

Introducción a la Bioinformática

Luis E. Garreta U

Pontificia Universidad Javeriana – Cali
Facultad de Ingeniería - Carrera de Biología

30 de julio de 2018

Las herramientas cuantitativas son indispensables en la biología moderna

Las investigaciones biológicas actuales implican la aplicación de algún tipo de matemáticas, estadística, o de herramientas computacionales.



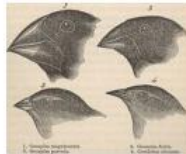
Sin embargo, el simple hecho de usar herramientas cuantitativas o computacionales en biología no necesariamente pueden ser considerados como parte de la bioinformática

Avances en Biología y Computación últimos 30-40 años

El desarrollo de la bioinformática es el resultado de los avances tanto en biología molecular como en ciencias computacionales a lo largo de los últimos 30-40 años.



Birth of Microbiology
~1650s – 1850s



Natural Selection
~1850s – 1900s

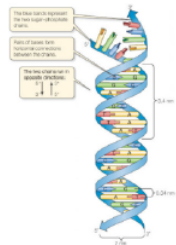
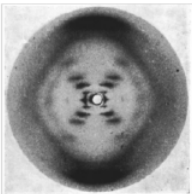


Molecular Biology
~1900s – 1950s

1950: Descubrimiento de la estructura 3D del ADN

ADN, la molécula responsable de transmitir la herencia genética

- ▶ Años 1950's: se logra comprender la estructura tridimensional del ADN,
- ▶ Mayo de 1952, Rosalind Franklin obtiene la famosa fotografía 51, una imagen del ADN obtenida mediante difracción de rayos X
- ▶ Abril de 1953, James D. Watson y Francis Crick reportan la estructura tridimensional del ADN (modelo de doble hélice)



1965: Primera Base de Datos Bioinformática

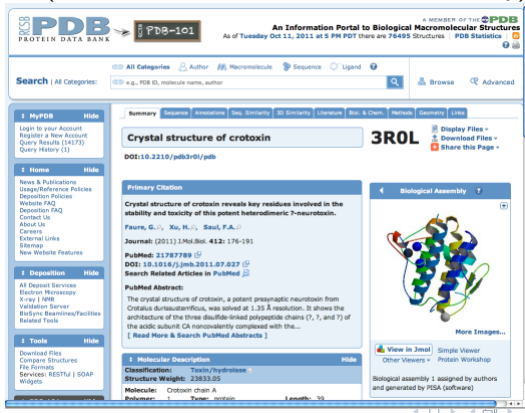
- ▶ Probablemente el primer proyecto Bioinformático.
- ▶ Llevado a cabo por Margaret Dayhoff en 1965
- ▶ Desarrolló la primera BD de secuencias de proteínas



| Amino Acid ⇄ | Short ⇄ | Abbrev. ⇄ | Amino Acid ⇄ | Short ⇄ | Abbrev. ⇄ |
|---------------|---------|-----------|--------------|---------|-----------|
| Alanine | A | Ala | Methionine | M | Met |
| Cysteine | C | Cys | Asparagine | N | Asn |
| Aspartic acid | D | Asp | Pyrrolysine | O | Pyl |
| Glutamic acid | E | Glu | Proline | P | Pro |
| Phenylalanine | F | Phe | Glutamine | Q | Gln |
| Glycine | G | Gly | Arginine | R | Arg |
| Histidine | H | His | Serine | S | Ser |
| Isoleucine | I | Ile | Threonine | T | Thr |
| Lysine | K | Lys | Valine | V | Val |
| Leucine | L | Leu | Tryptophan | W | Trp |
| | | | Tyrosine | Y | Tyr |

1970: Creación del Primer Banco de Datos de Proteínas: PDB

Al inicio de los años 1970, el Laboratorio Nacional de Brookhaven (USA) estableció el Banco de Datos de Proteínas para almacenar estructuras de proteínas en 3D (12 al inicio, más de 200.000 al día de hoy)



The screenshot displays the PDB website interface. At the top, the PDB logo and "Protein Data Bank" text are visible. A search bar is present with the text "Search | All Categories: e.g., PDB ID, molecule name, author". The main content area shows the entry for "Crystal structure of crotoxin" (PDB ID: 3ROL). The entry includes a primary citation: "Crystal structure of crotoxin reveals key residues involved in the stability and toxicity of this potent heterodimeric T-neurotoxin. Faure, G., Xu, H., Saul, F.A. Journal: (2011) J.Mol.Biol. 412: 176-191". The entry also includes a PubMed abstract: "The crystal structure of crotoxin, a potent presynaptic neurotoxin from Crotoalus durissus terrificus, was solved at 1.35 Å resolution. It shows the architecture of the three disulfide-linked polypeptide chains (7, 7, and 7) of the acidic subunit GA noncovalently complexed with the...". The entry is categorized under "Molecular Description" and "Biological Assembly". The entry is assigned by PISA (software).

1970: Primer Algoritmo de Alineamiento de Secuencias

- ▶ El primer algoritmo para alineamiento de secuencias fue desarrollado por Needleman y Wunsch en 1970
- ▶ Este fue un paso fundamental en el desarrollo del campo de la bioinformática
- ▶ Se abrió el camino para las comparaciones de secuencias y búsquedas en BD de rutina realizadas por los biólogos modernos.

| | | A | C | T | G | A | T | T | C | A |
|---|-----|-----|----|----|----|-----|-----|-----|-----|-----|
| | 0 | -2 | -4 | -6 | -8 | -10 | -12 | -14 | -16 | -18 |
| A | -2 | 2 | 0 | -2 | -4 | -6 | -8 | -10 | -12 | -14 |
| C | -4 | 0 | 4 | 2 | 0 | -2 | -4 | -6 | -8 | -10 |
| G | -6 | -2 | 2 | 1 | 4 | 2 | 0 | -2 | -4 | -6 |
| C | -8 | -4 | 0 | -1 | 2 | 1 | -1 | -3 | 0 | -2 |
| A | -10 | -6 | -2 | -3 | 0 | 4 | 2 | 0 | -2 | 2 |
| T | -12 | -8 | -4 | 0 | -2 | 2 | 6 | 4 | 2 | 0 |
| C | -14 | -10 | -6 | -2 | -4 | 0 | 4 | 2 | 6 | 4 |
| A | -16 | -12 | -8 | -4 | -5 | -2 | 2 | 1 | 4 | 8 |

1974: Primer Algoritmo para Predicción de Estructuras de Proteínas

El primer algoritmo para predicción de estructuras de proteínas fue desarrollado por Chou y Fasman en 1974

| Amino acid | α -helix (P α) | β -sheet (P β) | Turn (Pt) | |
|------------|-------------------------------|-----------------------------|-----------|-------------------------|
| Ala | 1.29 | 0.90 | 0.78 | Favor α -helices |
| Cys | 1.11 | 0.74 | 0.80 | |
| Leu | 1.30 | 1.02 | 0.59 | |
| Met | 1.47 | 0.97 | 0.39 | |
| Glu | 1.44 | 0.75 | 1.00 | |
| Gln | 1.27 | 0.80 | 0.97 | |
| His | 1.22 | 1.08 | 0.69 | |
| Lys | 1.23 | 0.77 | 0.96 | |
| Val | 0.91 | 1.49 | 0.47 | Favor β -sheets |
| Ile | 0.97 | 1.45 | 0.51 | |
| Phe | 1.07 | 1.32 | 0.58 | |
| Tyr | 0.72 | 1.25 | 1.05 | |
| Trp | 0.99 | 1.14 | 0.75 | |
| Thr | 0.82 | 1.21 | 1.03 | |
| Gly | 0.56 | 0.92 | 1.64 | Favor turns |
| Ser | 0.82 | 0.95 | 1.33 | |
| Asp | 1.04 | 0.72 | 1.41 | |
| Asn | 0.90 | 0.76 | 1.23 | |
| Pro | 0.52 | 0.64 | 1.91 | |
| Arg | 0.96 | 0.99 | 0.88 | |

1980's: Creación del GenBank y desarrollo de nuevos algoritmos

Los años 1980 vieron el establecimiento del GenBank y el desarrollo de rápidos algoritmos para búsquedas en BD como FASTA de William Pearson y BLAST de Stephen Altschul et al.

The screenshot shows the NCBI GenBank Overview page. At the top, the NCBI logo is on the left, and the title "GenBank Overview" is in the center. Below the title is a navigation bar with links: PubMed, Entrez, BLAST, OMIM, Books, Taxonomy, and Structure. Below this is a search bar with the text "Search Entrez" in a dropdown menu, followed by "for eat-4 elegans" in a text input field, and a "Go" button. On the left side, there is a vertical menu with links: NCBI, SITE MAP, Submit to GenBank, BankIt, and Sequin. The main content area on the right starts with the heading "What is GenBank?" followed by a paragraph: "GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences ([Nucleic Acids Research 2004 Jan 1;32\(1\):23-6](#)). There are approximately 37,893,844,733 bases in 32,549,400 sequence records as of February 2004 (see [GenBank growth statistics](#)). As an example, you may view the [record](#) for a *Saccharomyces cerevisiae* gene. The complete [release notes](#) for the current version of GenBank are available. A new release is made every..."

1980: Inicio del Proyecto Genoma Humano

El inicio del proyecto del genoma humano a finales de los años 1980 propició un rápido desarrollo de la bioinformática.



1990s: Uso Masivo de Internet

Por su parte el uso masivo de Internet en los años 1990 hicieron posible el acceso inmediato, intercambio y diseminación de datos biológicos

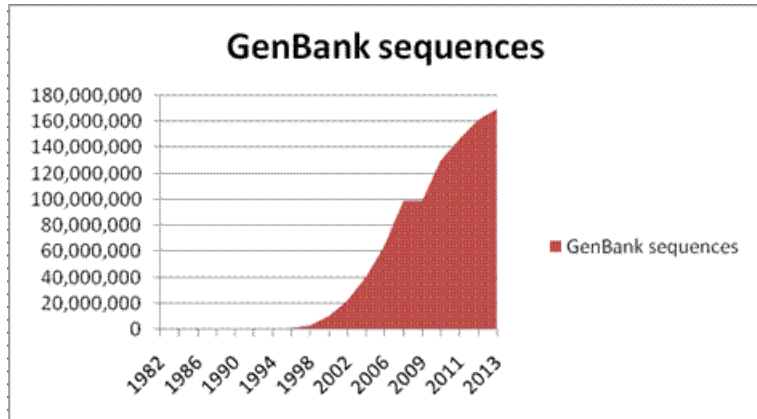
1990— : Producción Masiva de Datos Biológicos

- La bioinformática ganó prominencia como una disciplina debido al avance en los estudios sobre el genoma que produjeron grandes cantidades de datos biológicos



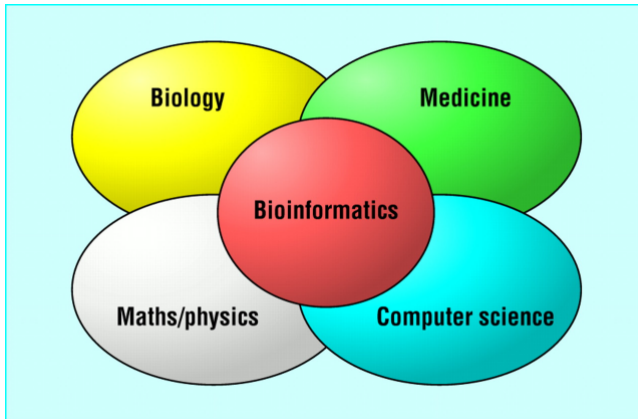
1990— : Crecimiento exponencial de Secuencias Biológicas

La explosión de información sobre secuencias genómicas generaron una demanda de herramientas computacionales eficientes para manejar y analizar los datos



Confluencia de muchas disciplinas

El desarrollo de estas herramientas dependió del conocimiento generado en disciplinas tan diversas como las matemáticas, la estadística, las ciencias computacionales, las tecnologías de la información y la biología molecular



Avances en la Biología Molecular

- ▶ En las dos últimas décadas se han producido importantes avances tecnológicos en el campo de la Biología Molecular.
- ▶ Estos avances han generado una enorme cantidad de datos experimentales y el nacimiento de nuevas áreas de conocimiento como:
 - ▶ la genómica,
 - ▶ la proteómica,
 - ▶ la transcriptómica,
 - ▶ la lipidómica,
 - ▶ la glicómica ,
 - ▶ la metabolómica y
 - ▶ la interactómica.

Necesidad del uso de la Tecnología y la Computación

Para almacenar, organizar, manejar y analizar toda esta información es necesario el uso de computadores.



Definición de Bioinformática

Por tanto, se puede definir la Bioinformática como una nueva área de la ciencia que combina la Biología con la Tecnología de la Información y de la Computación para responder a cuestiones biológicas.

Principales Objetivos de la Bioinformática

Organización de los datos y desarrollo de algoritmos:

- El desarrollo de nuevos algoritmos y estadísticas para establecer relaciones entre miembros de grandes grupos de datos.

Principales Objetivos de la Bioinformática

Organización de los datos y desarrollo de algoritmos:

- El desarrollo de nuevos algoritmos y estadísticas para establecer relaciones entre miembros de grandes grupos de datos.

Análisis e Interpretación de los tipos de Datos Biológicos:

- El análisis y al interpretación de varios tipos de datos incluyendo secuencias de nucleótidos y aminoácidos, dominios proteicos y estructuras de proteínas

Principales Objetivos de la Bioinformática

Organización de los datos y desarrollo de algoritmos:

- ▶ El desarrollo de nuevos algoritmos y estadísticas para establecer relaciones entre miembros de grandes grupos de datos.

Análisis e Interpretación de los tipos de Datos Biológicos:

- ▶ El análisis y la interpretación de varios tipos de datos incluyendo secuencias de nucleótidos y aminoácidos, dominios proteicos y estructuras de proteínas

Desarrollo de nuevas herramientas:

- ▶ El desarrollo y la implementación de herramientas que permitan acceso y manejo eficientes de diferentes tipos de información

Aplicaciones Clásicas de la Bioinformática

► Encontrar homólogos



► Diseño de fármacos

Otras Aplicaciones de la Bioinformática

- ▶ Anotación de genomas:
- ▶ Biología evolutiva computacional:
- ▶ Medición de la biodiversidad
- ▶ Análisis de la expresión génica
- ▶ Análisis de la regulación
- ▶ Análisis de la expresión de proteínas
- ▶ Análisis de mutaciones en el cáncer
- ▶ Predicción de la estructura de las proteínas
- ▶ Genómica comparativa
- ▶ Modelado de sistemas biológicos
- ▶ Análisis de imagen de alto rendimiento
- ▶ Acoplamiento proteína-proteína

Bioinformática y Experimentación

- ▶ La bioinformática tiene gran potencial pero no se debe sobreestimarla.
- ▶ La bioinformática y la biología experimental son actividades independientes, pero complementarias.
- ▶ La bioinformática depende de la ciencia experimental para producir datos primarios para el análisis.

Bioinformática y las Predicciones

- ▶ A su vez la bioinformática, proporciona la interpretación de los datos experimentales e importantes pistas para seguir la investigación experimental
- ▶ Las predicciones hechas en bioinformática no son pruebas formales de cualquier concepto
- ▶ La calidad de las predicciones de bioinformática depende de la calidad de los datos y la sofisticación de los algoritmos utilizados