemáticas de la computación' en la resolución de problemas inspirados por la biología

¿Qué es la biología computacional?



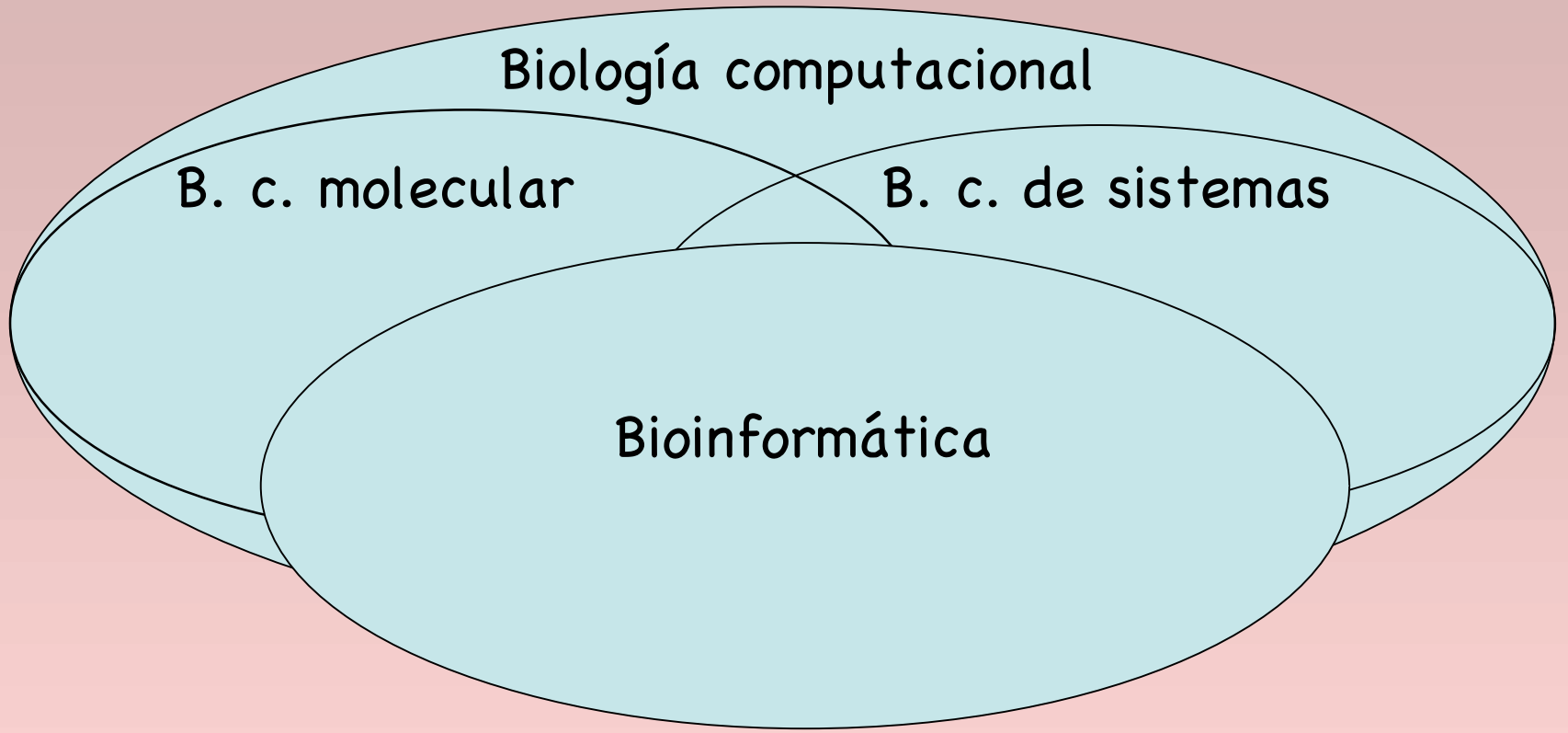
La biología computacional molecular trata problemas relacionados con el análisis y manejo de datos en biología molecular

¿Qué es la biología computacional?



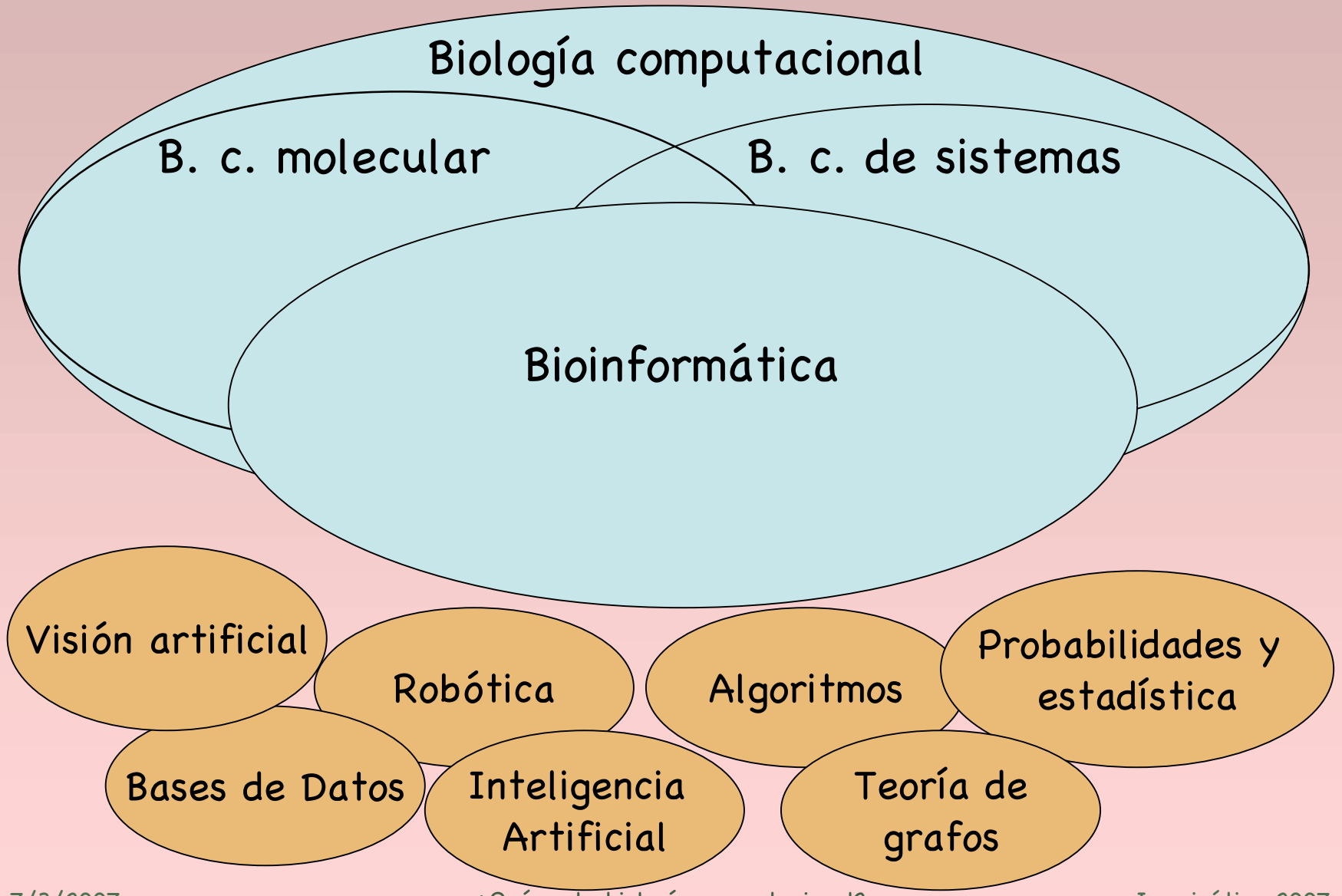
La biología computacional de sistemas se ocupa del modelado, estudio y simulación de redes de interacciones biológicas

¿Qué es la biología computacional?



La bioinformática es en su mayor parte el 'brazo armado' de la biología computacional

¿Qué es la biología computacional?



¿Qué es la biología computacional?

Quedemos en que:

- La biología computacional es el desarrollo de métodos computacionales para resolver problemas en biología (relacionados con ADN, ARN, proteínas, metabolismo)
- La bioinformática es la implementación informática de las herramientas desarrolladas por la b.c. y su aplicación en el manejo y análisis de datos biológicos reales
- Voy a juntarlas en un solo paquete

La bioinformática está de moda

(17/02/2007)

¿Por qué?

- Los avances de la biotecnología han llevado al crecimiento exponencial en el descubrimiento de información genómica y biomolecular
- Los biólogos necesitan métodos computacionales para analizar grandes conjuntos de datos
- Y que estos métodos estén implementados de manera conveniente

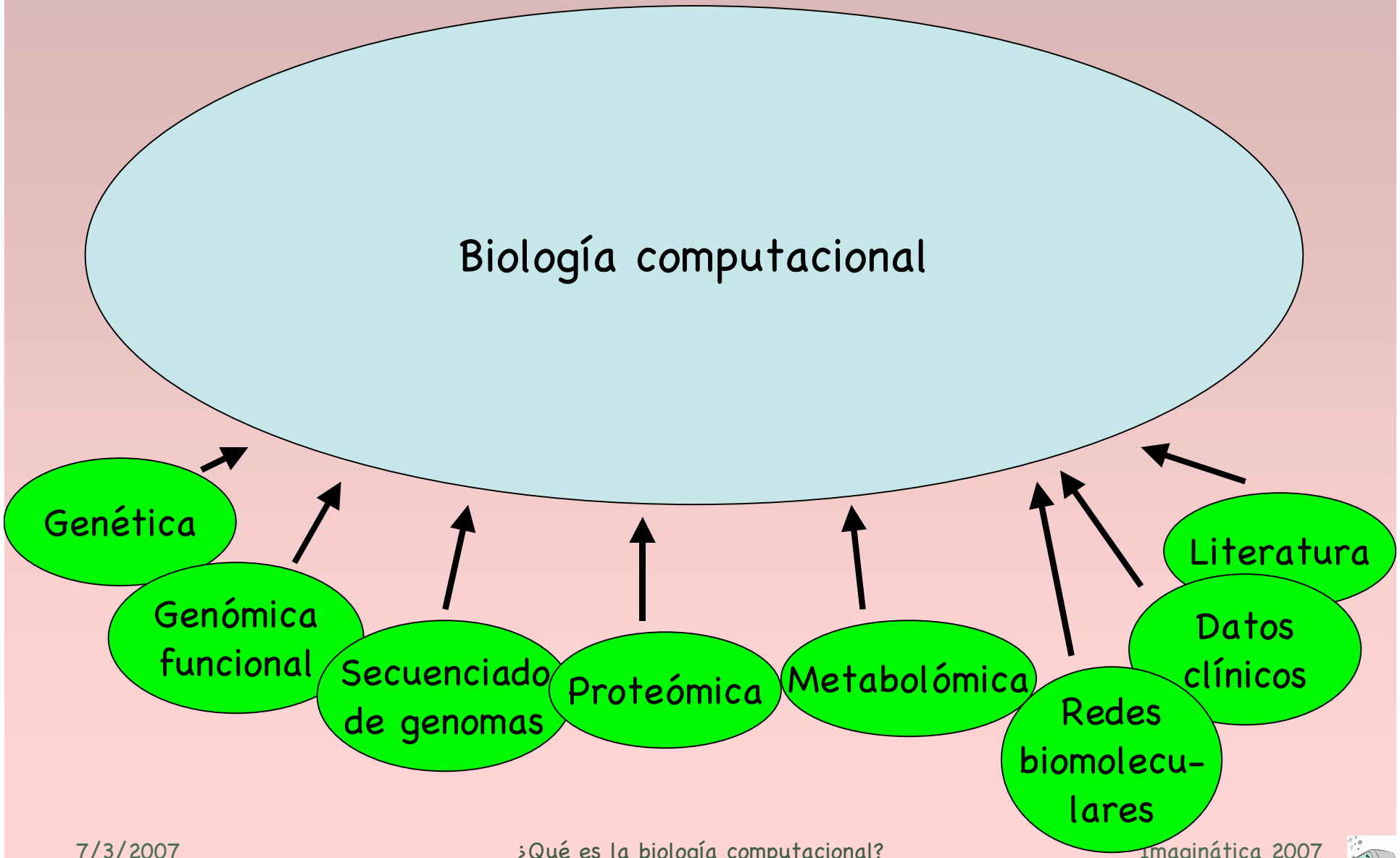
Crecimiento de los datos biomoleculares

En los últimos 15 años hemos vivido un tsunami de datos bioquímicos:

- Más de 400 genomas secuenciados
- Más de 100 Gb en secuencias de ADN
- Más de 300,000 secuencias de proteínas anotadas
- Más de 40,000 estructuras de proteínas
- Más de 47,000 rutas metabólicas

Fuerte marejada de tipos de datos

Variedad de los datos biomoleculares



El ADN

- El ADN (ácido desoxirribonucleico) es la molécula que contiene la información para el desarrollo y funcionamiento de los seres vivos
- Está formado por dos largas cadenas de nucleótidos o 'bases':
  - A: Adenina
  - C: Citosina
  - G: Guanina
  - T: Timina
- Estas cadenas forman una doble hélice

El ADN

- Estas dos cadenas son complementarias (salvo errores):

  A se pega a T

  C se pega a G

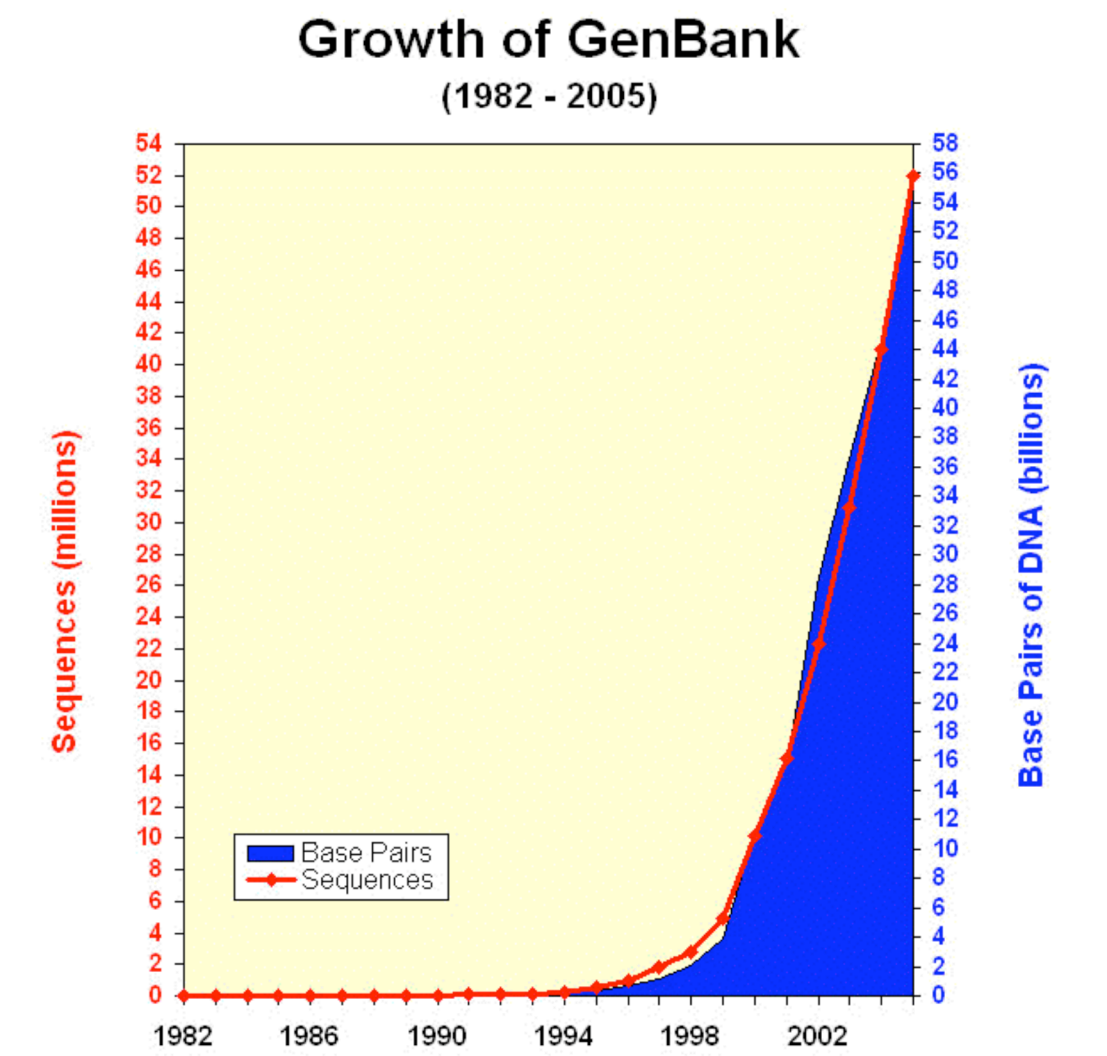
- Por tanto, podemos interpretar una molécula de ADN como una palabra sobre el alfabeto de las bases: A, C, G, T

  (Deformación profesional?)

Secuencia de ADN

…tgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaat
gcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgct
aagctgggatccgatgacaatgcatgcggctatgctaatgaatggtcttgggattt
accttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaat
ggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctgggat
ccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatccg
atgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctgggatc
cgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatcctg
cggctatgctaatgaatggtcttgggatttaccttggaatgctaagctgggatccg
atgacaatgcatgcggctatgctaatgaatggtcttgggatttaccttggaatatg
ctaatgcatgcggctatgctaagctgggaatgcatgcggctatgctaagctgggat
ccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatccg
atgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctcatgcg
gctatgctaagctgggaatgcatgcggctatgctaagctgggatccgatgacaatg
catgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgcta
agctgcggctatgctaatgcatgcggctatgctaagctcggctatgctaatgaatg
gtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggct
atgctaatgaatggtcttgggatttaccttggaatatgctaatgcatgcggctatg
ctaagctgggaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcg
gctatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctg…

Crecimiento de secuencias de ADN

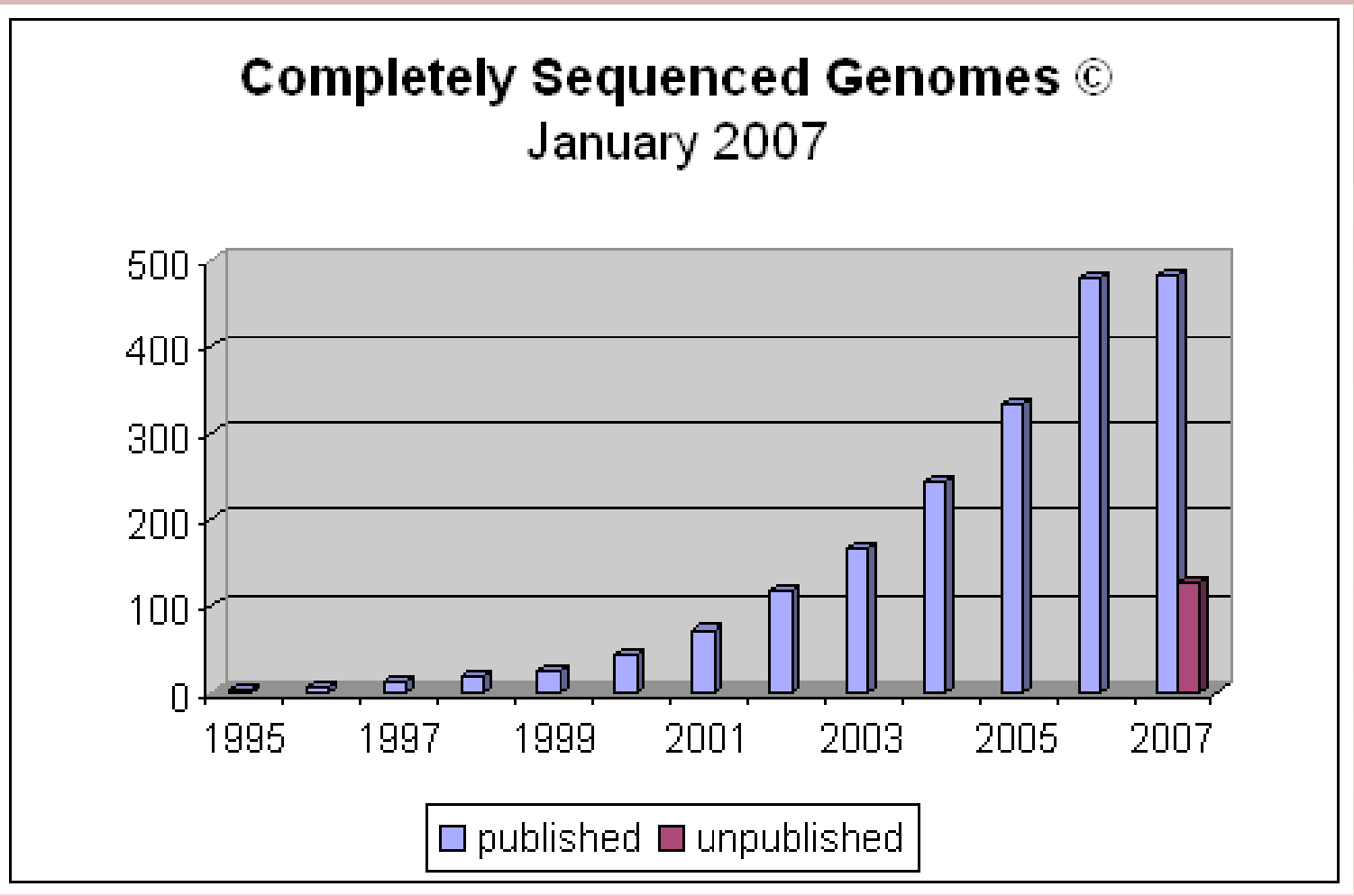


http://www.ncbi.nlm.nih.gov/Genbank/index.html

Genoma y genes

- Todas las células de un organismo tienen (salvo errores) el mismo ADN
- El genoma de un organismo es el conjunto completo de moléculas de ADN de una célula
- Los genes son las unidades físicas básicas de la herencia (definición poco clara). Son secuencias de ADN específicas que codifican las instrucciones para producir proteínas o ARN

Genomas secuenciados



http://www.genomesonline.org/

Tamaños de genomas

	Longitud	Genes
HIV	9,700	9
Myc. genitalum	580,073	483
Escherichia coli	4.6×10^6	5416
Arroz	3.9×10^8	37,544
Perros	2.4×10^9	19,300
Mosca de la fruta	1.3×10^8	13,600
Humanos	3.3×10^9	20,000-25,000
Algunas amebas	6.9×10^{11}	11,000

Apostad en Genesweep

http://www.genomesize.com

Proteínas

- Son parte esencial de los organismos vivos y participan en todos los procesos celulares
- Las proteínas son cadenas de aminoácidos. Hay 20 aminoácidos.
- Por tanto, una proteína es una palabra sobre un alfabeto de 20 letras:

  A,R,N,D,C,E,Q,G,H,I,L,K,M,F,P,S,T,W,Y,V

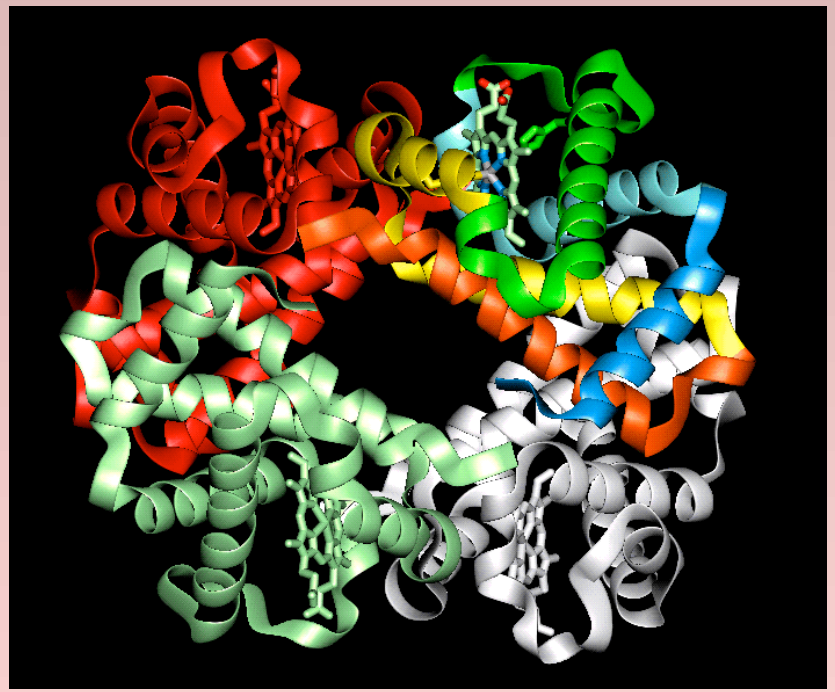
(A: Alanina, R: Arginina, N: Asparagina, etc.)

…KKTILAIAIPALFASAANAAVIYDKDGTTFDVYGRVQANYYGDTNEADSTAASGYK
DVDGELKGSSRLGWSGKIALNNTWSGIAKTEWQVSAENSANKFDSRHIYVGFDGTQY
GKVIFGQTDTAFYDVLEPTDIFNEWGSEGNFYDGRQEGQVIYSNAIGGFKGKVSYQT
NDDQAVKVADVAGGIKTTVFPDVKRKYAYAAAVGYDFDFGLGFNGGYAYNGKTTFE…

Estructuras de proteínas

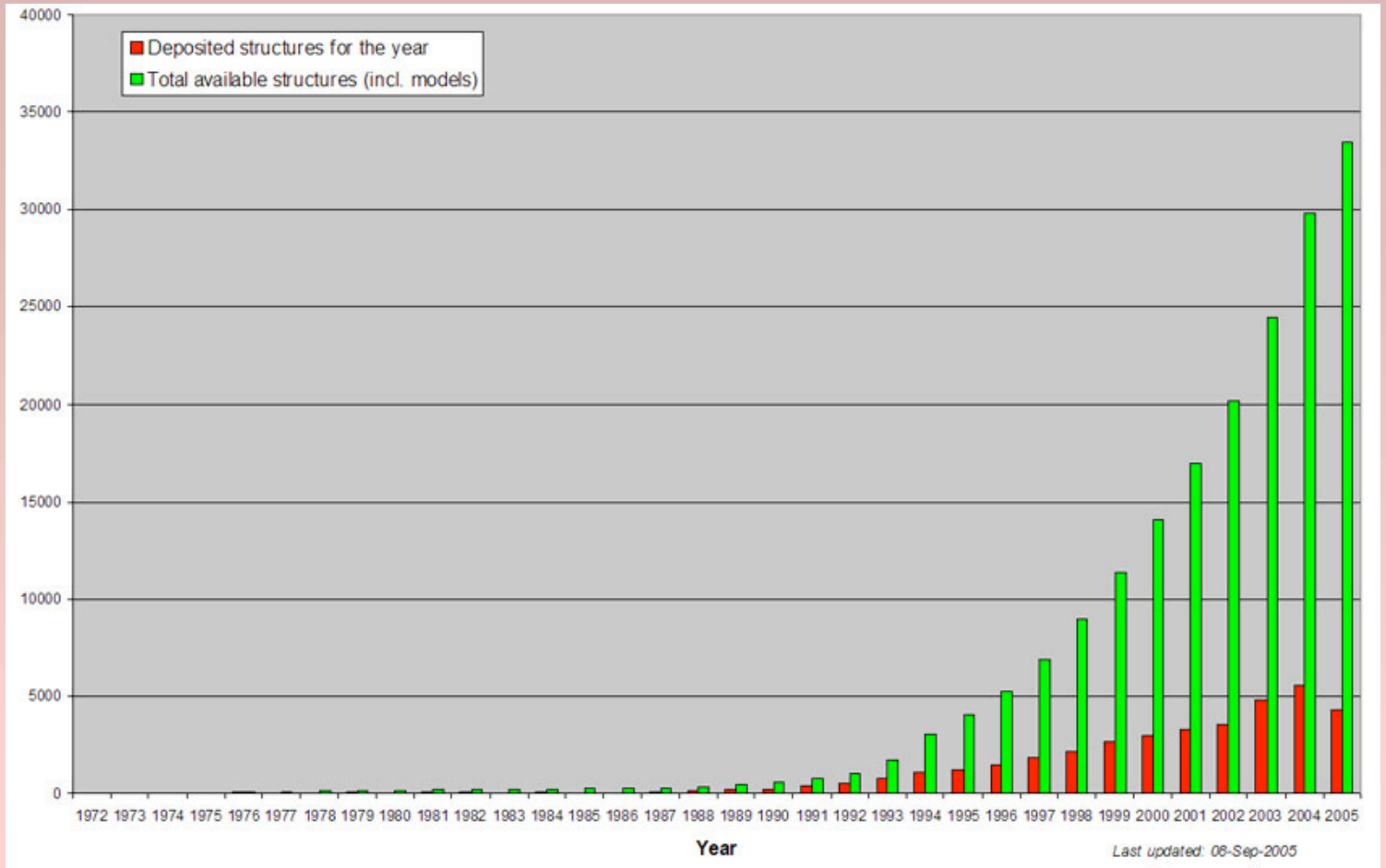
Las proteínas se pliegan dentro de las células en complicadas estructuras 3D que determinan su función

En general, esta estructura viene determinada por la secuencia de aminoácidos



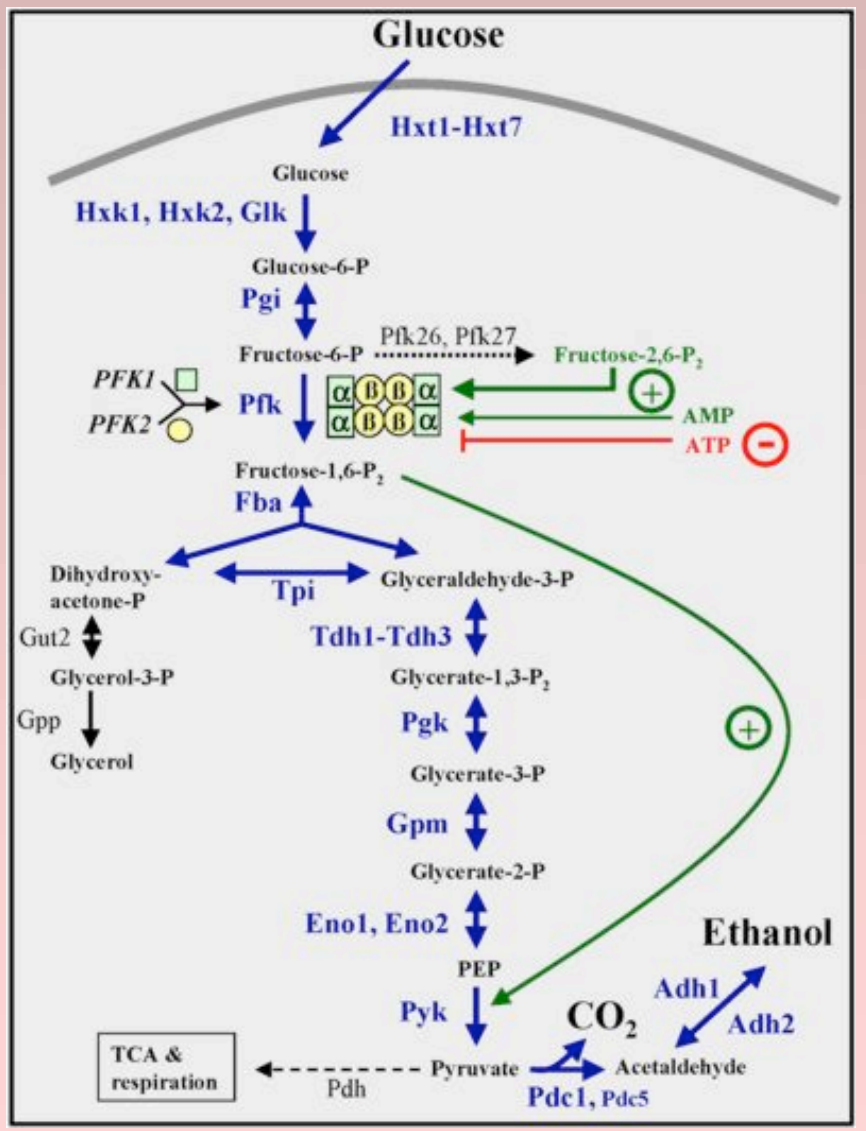
Estructura de una hemoglobina

Crecimiento de estructuras de proteínas



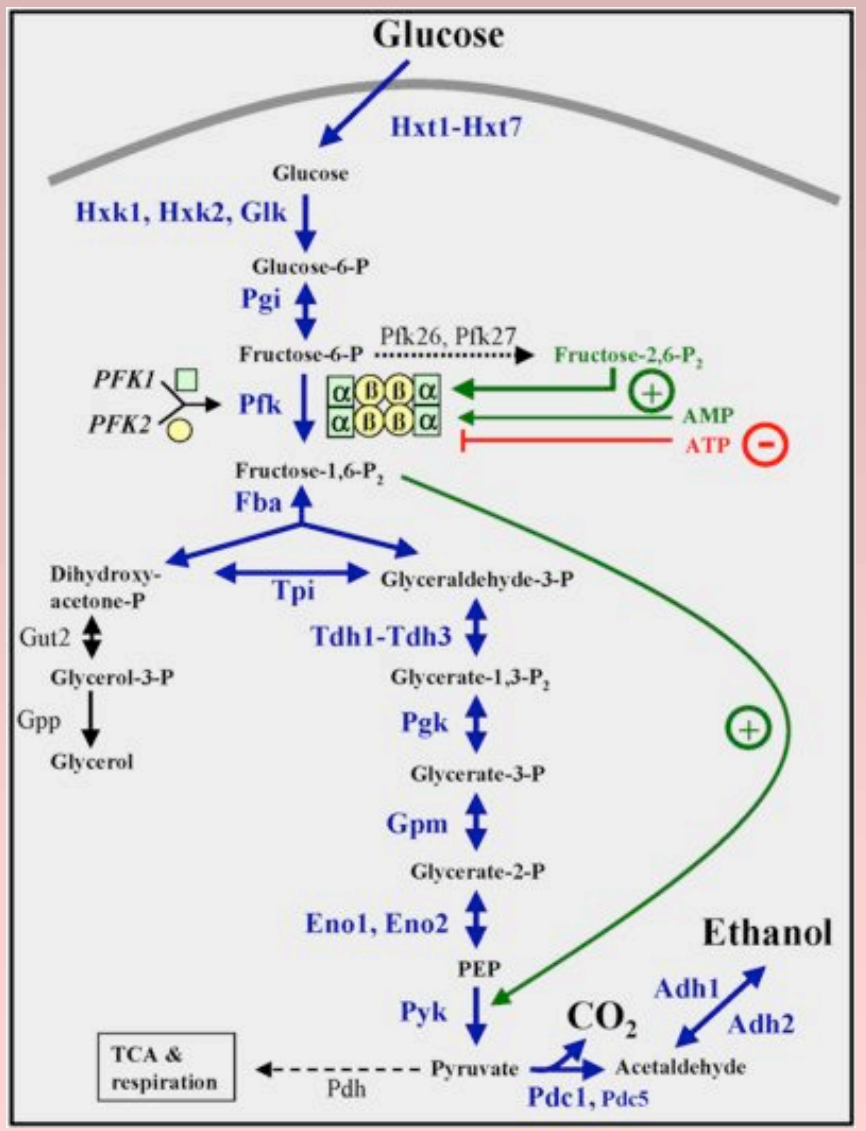
Rutas biomoleculares

Ruta metabólica:
Serie de procesos y reacciones bioquímicas catalizadas por enzimas y que producen compuestos (metabolitos) que son usados o almacenados por la célula



Ruta metabólica

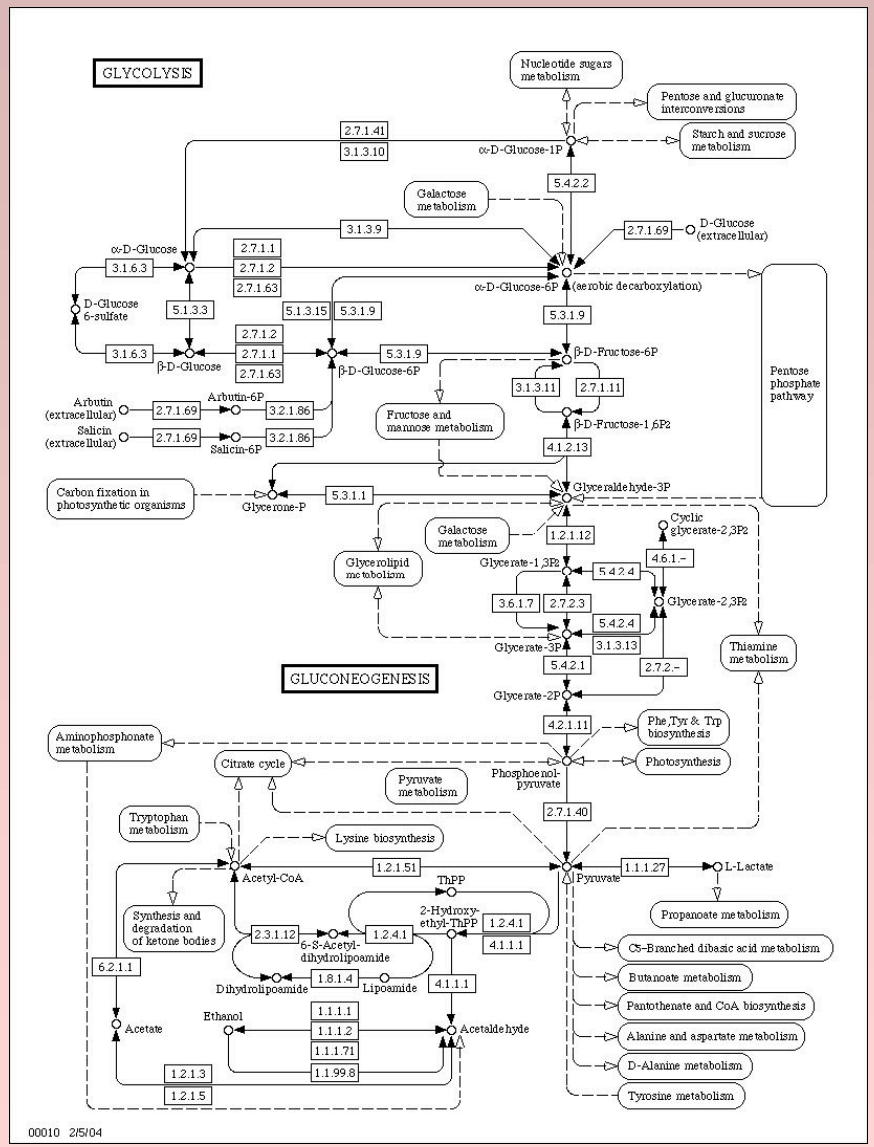
Por ejemplo, la glucólisis que oxida la glucosa y produce ATP, NADH (energía) y ácido pirúvico



Rutas metabólicas

Las rutas metabólicas se combinan y entrecruzan...

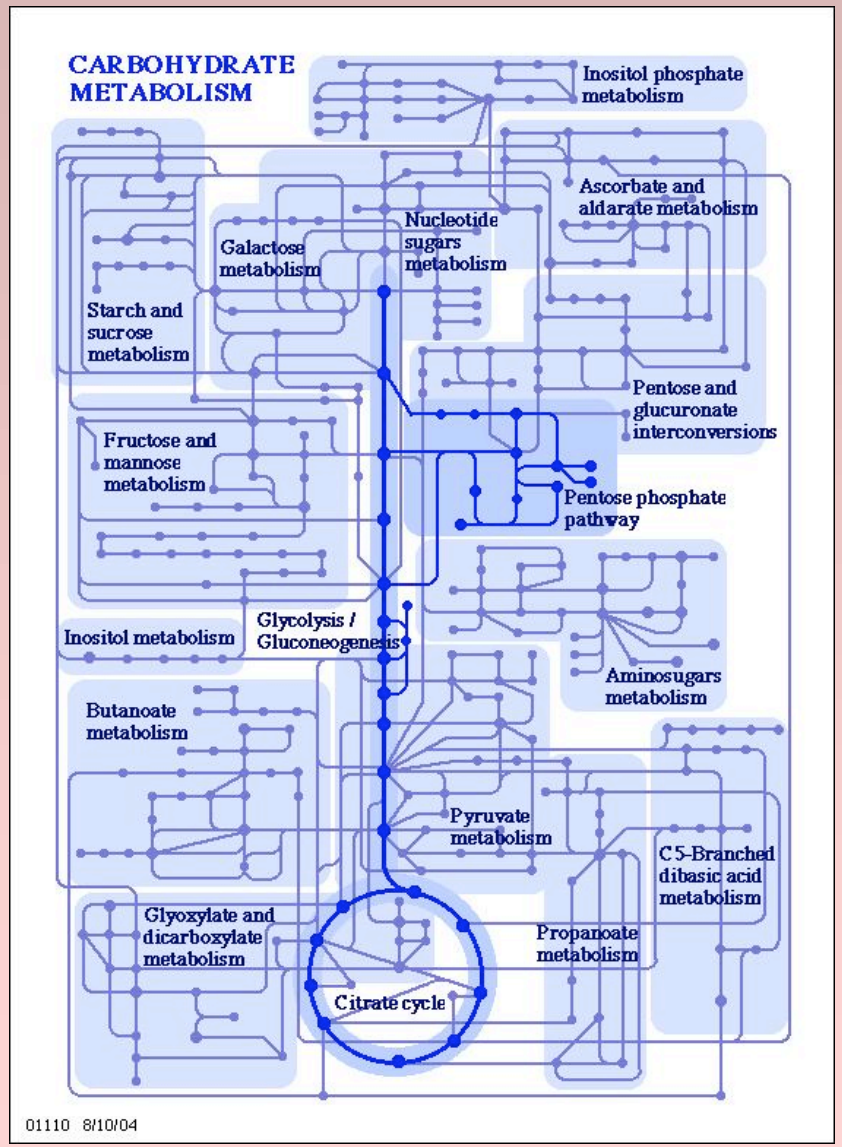
http://www.genome.ad.jp/kegg/pathway/sco/sco00010.html



Rutas metabólicas

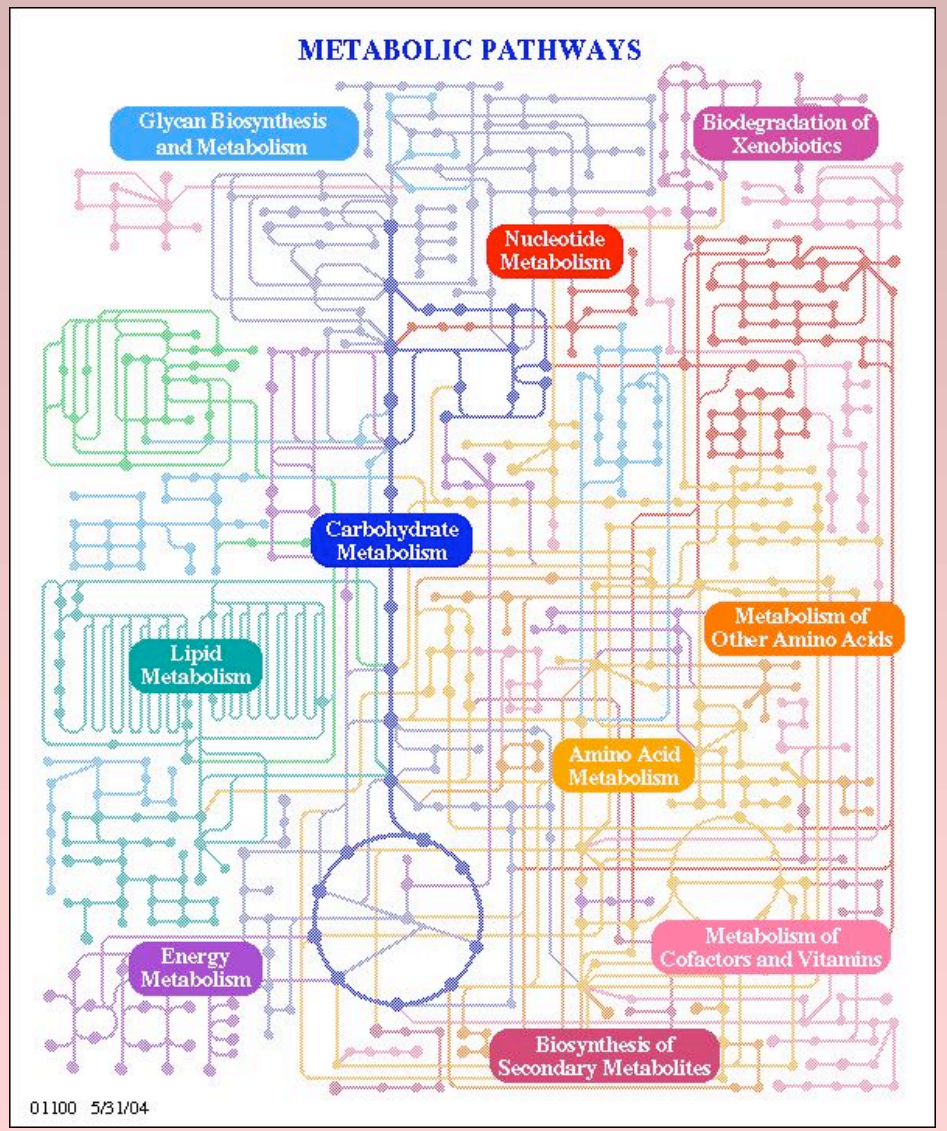
...y se combinan
y entrecruzan...

http://www.genome.ad.jp/kegg/pathway



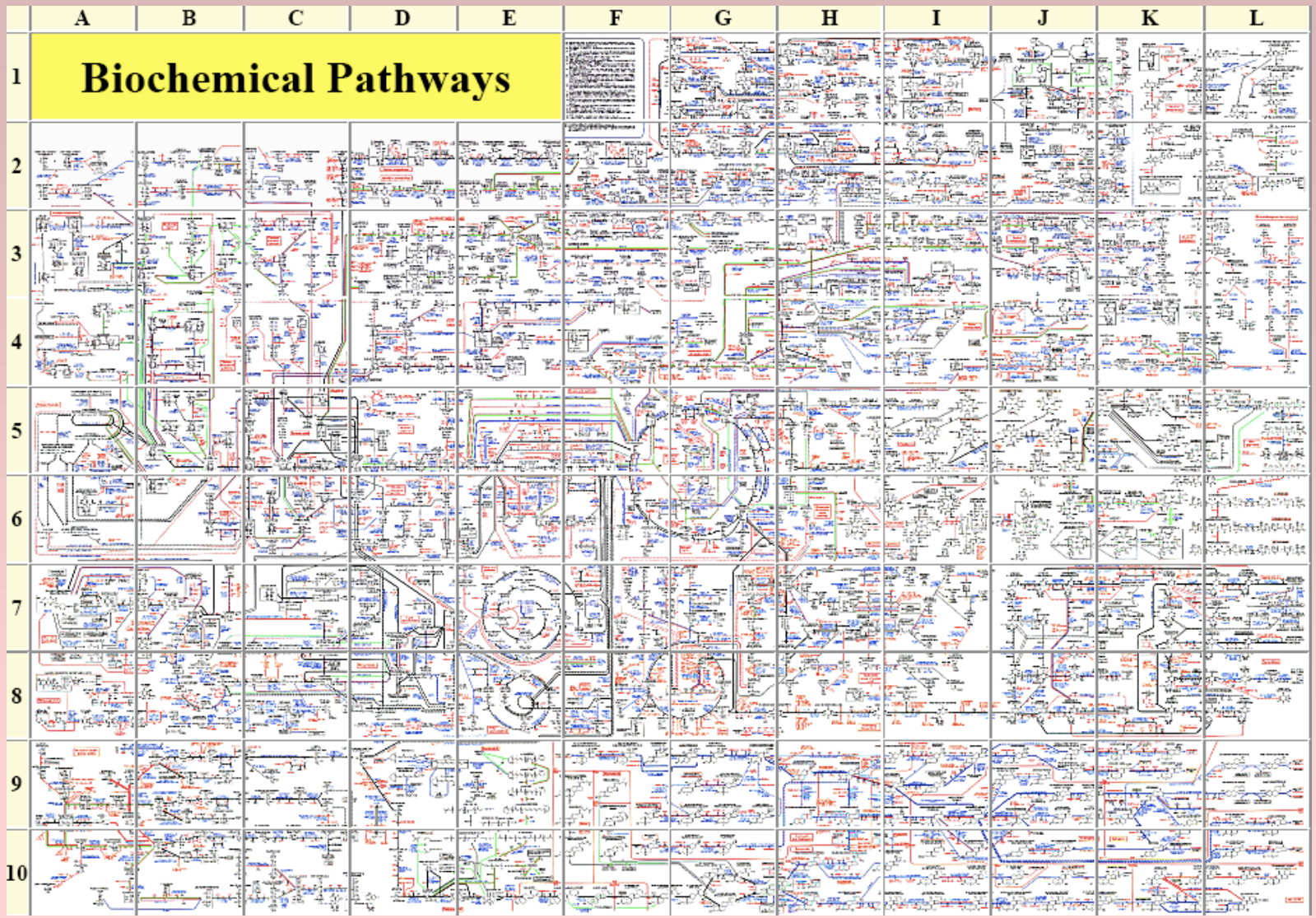
Red metabólica

... hasta formar una red metabólica: el conjunto de todas las rutas metabólicas de una célula.



http://www.genome.ad.jp/kegg/pathway

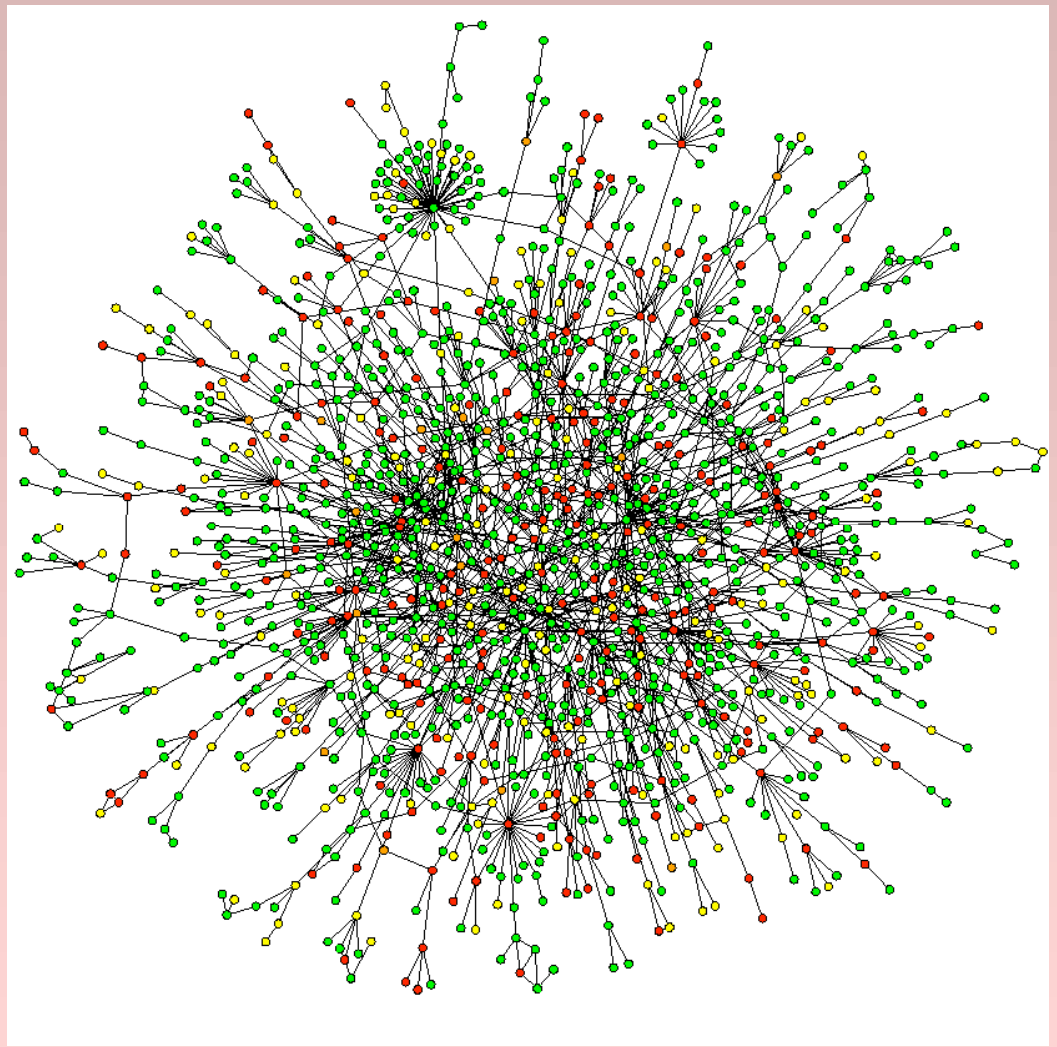
Red metabólica



https://www.roche-applied-science.com

Rutas biomoleculares

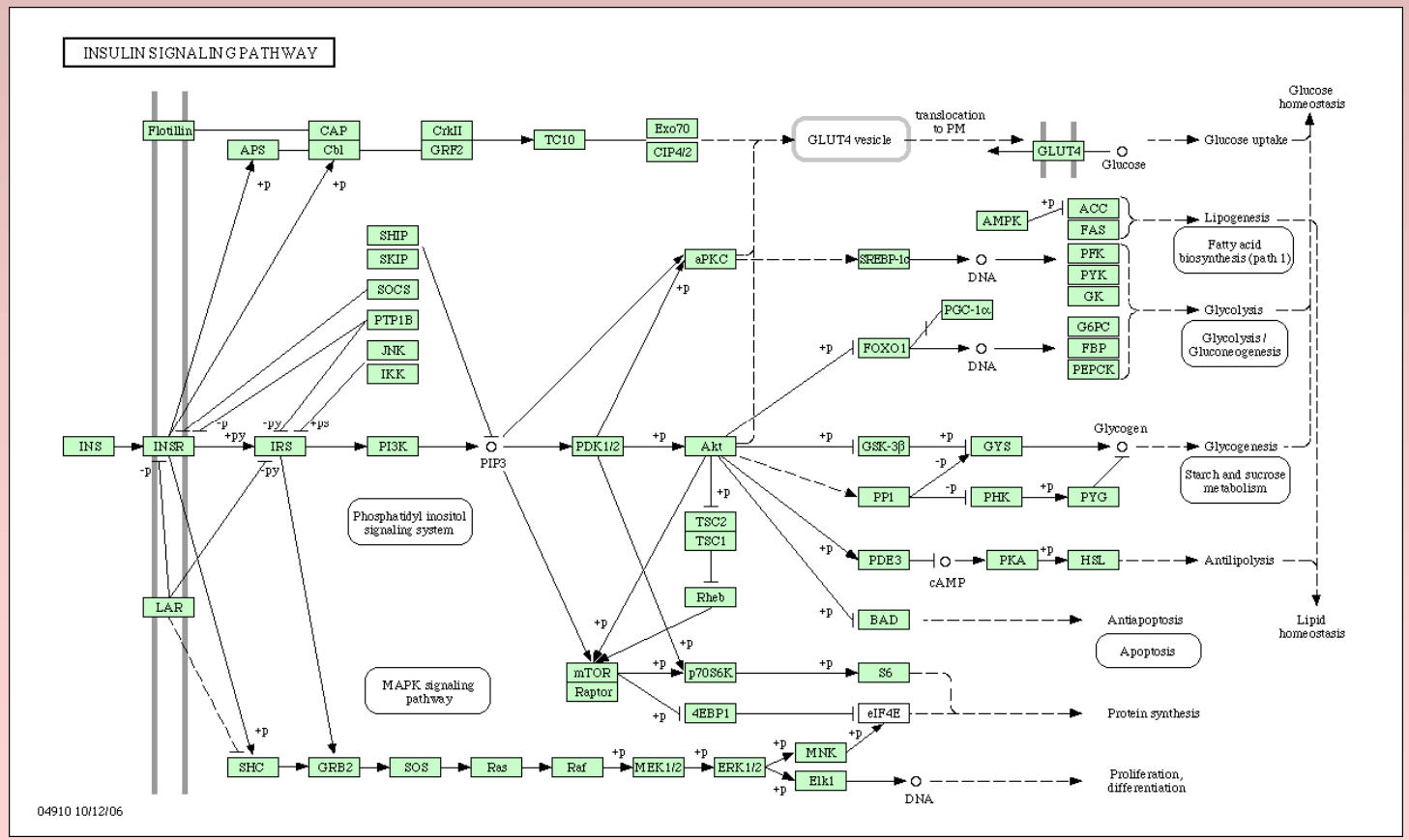
Red de proteínas: Conjunto de interacciones entre proteínas en una célula



Red de proteínas de la levadura
S. Wuchty et al. Nature Genetics (2003)

Ruta de señales

Ruta de señales: Serie de procesos y reacciones bioquímicas que traducen una señal bioquímica en una respuesta celular

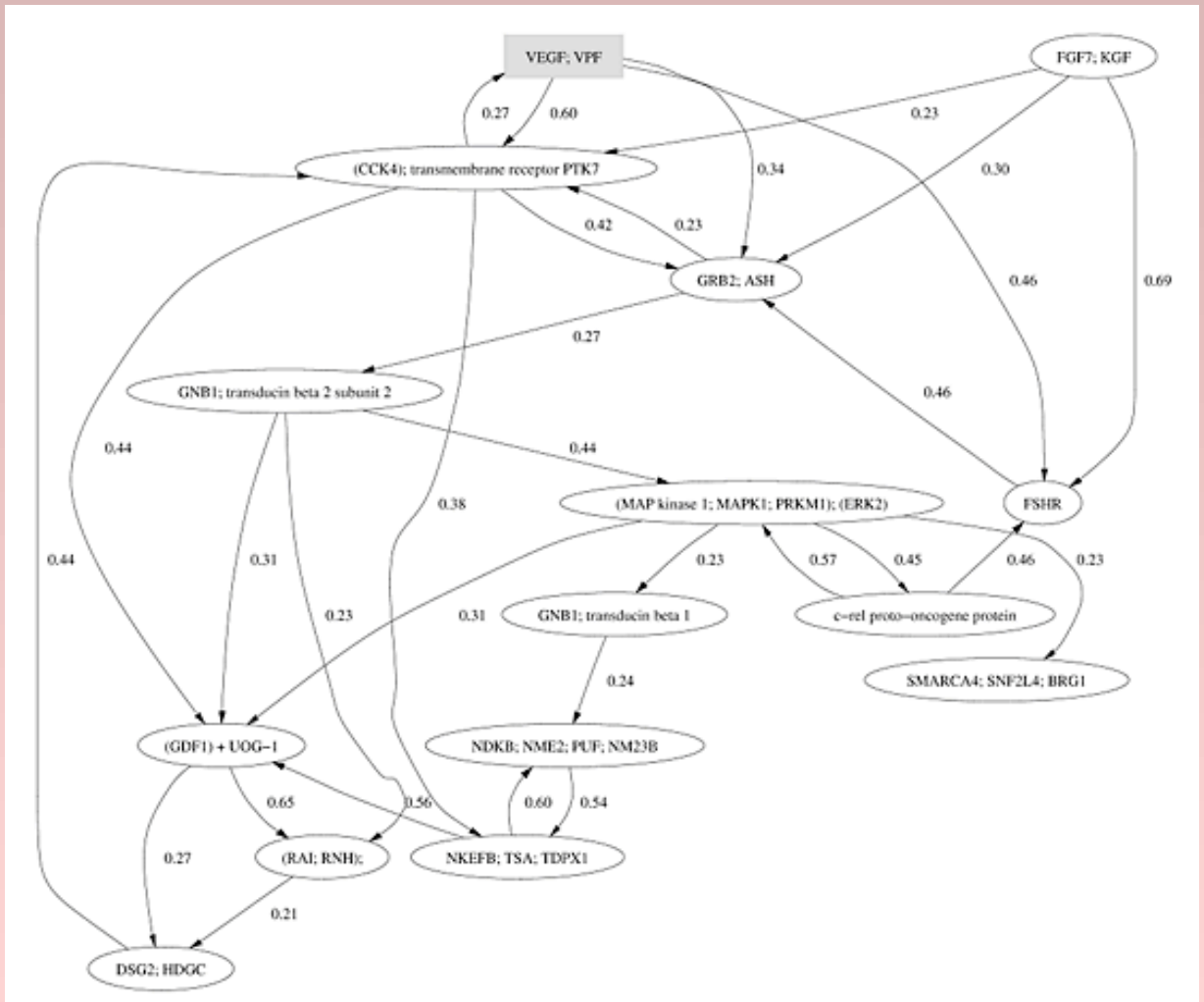


http://www.genome.ad.jp/kegg/pathway/hsa/hsa04910.html

Red genética

Red genética: Conjunto de las interacciones regulatorias entre genes y proteínas

Trozo de red de transcripción genética del glioma



http://www.systemsbiology.org/

Rutas biomoleculares

KEGG PATHWAY
http://www.genome.ad.jp/kegg/pathway.html

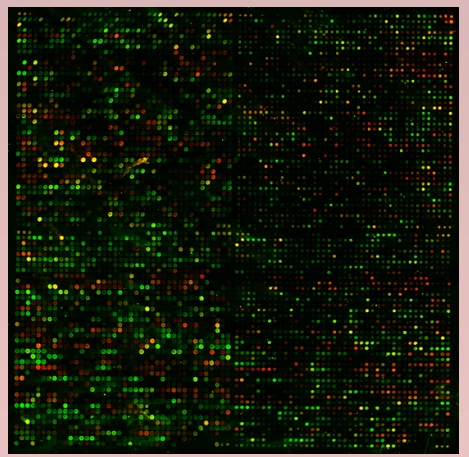
contiene 298 rutas biomoleculares de referencia para muchos organismos, en total 47,141

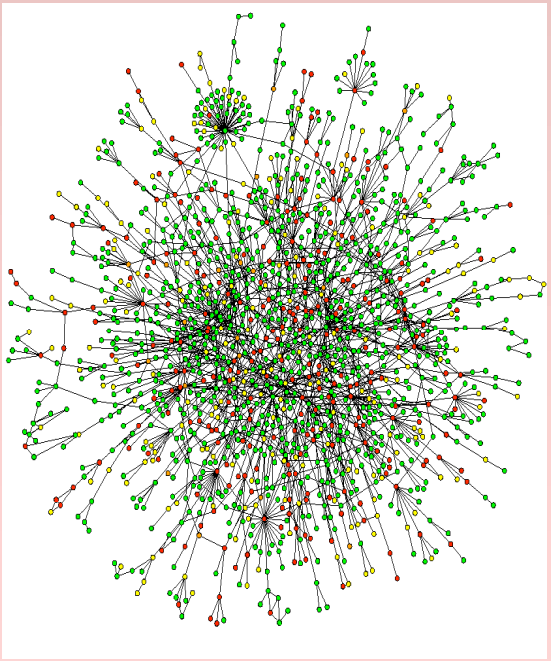
Necesidad de la bioinformática

- La naturaleza discreta de las secuencias de ADN, ARN y proteínas son ideales para su análisis empleando técnicas de matemática discreta (y probabilidades y estadística) sobre ordenadores digitales
- El tamaño y complejidad de los datos obtenidos en biología molecular y celular hacen imposible su análisis sin ordenadores

Cantidades ingentes de información

- Los chips de microarrays permiten observar la actividad de miles de genes, simultáneamente, bajo diferentes condiciones
- Nuevas tecnologías permiten examinar todas las proteínas presentes en una célula y su interacción bajo diferentes condiciones
- Hay que analizar y comparar esta información





Cantidades ingentes de información

Miles de bases de datos de información biológica para exprimir:

- DNA: GenBank, EMBL, DDBJ
- Genomas: Ensembl, SGD, UCSC
- Secuencias de proteínas: PiR, Swiss-Prot, PROSITE
- Estructuras de proteínas: PDB, CATH, SCOP
- Microarrays: ArrayExpress, GEO
- Interacciones de proteínas: BioGRID, STRING
- Rutas metabólicas: KEGG, BioCyc
- Literatura: MEDLINE, PubMed
- Metabases: Entrez, MetaDB (+1200 entradas de 60 categorías)

El reto de la bioinformática

Desarrollar técnicas computacionales para analizar estos datos que permitan extraer el máximo de información biológica y permitan a los biólogos:

- interpretar datos experimentales de manera efectiva
- desarrollar hipótesis rápidamente
- diseñar nuevos experimentos para validación o recolección
- modelar computacionalmente procesos biológicos
- predecir el comportamiento de un proceso biológico
- ……..

Los biólogos son diferentes

- Nada es completamente cierto o falso en biología, mientras que todo es completamente verdadero o falso en informática
- Los biólogos tienen dogmas, los informáticos teoremas
- Los biólogos tienen claro que en todos los datos hay errores, los informáticos lo odian

Los biólogos son diferentes

- Los biólogos contruyen a partir de datos, los informáticos a partir de conceptos abstractos
- Para trabajar en biología hay que saber muchas cosas, en informática hay que saber hacer muchas cosas
- Es más fácil estudiar por cuenta propia biología que matemáticas o informática

Los biólogos son diferentes

- En los laboratorios de biología hay una jerarquía muy estricta, los informáticos son mucho más democráticos
- El ideal platónico del biólogo es el director de un gran laboratorio con montones de gente trabajando para él; el ideal platónico del informático es el *hacker* trabajando en un sótano

¿Qué hace un bioinformático/ biólogo computacional?

MAYKO TRAN

Esta especialista en bioinformática nació en Vietnam, pero se crió en Canadá.

Allí estudió en la Universidad de la Columbia Británica y posteriormente cursó masters en Biología e Ingeniería Informática en la Universidad de Toronto y en Bioinformática en el prestigioso Massachussets Institute of Technology (MIT).

¿Qué hace un bioinformático/ biólogo computacional?

Según la International Society for Comput. Biology ISCB, un bioinformático:

- Crea nuevas formas de recolección, manejo y análisis de datos
- Desarrolla programas de ordenador que modelan procesos biológicos a nivel genético o molecular
- Asiste a un equipo en la interpretación de datos biológicos y médicos

http://www.iscb.org

¿Qué hace un bioinformático/ biólogo computacional?

Según la International Society for Comput. Biology ISCB, un biólogo computacional:

- Desarrolla nuevos modelos de procesos biológicos y métodos para su análisis
- Trabaja con modelos computacionales y bases de datos para predecir o explicar matemáticamente los resultados de una investigación biológica

http://www.iscb.org

¿Qué hace un bioinformático/ biólogo computacional?

La mayor parte de problemas en biomedicina

- Necesitan bioinformática/biol. comp.
- Pero necesitan sólo algo de bioinformática/ biol. comp.

Por tanto, el bioinformático:

- Suele trabajar en diferentes proyectos simultáneamente
- Es indispensable, pero no dirige los proyectos
- … salvo en empresas de bioinformática

http://www.iscb.org

¿Qué hace un bioinformático? Análisis funcional de genes

- Secuenciar una cadena larga de ADN
- Búsqueda de genes
- Traducción en proteína
- Búsqueda de homólogos
- Búsqueda de dominios y motivos en la proteína
- Análisis de la estructura secundaria de la proteína
- Predicción de la estructura 3D de la proteína
- Análisis de la expresión del gen

¿Cómo formarse en bioinformática?

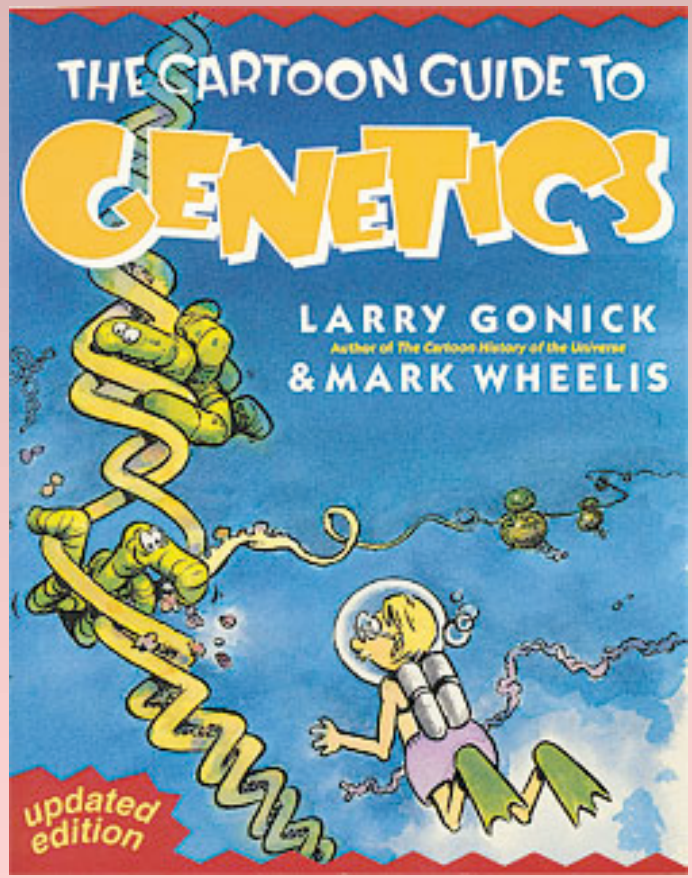
En la carrera:

- Aprovechad asignaturas de libre configuración para cursar algunas asignaturas de bioquímica, genética y biología celular (y colaborad con vuestros compañeros biólogos)
- Cursad alguna asignatura de Análisis de Datos o de Probabilidades y Estadística II
- Os presupongo la capacidad de aprender fácilmente lenguajes de programación nuevos

¿Cómo formarse en bioinformática?

Mis libros favoritos:

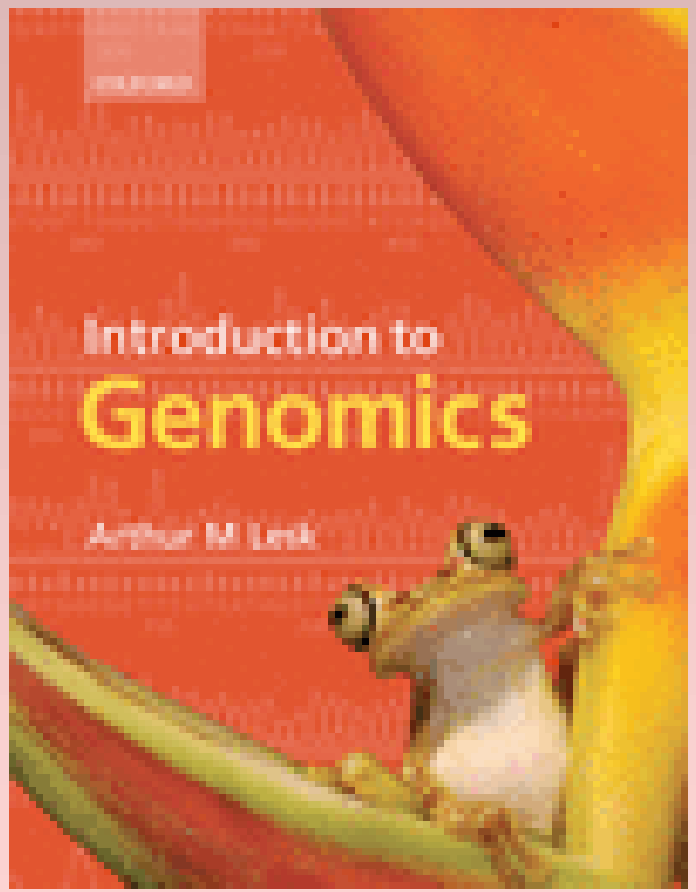
- De biología:
  - L. Gonick, M. Wheelis, "The cartoon guide to Genetics"



¿Cómo formarse en bioinformática?

Mis libros favoritos:

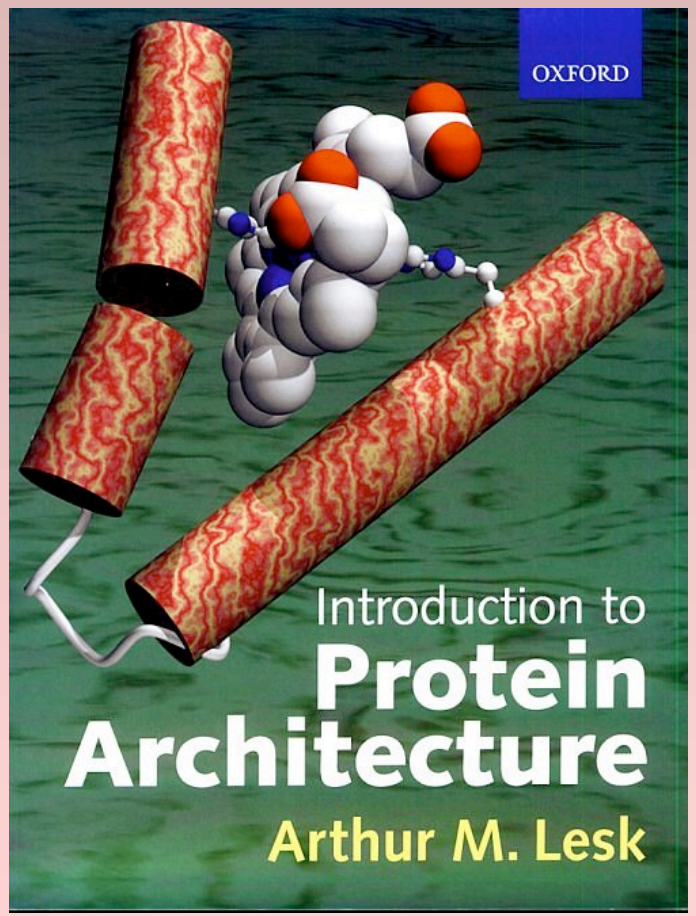
- De biología:
  - A. Lesk, "Introduction to genomics"



¿Cómo formarse en bioinformática?

Mis libros favoritos:

- De biología:
  - A. Lesk, "Introduction to protein architecture"



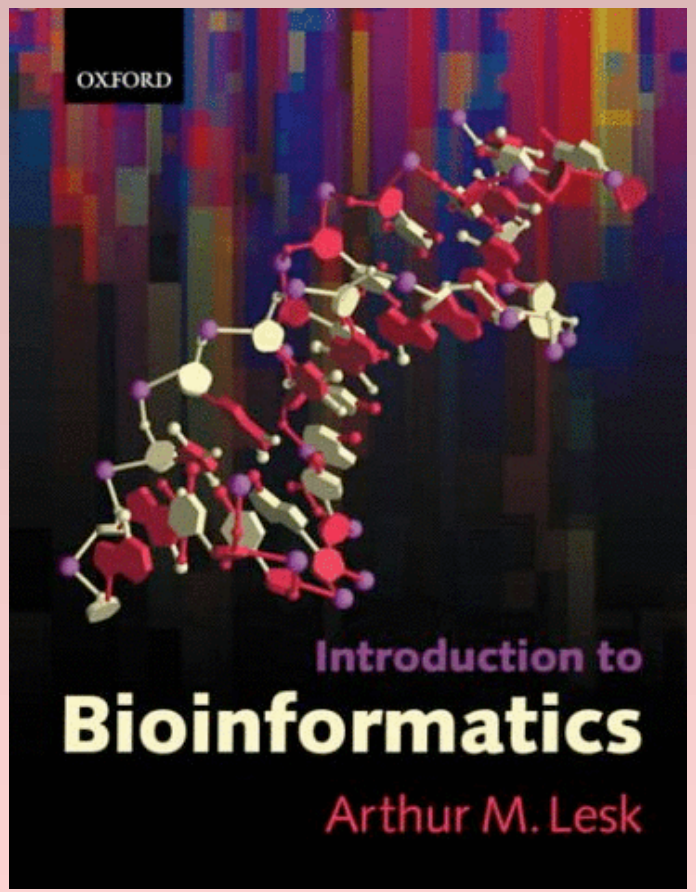
¿Cómo formarse en bioinformática?

Mis libros favoritos:

- De bioinformática:
  - A. Lesk, "Introduction to bioinformatics"

    (una buena introducción general)

    (¡No, no tengo comisión con los libros de Lesk!)

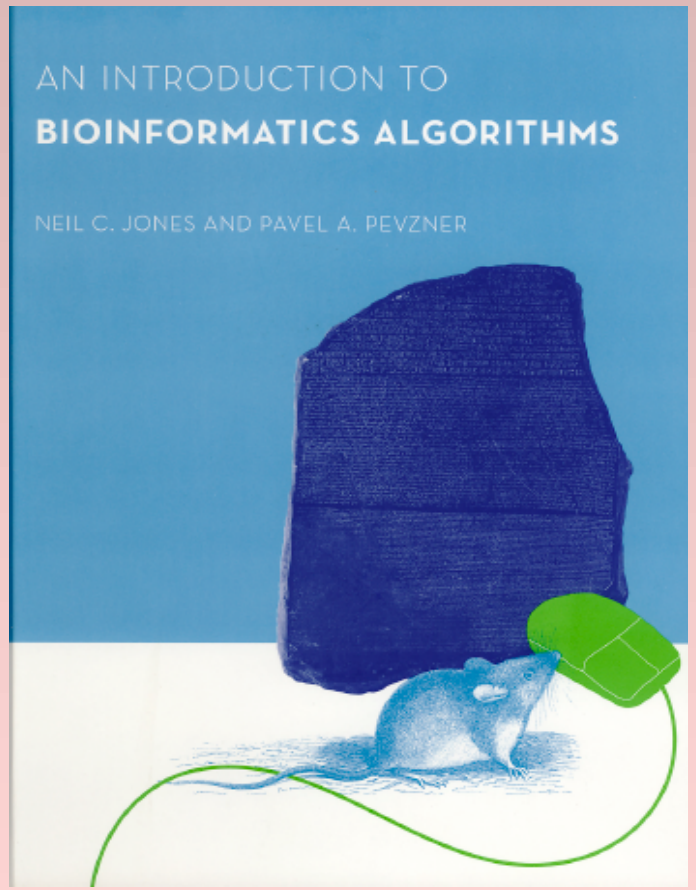


¿Cómo formarse en bioinformática?

Mis libros favoritos:

- De bioinformática:
  - N. Jones, P. Pevzner, "An introduction to bioinformatics algoritms"

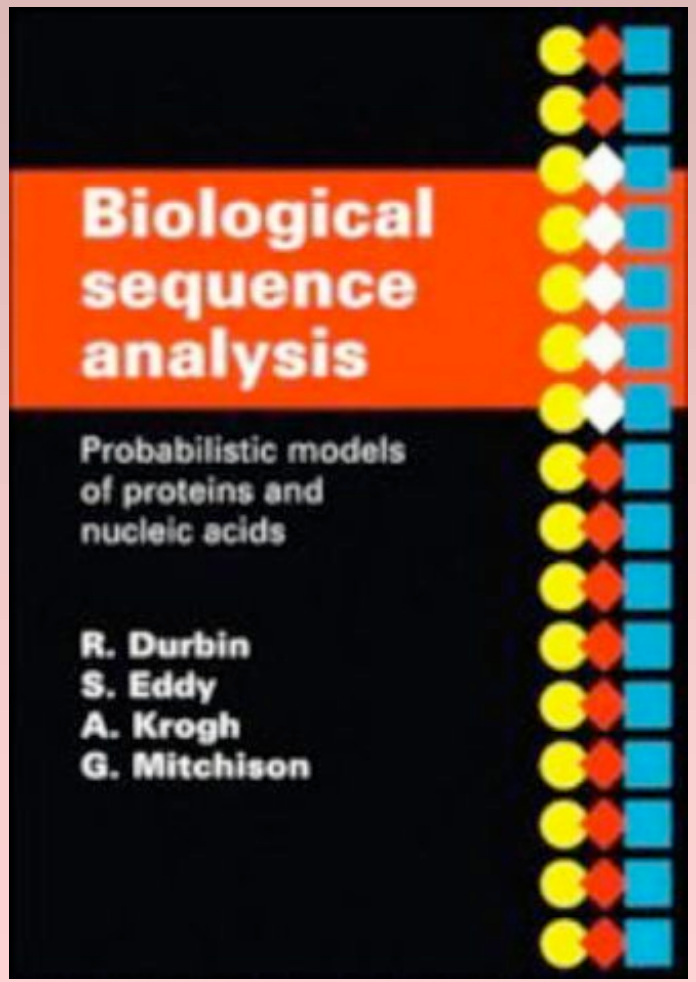
(algoritmia elemental)



¿Cómo formarse en bioinformática?

Mis libros favoritos:

- De bioinformática:
  - R. Durbin, S. Eddy et al, "Biological sequence analysis"



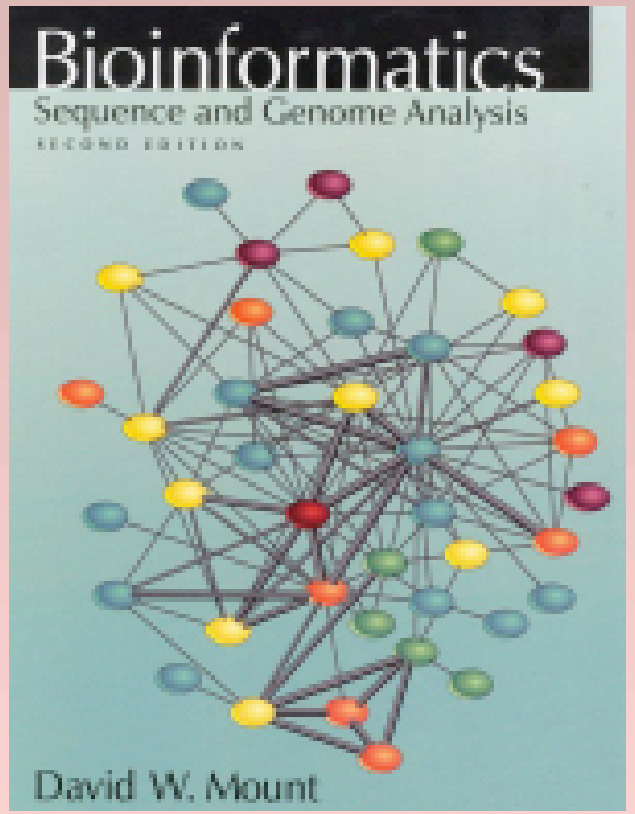
(métodos probabilísticos, muy fácil de estudiar)

¿Cómo formarse en bioinformática?

Mis libros favoritos:

- De bioinformática:
  - D. Mount, "Bioinformatics: Sequence and genome analysis" (2a edición)

(avanzado)

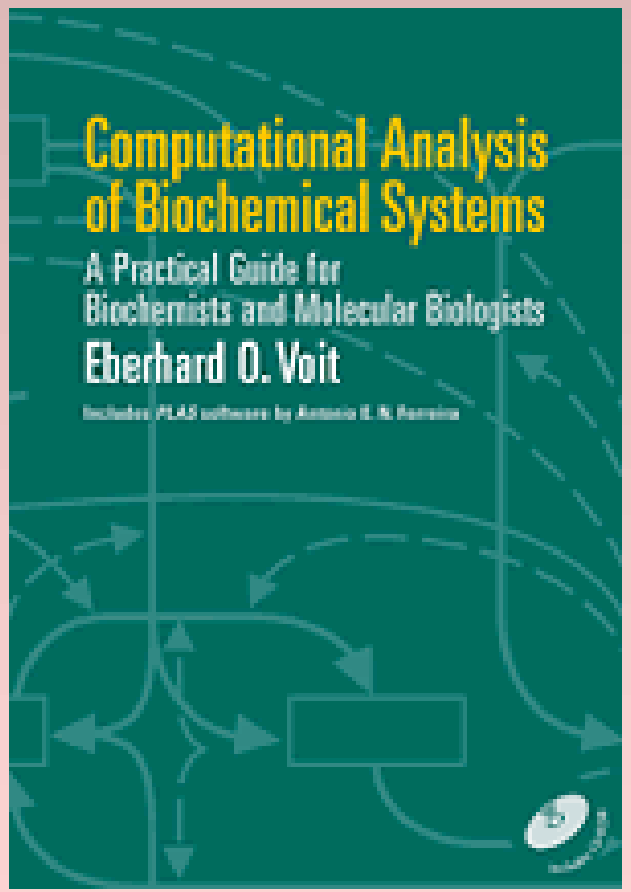


¿Cómo formarse en bioinformática?

Mis libros favoritos:

- De bioinformática:
  - E. Voit, "Computational analysis of biochemical systems"

    (introducción a la biología computacional de sistemas)



Algunos problemas en biología computacional

- Alineamiento de secuencias
- Secuenciado de ADN
- Búsqueda de genes
- Estructura 3D de las proteínas
- Metabolómica

Alineamiento de secuencias

Lewis Carroll (1878), juego del *doblete*: transformar una palabra en otra empleando número mínimo de cambios de una letra y de manera que las palabras intermedias tengan sentido

Alineamiento de secuencias

tonto

listo

Alineamiento de secuencias

tonto

tinto

pinto

pisto

listo

Alineamiento de secuencias

Juegos de palabras de Lewis Carroll (*A través del espejo*, 1936):

'Twas brillig and the slithy toves
Did gyre and gimble in the wabe:
All mimsy were the borogoves
And the mome raths outgrabe

"*Brillig* means four o'clock in the afternooon: the time when you begin *broiling* things for dinner"

Alineamiento de secuencias

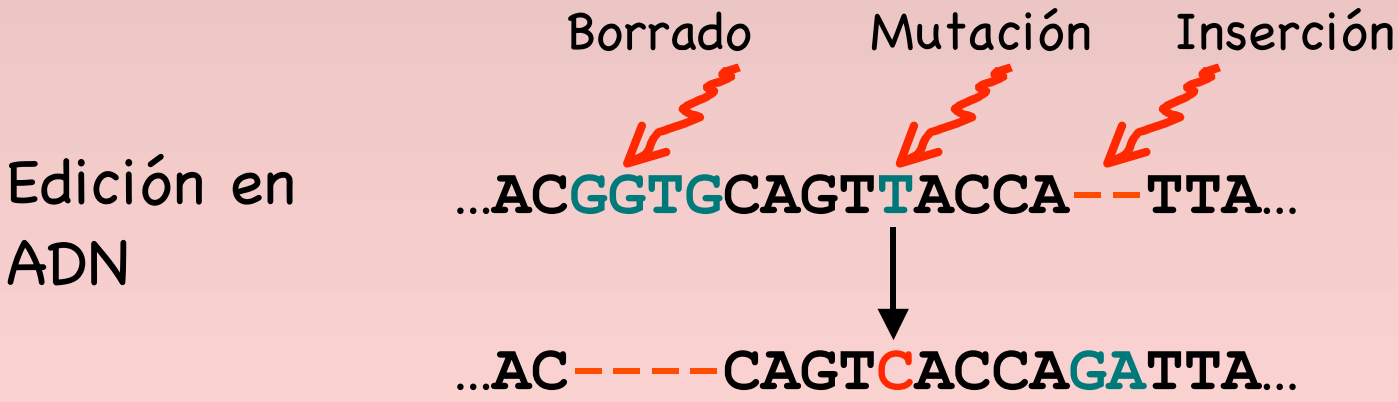
Un juego de palabras de Lewis Carroll (*A través del espejo*, 1936):

'Twas brillig and the slithy toves

```
BR-ILLI-G
|| ||  |  |
BROIL-ING
```


Alineamiento de secuencias

El alineamiento de secuencias busca correspondencia entre secuencias, en base a operaciones de edición



Importancia del alineamiento de secuencias

El alineamiento de dos secuencias permite medir su semejanza

Semejanza de secuencias puede indicar función similar, o cercanía evolutiva, o similaridad estructural, o alguna otra cosa importante...

Alineamiento de dos secuencias

Problema: Dadas dos secuencias

$$x = x_1 x_2 \ldots x_N, \qquad y = y_1 y_2 \ldots y_M$$

un alineamiento es una matriz de dos filas cuya primera (segunda) fila contiene los carácteres de x (de y) ordenados, separados por espacios en blanco y de manera que ninguna columna contenga dos espacios en blanco

```
BR-ILLI-G
BROIL-ING
```


Alineamiento de dos secuencias

AGGCTATCACCTGACCTCCAGGCCGATGCCC
TAGCTATCACGACCGCGGTCGATTTGCCCGAC

-AGGCTATCACCTGACCTCCAGGCCGA--TGCCC---
TAG-CTATCAC--GACCGC--GGTCGATTTGCCCGAC

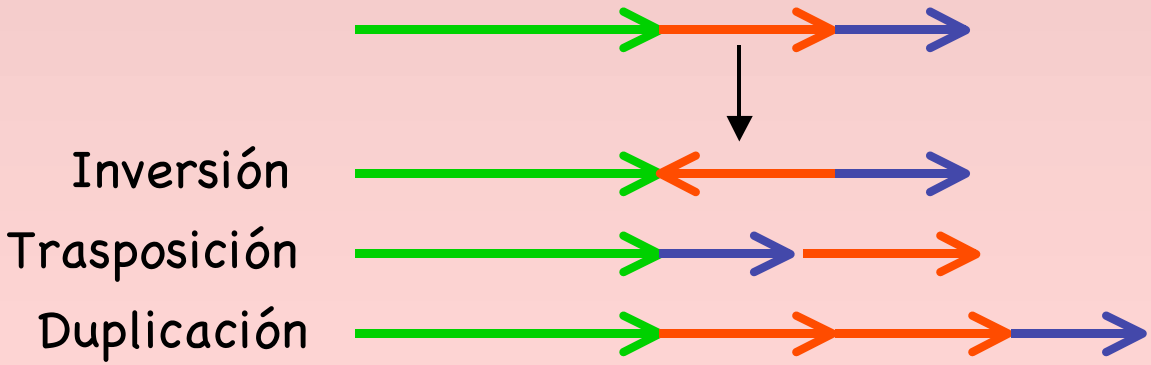
Alineamiento de dos secuencias

Problema: Encontrar el mejor alineamiento entre dos secuencias

Si se supone que dos secuencias provienen de un antepasado común por combinaciones de operaciones de edición, mejor tendrá que ver con menor coste de edición para transformar una en otra

Alineamiento de secuencias

- ¿Cuestan lo mismo todas las operaciones de edición?
- ¿Cuál es el significado de la semejanza detectada?
- ¿Qué quiere decir que dos genomas sean similares? Otras operaciones de edición:



La puntuación

- Se usan matrices de puntuación de mutaciones, y coste de inserción/borrado para evaluar la calidad de los alineamientos (se suman las puntuaciones de parejas y agujeros)
- Pueden usarse matrices diferentes en contextos diferentes
- Hay matrices de puntuación con significado biológico y estadístico
- No se conoce coste significativo para agujeros

Matrices BLOSUM (Henikoff[2] 1992)

- Se usa secuencias de proteínas alineadas de la base de datos BLOCKS
- Para cada par de aminoácidos A_i, A_j (i,j=1,...,20), se busca la frecuencia q_{ij} con que aparecen en el mismo lugar en diferentes secuencias
- Para cada aminoácido A_i, se busca la frecuencia p_i con que aparece
- Se normalizan los valores de q_{ij} y p_i (dividiendo por la suma de todos)

Matrices BLOSUM (Henikoff² 1992)

- Ahora tomamos las probabilidades esperadas de aparición: $e_{ii}=p_i^2$ $e_{ij}=2p_ip_j$
- Se toma $s_{ij}=\log_2(q_{ij}/e_{ij})$
- La matriz BLOSUM está formada por los valores $2s_{ij}$ redondeados
- Se calculan matrices BLOSUM-X a partir sólo de secuencias que tienen alineado más de un X%

Matrices BLOSUM (Henikoff² 1992)

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	**9**	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2
S	-1	**4**	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-3
T	-1	1	**4**	1	-1	1	0	1	0	0	0	-1	0	-1	-2	-2	-2	-2	-2	-3
P	-3	-1	1	**7**	-1	-2	-1	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4
A	0	1	-1	-1	**4**	0	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-2	-3
G	-3	0	1	-2	0	**6**	-2	-1	-2	-2	-2	-2	-2	-3	-4	-4	0	-3	-3	-2
N	-3	1	0	-2	-2	0	**6**	1	0	0	-1	0	0	-2	-3	-3	-3	-3	-2	-4
D	-3	0	1	-1	-2	-1	1	**6**	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4
E	-4	0	0	-1	-1	-2	0	2	**5**	2	0	0	1	-2	-3	-3	-3	-3	-2	-3
Q	-3	0	0	-1	-1	-2	0	0	2	**5**	0	1	1	0	-3	-2	-2	-3	-1	-2
H	-3	-1	0	-2	-2	-2	1	1	0	0	**8**	0	-1	-2	-3	-3	-2	-1	2	-2
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	**5**	2	-1	-3	-2	-3	-3	-2	-3
K	-3	0	0	-1	-1	-2	0	-1	1	1	-1	2	**5**	-1	-3	-2	-3	-3	-2	-3
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	**5**	1	2	-2	0	-1	-1
I	-1	-2	-2	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	**4**	2	1	0	-1	-3
L	-1	-2	-2	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	**4**	3	0	-1	-2
V	-1	-2	-2	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	**4**	-1	-1	-3
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	**6**	3	1
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	**7**	2
W	-2	-3	-3	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	**11**

Matriz BLOSUM62

¿Cómo se calcula un alineamiento?

- Fuerza bruta: coste $O(2^{N+M})$ impensable
- Programación dinámica: va ampliando alineamientos óptimos entre prefijos; coste $O(n^2)$, óptimo, pero impracticable en búsquedas en bases de datos
- Heurísticos: no buscan alineamiento óptimo, sino suficientemente bueno; se usan en bases de datos (BLAST, FASTA, ...)

Alineamiento local

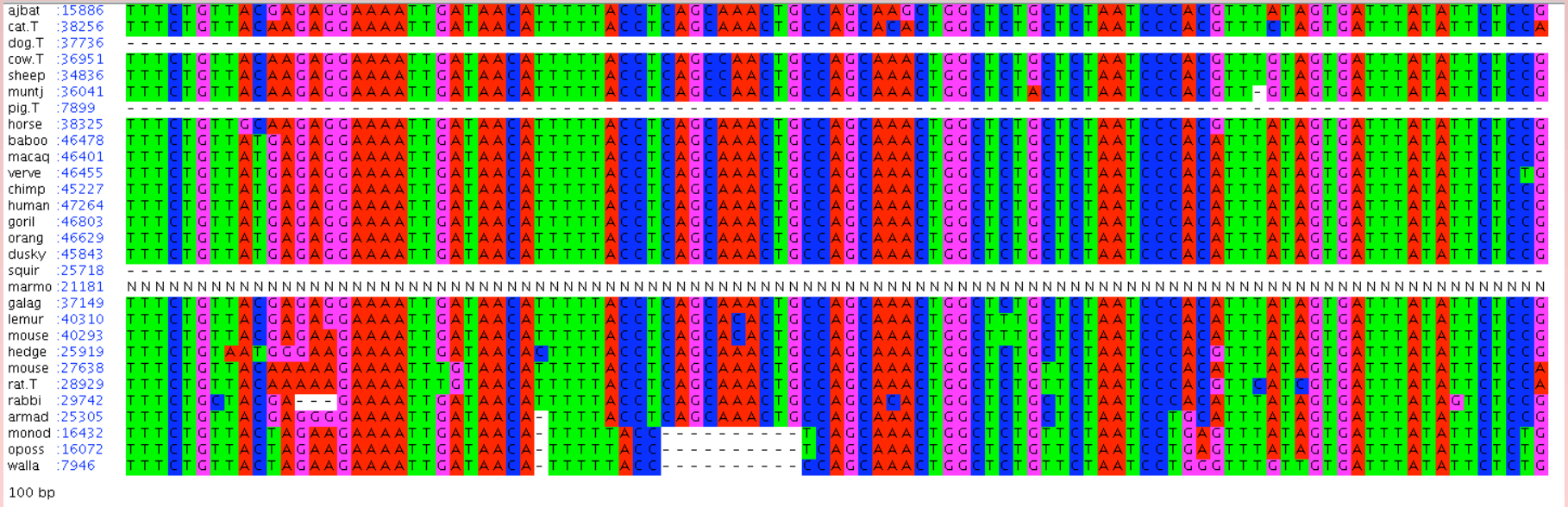
No se comparan dos palabras enteras, sino que se buscan la subpalabras que se parezcan más (coste de edición menor)

```
--T--CC-C-AGT--TATGT-CAGGGGACACG-A-GCATGCAGA-GAC
  |  || |  ||  | | | |||     || | | |  | ||||   |
AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG-T-CAGAT--C
```

```
                    TCCCAGTTATGTCAGGGGACACGAGCATGCAGAGAC
                       ||||||||||||
AATTGCCGCCGTCGTTTTCAGCAGTTATGTCAGATC
```


Alineamiento múltiple

- Utilidad del alineamiento de diversas secuencias:
  - Identificación de familias relacionadas de secuencias
  - Análisis filogenético



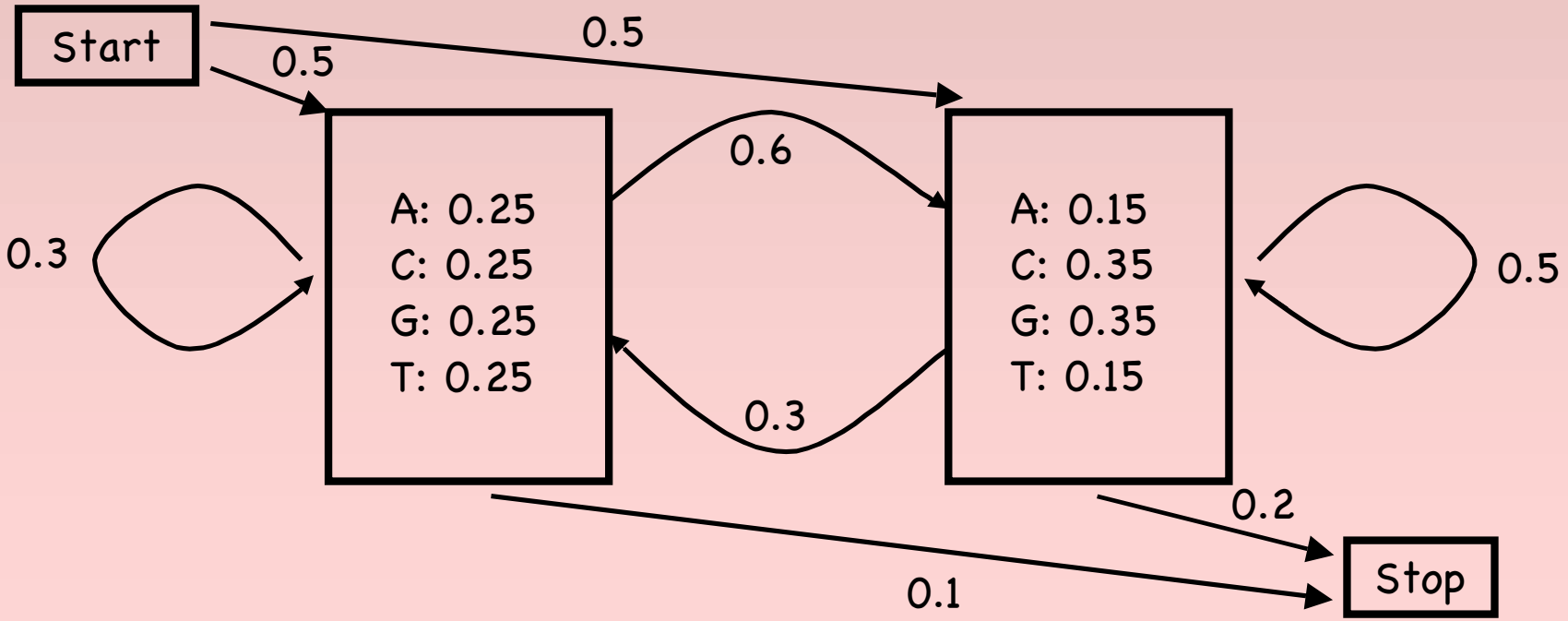
Alineamiento múltiple: métodos

- Programación dinámica: coste $O(n^M)$, impracticable para M grande
- Progresivos: añadiendo una cadena cada vez, coste $O(Mn^2)$, pero poco fiables
- Iterativos: a partir de alineamiento progresivo, ir mejorándolo hasta resultado aceptable
- Modelos de Markov Ocultos: muy populares, dan `alineamiento más probable según modelo'

Modelos de Markov Ocultos (HMM)

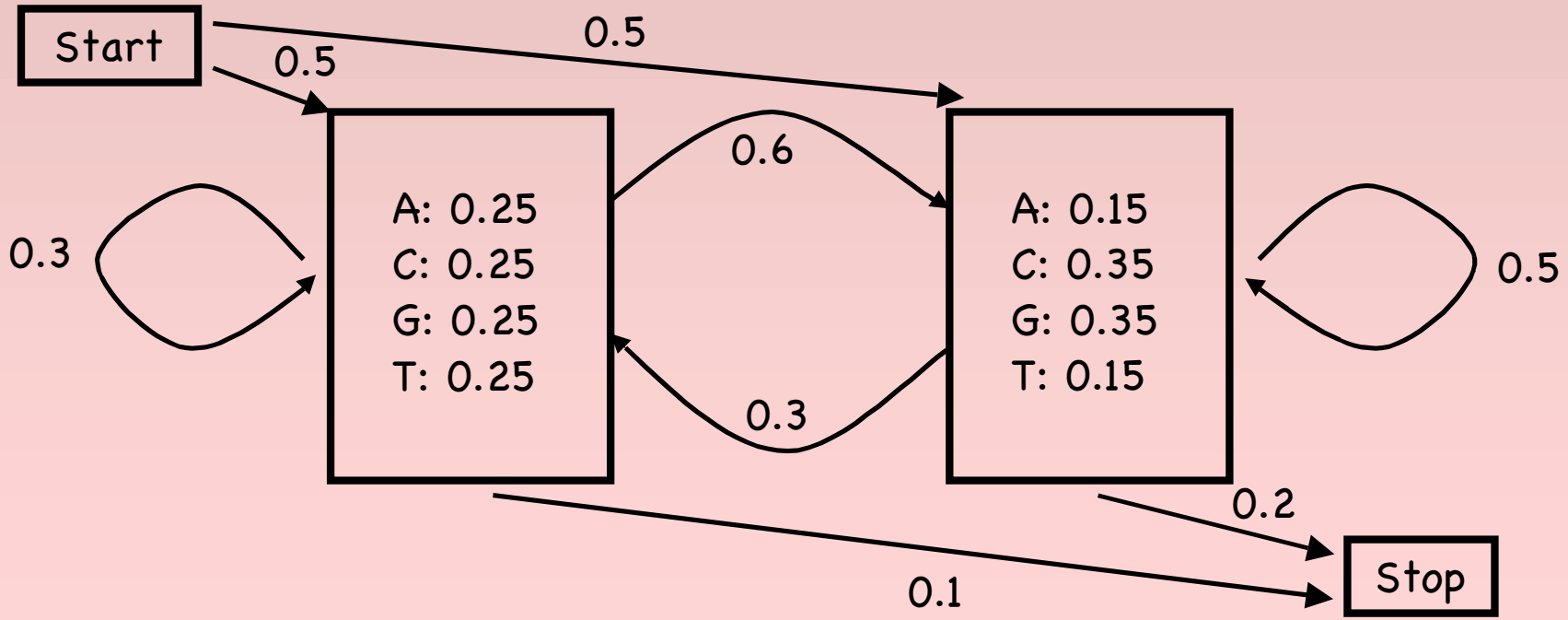
Un HMM es un autómata con probabilidades en las transiciones y donde los estados emiten símbolos con probabilidades

Emiten palabras con probabilidades



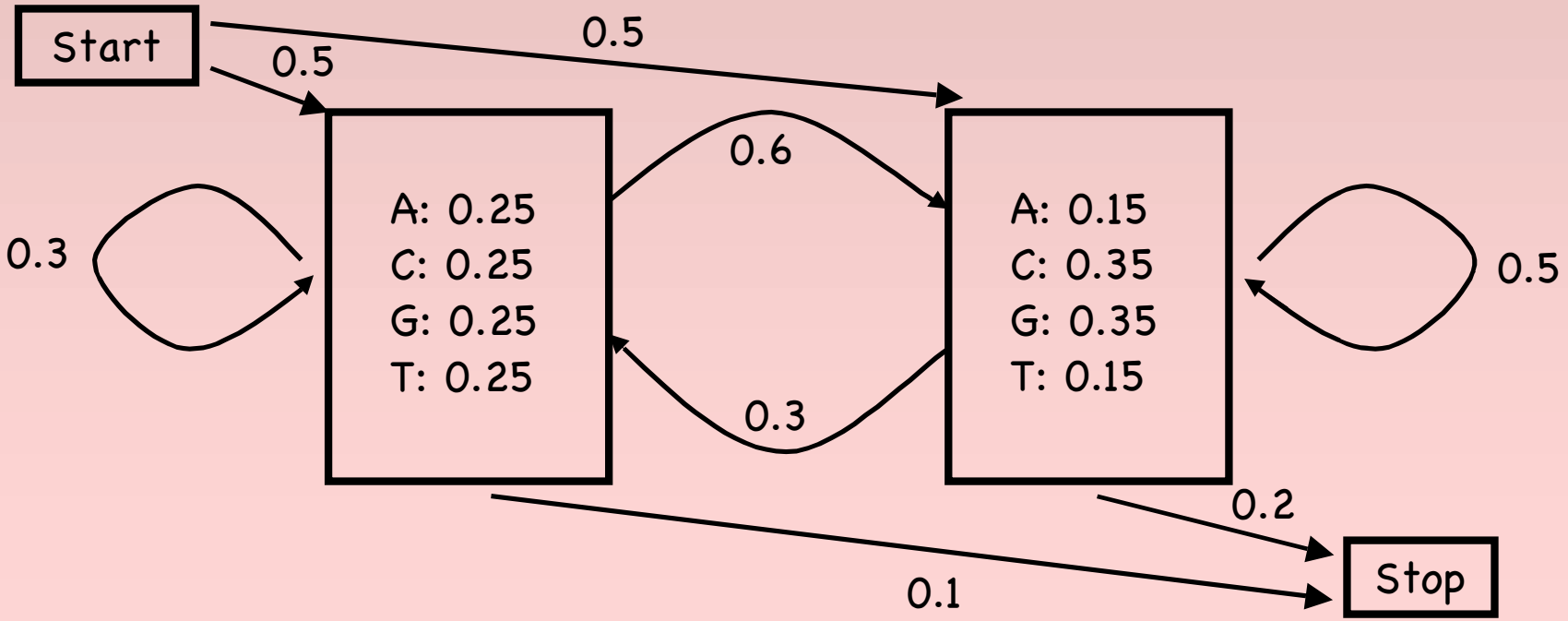
Modelos de Markov Ocultos (HMM)

Si se tienen palabras y los caminos de estados que las han producido, se pueden calcular las probabilidades más verosímiles (entrenar)



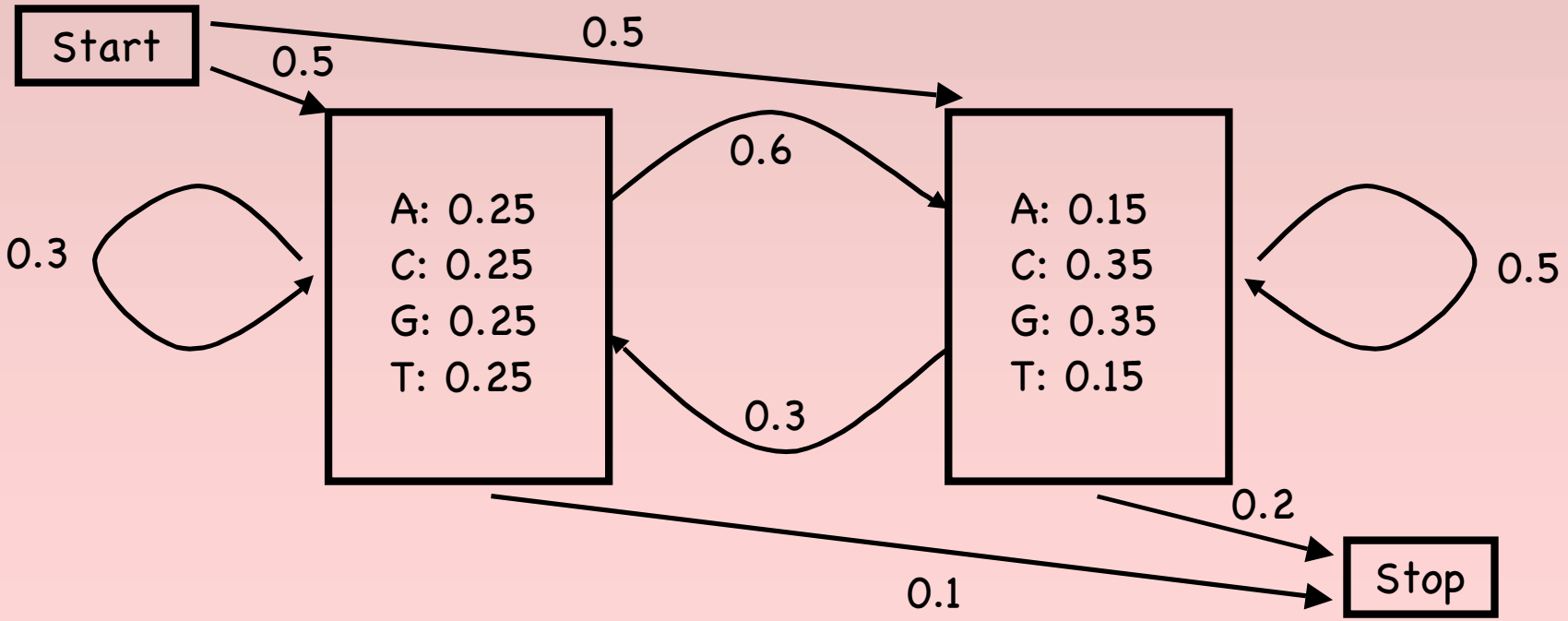
Modelos de Markov Ocultos (HMM)

Se puede entrenar un HMM a partir de un conjunto de secuencias sin conocer qué estados han emitido los símbolos



Modelos de Markov Ocultos (HMM)

Alineamiento de secuencias con HMM: se hacen corresponder ordenadamente las letras emitidas por los mismos estados

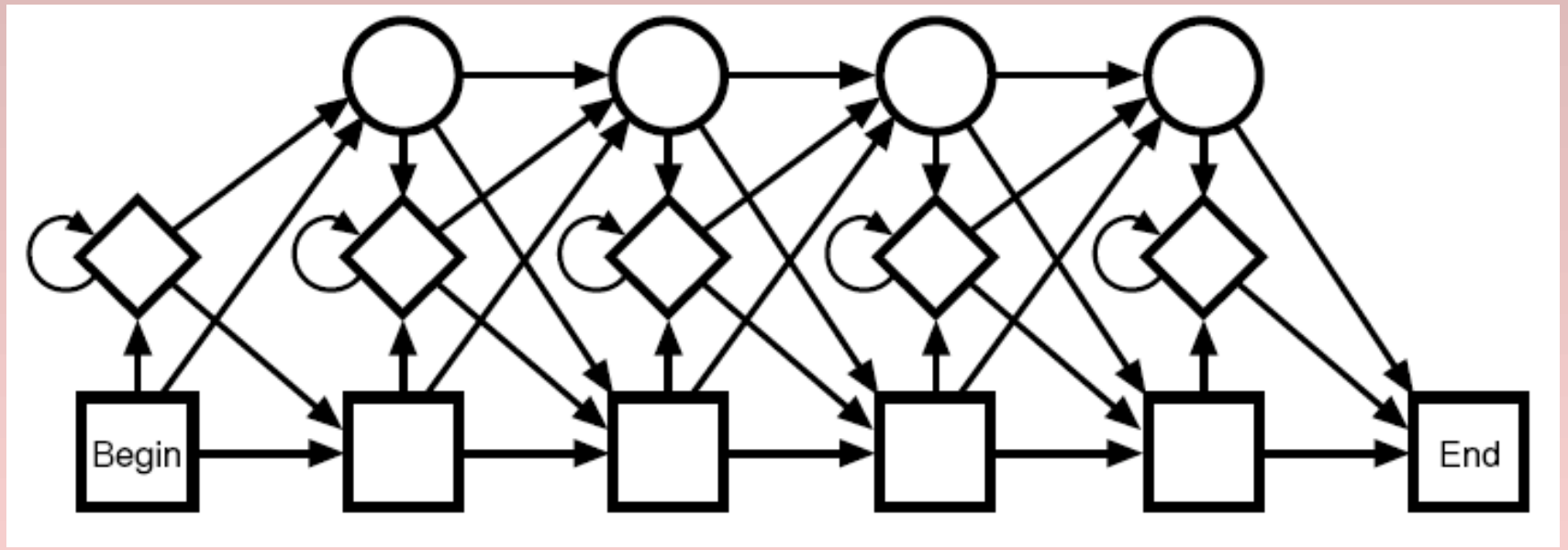


Modelos de Markov Ocultos (HMM)

Cómo alinear palabras con un HMM:

- Se diseña la estructura del HMM
- Se entrena con las palabras dadas
- Se alinean las palabras por los estados que las han emitido
- Además se tiene la probabilidad de que hayan sido producidas de esa manera: significado del alineamiento

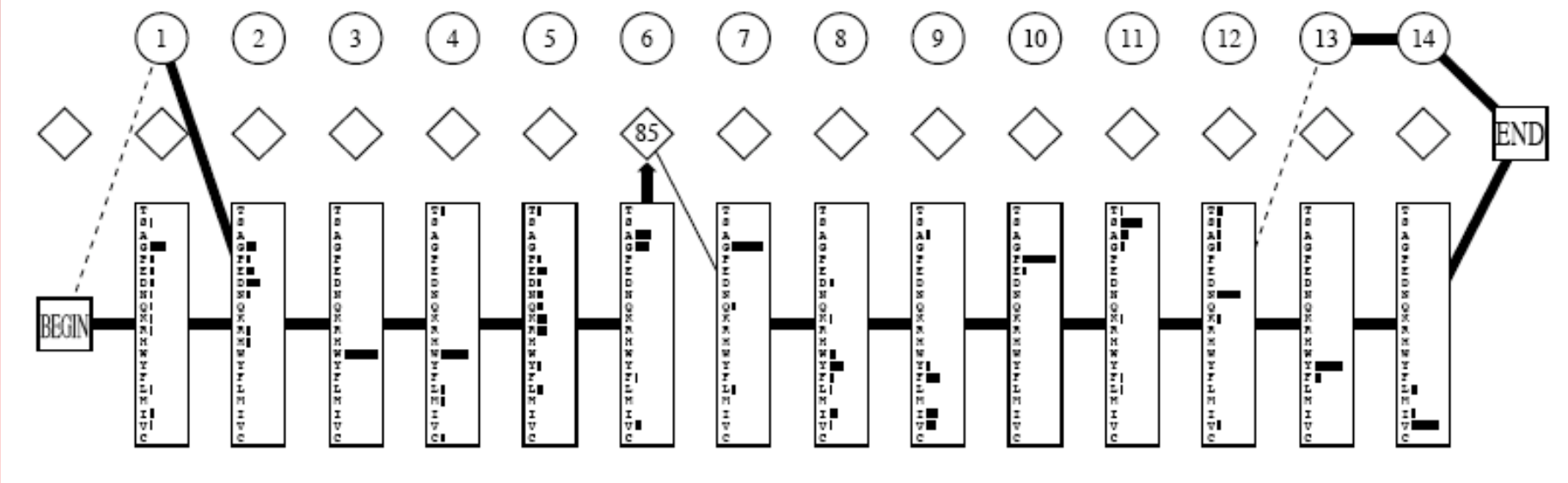
Modelos de Markov Ocultos (HMM)



HMM para alineamientos de ADN

Modelos de Markov Ocultos (HMM)

```
GGWWRGdy.ggkkqLWFPSNYV
IGWLNGynettgerGDFPGTYV
PNWWEGql..nnrrGIFPSNYV
DEWWQArr..deqiGIVPSK--
GEWWKAqs..tgqeGFIPFNFV
GDWWLArs..sgqtGYIPSNYV
GDWWDAel..kgrrGKVPSNYL
-DWWEArslssghrGYVPSNYV
GDWWYArslitnseGYIPSTYV
GEWWKArslatrkeGYIPSNYV
GDWWLArslvtgreGYVPSNFV
GEWWKAkslssKreGFIPSNYV
GEWCEAqt.knngq.GWVPSNYI
SDWWRVvnlttrqeGLIPLNFV
LPWWRArd.kngqeGYIPSNYI
RDWWEFrsktvytpGYYESGYV
EHWWKVkd.algnvGYIPSNYV
IHWWRVqd.rngheGYVPSSYL
KDWWKVev..ndrqGFVPAAYV
VGWMPGlner trqrGDFPGTYV
PDWWEGel..ngqrGVFPASYV
ENWWNGei..gnrkGIFPATYV
EEWLEGec..kgkvGIFPKVFV
GGWWKGdy.gtriqQYFPSNYV
DGWWRGsy..ngqvGWFPSNYV
QGWWRGel..ygrvGWFPANYV
GRWWKArr.angetGIIPSNYV
GGWTQGel.ksgqkGWAPTNYL
GDWWEArsn.tgenGYIPSNYV
NDWWTGrt..ngkeGIFPANYV
```



Secuenciado de genomas

Secuenciado de genomas

© Sidney Harris

Secuenciado de genomas

- Las secuenciadoras de ADN leen cadenas de 500-1000b
- El genoma humano tiene 3,300,000,000b
- Está claro que hacen falta ordenadores para ensamblar las cadenas obtenidas de un genoma

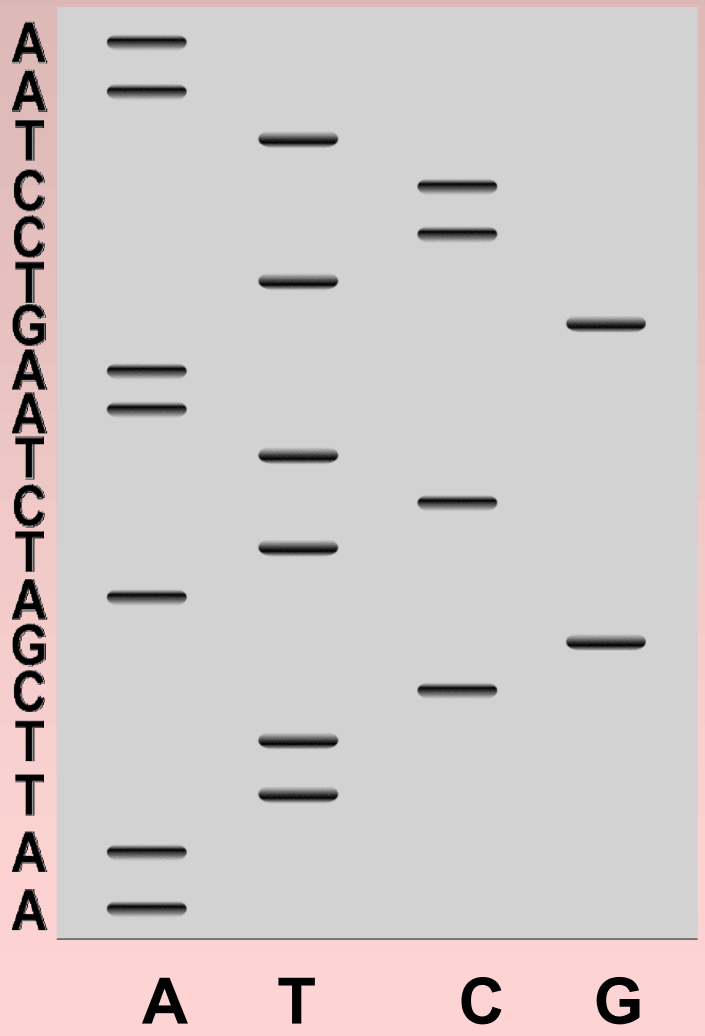
El secuenciado de ADN

- Dos métodos diferentes de secuenciado de cadenas cortas de ADN fueron inventados en 1977 por Fred Sanger y Walter Gilbert
- El método más popular hoy en día es el Sanger, y permite secuenciar cadenas de 500-1000b
- Los dos métodos se basan en romper la cadena de ADN en posiciones que marcan la aparición de una base concreta

El método Sanger

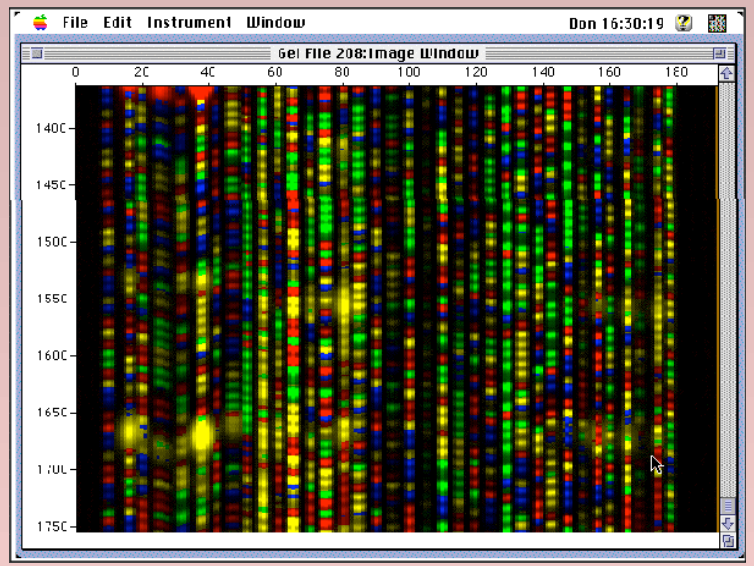
- Se mezclan muchas copias de la cadena a secuenciar, bases A, C, G, T, y versiones no funcionales y 'coloreadas' A*, C*, G*, T* que paran la replicación de ADN
- Con tiempo, la replicación se para en todas las posiciones
- Por la longitud de cada resultado y la base con la que acaba, se sabe qué base hay en cada posición

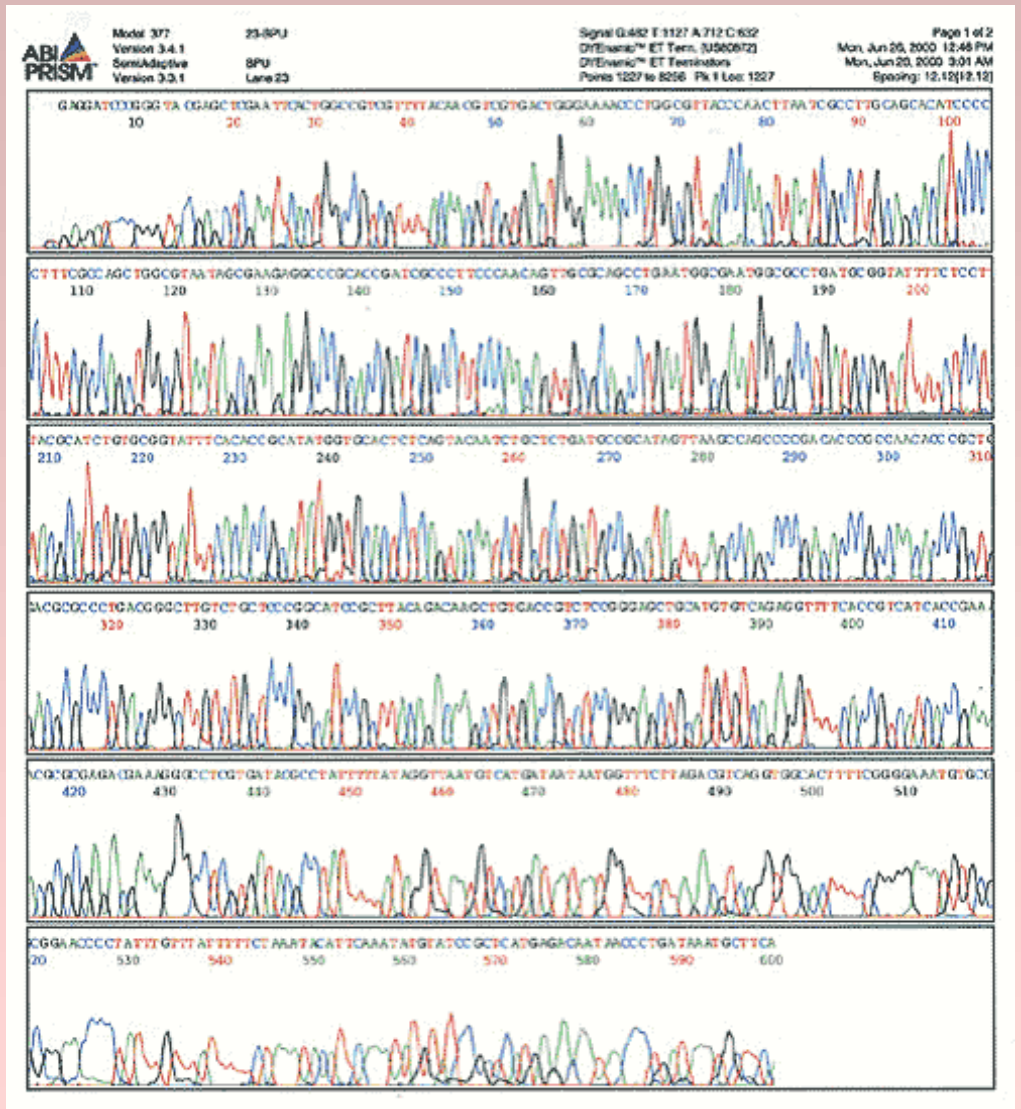
El método Sanger



AATTCGATCTAAGTCCTAA*
AATTCGATCTAAGTCCTA*
AATTCGATCTAAGTCCT*
AATTCGATCTAAGTCC*
AATTCGATCTAAGTC*
AATTCGATCTAAGT*
AATTCGATCTAAG*
AATTCGATCTAA*
AATTCGATCTA*
AATTCGATCT*
AATTCGATC*
AATTCGAT*
AATTCGA*
AATTC G*
AATTC*
AATT*
AAT*
AA*
A*

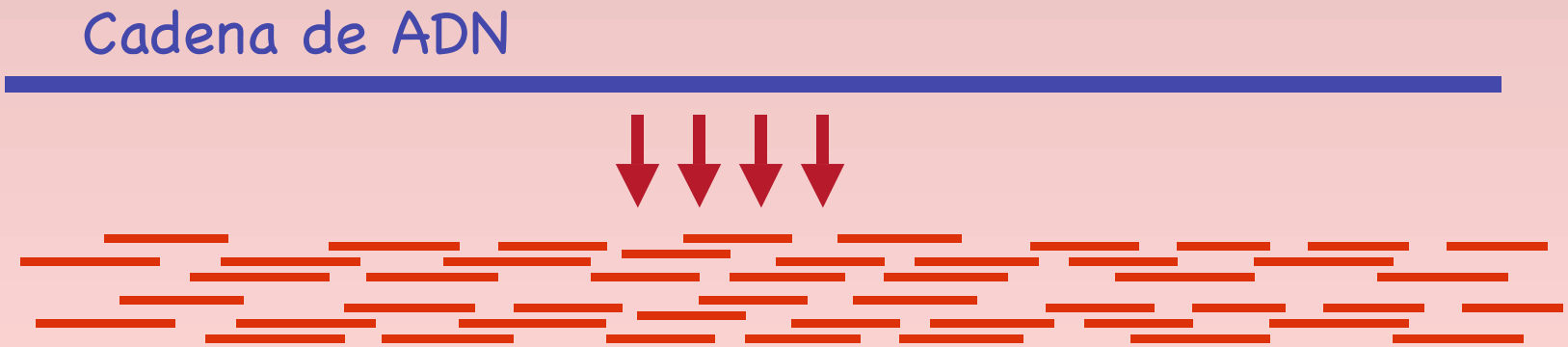
El método Sanger





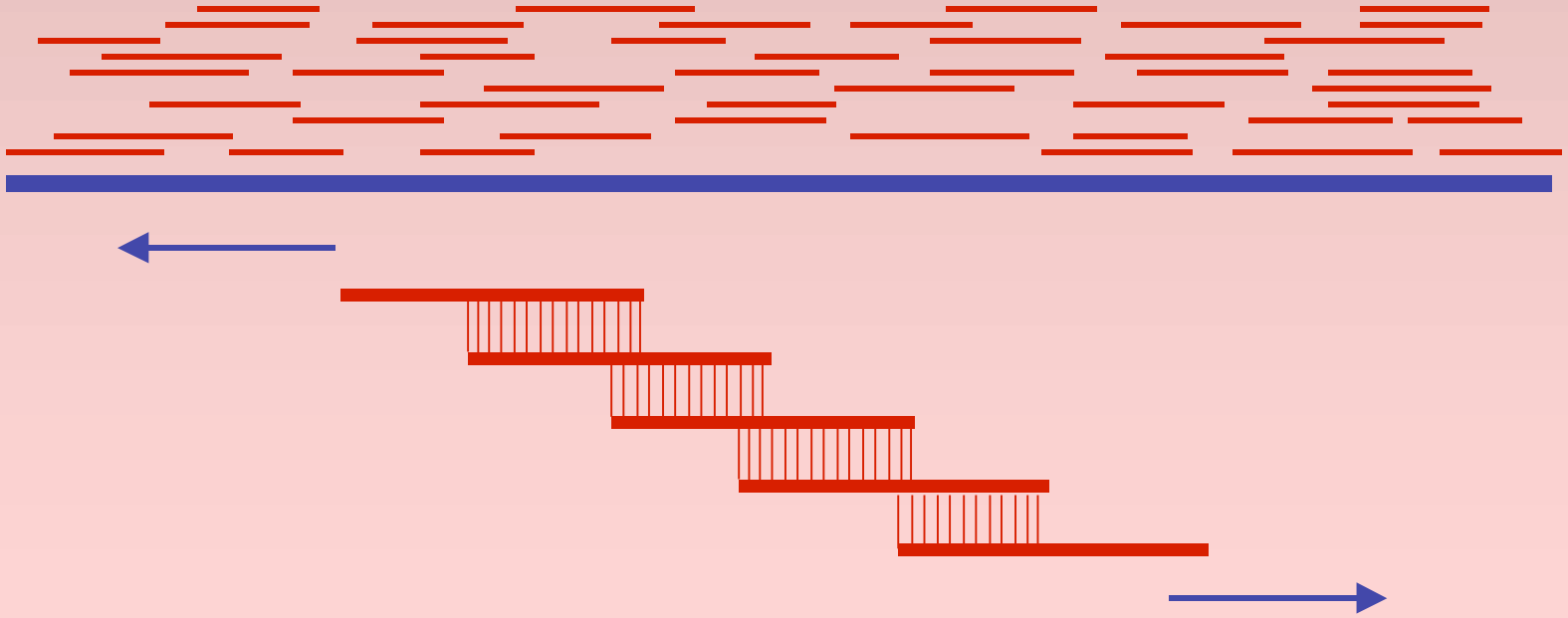
Secuenciado de cadenas largas

- Muchas copias de la cadena larga de ADN que queremos secuenciar se rompen al azar en fragmentos de longitud fija y secuenciable
- Estos fragmentos se secuencian



Secuenciado de cadenas largas

- Ensamblando estas subsecuencias, queremos reproducir la secuencia de bases de la cadena de partida



Secuenciado de cadenas largas

Una metáfora:

- Escribimos el Quijote en una sola línea
- Hacemos muchas copias
- Las cortamos al azar en trozos de 500 car.
- Mezclamos
- Cogemos un gran montón de estos fragmentos
- Queremos reconstruir el texto del Quijote

Secuenciado de cadenas largas

El secuenciado de ADN es aún más difícil:

- Errores en las copias
- Existencia de trozos que se repiten (hasta un 50% del genoma en mamíferos) difíciles de identificar y reconstruir

Secuenciado de genomas

- Extensión: se secuencia una cadena corta, se busca un fragmento que la continue, se secuencia este trozo y así sucesivamente...
- *Shotgun* (perdigonazo): se fragmenta la cadena en muchos fragmentos cortos, se hacen copias, y se ensamblan buscando solapamientos entre todos
- *Mapping* (mapeado): se usa *shotgun* sobre cadenas medianas, y se ensamblan sobre un andamio preconstruido sobre el genoma

Algo de historia

- 1982: F. Sanger usa *shotgun* para secuenciar el genoma del fago λ; 48Kb
- 1995: C. Venter usa *shotgun* para secuenciar el genoma de *H. influenziae*; 1.8×10^6 Kb (20,000 fragmentos)

  Usa *software* específico: TIGR Assembler

  Aparecen las primeras ofertas de trabajo para bioinformáticos

Algo de historia

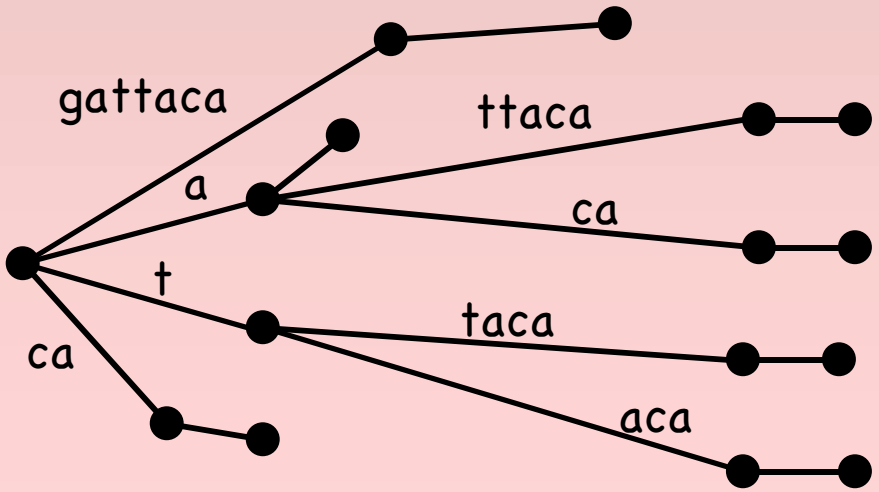
- 1990: DOE y NIH americanos ponen en marcha el Proyecto Genoma Humano (HGP); fecha objetivo, 2005; usarán *mapping* y la colaboración de muchísimos laboratorios
- 1998: C. Venter funda Celera Genomics y anuncia que secuenciará el genoma humano en tres años; usará *shotgun*
- 2001: HGP y Celera publican sus primeras versiones del genoma humano completo

Método *shotgun*

- El secuenciado *shotgun* del genoma humano por Celera usó ~28.4x10^6 fragmentos de ~550b
- La búsqueda eficiente de solapamientos entre estos fragmentos fue un punto clave de su éxito
- Uso de estructuras de datos y algoritmos adecuados

Arboles de sufijos

- El árbol de sufijos de una palabra captura todos los sufijos de ésta como caminos de la raíz a hojas
- Se almacena en espacio lineal, se puede construir en tiempo lineal



Arbol de sufijos de GATTACA

Arrays de sufijos

- Se ordenan alfabéticamente todos los sufijos de una palabra
- La búsqueda de subpalabras es una búsqueda binaria (toda palabra es prefijo de un sufijo), se hace en tiempo logarítmico

7: a
5: aca
2: attaca
6: ca
1: gattaca
4: taca
3: ttaca

Array de sufijos de GATTACA

Arrays de sufijos

- Se almacenan en espacio lineal: basta almacenar la palabra original y las posiciones de inicio de cada sufijo
- Se calculan en tiempo lineal a partir de árbol de sufijos

	gattaca
7: a	(1,7)
5: aca	(2,5)
2: attaca	(3,2)
6: ca	(4,6)
1: gattaca	(5,1)
4: taca	(6,4)
3: ttaca	(7,3)

Array de sufijos de GATTACA

Arrays de sufijos y secuenciado

- El uso astuto de arrays de sufijos permite construir el grafo de solapamientos entre fragmentos en tiempo casi lineal
- Si dos fragmentos se solapan, un prefijo de uno es igual a un sufijo del otro
- Se construye el array de todos los sufijos de todos los fragmentos y de sus inversos-complementarios: los fragmentos que se solapan estarán cerca
- Se buscan solapamientos suficientemente largos
- Problemas con errores de secuenciado

Nuevo secuenciado de genomas

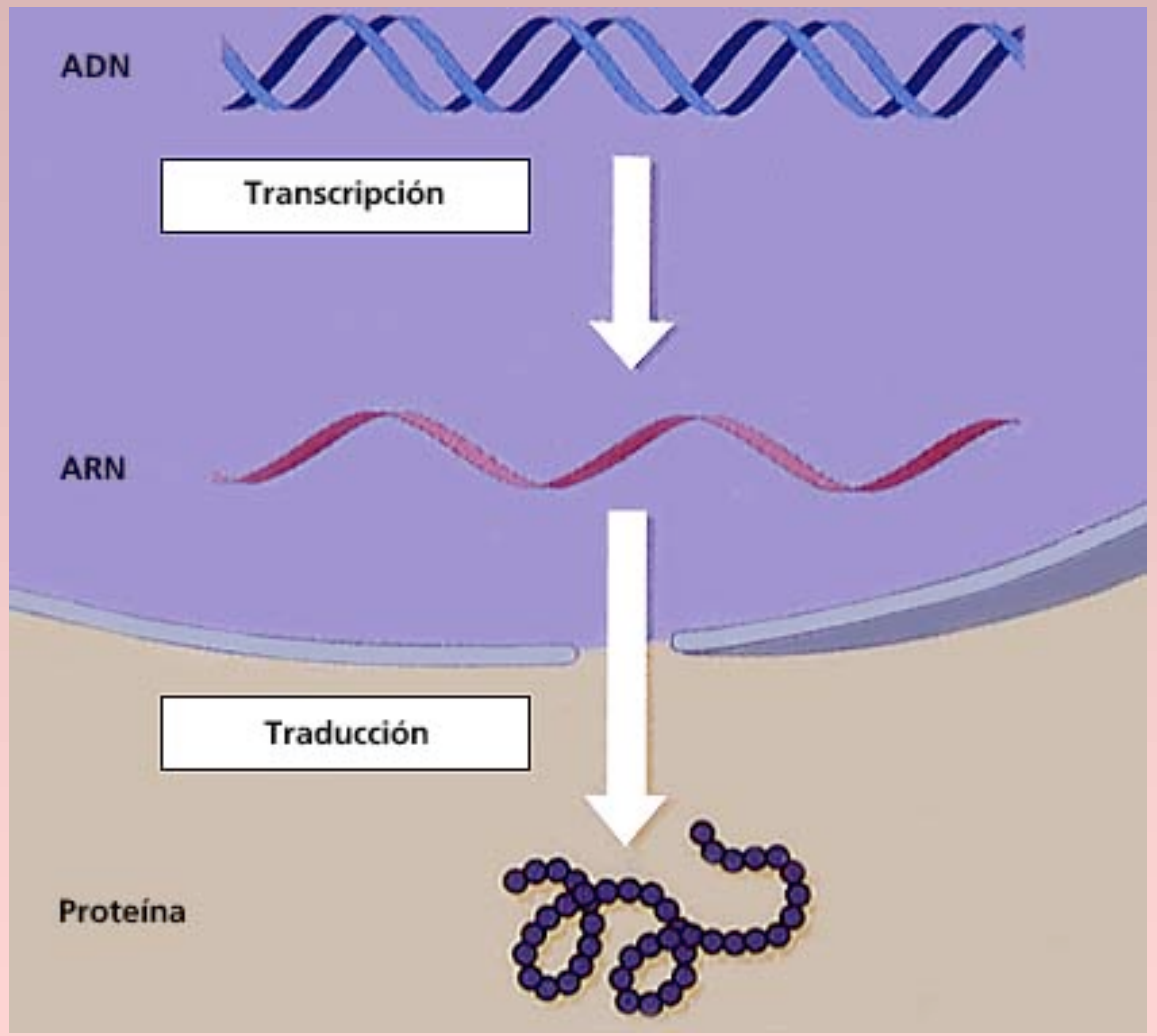
- Se han desarrollado nuevos métodos de secuenciación de cadenas muy cortas de ADN: se obtienen muy rápidamente millones de cadenas de 25b
- Estos nuevos métodos requieren nuevos algoritmos y programas de secuenciación de cadenas largas (mayor problema con las repeticiones, otras estructuras de datos)
- E.g: SSAKE
  http://www.bcgsc.ca/platform/bioinfo/software/ssake

Búsqueda de genes

- Los genes son las unidades físicas básicas de la herencia. Son secuencias de ADN específicas que codifican las instrucciones para producir proteínas
- No hay una caracterización precisa de las secuencias que son genes

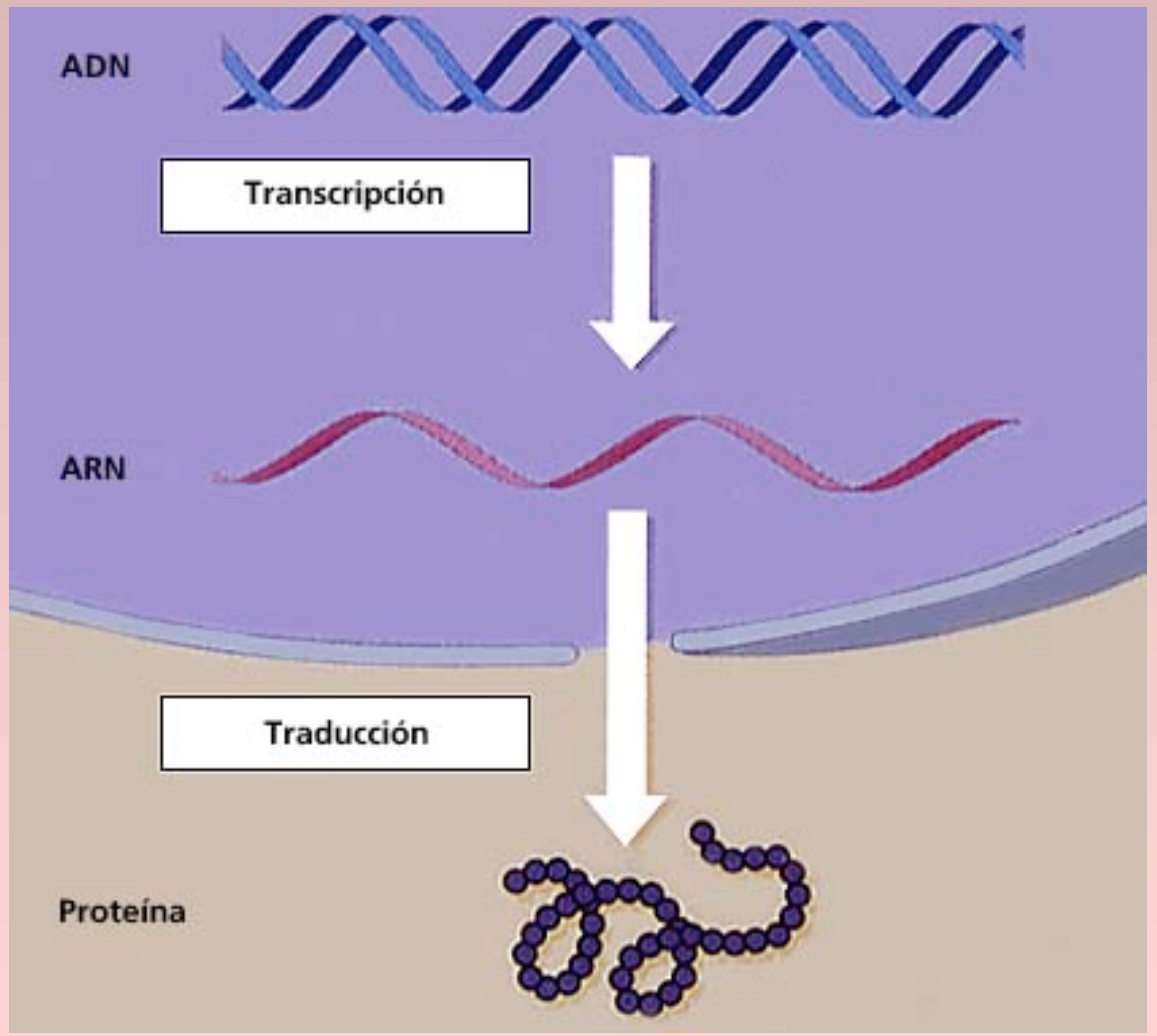
El dogma central v1.0

- Un gen se transcribe en ARN base a base por complementariedad:
  - A en U (Uracilo)
  - C en G
  - G en C
  - T en A



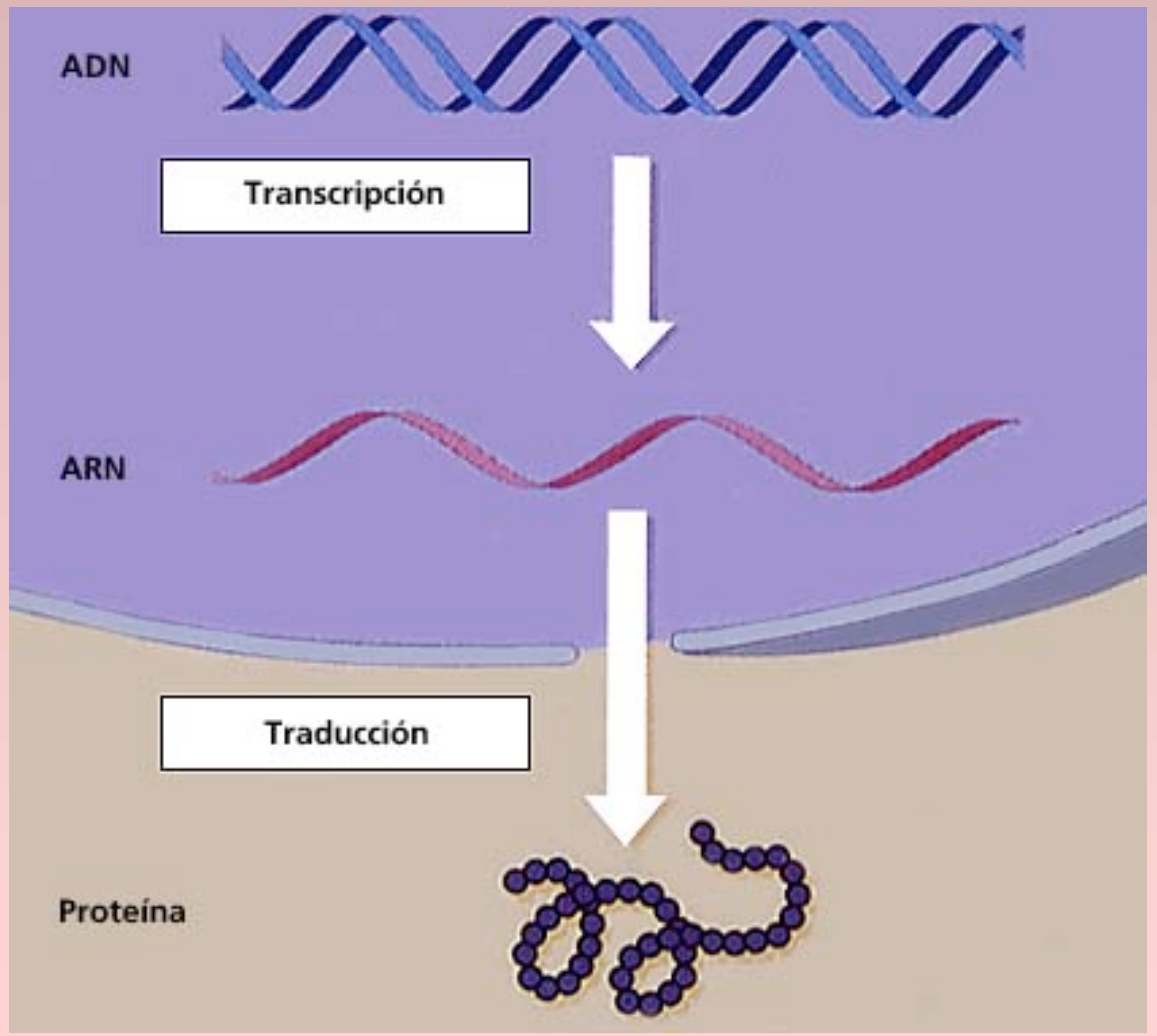
El dogma central v1.0

- Una molécula de ARN se traduce en proteína codón a codón
- Codón: tres bases consecutivas
- Cada codón codifica un aminoácido según el código genético universal

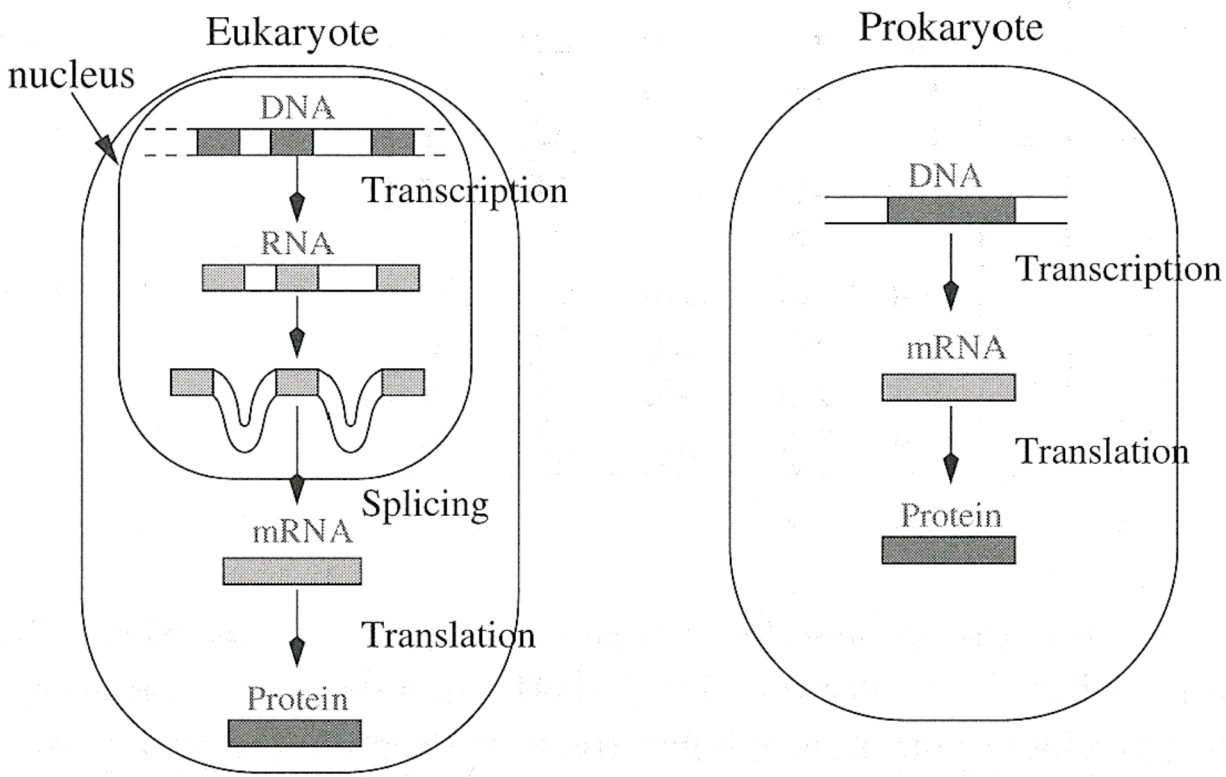


El dogma central v1.0

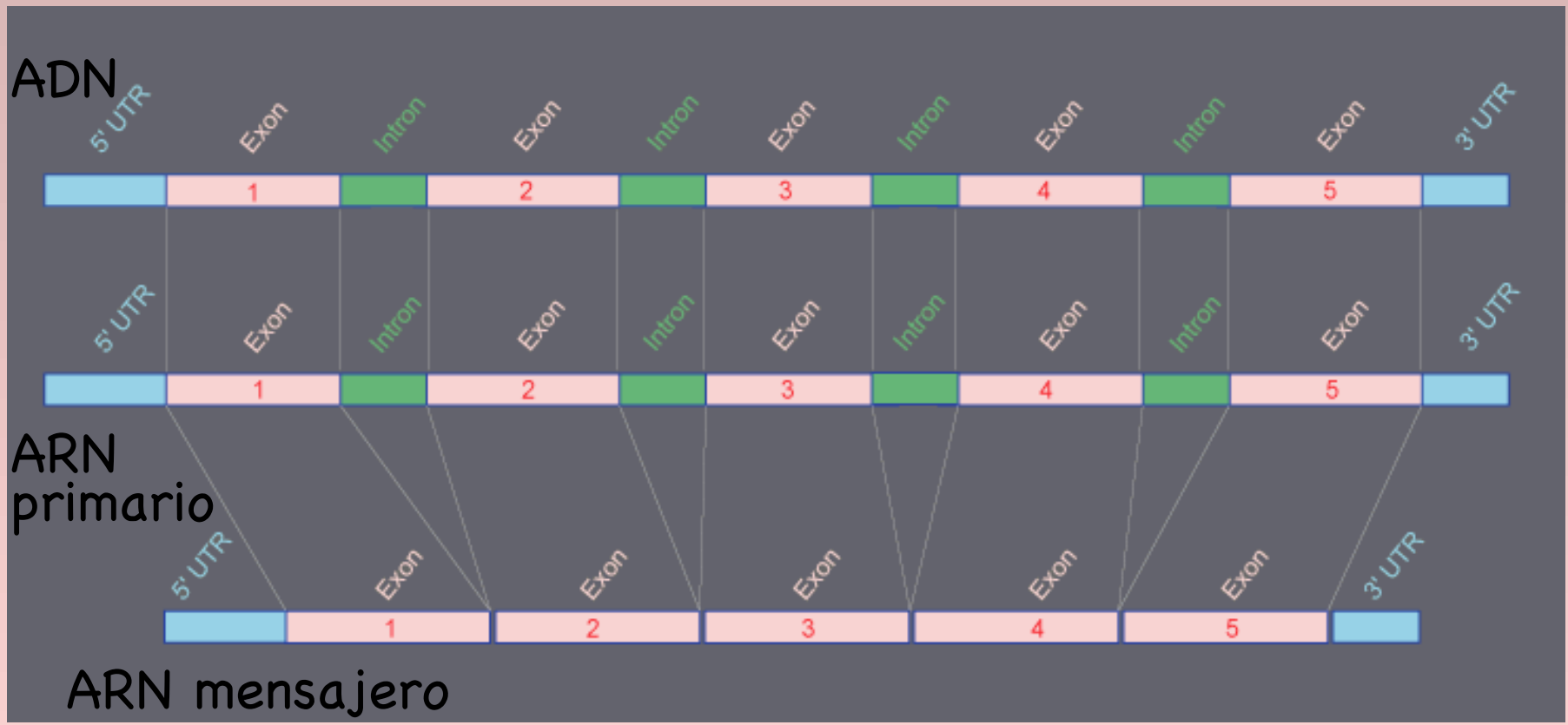
- La expresión de un gen es la producción de la proteína que codifica



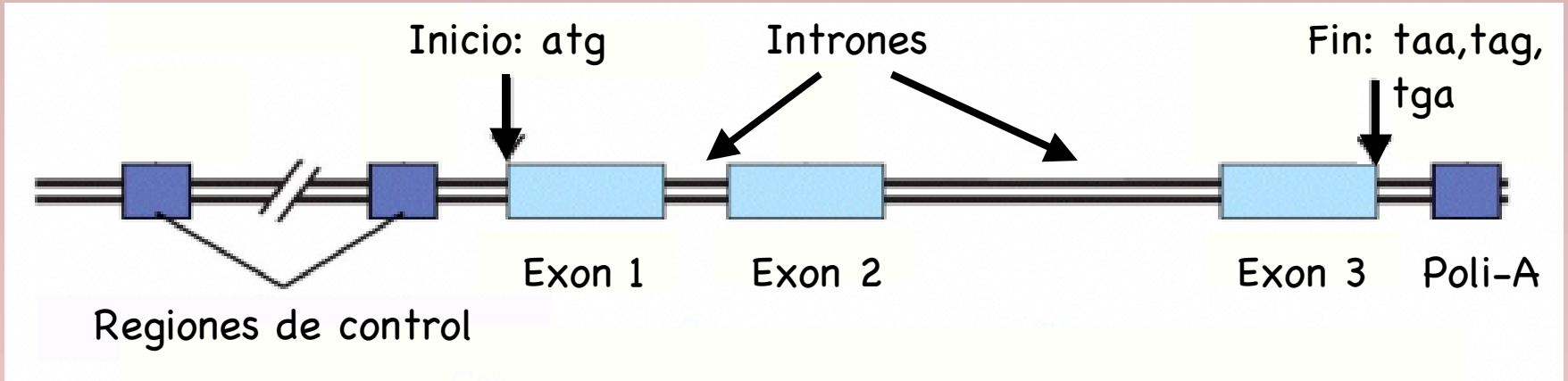
El dogma central v1.1



El dogma central v1.1



Estructura del gen eucariota



- Regiones de control: hasta a 50kb antes del inicio
- Exones: entre 1 y 178, de longitud 8b-17kb cada uno; código de proteína y UTR
- Intrones: alrededor de 1kb-50kb cada uno; estructura típica gt.....ag
- Los genes pueden solaparse, incluirse, ...

¿Cómo buscar genes?

- No hay una caracterización precisa de cuándo una secuencia de ADN es un gen, ni de qué trozos son exones
- Uno de los problemas más importantes en biología computacional es el de la Predicción de Genes: encontrar todos los genes contenidos en una secuencia de ADN

¿Cómo buscar genes?

Una metáfora:

En un libro escrito al azar, hay que encontrar un mensaje corto escrito en un idioma extranjero que no hablamos y roto en fragmentos dispersados por el folleto

¿Cómo buscar genes?

…tgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaat
gcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgct
aagctgggatccgatgacaatgcatgcggctatgctaatgaatggtcttgggattt
accttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaat
ggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctgggat
ccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatccg
atgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctgggatc
cgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatcctg
cggctatgctaatgaatggtcttgggatttaccttggaatgctaagctgggatccg
atgacaatgcatgcggctatgctaatgaatggtcttgggatttaccttggaatatg
ctaatgcatgcggctatgctaagctgggaatgcatgcggctatgctaagctgggat
ccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatccg
atgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctcatgcg
gctatgctaagctgggaatgcatgcggctatgctaagctgggatccgatgacaatg
catgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgcta
agctgcggctatgctaatgcatgcggctatgctaagctcggctatgctaatgaatg
gtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggct
atgctaatgaatggtcttgggatttaccttggaatatgctaatgcatgcggctatg
ctaagctgggaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcg
gctatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgt

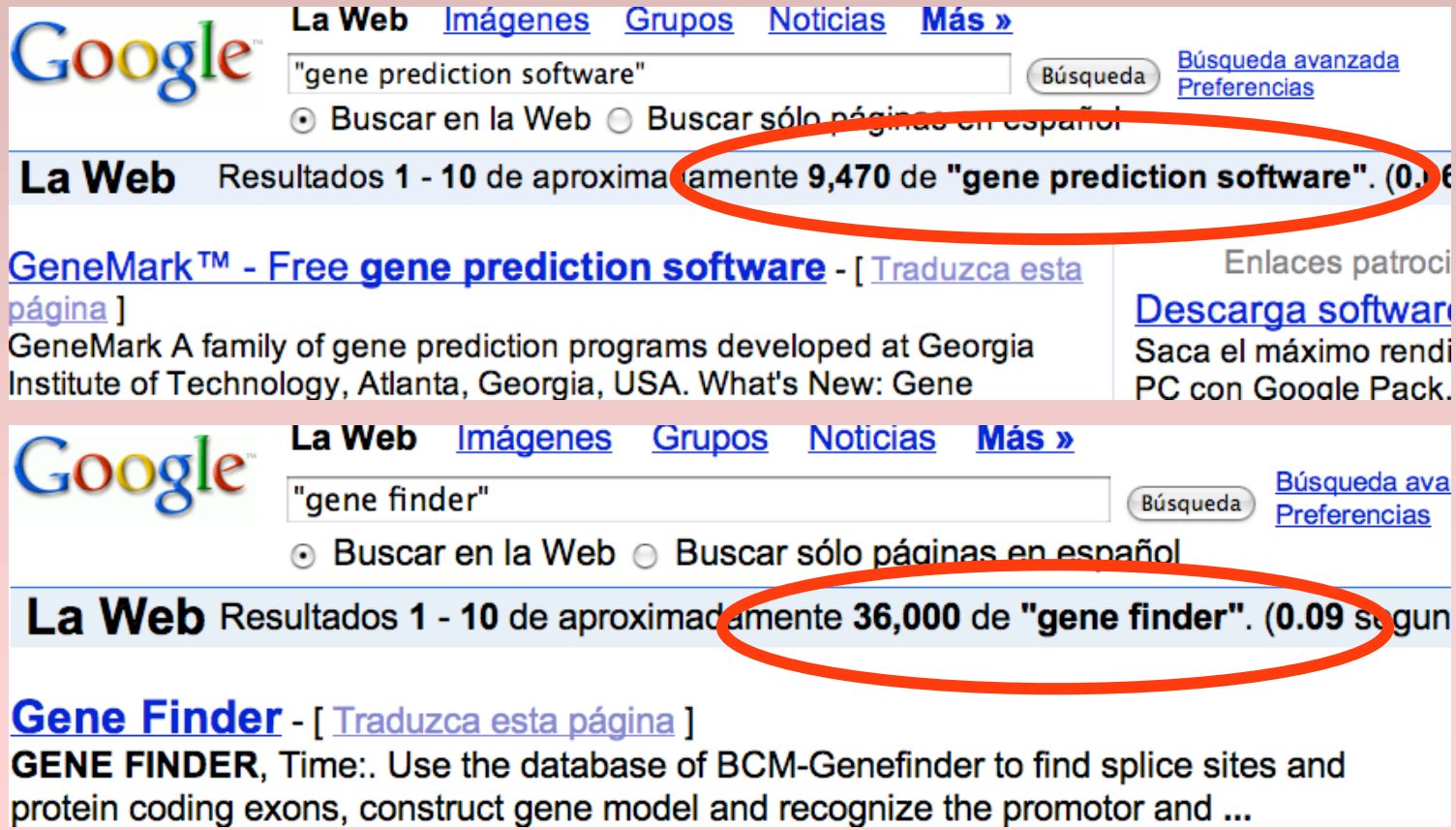
¿Cómo buscar genes?

…tgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaat
gcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgct
aagctgggatccgatgacaatgcatgcggctatgctaatgaatggtcttgggattt
accttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaat
ggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctgggat
ccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatccg
atgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctgggatc
cgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatcctg
cggctatgctaatgaatggtcttgggatttaccttggaatgctaagctgggatccg
atgacaatgcatgcggctatgctaatgaatggtcttgggatttaccttggaatatg
ctaatgcatgcggctatgctaagctgggaatgcatgcggctatgctaagctgggat
ccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatccg
atgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctcatgcg
gctatgctaagctgggaatgcatgcggctatgctaagctgggatccgatgacaatg
catgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgcta
agctgcggctatgctaatgcatgcggctatgctaagctcggctatgctaatgaatg
gtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggct
atgctaatgaatggtcttgggatttaccttggaatatgctaatgcatgcggctatg
ctaagctgggaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcg
gctatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgt

http://www.bioalgorithms.com

¿Cómo buscar genes?

Hay miles de programas, públicos o comerciales, generales o específicos, para buscar genes, ninguno es perfecto



(17/2/2007)

¿Cómo buscar genes?

Hay diversos métodos:

- Búsqueda de ORFs
- Métodos estadísticos
- Reconocimiento de pautas
- Comparación con proteínas
- Comparación con otros genes

Búsqueda de ORFs

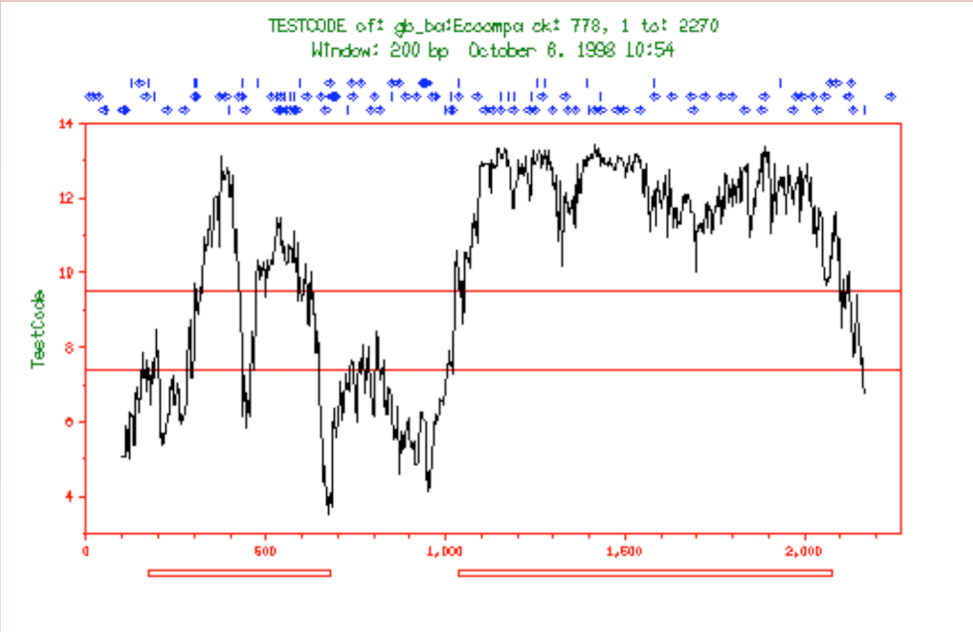
- ORF (Open Reading Frame): secuencia de ADN que empieza con codón de inicio y termina con codón de fin, sin otro codón de fin en medio
- Para buscar secuencias concretas se usan algoritmos de búsqueda eficientes
- Un ORF largo suele ser un gen
- Un ORF con frecuencias de codones adecuadas suele ser un gen
- Los buscadores de ORFs suelen ser útiles en procariotas (sin intrones-exones)

Métodos estadísticos

- Se buscan señales (inicios y finales de intrones, inicios y finales de genes, regiones de control...) mediante características estadísticas de su composición
- Se buscan secuencias con patrones estadísticos de subpalabras determinados

TestCode

- Se basa en propiedades estadísticas de las regiones que codifican
- Indica la probabilidad de que una zona de 200b sea codificadora o no



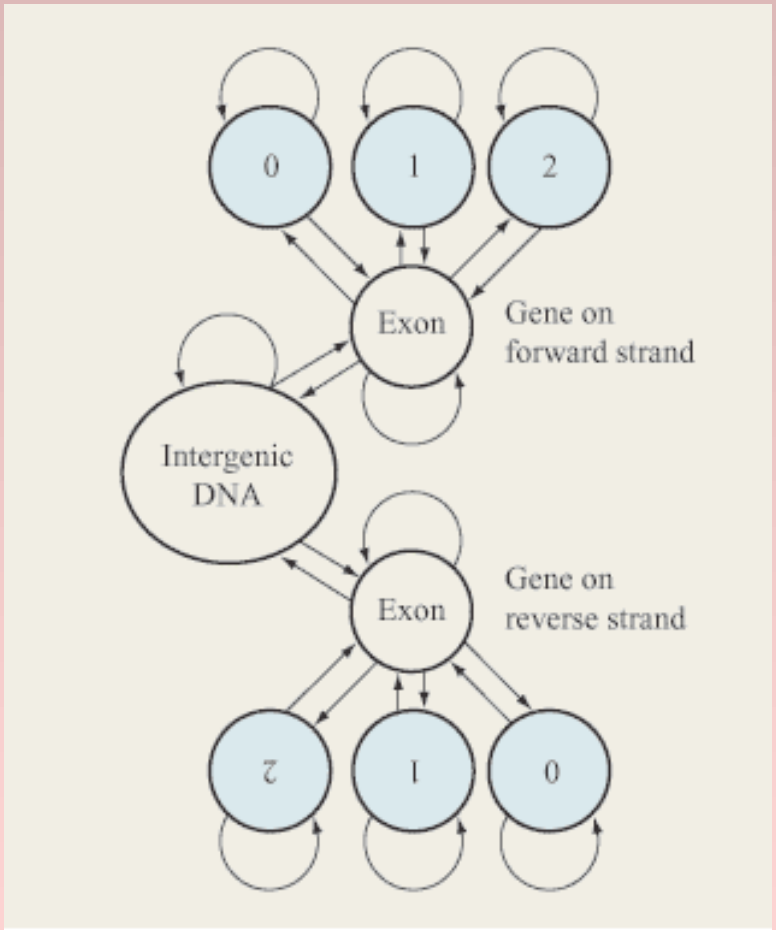
Zona codificadora

No sabe, no contesta

Zona no codificadora

Modelos de Markov Ocultos (HMM)

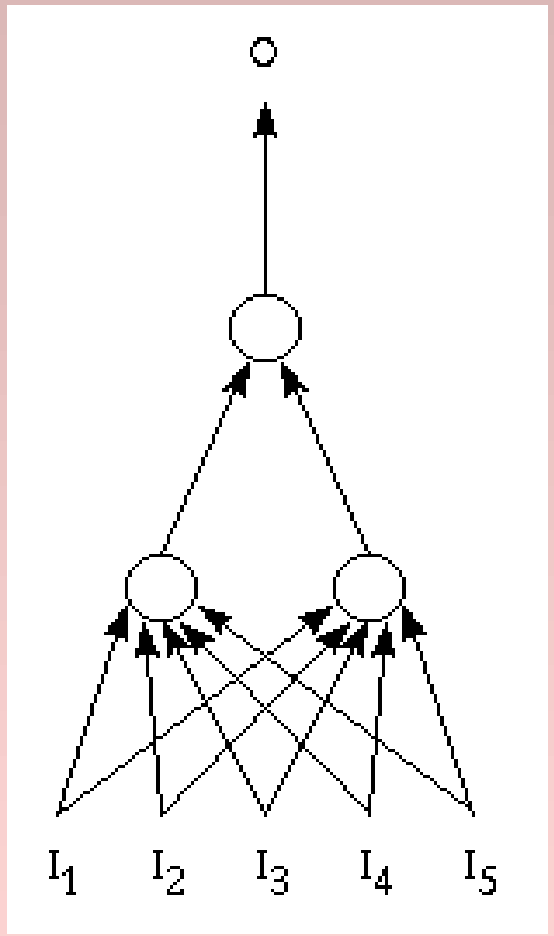
- Un HMM es un autómata finito con probabilidades en las transiciones y los estados emiten símbolos con probabilidades
- Se pueden entrenar incluso sin conocer la estructura de las secuencias de entrenamiento
- Genscan, Twinscan



http://www.research.ibm.com/journal/rd/453/birney.html

Redes neuronales

- Modelo del funcionamiento de las neuronas
- Los pesos de las conexiones entre neuronas varía y es `aprendido' por entrenamiento sobre pautas
- La red entrenada reconoce la pauta que ha aprendido en nuevas entradas

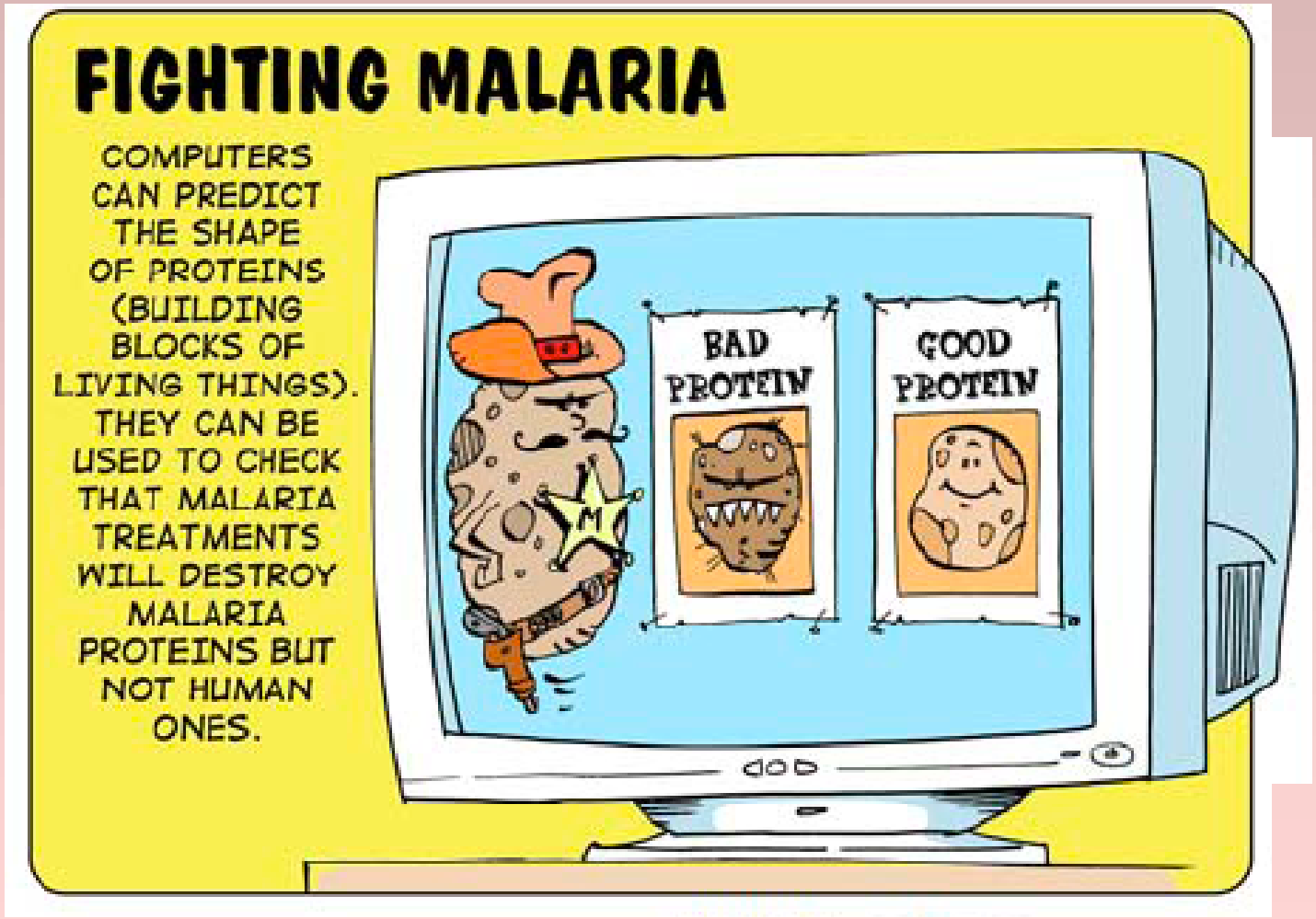


http://compbio.ornl.gov/grailexp/

Métodos de comparación

- Comparación con genes de otras especies (genes relacionados en diferentes organismos serán similares)
- Comparación con traducción ADN de proteínas (problema: más de un codón por aminoácido)
- Comparación con traducción ADN de ARNm
- En todos los casos se trata de búsqueda de conjuntos de alineamientos locales por medio de algoritmos específicos

Estructuras de proteínas



http://www.pub.ac.za/resources/teach.html

Estructuras de proteínas

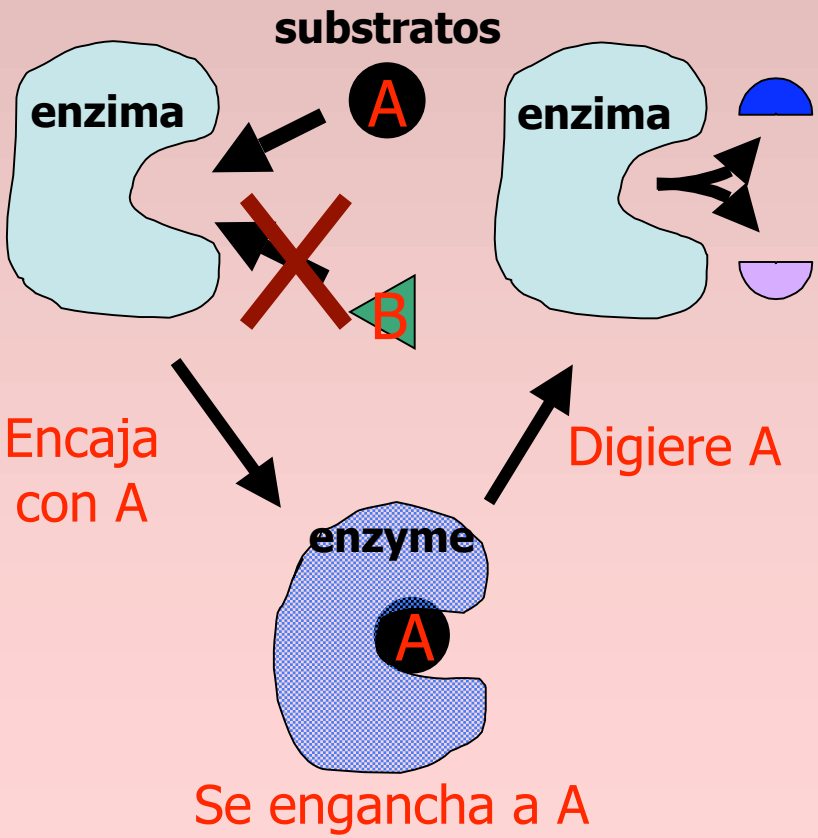
- Las proteínas son cadenas de aminoácidos
- Las proteínas se pliegan dentro de las células en complicadas estructuras 3D que determinan su función
- Esta estructura viene determinada por la secuencia de aminoácidos

Estructura de la proteína producida por el oncogen Ras

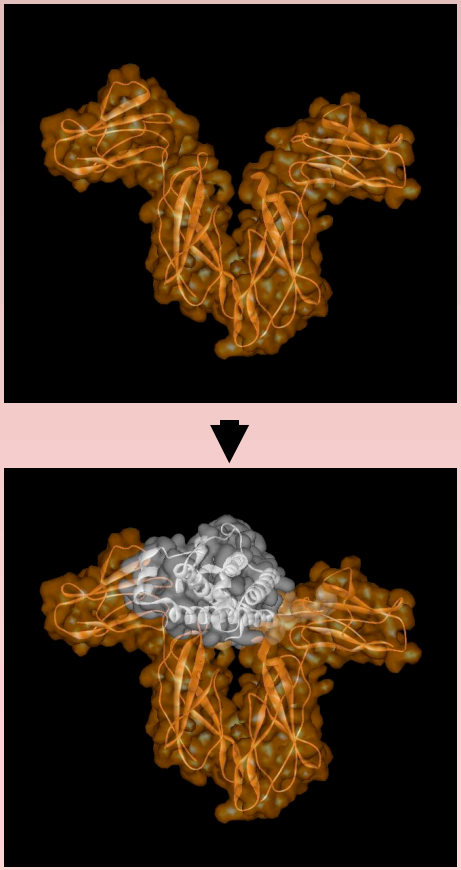
Estructura 3D de proteínas

- La estructura 3D determina la función de la proteína

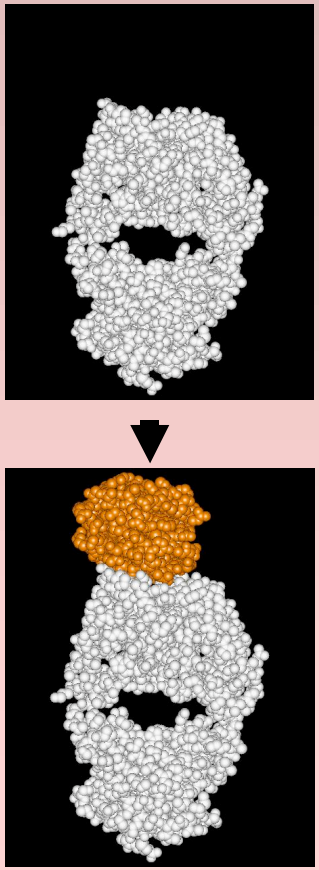
Ejemplo de reacción enzimática



Receptor hormonal



Anticuerpo



Jerarquía de las estructuras

Estructura primaria

(secuencia de aminoácidos)

↓

Estructura secundaria

(estructuras regulares: α-hélices, β-hojas,...)

↓

Estructura terciaria

(estructura 3D formada por ensamblado de estructuras secundarias)

↓

Estructura cuaternaria

(estructura formada por más de una cadena)

Determinación de la estructura 3D

La estructura 3D se puede medir directamente:

- Cristalografía de rayos X (XrC)

  La posición de los átomos se reproduce a partir de cómo difractan los rayos X

- Espectroscopia por Resonancia Magnética Nuclear (NMRS)

  La posición de los átomos se reproduce a partir de cómo absorben radiación electromagnética

- Microscopía crioelectrónica (CM)

Determinación de la estructura 3D

Todas estas técnicas tienen peros:

- No se revela la estructura completa (XrC)
- La proteína preparada para su estudio puede tener estructura diferente de la *in vivo* (XrC)
- Dificultad de preparación (NMRS)
- Baja resolución (CM)
- No sirve para todas las proteínas (todos)
- Caro y lento (todos)

Así y todo, se han determinado unas 15,000 estructuras (sólo 5000 realmente diferentes)

¿De qué depende la estructura?

En general, la estructura primaria de la proteína determina la estructura 3D... salvo excepciones:

- depende del ambiente
- hay proteínas que toman diferentes estructuras
- los priones cambian la estructura de otras proteínas

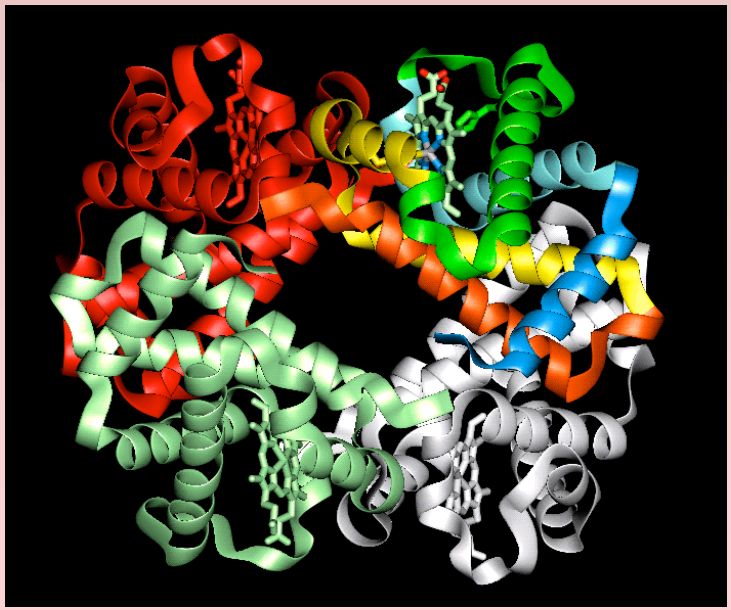
Predicción de la estructura 3D

Problema:

- Predecir (o deducir información sobre) la estructura 3D de una proteína a partir de la secuencia de aminoácidos

```
...KKTILAIAIPALFA
SAANAAVIYDKDGTTFD
VYGRVQANYYGDTNEAD
STAASGYKDVDGELKGS
SRLGWSGKIALNNTWSG
IAKTEWQVSAENSANKF
DSRHIYVGFDGTQY...
```

?



Predicción de la estructura 3D

Dos problemas:

- Predecir (o deducir información sobre) la estructura 3D de una proteína a partir de la secuencia de aminoácidos
- Predecir la función de la proteína a partir de (lo que sepamos de) su estructura 3D

Predicción de la estructura 3D

Tres problemas:

- Predecir (o deducir información sobre) la estructura 3D de una proteína a partir de la secuencia de aminoácidos
- Predecir la función de la proteína a partir de (lo que sepamos de) su estructura 3D
- Comparar dos estructuras 3D de proteínas

¿De qué depende la estructura?

El plegado de la proteína depende de:

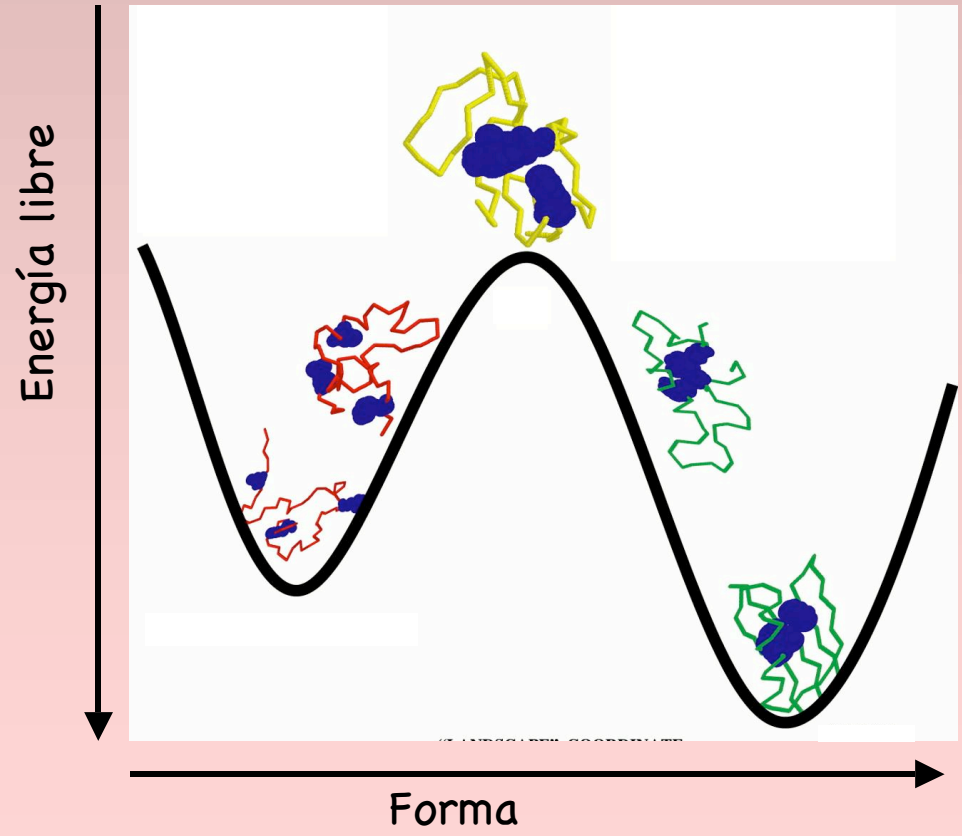
- La rigidez del esqueleto
- Interacciones entre aminoácidos:
  - puentes de hidrógeno
  - fuerzas de Van der Waals
  - atracciones electrostáticas
- Fuerzas externas
  - interacciones de aminoácidos con agua
  - interacción con componentes de membrana celular

Predicción de la estructura 3D

En teoría, la estructura 3D de una proteína se puede determinar computacionalmente:

Dogma: Una proteína toma una estructura que minimiza su energía potencial libre

(Aunque el proceso de plegado aún no se comprende del todo)



Predicción de la estructura: a lo bruto

El problema puede plantearse como un problema de optimización:

- Espacio de todos los plegados, se calcula la energía libre de cada uno (uhm...), se busca el plegado de energía mínima (más uhm...)
- Este espacio de búsqueda es enorme
- El número de mínimos locales crece exponencialmente con la longitud de la proteína

Computacionalmente impracticable (NP-completo)

Predicción de la estructura: *Threading*

Se busca un alineamiento 'optimo' entre la secuencia de la proteína y una estructura del PDB (Protein Data Bank)

(Se estima que hay un 60% de probabilidad que la estructura aproximada de la proteína ya exista en el PDB)

PROSPECT, 3D-PSSM, THREADER,...

Predicción de la estructura: Homología

Se busca proteína P' que sea muy similar (por ejemplo, por alineamiento o semejanza de composición) a la proteína P dada y que tenga estructura conocida

Si se ha dado un alineamiento, éste asignará las posiciones de P a la estructura de P'

Se usan herramientas de alineamiento de bases de datos de secuencias de proteínas

Predicción de la estructura: Otros problemas

- Predicción de las estructuras secundarias: determinar cada amonoácido a qué tipo de estructura secundaria pertenece
- Predicción de modelos simplificados: determinar sólo información sobre proximidad de aminoácidos o sobre proximidad de estructuras secundarias, y no la estructura geométrica 3D

Predicción de la estructura: CASP

Critical Assessment of Techniques for Protein Structure Prediction

Competición bianual de predicción de estructura 3D de proteínas a partir de las secuencias

CASP7 (2006): más de 300 participantes

http://predictioncenter.gc.ucdavis.edu/

Comparación de estructuras

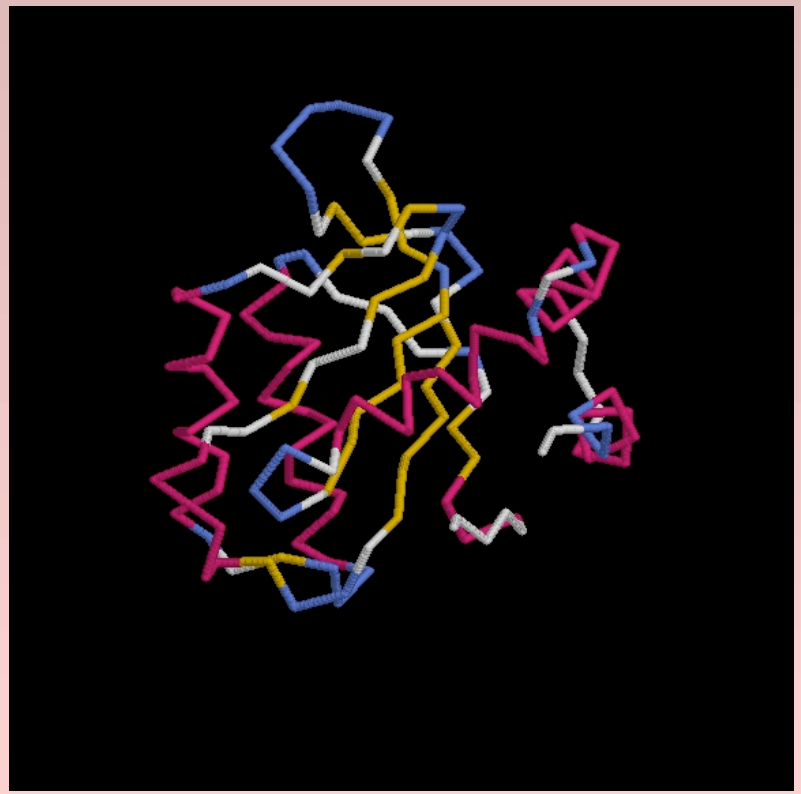
El objetivo es alinear dos estructuras 3D de proteínas

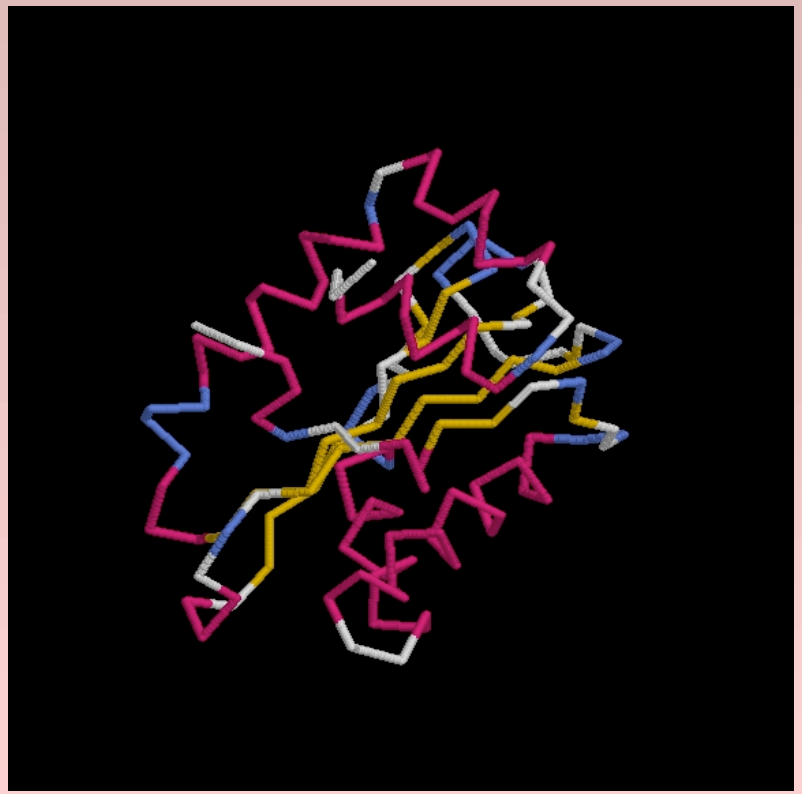
A diferentes niveles de detalle:

- átomos
- carbonos α
- aminoácidos
- estructuras secundarias

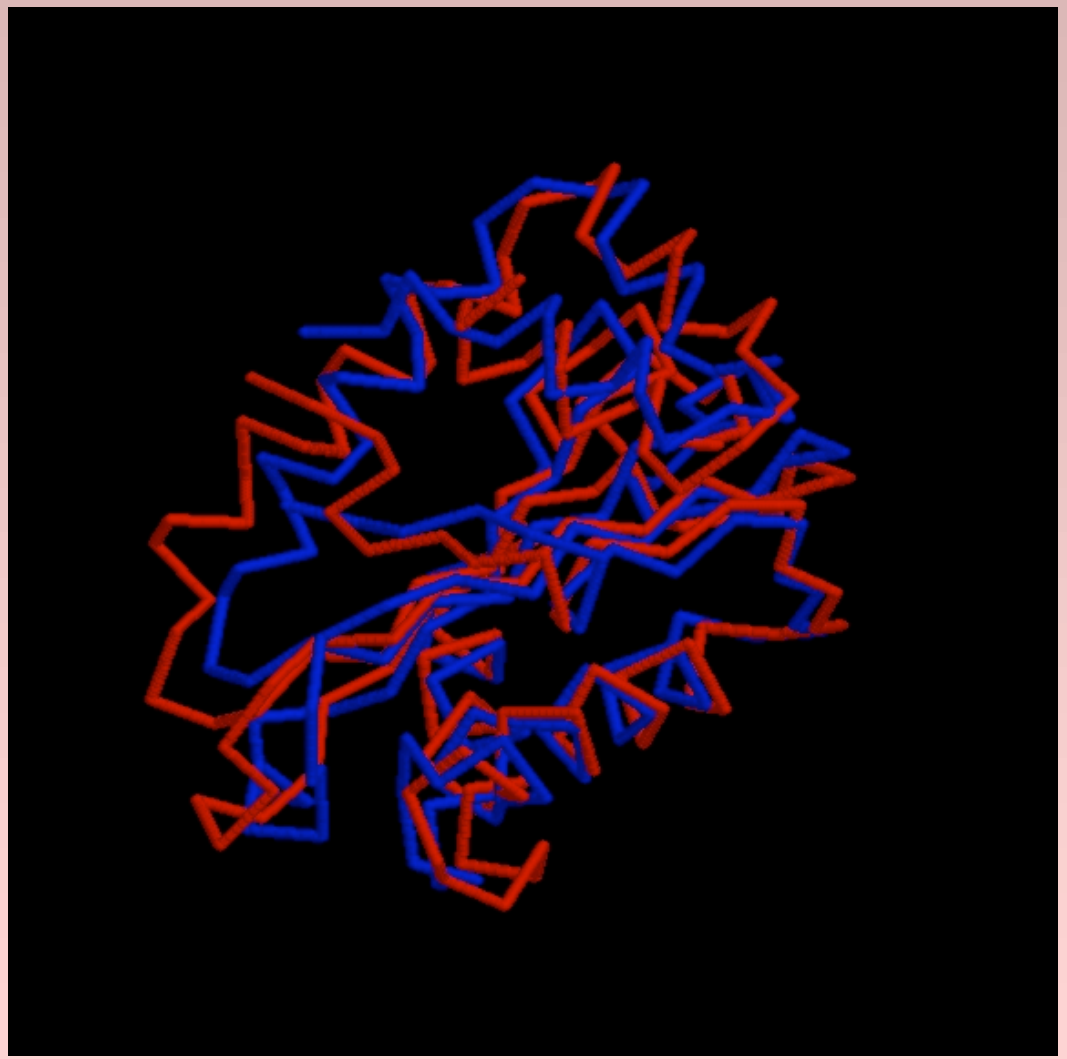
Global (estructuras enteras) o local (buscar trozos similares)

Comparación de estructuras





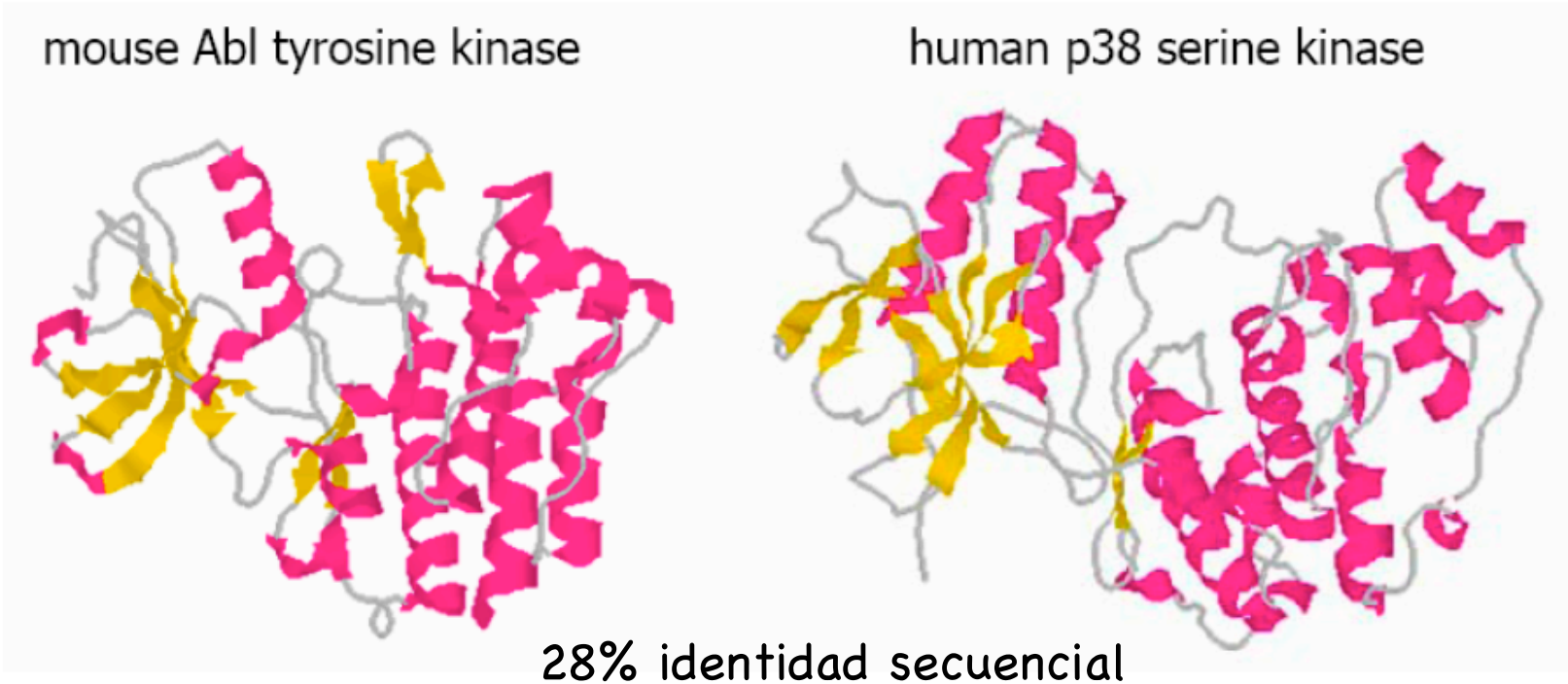
Comparación de estructuras



Comparación de estructuras: ¿Por qué?

- Identificar parientes lejanos de proteínas (con poca semejanza secuencial)
- Predecir función de proteínas
- Diseñar proteínas con funciones específicas
- Evaluar métodos de predicción de estructuras
- Agrupar proteínas en familias estructurales

Comparación de estructuras: ¿Por qué?



La semejanza estructural se conserva evolutivamente más que la secuencial, y no implica ésta

Comparación de estructuras

- Programación dinámica SSAP
- Comparación de las matrices de distancias DALI, MATRAS
- Heurística STRUCTAL
- Detección de subgrafo común maximal y expansión del alineamiento local
- Búsqueda de movimiento rígido que minimiza la RMSD (raíz de la media de las distancias al cuadrado)

Comparación de estructuras: Problemas

- No existe ningún algoritmo que detecte las semejanzas que detectan expertos humanos
- El coste computacional es muy elevado en métodos no probabilísticos
- No está aún muy claro el significado estadístico de la semejanza que dan algunos programas
- De hecho, no existe una noción clara de semejanza de estructuras de proteínas

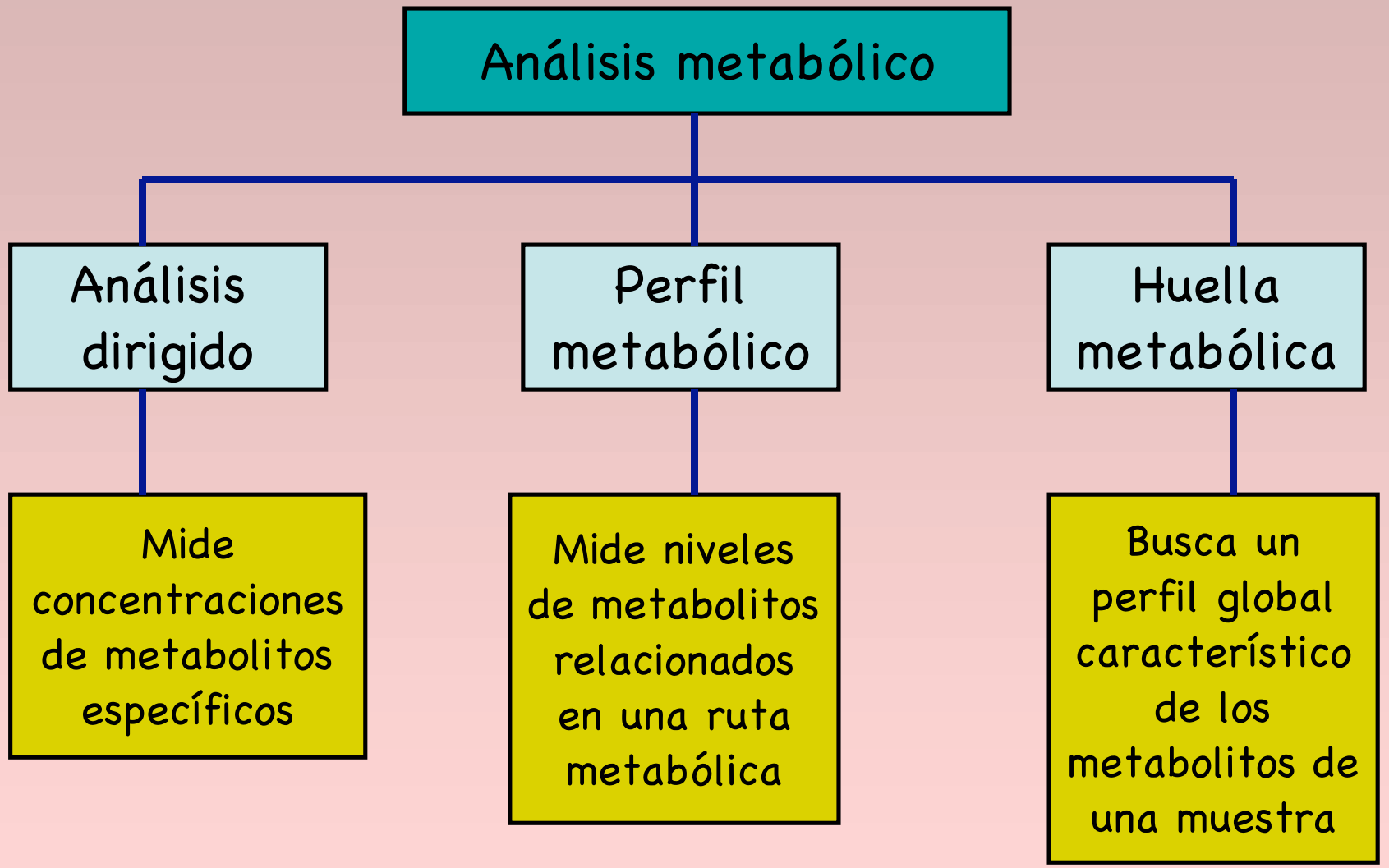
Metabolómica

More than1000 metabolites(代謝物質)
in 30 minutes, 45 samples a day

Metabolómica

- La metabolómica es el análisis global de todos o muchos metabolitos celulares
- Los metabolitos son las 'otras' moléculas involucradas en reacciones bioquímicas
- El metaboloma (conjunto de todos los metabolitos) humano ha sido completado en enero 2007: 2500 metabolitos
- Es el último gran reto de la bioinformática

Obtención de datos metabólicos



Técnicas de obtención de datos

- Resonancia magnética nuclear (NMR)
- Espectrometría de masas (MS)
- Cromatografía de gases+Espectrometría de masas (GC-MS)
- Cromatografía líquida+Espectrometría de masas (LC-MS)
- Electroforesis por capilaridad+Espectrometría de masas (CE-MS)
- TLC, HPLC, PDA, FT-IR,...

GC-MS

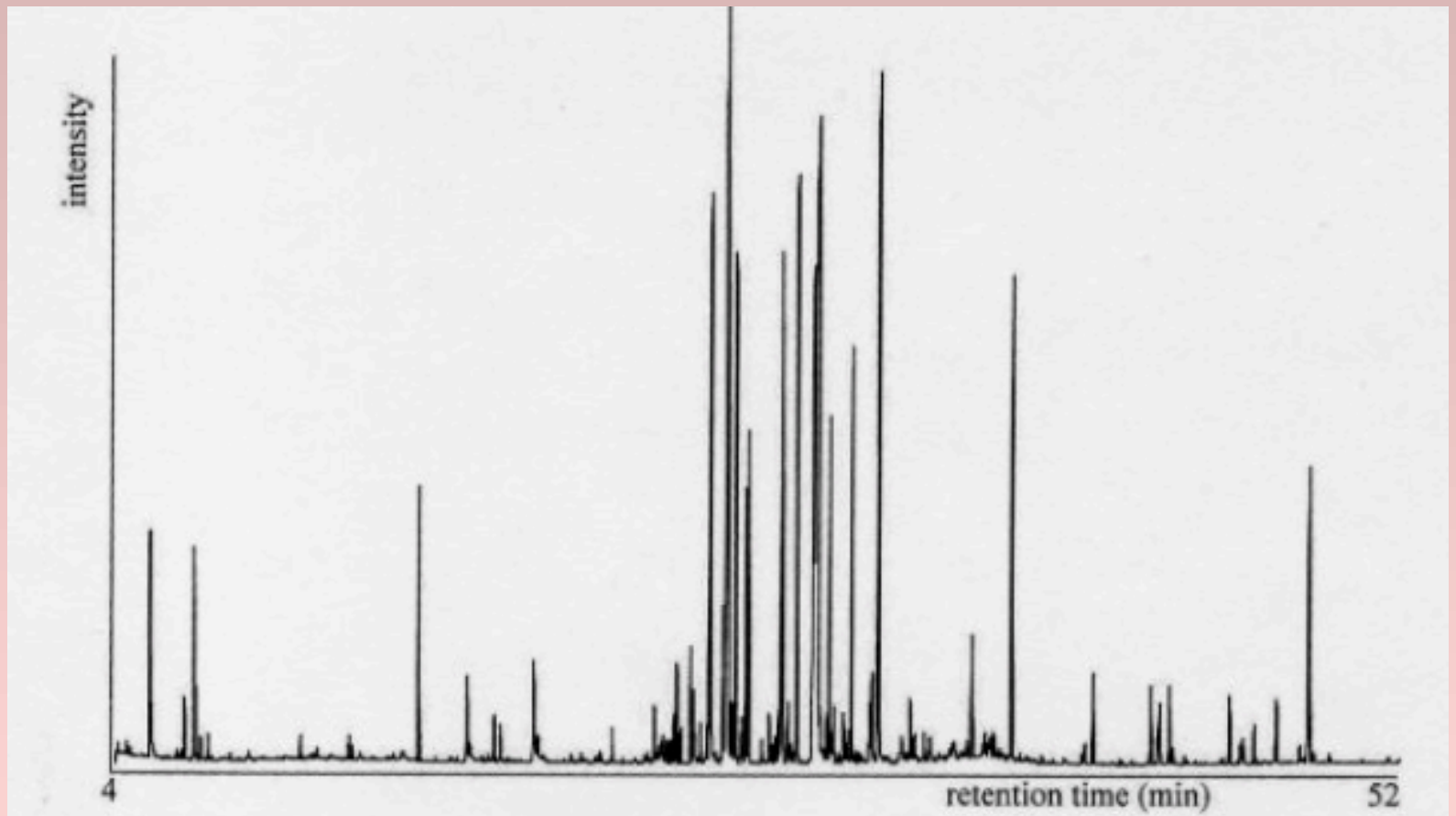


Figure 1. Lipophilic phase of *Arabidopsis thaliana* leaves analysed by GC/quadrupole MS (unpublished results). Inspection of peaks apparent in the base peak chromatogram results in 160 distinct metabolites. Abundant peaks in the middle of the chromatogram are methylated fatty acids. At the end of the chromatogram, trimethylsilylated sterols are eluted.

O. Fiehn et al, Comp. Func. Genom (2001)

LC-MS

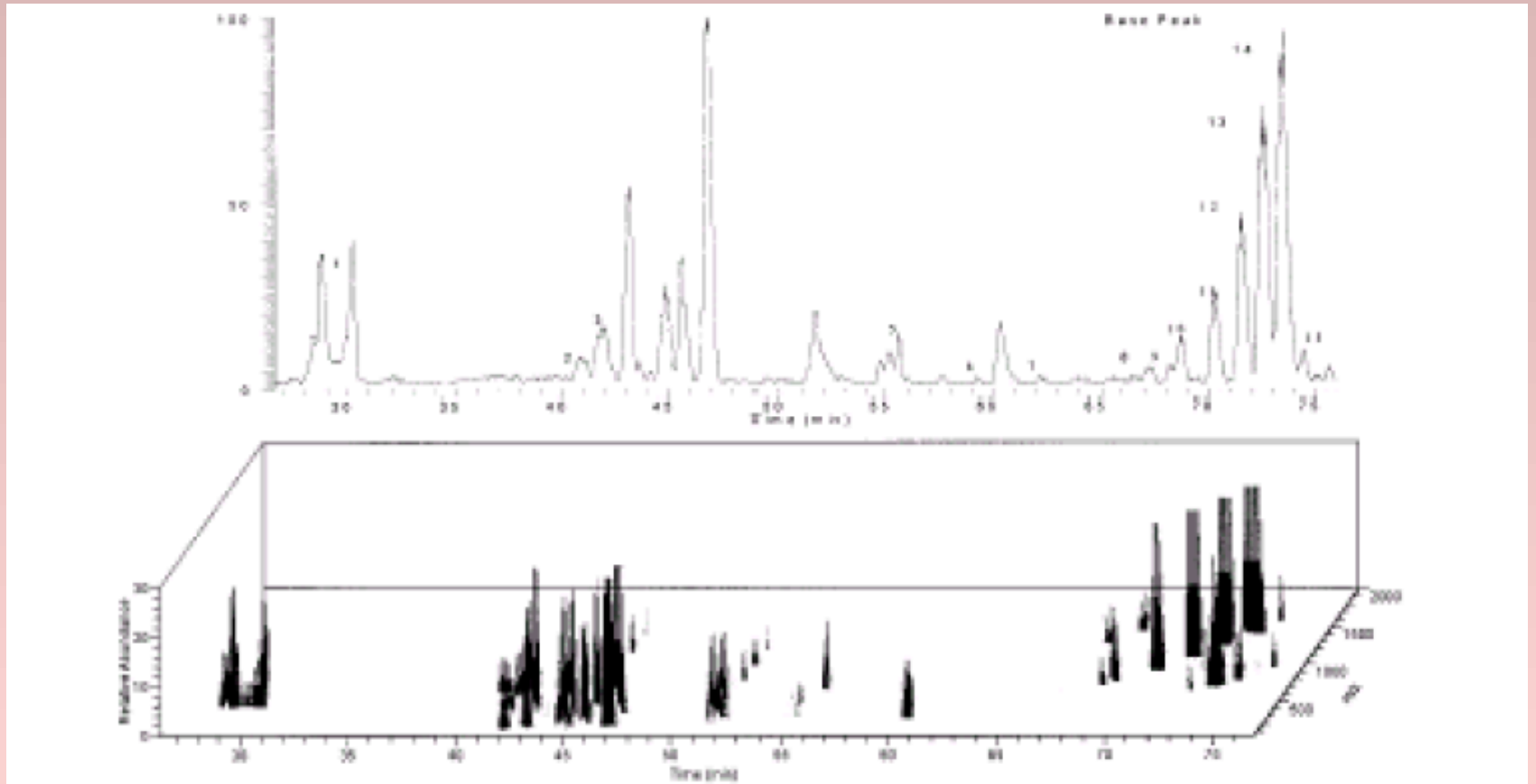


FIG. 4. HILIC/MS base peak chromatogram of *Cucurbita maxima* phloem, presented in a map view (TSK Gel Amide 80, gradient to 60% B was completed at 90 min). Identified peaks are (1) 3-*O*[β-D-xylopyranosyl-(1-6)-β-glucopyranosyl]-(3*R*)-octen-3-ol, (2) UDP(NAc)Gal, (3) UDPGlc, (4) DAB, (5) stachyose, (6) verbascose, (7–15) *O*-glycans.

V. Tolstikov et al, Annal. Biochem. (2002)

CE-MS

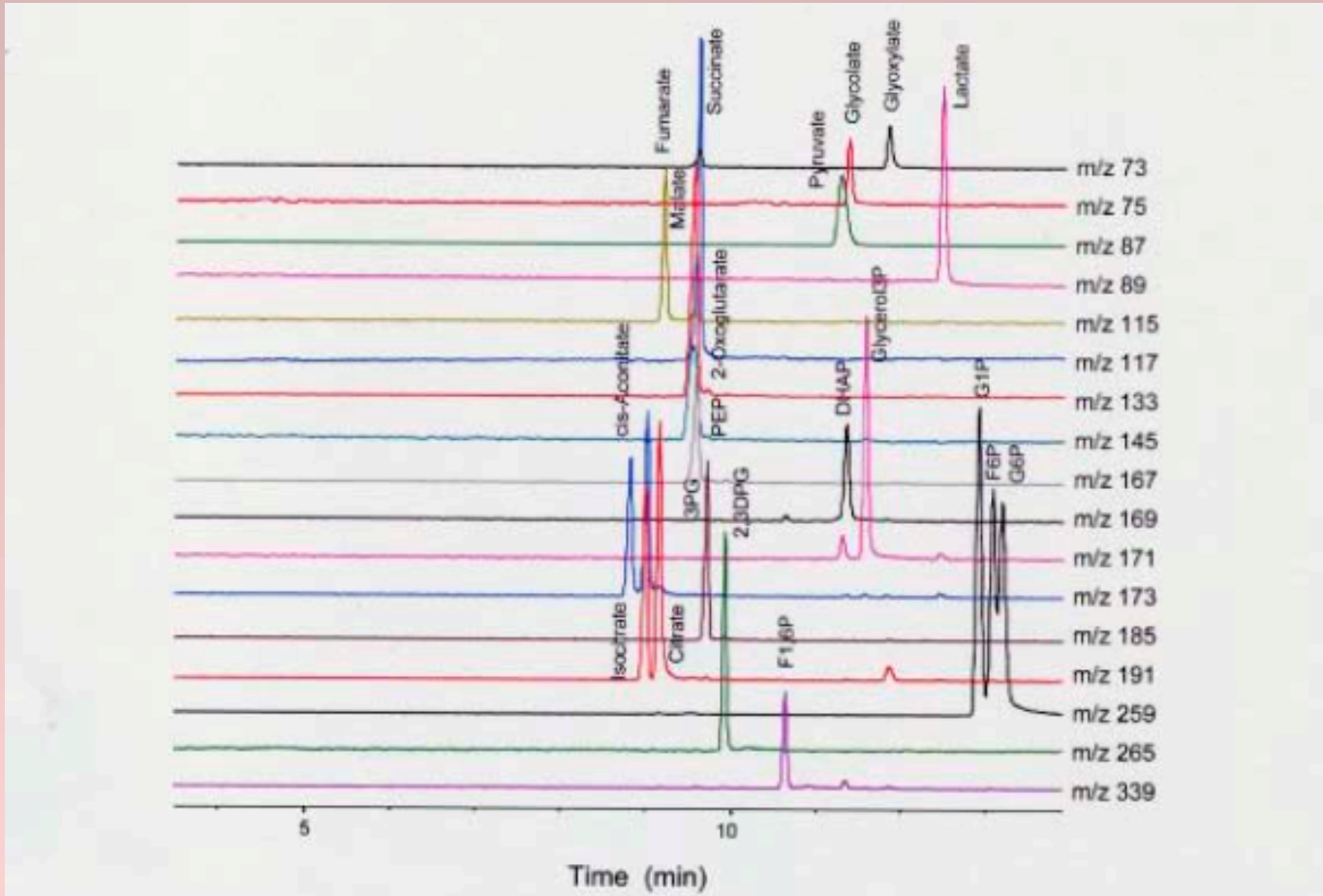


Figure 2. CE–ESI-MS selective ion electropherograms for a standard mixture of 20 metabolites of glycolysis and the TCA cycle. Experimental conditions: sample concentration, 100 μmol/L each; capillary, SMILE(+) 50 μm i.d. × 100 cm; electrolyte, 50 mM ammonium acetate, pH 9.0; applied potential, −30 kV; injection, 30 s at 50 mbar; temperature, 20 °C; sheath liquid, 10 μL/min 5 mM ammonium acetate in 50% (v/v) methanol–water.

T. Soga et al, Annal. Chem. (2002)

El análisis metabólico es más difícil

- ADN/ARN 4 bases
- Proteínas 20 aminoácidos
- Secuenciado lineal posible
- Extracción y análisis fáciles de automatizar

- Muchísimos metabolitos
- Naturaleza y estructuras muy diferentes
- Algunos aún no identificados
- No hay un método único de extracción de información

Necesidad de la bioinformática

- El análisis metabólico produce cantidades aún mayores de datos
- El manejo, procesado y análisis de estos datos requiere herramientas matemáticas e informáticas especializadas:
  - Reconocimiento de patrones en datos complejos
  - Relación entre instantáneas de concentraciones y rutas metabólicos
  - Predicción de consecuencias metabólicas de tratamientos

Técnicas usadas

- Procesado de datos: requiere reducción de ruído, detección e integración de picos, alineamiento de cromatogramas, identificación de compuestos, deconvolución del espectro,...

  Existen herramientas comerciales y públicas (AMDIS para GC-MS, MZmine para LC-MS,...), pero ninguna resuelve todos los problemas ni se aplica a todos los tipos de datos

Técnicas usadas

- Análisis de datos: requiere el manejo de grandes cantidades de datos infradeterminados (más variables que muestras), por lo que hay que reducir el número de variables

  Se usan métodos no supervisados (agrupamiento jerárquico, análisis de componentes principales) y supervisados (ANOVA, análisis discriminante de funciones), y algoritmos genéticos para eliminar variables

Técnicas usadas

- Bases de datos: deberían recoger metadatos (diseño del experimento, naturaleza de las muestras, tratamiento previo, información sobre técnica analítica,...), datos sin procesar, datos procesados, resultados estadísticos,...

  Hay muchas bases de datos relacionadas: KEGG, MetaCyc, AraCyc, MetNet; la más completa és DOME

La herramienta bioinformática que aún no existe

- Se recogen datos en grandes cantidades y producidos por diferentes técnicas
- Los datos se almacenan automáticamente en formato adecuado en una base de datos
- Programas que transforman y analizan los datos automáticamente, escogiendo los métodos más adecuados
- Sistema ampliable fácilmente
- El sistema informa de diferencias/semejanzaes estadísticas relevantes en formato fácil de entender

V. Shulaev, Brief. Bioinf. (2006)


LOONEY TUNES
"That's all Folks!"
A WARNER BROS. CARTOON

La bioinformática está de moda

año	número de publicaciones
Hasta 1990	1
1991-1995	33 (6.6 pub/año)
1996-2000	1411 (282.2 pub/año)
2001-2005	8263 (1653 pub/año)
2006	1857

Búsqueda por palabras clave "bioinformatics" o "computational biology" en Medline/PubMed (17/2/2007)
http://www.pubmed.gov

La bioinformática está de moda

	Google	Google académico	Amazon
Bioinformatics	15,200,000	205,000	4610
"Computational biology"	1,440,000	26,300	1607
"Electron microscopy"	20,000,000	1,040,000	16,663

(17/2/2007)

La bioinformática está de moda

	Google	Google académico	Amazon
Bioinformatics Inventada ~1965	15,200,000	205,000 G/G.a=74.1	4610 G/Am=3297.2
"Comp. biol." Inventada ~1970	1,440,000	26,300 G/G.a=54.7	1607 G/Am=896.1
"Electr. Micr." Inventada ~1930	20,000,000	1,040,000 G/G.a=19.2	16,663 G/Am=1200.2

(17/2/2007)

¿La bioinformática está de moda?

	Google	Google académico	Amazon La casa del libro
Bioinformatics Bioinformática	15,200,000 409,000	205,000 955	4610 2
"Comp. biology" "Biología comput."	1,440,000 20,700	26,300 70	1607 0
"Electr. Micr." "Microscopía Electr."	20,000,000 244,000	1,040,000 4,740	16,663 3

(17/2/2007)