

# BIOLOGÍA Y COMPUTACIÓN

JOSÉ ANTONIO PEREIRO MOREJÓN<sup>1\*</sup>

15 de abril de 2022

## ÍNDICE

1. Introducción	1
2. Primera mitad del siglo XX	2
3. Segunda mitad del siglo XX	3
4. Años 90's: Despegue de las ómicas	6
5. Siglo XXI	7
6. Biología “quo Vadis” †	10

## 1 INTRODUCCIÓN

Al ser esta la primera asignatura en la carrera con carácter explícitamente computacional, hemos querido comenzar agregando un poco de contexto sobre la relación existente entre la Biología y la Computación. Para ello, haremos una revisión muy breve sobre la historia de cómo la primera ha ido encontrando en esta última una herramienta cada vez más útil, incluso imprescindible, a la hora de resolver las cuestiones fundamentales en el estudio de la vida. En su elaboración, el mayor reto al que nos enfrentamos es la diversidad de temas de interés. La Biología es una ciencia muy amplia, que va desde los estudios más holísticos en la ecología y la conservación, hasta los más reduccionistas como la caracterización de una proteína. Para cada una de estas áreas se pudiera hacer un relato independiente pero en casi todos encontraríamos un patrón similar. La fracción de cualquier estudio biológico dependiente de las habilidades computacionales del investigador ha ido creciendo sostenidamente; ya sea durante la recolección de los datos, su análisis, su almacenamiento o su publicación [1, 2].

Por problemas de tiempo, y aceptando que dejaremos fuera muchos casos interesantes, en este documento nos centraremos en una historia en particular: el desarrollo y consolidación del Dogma Central de la Biología Molecular. La justificación de nuestra selección es que dicho principio es uno de los conceptos fundamentales de la Biología moderna y encuentra aplicación en todas las áreas de la misma. A grandes rasgos, este establece la popular noción de que en los sistemas vivos el flujo de información comienza típicamente por la secuencia de ADN (ácido desoxirribonucleico) y luego se ramifica escalonadamente por varios niveles. Otro rasgo importante es que este proceso es unidireccional. O sea, los cambios fenotípicos no se propagan hacia la cumbre de la jerarquía afectando a la secuencia de ADN [3]. Posteriormente profundizaremos más sobre el tema pero por ahora, comenzaremos un poco antes de que todo ese conocimiento se estableciera.

---

<sup>1</sup> Grupo de Computación, Facultad de Biología, Universidad de la Habana  
\* apereiro@fbio.uh.cu

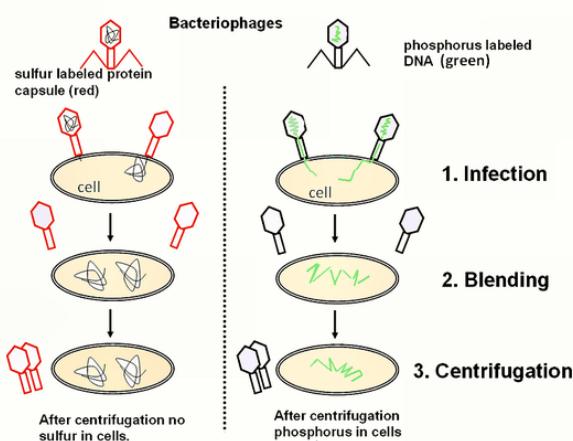
## 2 PRIMERA MITAD DEL SIGLO XX

Es increíble lo mucho que la ciencia ha avanzado en los últimos 120 años. Como referencia, tengan en cuenta que al tiempo que Albert Einstein cursaba sus estudios universitarios todavía había un fuerte debate en la comunidad científica sobre lo que hoy asumimos como una verdad elemental: la materia está compuesta por átomos. De hecho, en 1905 el propio Albert Einstein contribuyó al debate tratando de responder algo tan básico como: ¿de qué tamaño son los átomos? [4]. Con el establecimiento de la teoría atómica y el posterior desarrollo de las Mecánicas Cuánticas y Estadísticas, se consolidaron y extendieron los resultados de otras ciencias más tradicionales como la Química y la Termodinámica. Paralelo, en la Biología comienzan a enraizar corrientes que describen la vida como un proceso químico-físico, por ejemplo, la Bioquímica y la Biología Molecular. Estas ciencias definen a las células como un pequeño reactor químico compuesto principalmente de agua y de un grupo de compuestos orgánicos: los carbohidratos, las proteínas, los ácidos nucleicos y los lípidos.

Aunque se progresaba sostenidamente, estos pudieran ser considerados todavía “tiempos oscuros” para la Biología. Por ejemplo, resultados como que algunas proteínas tienen capacidad enzimática no fueron establecidos hasta 1926 [5]. Además, otras preguntas fundamentales seguían abiertas. A pesar que las teorías de la Herencia y Evolución estaban ya acompañadas de evidencias experimentales sólidas, como las aportadas por los registros fósiles, la Biología Comparada y experimentos de cruzamiento [6, 7], sus bases moleculares seguían siendo un misterio.

El descubrimiento de que el ADN es el portador de la información genética fue uno de los logros más importantes de la Biología Molecular durante la década de 1950. Para nada este fue un tema libre de controversias, muchos investigadores apostaban por las proteínas como el mejor candidato [8]. Esto se debía a que para la época, el ADN era visto solo como una “aburrida” sustancia constituida por grandes cantidades de cuatro subunidades diferentes (bases nitrogenadas), las cuales se agrupaban en pares debido a que se encontraban en igual proporción. Había discusiones sobre si una sustancia tan simple químicamente era capaz de encerrar la complejidad necesaria para describir a un organismo completo. Las proteínas, por otro lado, son más diversas al estar compuestas por veinte subunidades (aminoácidos) y hay miles diferentes en cada célula. Esta discusión fue concluyendo ante los resultados de una serie de experimentos claves.

En 1944, ya se había observado que el ADN extraído de una cepa de bacteria podía transferir su virulencia a otra cepa anteriormente inofensiva [10]. Pero los ex-



**Figura 1:** (Figura en inglés) Esquema del experimento de Martha Chase y Alfred Hershey. Más detalles en el texto principal. Fuente [9]



Figura 2: Watson and Crick frente al modelo de doble hélice. Fuente [12]

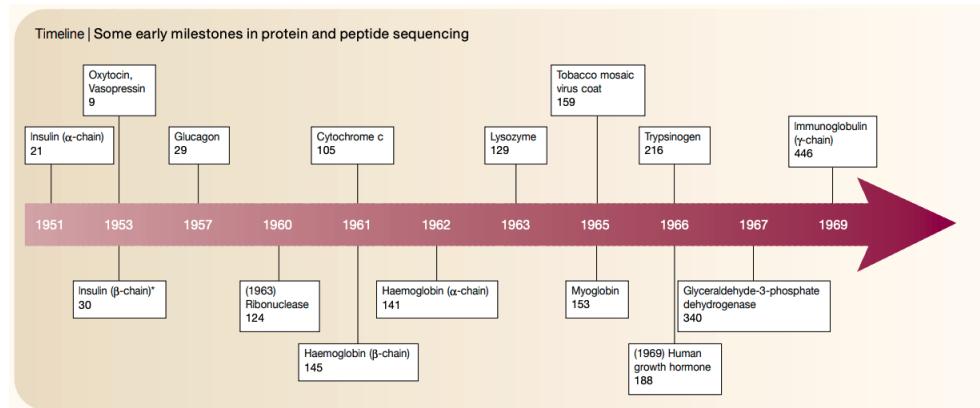
perimentos decisivos se llevaron a cabo en 1952 por Martha Chase y Alfred Hershey. En este experimento [11] los autores emplearon bacteriófagos (virus compuestos solo de ADN y proteínas) para dilucidar qué sustancia penetraba la célula (figura 1). Resulta que el ADN carece de azufre en su composición, mientras que las proteínas del virus sí presentan este elemento. Por otro lado, el ADN es rico en fósforo, mientras que las proteínas en cuestión carecían de este. Esta diferencia permitió marcar ambas sustancias de manera independiente y al infestar las bacterias, solo el fósforo marcado fue detectado su interior. El resultado era claro, la única sustancia proveniente de los virus que penetraba la célula era el ADN, llevando con ella la información necesaria para producir nuevos bacteriófagos.

Casi al unísono, en 1953, otro importante descubrimiento fue publicado: la estructura del ADN [13]. Usando cristalografía de rayos X James Watson and Francis Crick (figura 2) establecieron que el ADN es un polímero con una conformación en doble hélice donde las bases nitrogenadas están apareadas por afinidad química. Estos resultados sugerían que la información genética estaba codificada en la larga secuencia de esos pares. Toda la diversidad de la vida “escrita” en un mismo lenguaje. Había nacido la Genética Molecular.

También de relevante en ese período fue el trabajo de Frederick Sanger, quien publicó en 1951 la secuencia de una de las cadenas de la Insulina [14]. Estos resultados terminaron de imponer la teoría de que las proteínas son polímeros de aminoácidos con una secuencia lineal definida. El vínculo entre la estructura de las proteínas y la secuencia de ADN era más evidente que nunca antes. Comenzaban así a dibujarse los primeros bosetos del dogma central de la Biología Molecular.

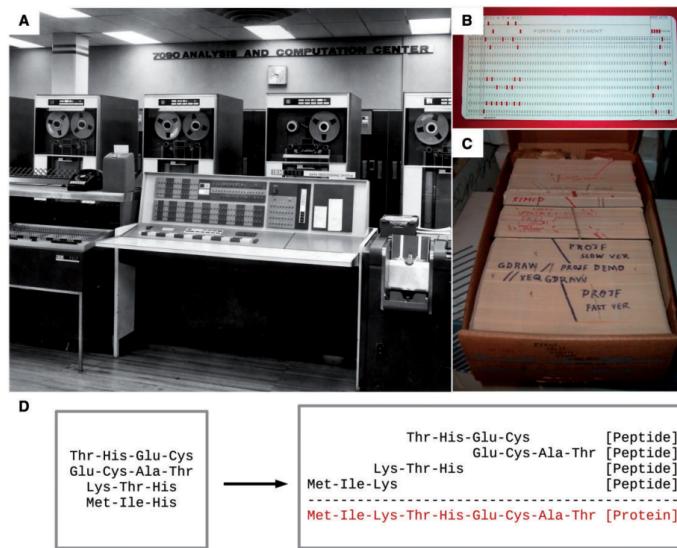
### 3 SEGUNDA MITAD DEL SIGLO XX

Hasta ahora no hemos mencionado ningún ejemplo sobre el uso de la Computación en las investigaciones biológicas. Esto tiene una explicación muy simple, durante la primera mitad del siglo XX el acceso a medios computacionales era escaso. Como referencia, fue en 1946 que se terminó de construir la primera computadora electrónica de uso general, la ENIAC. En estas etapas iniciales solo las instituciones más grandes (como los gobiernos y los ejércitos) tenían acceso a estas máquinas. Pero a finales de los 50's esta tendencia fue cambiando, ya comenzaba a ser común para algunas universidades disponer de estos medios.



**Figura 3:** (Figura en inglés) Momentos claves iniciales en el proceso de secuenciación de proteínas. Los números debajo de los nombres de las proteínas indican el número de aminoácidos de su secuencia (ej: Glucagón: 29). Fuente [15]

Mientras tanto, el primer enunciado formal del dogma central de la Biología Molecular fue expuesto por el propio Francis Crick durante una conferencia en 1957 [16]. En este, las secuencias de ADN juegan el papel central como punto de partida del flujo de la información que caracteriza el desarrollo de los organismos. A pesar de esto, por limitaciones experimentales de la época, son las secuencias proteicas las que son más accesibles. Luego de que las primeras proteínas fueran resueltas "a mano", desde principios de los 60's se comienza a automatizar el proceso. Proteínas cada vez más "largas" se secuencian en cada vez menos tiempo (figura 3). A finales de los 60's, Pehr Edman diseñó el *sequenator* [17], un equipo completamente automático. Cada vez más laboratorios se aventuraban a secuenciar proteínas de interés y así surgieron los primeros repositorios de secuencias [15].



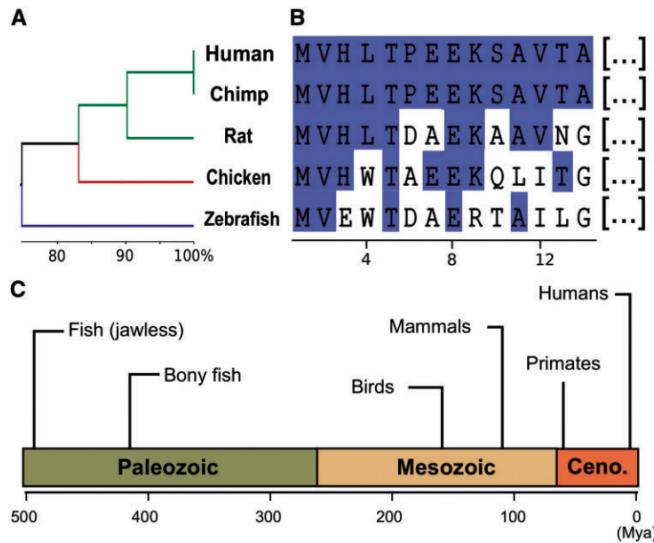
**Figura 4:** Lo que es considerado como uno de los primeros sistemas bioinformáticos. **Panel A:** una computadora de la época (1962). **Panel B:** una tarjeta perforada (un medio de almacenamiento digital "antiguo"). **Panel C:** conjunto de miles de tarjetas perforadas que contienen el programa ensamblador. **Panel D:** representación del proceso de ensamblaje de la secuencia completa a partir de los fragmentos secuenciados. Fuente [1]

El uso de las computadoras durante el proceso de secuenciación fue imprescindible desde bien temprano (figura 4) [18]. El problema principal radicaba en que los métodos bioquímicos de la época solo podían resolver en cada corrida secuen-

cias de tamaño limitado. Por ello, para secuenciar proteínas de mayor tamaño, estas eran divididas en fragmentos manejables que se procesaban individualmente. Luego, la secuencia completa de la proteína era reconstruida aprovechando secciones redundantes entre los fragmentos adyacentes. La complejidad del proceso reconstructivo aumenta considerablemente con el tamaño de las proteínas y la cantidad de fragmentos en cuestión, por lo que rápidamente se hace impráctico realizarlo manualmente.

Los años 50's y 60's son etapas muy tempranas en el desarrollo de la Computación. El hecho de que ya desde esta época la Biología asimilara su uso pone de manifiesto un patrón universal en la ciencia. No existen herramientas únicas de un campo o disciplina, o herramientas que no sean de su interés. Esas ideas son, en el mejor de los casos, solo una ilusión temporal. Lo que existen son problemas, cuestiones que se quieren resolver. El cómo atacarlos cambia dependiendo de qué recursos estén disponibles. Seguramente Albert Einstein no estudió programación en la universidad (las computadoras no eran muy populares por esa época), sin embargo, pero al tiempo de su muerte en 1955, los físicos ya habían incorporado esta a su caja de herramientas.

Sin embargo, el impacto del uso de la Computación en las investigaciones biológicas solo comenzaba. La acumulación de secuencias de proteínas potenció el desarrollo de nuevas ramas dentro de esta ciencia. Una de las más tempranas fue el campo de la Evolución Molecular. El dogma central tiene consecuencias muy interesantes. Primero, al ser el ADN el principal portador de la información genética, la evolución puede ser vista como la acumulación de cambios en dicha secuencia y su posterior discriminación por selección natural. De esta manera, una comparación de secuencias de diferentes especies debería arrojar luz sobre su pasado evolutivo. Segundo, que se puede aprender sobre las secuencias de ADN partiendo del estudio de las secuencias de las proteínas. Así, los estudios evolutivos también se podían llevar a cabo empleando las secuencias de estas últimas.



**Figura 5:** (Figura en inglés) Panel A y B: el árbol filogenético (izquierdo) y un fragmento de la comparación (derecha) entre las secuencias de hemoglobina que se emplearon en su elaboración (cada letra representa un aminoácido). Panel C: cronología de los fósiles más antiguos encontrados de las especies/grupos estudiados (Mya Millones de años). Fuente [1]

La hipótesis era simple. Si dos especies divergieron recientemente en la historia evolutiva, las secuencias de proteínas comunes a ambas (denominadas homólogas) deberían ser similares. Por el contrario, si las especies están muy poco emparentadas, esta similitud se espera sea menor. De esa forma, ordenando un grupo de



**Figura 6:** Secuenciación a escala industrial durante Proyecto del Genoma Humano. Fuente [1]

especies con respecto a la “similitud” entre sus secuencias homólogas, se podría construir su árbol filogenético. Por ejemplo, si se compara la secuencia de la hemoglobina humana con la del chimpancé se aprecian menos diferencias que entre un humano y un ratón. Los árboles filogenéticos obtenidos de esta forma fueron luego comparados, con muy buenos resultados, con los obtenidos mediante los métodos tradicionales, como el registro fósil (figura 5) [19].

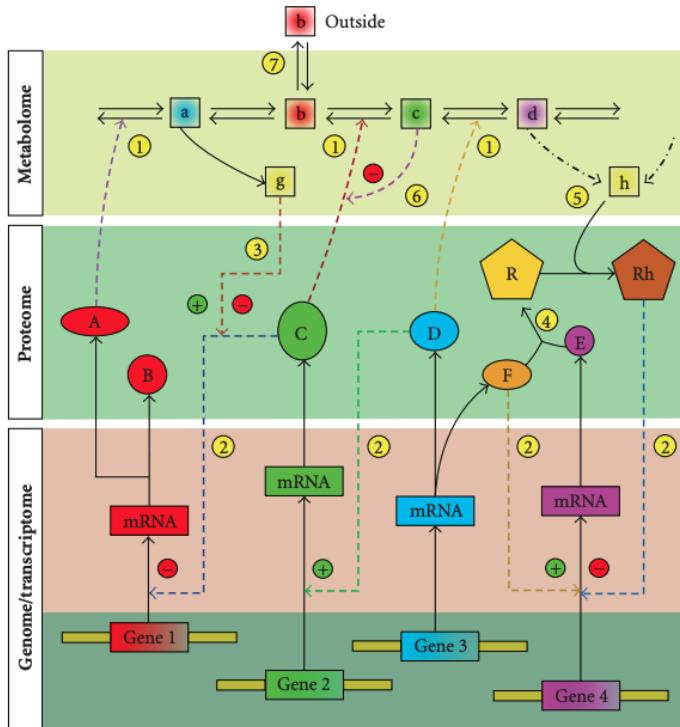
Una vez más, la Computación fue clave en estas investigaciones. Mientras más secuencias de más especies se acumulaban, realizar el alineamiento necesario para determinar su similitud se hacía demandante, incluso, para las computadoras más avanzadas; sobre todo si las especies estudiadas estaban muy poco emparentadas y las diferencias entre las secuencias eran pronunciadas o implicaban adiciones o eliminaciones. Como referencia, comparar solo 10 secuencias de 100 aminoácidos de extensión cada una tomaba como promedio  $10^{18}$  operaciones [20]. Estas investigaciones no solo estaban en la vanguardia de la Biología, sino también de la Computación. Como veremos a continuación esta tendencia llega hasta hoy día. La relación entre ambas ciencias siempre ha sido de beneficio mutuo y con una clara inclinación a profundizarse.

## 4 AÑOS 90'S: DESPEGUE DE LAS ÓMICAS

Aunque las secuencias de proteínas fueron un punto de partida dada su temprana disponibilidad, la secuenciación de ADN también fue avanzando aceleradamente y ya en la década de los 90's tenía gran desarrollo. Ejemplo de esto fue la secuenciación del primer genoma completo de un organismo independiente (no parásito) en 1995, *Haemophilus influenzae* [21]. Pero aún más importante fue el lanzamiento de unos de los proyectos más ambiciosos en la historia de Biología: el Proyecto del Genoma Humano.

El proyecto, financiado principalmente por el gobierno de los Estados Unidos, comenzó en 1991 (figura 6). Una década, 2.7 miles de millones de dólares y 3 mil millones de pares de bases después [22], se presentaba la primera versión de un genoma humano de referencia [23]. Este acontecimiento es considerado el punto de partida de lo que se conoce como la era de las ómicas en la Biología.

Las ómicas son un conjunto de disciplinas que se especializan en el estudio de los componentes o relaciones que conforman un organismo vivo (figura 7). Estas disciplinas se caracterizan por su carácter total, no se estudia un componente en particular, se intenta alcanzar a todos en el organismo en cuestión. Entre ellas están



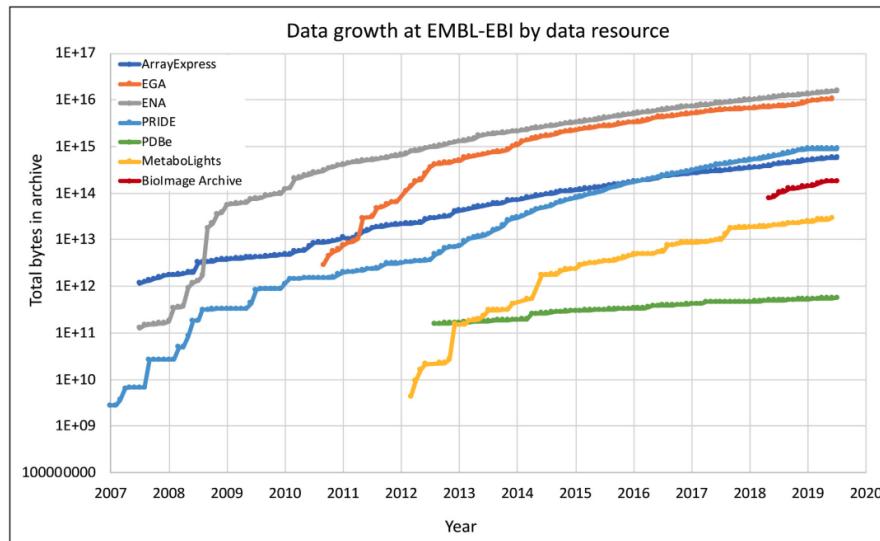
**Figura 7:** (Figura en inglés) El Dogma Central de la Biología Molecular y las ómicas. Con el desarrollo de las ómicas se pretende construir un mapa detallado que comprenda todos los componentes celulares y sus interrelaciones, cientos de miles de componentes y millones de interacciones. Fuente [25]

las que estudian los diferentes componentes moleculares, la genómica (el ADN), la transcriptómica (el ARN), la proteómica (las proteínas), la metabolómica (los intermediarios metabólicos), etc. Además están las que estudian otros aspectos como la epigenética (estudio de los patrones de metilación del ADN), la fenómica (estudio de todos los fenotipos mutantes), la regulómica (los componentes moleculares regulatorios) y muchas más [24]. El patrón es simple, “algo biológicamente interesante” + “ómica”. Estas, casi siempre surgen cuando se desarrolla alguna técnica experimental que permite obtener los datos requeridos a gran escala.

En estas disciplinas los conceptos computacionales no son solo fundamentales, sino nativos. Junto con los elaborados procedimientos experimentales, nuevas tecnologías y algoritmos deben ser creados para hacer frente a la cantidad de datos sin precedentes que se generan. Un papel muy importante en la etapa inicial de desarrollo de las ómicas lo jugó la evolución de la red mundial de computadoras [1]. A principios de los 90's Tim Berners-Lee crea la *World Wide Web*, un sistema internacional de documentos interconectados, lo que se considera como el comienzo de la internet [26]. Esta red ha jugado, y juega, un papel esencial a la hora de facilitar la colaboración a gran escala necesaria en proyectos de gran magnitud. Desde el almacenamiento hasta la publicación y acceso de los resultados, la Biología es “subida a la nube”.

## 5 SIGLO XXI

En las últimas dos décadas la Biología se ha posicionado como una de las ciencias que genera mayor cantidad de datos. Por ejemplo, en el 2015, la cantidad de genomas humanos secuenciados se duplicaba cada 7 meses. O sea, si a comienzos de los 2000 tras una década de trabajo se publicaba el primer genoma de humano, veinte años después se acumulan más de un millón. ¿Y es eso mucho? Como refe-

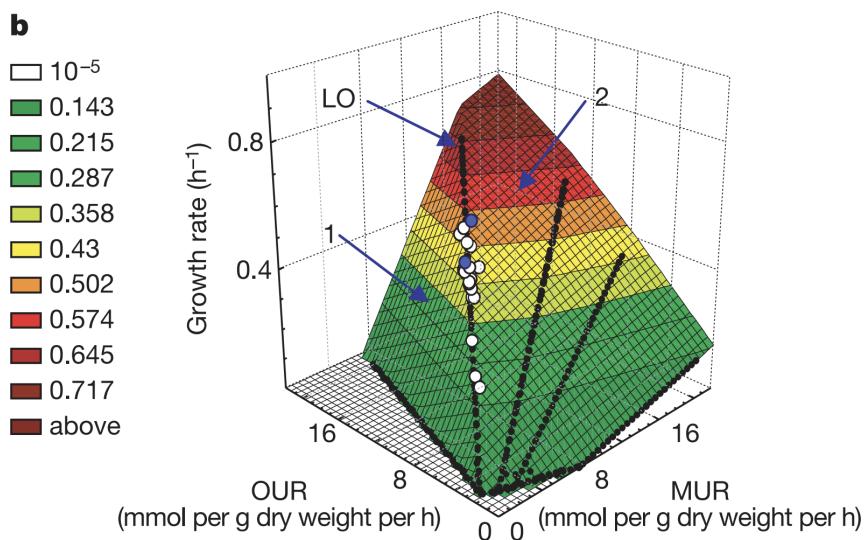


**Figura 8:** (Figura en inglés) Crecimiento exponencial de los datos manejados por el Instituto Europeo de Bioinformática (EMBL-EBI). Note que un Terabyte son  $10^{12}$  bytes. En la gráfica se muestra la evolución de diferentes bases de datos, desde genómicas hasta imágenes de microscopía. Para más detalles ir a la Fuente [27]

rencia tomen que, en 2015, youtube generó entre 1-2 exabytes (millones de terabytes) de nuevo contenido, mientras que se estima que la genómica como campo pudo generar hasta 40 [28]. Esta tendencia no es única para la genómica. Cada elemento de interés biológico está siendo estudiado y recopilado en enormes bases de datos. ¡Hay curvas de crecimiento exponencial por todos lados! (figura 8).

A pesar de esto, aunque es un gran reto, el almacenamiento no es el mayor de los problemas a enfrentar en la actualidad. Más agudo aún es el de extraer de los datos la información relevante. Por ejemplo, si reevaluamos el alineamiento de secuencias en el contexto contemporáneo veremos la magnitud del reto. El alineamiento de un genoma completo de humano con otro de ratón, usando los últimos algoritmos, toma alrededor de 100 horas (~ 4 días) en un CPU moderno. Ahora, si escalamos este análisis para todos los genomas de las más de 2.5 millones de especies que se espera tener secuenciadas para el 2025, se necesitaría toda la capacidad actual de cómputo de la humanidad multiplicada por un millón [28]. ¿Pero acaso la humanidad no está aumentando su poder de cómputo exponencialmente? ¿No se puede simplemente esperar hasta que sea suficiente? Pues no es tan simple, mientras que la evolución del poder de cómputo ha seguido la tendencia de duplicarse cada 2 años (llamada Ley de Moore), incluso aunque esta se mantenga, la generación de datos biológicos es simplemente mucho mayor (es duplicada en menos de un año) [28]. A esto se le suma el hecho de que hay otras áreas de la actividad humana generando cantidades similares de datos, desde la Astrofísica hasta las redes sociales. No todas las computadoras son para nosotros.

La discusión anterior ilustra sólo una parte del problema: el enorme volumen de datos dificulta incluso la aplicación de técnicas ya establecidas. Pero otro igual de significativo es el de generar nuevas formas de análisis. Por ejemplo, en la era pre-óMICAS, las áreas de la Biología que intentaban construir modelos detallados de sistemas completos no atraían mucho interés. Los sistemas vivos son simplemente muy complejos. Estos pueden comprender miles de componentes por cada célula, miles de tipos celulares por cada organismo y un sin número de especies diferentes por ecosistema. Todo ello extremadamente interconectado y sometido a cambios constantes. Sin los datos suficientes, la creación de modelos capaces de capturar tal complejidad estaba fuera del alcance de los investigadores. Pero en los últimos años la situación es diferente, como se evidencia con la consolidación de la Biología de Sistemas [25]. Esta disciplina intenta explicar detalladamente la relación que

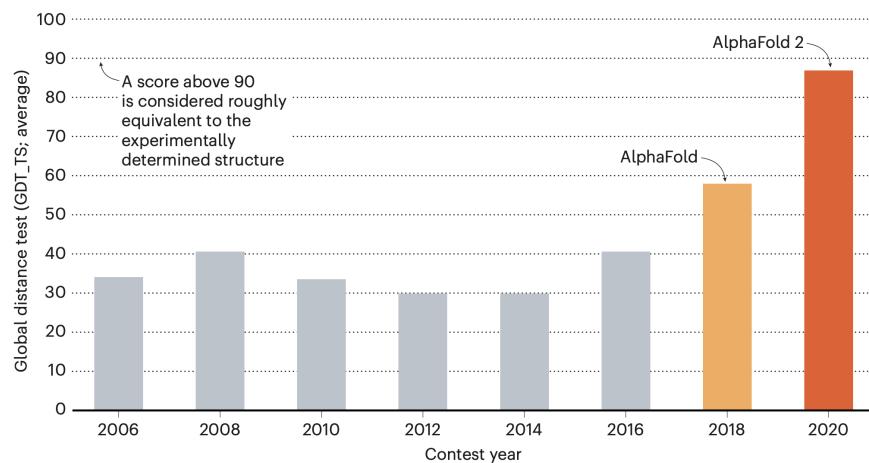


**Figura 9:** (Figura en inglés) En la figura se muestra cómo dadas las disponibilidades de matalo (MUR) y oxígeno (OUR), principales nutrientes limitantes en el cultivo, el crecimiento de la bacteria (círculos blancos y azules) sigue la tendencia esperada descrita por el modelo (puntos negros). Para más detalles ir a [29]

enuncia el dogma central y sus consecuencias: ¿cómo el genoma determina función? Este nuevo enfoque mueve el centro de estudio desde el *elemento* a la *red de elementos*. O sea, se trata de explicar cómo una determinada función emerge de la interacción de las partes del sistema. Lo cual es posible solo si se tienen datos lo suficientemente completos para representar una parte significativa del mismo.

Por ejemplo, hoy se dispone de redes que comprenden todas las reacciones metabólicas que se han identificado en el genoma de *Escherichia coli*. Con estas redes se pueden construir modelos para simular el crecimiento de dicha bacteria en un medio dado. A medida que más datos se le han ido incorporando a estos modelos, estos han ido mejorando su poder predictivo [29]. Además de las aplicaciones ingenieriles que se pueden derivar, la modelación también contribuye a generar conocimientos biológicos. En los resultados mostrados en la figura 9, por ejemplo, los valores experimentales (círculos blancos y azules) se ubican en el máximo posible dada las restricciones que establece el modelo. O sea, las bacterias están creciendo tanto como pueden dada la disponibilidad de nutrientes y su capacidad metabólica, lo cual pudiera ser referente de la presión adaptativa a la que se enfrentan típicamente en su ambiente natural. En general, avances como este hacen que la Biología se mueva cada vez más de su enfoque experimental de “prueba y error” a otro mucho más teóricamente intenso.

Por otro lado, la disponibilidad de grandes cantidades de datos sin analizar no haya pasado inadvertida para otras disciplinas. Físicos, ingenieros, científicos de la Computación y muchos otros especialistas están tratando de modificar las técnicas y habilidades que emplean en sus campos para analizar los datos biológicos. Un ejemplo reciente es el caso de *DeepMind*, una subsidiaria de *google*. Anteriormente, en el año 2017, esta empresa de inteligencia artificial logró el hito de crear un programa para jugar *Go* que pudo vencer al jugador humano con el primer puesto en el ranking de ese deporte. Pero en el año 2020, *DeepMind* volvió a ser noticia. Esta vez fue por usar su enorme capacidad para atacar lo que es considerado uno de los problemas computacionales más grandes de la Biología: predecir la estructura 3D de las proteínas a partir de su secuencia lineal [30]. En particular, *DeepMind* superó a otros 100 equipos en un concurso bianual llamado CAPS. En este concurso alrededor de 100 secuencias de proteínas son publicadas. Las estructuras 3D de esas proteínas son determinadas a partir de técnicas experimentales ya establecidas



**Figura 10:** (Figura en inglés) *AlphaFold*, las redes de *DeepMind*, han revolucionado el campo de la predicción de la estructura 3D de las proteínas a partir de su secuencia lineal. Sus predicciones están cerca de tener un 90 % de similitud con los resultados experimentales. Fuente [30]

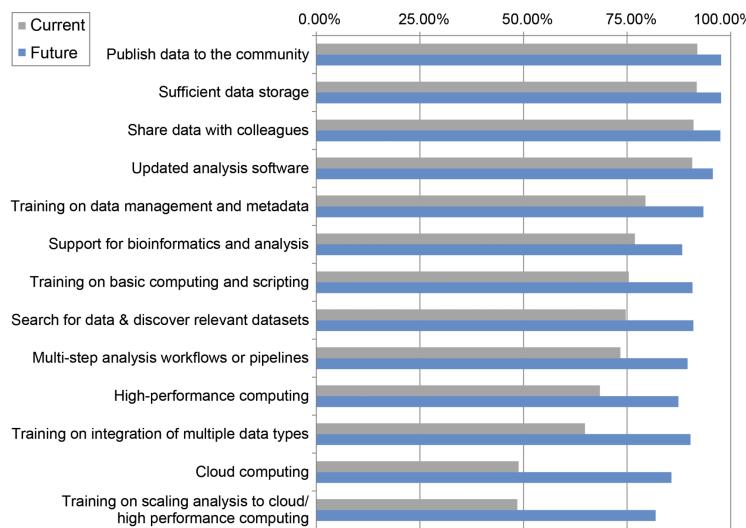
pero esta información no es pública hasta después del concurso. Así, los equipos tienen dos años para proponer sus propias estructuras que luego son comparadas con las experimentales. El uso masivo por parte de *DeepMind* de nuevas técnicas de *Machine Learning* les permitió obtener resultados revolucionarios en esta área (figura 10) [30].

## 6 BIOLOGÍA “QUO VADIS” †

Como se mencionó al comienzo, es imposible abarcar en un documento tan breve el desarrollo de una ciencia tan amplia como la Biología. Por ello, inevitablemente, se han quedado fuera muchas otras disciplinas que son ejemplos de la profunda interrelación entre la Computación y esta rama de la ciencia. Entre ellas podemos mencionar a las Neurociencias, la Biología Celular, la Fisiología, la Ecología y otras que se han beneficiado, por ejemplo, del desarrollo de las tecnologías de imágenes digitales y su procesamiento. Desde imágenes satelitales que son empleadas para evaluar la progresión de un daño ecológico o los patrones de migración de especies amenazadas [32], hasta el entrenamiento de redes neuronales para el temprano diagnóstico de cáncer a partir de imágenes de muestras de tejido [33]. Otras como la Bioacústica y la Biofísica se han beneficiado grandemente del desarrollo del procesamiento de señales en general. Por ejemplo, investigadores graban el sonido ambiente en los hábitats y luego lo procesan para estimar realizar el tamaño de poblaciones de diversos animales simultáneamente. La identificación de las especies y la cantidad de individuos se puede estimar mediante *Machine Learning* [34]. Al igual que las ómicas, todas estas disciplinas vienen acompañada de grandes acumulaciones de datos.

Por otro lado, otro fenómeno significativo es la democratización del acceso a las tecnologías de la información. Movimientos como el *Open Software* y *Open Hardware* han abierto un sinfín de nuevas posibilidades [35], particularmente en los países en vías de desarrollo. Hoy en día es posible fabricar “en casa”, usando tecnologías como la impresión 3D y microcontroladores, los sensores e instrumentos necesarios para la investigación. Esto no solo puede abaratar los costos, también permite que los investigadores diseñen los artefactos acorde a sus necesidades más particulares. Nunca ha habido más flexibilidad en el aspecto experimental.

† *quo Vadis*: del latín “¿A dónde vas?”



**Figura 11:** (Figura en inglés) Necesidades computacionales actuales y futuras identificadas en una encuesta realizada a cientos de investigadores. Noten que las necesidades futuras son siempre mayores que las actuales, en algunos casos muy marcadamente. Fuente [31]

La Biología se encuentra hoy en una posición muy diferente si es comparada con los comienzos de este siglo. La disponibilidad de grandes cantidades de datos ha aumentado considerablemente la necesidad de introducir nuevas habilidades en el currículum de los estudiantes e investigadores. En una encuesta reciente realizada a varios cientos de investigadores de alto perfil, se identificó como una de las principales áreas a priorizar, el entrenamiento de las nuevas generaciones en las habilidades requeridas para la manipulación, publicación y procesamiento de datos [31] (Figure 11). Esto avizora un futuro donde un sólido entrenamiento en computación no sea ya una habilidad excepcional, sino, esperada para un graduado de una especialidad biológica. Como mencionamos anteriormente, no existen herramientas propias de una disciplina, ni herramientas que estén fuera de su interés. Existen problemas fundamentales, y cada disciplina deberá considerar suyas todas aquellas que les permitan abordar los propios, la Biología no es excepción.

## REFERENCIAS

- [1] Jeff Gauthier, Antony T. Vincent, Steve J. Charette, and Nicolas Derome. A brief history of bioinformatics. *Briefings in Bioinformatics*, 20(6):1981–1996, November 2019.
- [2] D. Ewen Cameron, Caleb J. Bashor, and James J. Collins. A brief history of synthetic biology. *Nature Reviews Microbiology*, 12(5):381–390, may 2014.
- [3] Francis Crick. Central Dogma of Molecular Biology . *Nature*, 227(5258):561–563, aug 1970.
- [4] Albert Einstein. *Einstein's Miraculous Year : Five Papers That Changed the Face of Physics* . Princeton University Press , may 2021.
- [5] James B. Sumner. THE ISOLATION AND CRYSTALLIZATION OF THE ENZYME UREASE : PRELIMINARY PAPER . *Journal of Biological Chemistry*, 69(2):435–441, aug 1926.

- [6] Charles Darwin. *On the Origin of Species by Means of Natural Selection , Or The Preservation of Favoured Races in the Struggle for Life* . John Murray , Albemarle Street. , 1859.
- [7] Scott Abbott and Daniel J Fairbanks. Experiments on Plant Hybrids by Gregor Mendel . *Genetics*, 204(2):407–422, oct 2016.
- [8] A.J.F. Griffiths, S.R. Wessler, R.C. Lewontin, W.M. Gelbart, D.T. Suzuki, and U.J.H. Miller. *Introduction to Genetic Analysis* . Macmillan Higher Education , 2006.
- [9] Wikipedia contributors. Hershey–chase experiment — Wikipedia, the free encyclopedia, 2022. [Online; accessed 8-April-2022].
- [10] Oswald T. Avery, Colin M. MacLeod, and Maclyn McCarty. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES . *The Journal of Experimental Medicine*, 79(2):137–158, feb 1944.
- [11] A. D. Hershey and M. Chase. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *The Journal of General Physiology*, 36(1):39–56, may 1952.
- [12] Stephen S. Hall. Old School Ties : Watson , Crick , and 40 Years of DNA . *Science*, 259(5101):1532–1533, mar 1993.
- [13] J. D. Watson and F. H. C. Crick. Molecular Structure of Nucleic Acids : A Structure for Deoxyribose Nucleic Acid . *Nature*, 171(4356):737–738, apr 1953.
- [14] F. Sanger and H. Tuppy. The amino-acid sequence in the phenylalanyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. *Biochemical Journal*, 49(4):463–481, sep 1951.
- [15] Joel B. Hagen. The origins of bioinformatics. *Nature Reviews Genetics*, 1(3):231–236, 2000.
- [16] Matthew Cobb. 60 years ago , Francis Crick changed the logic of biology. *PLoS Biology*, 15(9):e2003243, sep 2017.
- [17] P. Edman and G. Begg. A protein sequenator. *European Journal of Biochemistry*, 1(1):80–91, mar 1967.
- [18] Margaret Oakley Dayhoff and Robert S. Ledley. Comprotein: A computer program to aid primary protein structure determination. In *Proceedings of the December 4-6, 1962, Fall Joint Computer Conference, AFIPS '62 (Fall)*, page 262–274, New York, NY, USA, 1962. Association for Computing Machinery.
- [19] W. M. Fitch and E. Margoliash. Construction of phylogenetic trees. *Science (New York , N.Y.)*, 155(3760):279–284, jan 1967.
- [20] L. Wang and T. Jiang. On the complexity of multiple sequence alignment. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, 1(4):337–348, 1994.
- [21] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd . *Science (New York , N.Y.)*, 269(5223):496–512, jul 1995.
- [22] Human genome project faq. [Online; accessed 8-April-2022].

- [23] International Human Genome Sequencing Consortium . Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, oct 2004.
- [24] Glen A. Evans. Designer science and the “omic” revolution. *Nature Biotechnology*, 18(2):127–127, feb 2000.
- [25] Vladimir A. Likić, Malcolm J. McConville, Trevor Lithgow, and Antony Bacic. Systems Biology : The Next Frontier for Bioinformatics . *Advances in Bioinformatics*, 2010:1–10, feb 2010.
- [26] The birth of the web. [Online; accessed 8-April-2022].
- [27] Charles E. Cook, Oana Stroe, Guy Cochrane, Ewan Birney, and Rolf Apweiler. The European Bioinformatics Institute in 2020: Building a global infrastructure of interconnected data resources for the life sciences. *Nucleic Acids Research*, 48(D1):D17–D23, jan 2020.
- [28] Zachary D. Stephens, Skylar Y. Lee, Faraz Faghri, Roy H. Campbell, Chengxiang Zhai, Miles J. Efron, Ravishankar Iyer, Michael C. Schatz, Saurabh Sinha, and Gene E. Robinson. Big Data : Astronomical or Genomical ? *PLoS Biology*, 13(7), jul 2015.
- [29] Rafael U. Ibarra, Jeremy S. Edwards, and Bernhard O. Palsson. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature*, 420(6912):186–189, nov 2002.
- [30] Ewen Callaway. ‘ It will change everything’: DeepMind ’s AI makes gigantic leap in solving protein structures. *Nature*, 588(7837):203–204, nov 2020.
- [31] Lindsay Barone, Jason Williams, and David Micklos. Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators. *PLoS Computational Biology*, 13(10):1–8, 2017.
- [32] Roberta Kwok. Ecology’s remote-sensing revolution. *Nature*, 556(7699):137–138, apr 2018.
- [33] Zeinab Mohammadzadeh, Reza Safdari, Marjan Ghazisaeidi, Somayeh Davoodi, and Zahra Azadmanjir. Advances in Optimal Detection of Cancer by Image Processing ; Experience with Lung and Breast Cancers . *Asian Pacific journal of cancer prevention: APJCP*, 16(14):5613–5618, 2015.
- [34] Stefan Kahl, Connor M. Wood, Maximilian Eibl, and Holger Klinck. BirdNET : A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61:101236, mar 2021.
- [35] Sandeep Ravindran. How DIY technologies are democratizing science. *Nature*, 587(7834):509–511, nov 2020.