

## Módulo 1

### 1.1. La Bioinformática. ¿Qué, para qué y cómo?

Por **Coral del Val Muñoz**

Titular de Universidad de Granada. Departamento de Ciencias de Computación e Inteligencia Artificial (DECSAI).

Por **Carlos Cano Gutiérrez**

Titular de Universidad de Granada. Departamento de Ciencias de Computación e Inteligencia Artificial (DECSAI)

---

#### 1. EL CONTEXTO: TSUNAMI DE DATOS

Desde sus orígenes, las investigaciones en Biología y la Medicina han tenido por objetivo unir datos y evidencias para tratar de entender el funcionamiento de los sistemas biológicos y las causas de las enfermedades. Un sistema biológico es una red de entidades biológicas que interactúan, y según la escala o resolución de nuestro estudio, las entidades biológicas de interés pueden ser una única célula, órganos y tejidos de un organismo, conjuntos de organismos o incluso ecosistemas completos. De este modo, estas áreas proponen investigaciones de alto interés en cuestiones relacionadas con Salud, Bienestar, Ecología, Energía, etc. A modo ilustrativo, algunas preguntas que se abordan en este tipo de investigaciones son:

- ¿cómo puede tratarse un suelo quemado para acelerar su recuperación?
- ¿es posible utilizar bacterias para generar y almacenar energía?
- ¿Cuál es la composición de especies de microorganismos en la flora intestinal humana? ¿Varía significativamente entre individuos? ¿Tiene relación con enfermedades?
- ¿Cómo se propaga un virus como SARS-Cov-2 y cómo afecta el COVID-19 al organismo infectado?
- ¿Es posible un diagnóstico precoz de la enfermedad de Alzheimer's? ¿y del cáncer de colon? ¿Es posible prevenirlas?

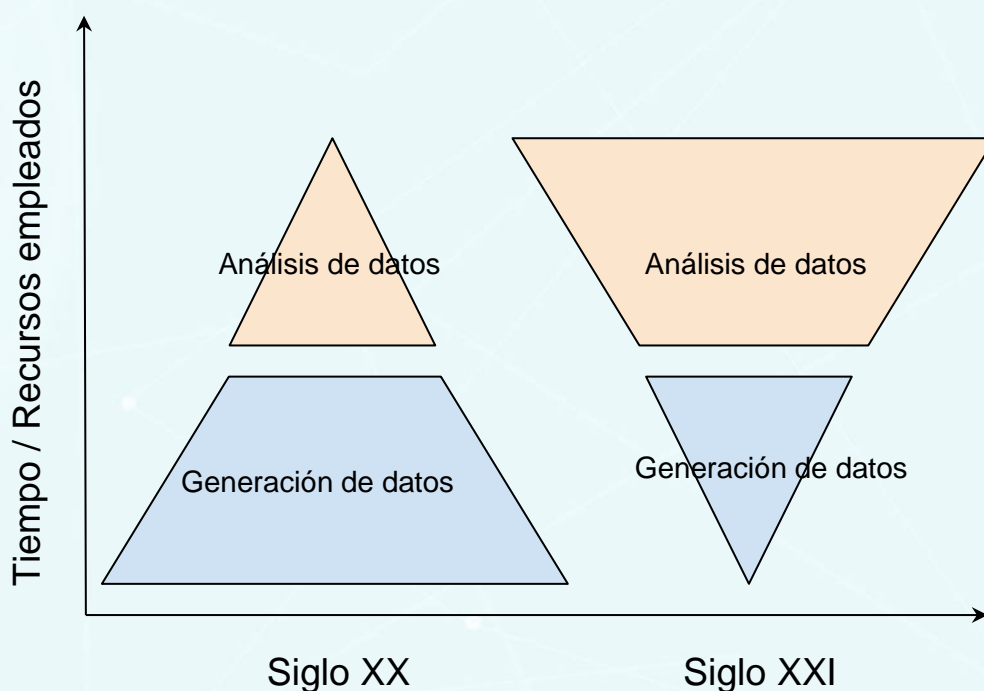
En el S.XX, la principal dificultad que encontraban los investigadores para hacer avances en estos campos era la falta de datos. Así, la mayor parte del esfuerzo de las investigaciones se dedicaba a generar datos, el volumen de datos generado resultaba manejable, y el análisis de

# MOOC MACHINE LEARNING Y BIG DATA PARA LA BIOINFORMÁTICA

los mismos se llevaba a cabo de forma manual o con herramientas informáticas y estadísticas básicas, como una hoja de cálculo.

Sin embargo, en las últimas dos décadas, se están produciendo enormes avances tecnológicos que están provocando una auténtica revolución en estas disciplinas. La tecnología está permitiendo observar o medir los sistemas biológicos con una precisión nunca antes vista y a un coste asequible y que sigue reduciéndose. Estos avances tecnológicos nos permiten, por ejemplo, medir la abundancia de distintas moléculas dentro de una única célula, identificar las especies de microbios de una muestra del entorno o monitorear con drones la expansión de una especie vegetal invasiva en un ecosistema.

Para ilustrar esta drástica reducción de costes alcanzada gracias a los avances tecnológicos, valga el siguiente ejemplo. En el año 2001 se completó la secuenciación del primer genoma humano tras un esfuerzo internacional de más de 300 millones de dólares. Hoy día, secuenciar un genoma humano cuesta en torno a los 1,000 dólares.



(Figura 1. Distribución del tiempo y recursos en investigación en BioCiencias y BioMedicina en el S.XX y S.XXI.)

Este abaratamiento de costes y el aumento del rendimiento de la tecnología está, provocando un auténtico *tsunami* de datos que está posibilitando enormes avances en estas disciplinas en unos pocos años. Sin embargo, las cantidades masivas de datos generadas para un mismo estudio (ipetabytes y hasta tera bytes!) son imposibles de analizar utilizando los medios tradicionales, lo que traslada la presión o el cuello de botella sobre el análisis. Y dado el enorme volumen de datos y su

complejidad, es obvio que estos análisis tienen que ser computacionales (utilizando ordenadores) y, además, requieren técnicas de análisis basadas en *Machine Learning* y *Big Data* (justamente, los contenidos que se tratan en este curso).

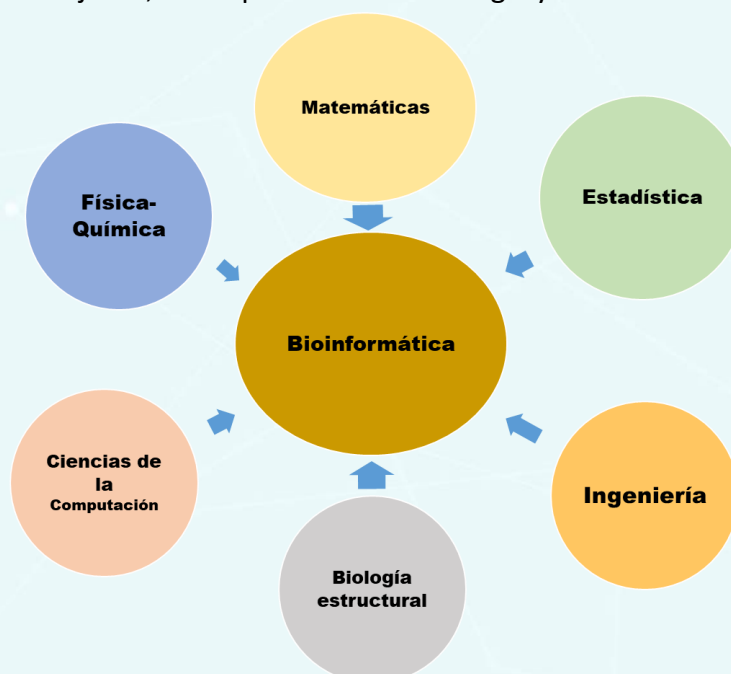
La Bioinformática surge en este contexto de revoluciones tecnológicas y con la necesidad de aplicar técnicas de computación eficientes y procesos automatizados al análisis de cantidades masivas de datos biológicos.

## 2. ¿QUE ES LA BIOINFORMÁTICA?

La bioinformática es un campo muy extenso y existen multitud de definiciones, por ejemplo, en Wikipedia se define como:

*“Campo interdisciplinar para el desarrollo de métodos y software para entender datos biológicos. La Bioinformática combina Biología, Ciencias de Computación, Ingeniería, Matemáticas y Estadística para analizar e interpretar datos biológicos”.*

Esta definición hace hincapié en el carácter multidisciplinar de la Bioinformática, que aúna distintas disciplinas tradicionalmente alejadas, como pueden ser la Biología y las Ciencias de Computación.



(Figura 2. Fuente: Elaboración propia)

# MOOC MACHINE LEARNING Y BIG DATA PARA LA BIOINFORMÁTICA

La definición también propone el principal objetivo de la Bioinformática: “desarrollo de métodos y software para *entender* datos biológicos” o “para *analizar* e *interpretar* datos biológicos”.

Una definición un poco más amplia es la que propone la revista *Nature*:

*“Campo de estudio que usa la computación para extraer conocimiento de datos biológicos. Incluye la adquisición, almacenamiento, recuperación y modelado para el análisis, visualización o predicción mediante el desarrollo de algoritmos o software”.*

Esta definición incide en los distintos procesos computacionales involucrados: *adquisición, almacenamiento, recuperación, modelado, análisis, visualización* y *predicción* de información o conocimiento.

Otra definición de Bioinformática, ésta muy visual, se muestra en la Figura 3: Bioinformática es la confluencia de las disciplinas: Ciencias de Computación, Biología y Matemáticas.



(Figura 3. Bioinformática es la confluencia simultánea de Matemáticas, Ciencias de Computación y Biología)

En cualquier caso, como en todo campo interdisciplinar, hay varias vertientes según el prisma con el que se mire. Así, desde el prisma de la Biología, la Bioinformática es “**Biología Computacional**”, es decir, una disciplina que implica el uso de ordenadores para analizar cualquier tipo de información biológica (p.ej. secuencias, imágenes de rayos X, mediciones clínicas, etc.). Mientras, desde el prisma de la Informática, la Bioinformática es “**Informática biológica**”, concepto que pone mucho más énfasis en la adquisición, almacenamiento y tratamiento de la información y el análisis de estos grandes volúmenes de datos.

A grandes rasgos, podemos definir los objetivos principales de la bioinformática como:

- **Organizar de manera eficiente grandes cantidades de datos** de biología, medicina, etc. (biociencias y biosalud en general). Este objetivo engloba la creación de bases de datos y repositorios de información sobre secuencias de ADN y genomas (como [Ensembl](#), [UCSC](#) o [GenBank](#)); proteínas ([UniProt](#)); transcriptomas ([GEO](#)); redes metabólicas ([KEGG](#)); bases de datos multifactoriales sobre enfermedades concretas ([The Cancer Genome Atlas](#)) y un larguísimo etcétera.
- **Diseñar y desarrollar herramientas y algoritmos para el análisis de dichos datos.** Uno de los ejemplos más exitosos de software bioinformático es [BLAST](#), utilizado para el análisis de similitud entre secuencias. Existen miles de herramientas y algoritmos desarrollados para el análisis de datos en **Biología** (biología molecular, agricultura, procesos farmacológicos, biotecnología de biomasa, energía renovables, desarrollo de vacunas, inmunología, microbiología, biotecnología alimentaria, fitomejora, impacto ambiental, mejora de la producción animal, ciencias forestales, etc.), **Química** (biotecnología química, toxicología, desarrollo de pesticidas) y **Medicina** (neurociencias, psiquiatría, nutrición, cálculo biomédico, cálculo de riesgo y diagnóstico de enfermedades, cáncer).
- **Desarrollar modelos que expliquen el funcionamiento de sistemas biológicos complejos.** Según la R.A.E., un modelo es una representación *teórica, generalmente en forma matemática, de un sistema o de una realidad compleja que se elabora para facilitar su comprensión y el estudio de su comportamiento*. Así, un objetivo de la bioinformática es inferir modelos, a partir de datos y algoritmos, para entender los mecanismos de regulación de estos sistemas para, por ejemplo, diagnosticar, prevenir o tratar una enfermedad.
- **Descubrir nuevo conocimiento** a partir de estos modelos y herramientas. Un ejemplo es el descubrimiento de nuevos marcadores (genes) capaces de diagnosticar de forma precoz un tipo de cáncer o la respuesta del paciente a un tratamiento.

- **Facilitar la interpretación precisa y significativa de resultados por parte de los expertos.** Este objetivo hace hincapié en la bioinformática como *habilitador* o asistente en la toma de decisiones. Este objetivo incide en la importancia de justificar el nuevo conocimiento descubierto y facilitar la interpretación y aplicación del mismo por parte de los expertos, por ejemplo, mediante visualizaciones que faciliten la interpretación del mismo.

### 3. LA MULTIDISCIPLINARIEDAD O QUE APRENDERÁS EN ESTE CURSO

Una conclusión evidente de todas las definiciones de Bioinformática es que se trata de un área multidisciplinar. Es decir, para formarse en éste área es importante tener nociones sobre los distintos *idiomas* implicados: principalmente Biología, Ciencias de Computación y Matemáticas. Nuestra experiencia en el área nos indica que estos perfiles profesionales y académicos mixtos son los más demandados y los que más enriquecen los equipos de trabajo en Bioinformática.

#### ¿Por qué?

A medida que avancemos contenidos en este curso, esperamos que esto resulte más y más evidente a tus ojos, pero, de inicio, piensa por un momento en lo que supone trabajar en un entorno científico multidisciplinar con perfiles de Biología, Medicina, Bio-Ciencias (Química, Ingenierías, Farmacia, Medio Ambiente, etc.), Matemáticas e Informática. Existe una clara *barrera terminológica* entre estas disciplinas. Uno de los principales retos en estos entornos multidisciplinares consiste precisamente en **formular necesidades de conocimiento de un área de forma que sean correctamente interpretadas por los profesionales de otras áreas**, que a su vez, tienen competencias que les permiten contribuir a satisfacer estas necesidades. Éstos últimos, a su vez, deben **comprender plenamente la naturaleza del problema** para identificar qué técnicas les permitirán aportar valor a la solución del mismo. Todo este proceso requiere lógicamente habilidades comunicativas en estos equipos, y es en estos entornos donde resultan especialmente valiosas las personas con capacidad para **hablar distintos idiomas, manejar la terminología que les permita formular cuestiones de interés en las distintas áreas, y conocer lo suficiente de las otras disciplinas para identificar y formular claramente qué pueden contribuir cada una en la resolución del problema.**

Este curso te permitirá adentrarte en la Bioinformática con énfasis en el *idioma* de las Ciencias de Computación, en particular del *Machine Learning* y del *Big Data*, pero sin alejarse de los problemas y aplicaciones reales del mundo de la Biología, Medicina y otras áreas de aplicación. Este curso te permitirá **descubrir lo que las Ciencias de Computación (*Machine Learning* y técnicas de *Big Data*) pueden contribuir en Bioinformática, qué tipo de problemas puedes resolver con estas técnicas y cómo hacerlo.**

Nuestro objetivo, por tanto, es darte unas primeras nociones para aprender a *hablar* Ciencias de Computación en Bioinformática. Nuestro deseo es que este curso te inicie o aumente tu formación multidisciplinar en Bioinformática para añadir valor a tu perfil en este campo. Además, en la bibliografía incluimos algunos recursos que te permitirán ampliar tu formación una vez terminado el curso.

#### 4. ORÍGENES DE LA BIOINFORMÁTICA: ANÁLISIS DE SECUENCIAS

En sus orígenes, la bioinformática surgió de la necesidad de comprender el código genético de los seres vivos con el objetivo de descubrir los mecanismos moleculares involucrados en los distintos procesos de desarrollo. Con el paso del tiempo, además del interés en las secuencias de moléculas como ADN, ARN o proteínas, surgió el interés en su estructura, sus interacciones con otras moléculas, los mecanismos de regulación entre las mismas, etc. Las herramientas bioinformáticas desarrolladas se han convertido en imprescindibles en numerosas disciplinas (Medicina, Agricultura, Nutrición, etc.). A continuación, se realiza un breve repaso por el origen de esta área y algunos de los problemas bioinformáticos más relevantes dentro de la Biología.

Como ya se ha adelantado, la Bioinformática surge originariamente de la necesidad de comprender el código genético. Esto implica que la Biología Molecular fue el primer *cliente* de la Bioinformática: la primera disciplina que generó tal volumen de datos que requirió de un nuevo *socio* para el análisis. Y el análisis de secuencias fue el primer problema que requirió esta necesidad de capacidades superiores de análisis y, por tanto, la primera aplicación de la Bioinformática.

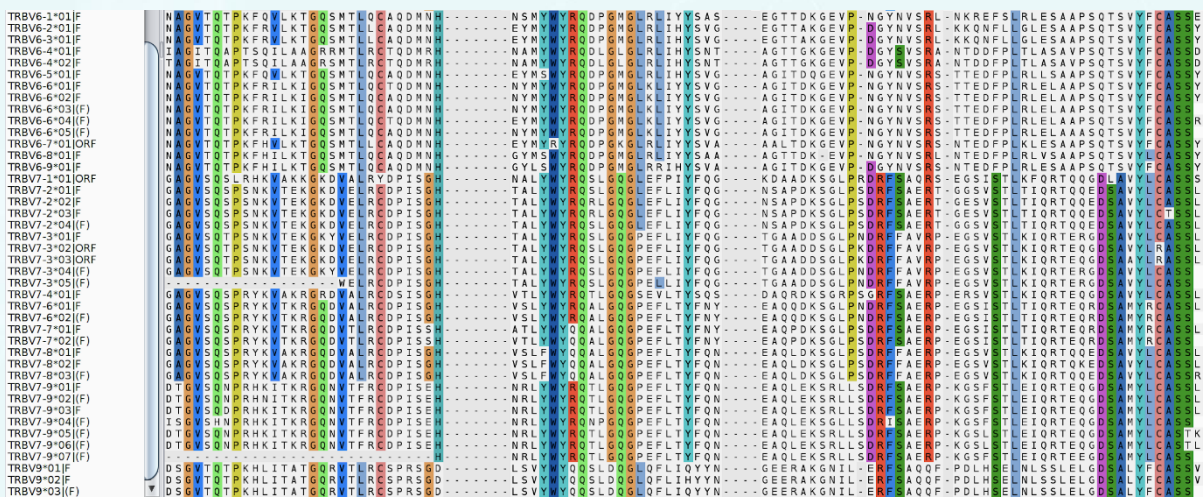
Para comprender el origen de la Bioinformática y estas primeras aplicaciones, es necesario contar con algunas nociones básicas de Biología Molecular. Por ejemplo, es importante conocer el denominado dogma central de la biología molecular. De forma simplificada, este dogma estipula que el flujo de la información genética fluye de ADN a ARN y a proteína. La información almacenada en el genoma (moléculas de ADN), en el núcleo de las células de un organismo, está codificada en secuencias de nucleótidos (Adenina, Citosina, Guanina y Timina -- A,C,G,T), es decir, en un lenguaje que utiliza un alfabeto de 4 símbolos. En un proceso biológico denominado transcripción, esta información se codifica en otro lenguaje, el de las secuencias de ARN (también nucleótidos, pero, en este caso, Adenina, Citosina, Guanina y Uracilo -- A,C,G,U). Las moléculas de ARN salen del núcleo de la célula y pasan al citoplasma, donde interactúan entre sí y con otras moléculas. En particular, algunas de estas secuencias (ARN mensajero o mRNA) interactúan con los Ribosomas, que las traducen a un nuevo tipo de moléculas, las proteínas, que codifican la información utilizando otro lenguaje: el de los aminoácidos, de 20 símbolos. Estas secuencias de aminoácidos adquieren una estructura tridimensional, que depende de la secuencia, y que les confiere una función dentro de la célula.

# MOOC MACHINE LEARNING Y BIG DATA PARA LA BIOINFORMÁTICA

Este proceso de transcripción y traducción de genes en proteínas debe entenderse como un proceso cuantitativo y sometido a constante regulación. De este modo, la presencia o abundancia de ciertas moléculas en la célula provocará / detendrá, la expresión de ciertos genes para producir cierto tipo de proteínas. Este proceso se denomina de forma genérica regulación de la expresión genética, y será el foco de atención en algunos programas de este curso.

Pues bien, la aparición de la bioinformática se remonta a la década de los 60, cuando Sanger y sus colaboradores desarrollaron un método para secuenciar unas de las moléculas participantes de este proceso: las proteínas. De ahora en adelante utilizaremos habitualmente el término *Secuenciar*, se trata, simplemente, de *determinar la secuencia* de cierto tipo de moléculas. De este modo, Sanger y colaboradores desarrollaron un método que permitía conocer la secuencia exacta de aminoácidos que conforman una proteína. Este avance experimental generó la necesidad de desarrollar nuevos algoritmos que permitieran analizar y comparar distintas secuencias de proteínas de distintos organismos, porque el volumen de secuencias disponibles impedía realizarlo manualmente. Así aparecieron los algoritmos de alineamiento múltiple, que buscan si dos secuencias son o no similares y qué regiones tienen en común, y se crearon las principales bases de datos como la ya mencionada

[GenBank](http://GenBank).

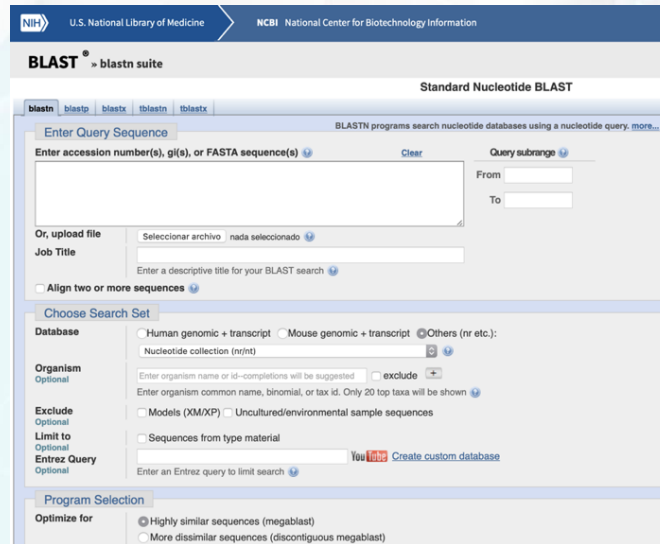


(Figura 4. Alineamiento múltiple de secuencias de aminoácidos. En un alineamiento múltiple, cada proteína se dispone en una fila y su secuencia se detalla añadiendo gaps (-) para destacar los aminoácidos que tienen las secuencias en común. Los aminoácidos de mayor consenso en cada posición (columna) se destacan con un código de colores)



# MOOC MACHINE LEARNING Y BIG DATA PARA LA BIOINFORMÁTICA

El despegue del análisis de secuencias coincidió con la aparición de internet en los años 90, lo que permitió la distribución más rápida de programas y descubrimientos, así como la aparición de páginas web que ofrecen a la comunidad estos análisis. Un ejemplo es la web del Centro Nacional de Información Biotecnológica de EEUU (NCBI).



(Figura 5. Página de BLAST, algoritmo que analiza millones de secuencias por segundo para encontrar la secuencia más similar a una secuencia dada)

Desde entonces, ha habido incontables avances en los métodos de análisis, comparación y visualización de secuencias moleculares. Sin embargo, fue la culminación del “Proyecto Genoma Humano” en 2001-2003 lo que supuso un punto de inflexión, al poner a disposición de la comunidad científica la primera secuencia completa de un genoma humano. Este acontecimiento dio inicio a la denominada era post-genómica.



(Figura 6. Portadas de Science y Nature sobre la publicación del primer draft del genoma humano en 2001)

## 5. LA ERA POSTGENÓMICA Y LAS CIENCIAS ÓMICAS

El hecho de tener por primera vez una secuencia de referencia (assembly) del genoma humano supuso el comienzo de la llamada era post-genómica. Esta era ha estado caracterizada por el advenimiento de nuevas tecnologías de secuenciación más baratas que han puesto al alcance de muchos la posibilidad de estudiar genomas de todo tipo (genómica): plantas, animales, microbios y humanos (e.g. 1000 Genome Project). Esta revolución técnica ha supuesto la generación de nuevo conocimiento sobre la estructura y funcionamiento de los genomas.

Este nuevo conocimiento junto con avances biotecnológicos ha propiciado la aparición y desarrollo de otras **ciencias ómicas**. Las distintas ómicas se centran en la caracterización y cuantificación de un gran número de moléculas (e.g. lípidos, microRNAs, proteínas, metabolitos, ect...) agrupadas de acuerdo a sus características biológicas, estructurales y/o funcionales.

Algunas de estas ciencias Ómicas y sus respectivos objetos de estudio se describen brevemente a continuación:

- *Genómica*: estudio de la estructura y función de los genomas. Por ejemplo, comprende el estudio de secuencias, mutaciones, variaciones en el número de copias, inserciones y deleciones, variaciones estructurales como translocaciones de trozos de cromosomas, etc.
- *Transcriptómica*: estudio de la expresión de todas las moléculas de ARN en una célula o colección de células bajo unas circunstancias concretas. Por ejemplo, comprende el estudio de expresión diferencial de genes, fusión de genes, splicing alternativo, edición de ARN, expresión de genes codificadores de proteínas y de genes de ARN reguladores, etc.
- *Epigenómica*: estudio de cambios químicos en el ADN y en las histonas (proteínas responsables de la compactación del ADN). Por ejemplo, esta disciplina comprende el estudio de la metilación del ADN, modificación de histonas por de/acetilación, unión de factores de transcripción, etc. Podemos decir que el epigenoma añade marcas reversibles sobre el genoma y que estas marcas afectan a la activación/desactivación de la expresión de los genes. La epigenética pone el foco en estas marcas.
- *Proteómica*: colección de proteínas en una célula, tejido u organismo.
- *Metaboloma*: colección de pequeñas moléculas químicas, llamadas metabolitos (por ejemplo, hormonas) en una célula, tejido u organismo. Estos metabolitos son productos intermediarios encargados de funciones de señalización, estimulación, inhibición de enzimas o interacción con otros organismos (ej. pigmentos y feromonas).
- *Microbioma*: colección completa de microbios presentes en un organismo.
- *Metagenoma*: material genético completo obtenido de una muestra medioambiental (ej. laguna, nieve, intestino, piel, mucosa bucal).

# MOOC MACHINE LEARNING Y BIG DATA PARA LA BIOINFORMÁTICA

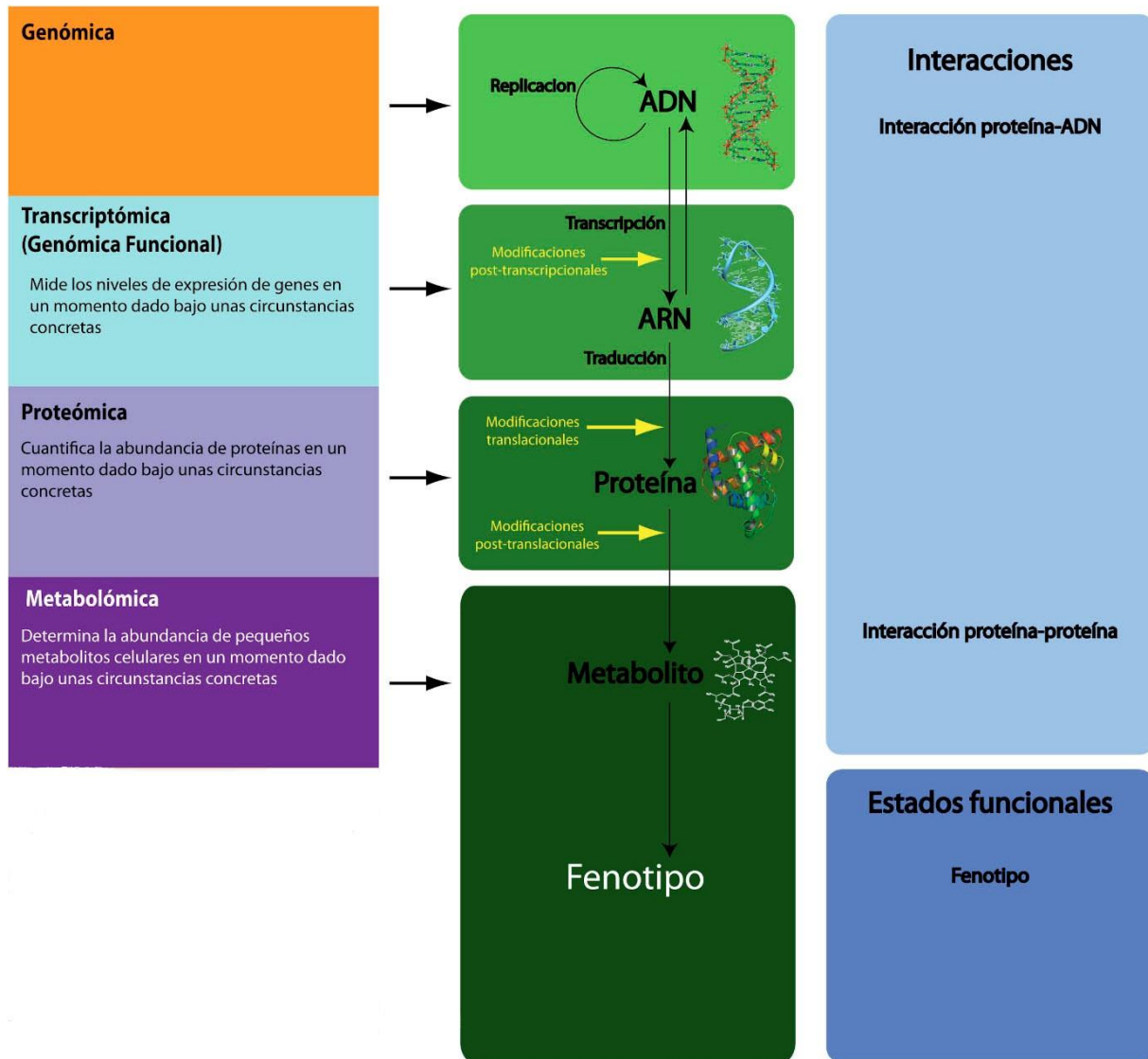
- **Lipidómica:** es la investigación de los lípidos celulares en sistemas biológicos. El lipidoma es parte del metaboloma y utiliza las mismas herramientas que la metabolómica, pero es especialmente importante en enfermedades cuya patogénesis está relacionada con el metabolismo lipídico como la obesidad o la hipertensión.
- **Glicómica:** es el estudio de los glúcidos (azúcares). Estos compuestos se generan en rutas metabólicas muy complejas y suelen unirse a otros elementos para formar glucoproteínas (importantes para el reconocimiento de célula a célula) o glucolípidos (importantes para la estabilidad celular). Distintos cánceres tienen distintos perfiles de glúcidos.
- **Fenotipo:** colección de características observables o medibles de un determinado organismo. El fenotipo es fruto de la expresión del genotipo en un determinado ambiente.

Las ómicas y sus datos (omas) han permitido tener una fotografía más completa y compleja de la regulación genética que la que ofrecía el dogma central de la biología molecular. En efecto, los avances en estas disciplinas han mostrado que no sólo las proteínas interactúan entre sí, con el ARN y con el ADN, para regular la transcripción, sino que el ARN también tiene un papel directo en la regulación génica. Incluso se ha puesto de manifiesto que el flujo de la información no sólo va de ADN a ARN, sino también en sentido contrario. Ha permitido el descubrimiento de nuevos elementos clave en el funcionamiento de la expresión génica: moléculas de ARN desconocidas hasta el momento como ARNs largos no codificantes (lncRNAs), micro ARNs (miRNAs), o activadores en *cis* de la transcripción como los *enhancers*.

Las ciencias ómicas han permitido la generación de una cantidad de datos que junto a avances importantes en matemáticas computacionales han hecho posible por primera vez el uso de estrategias de aprendizaje automático (machine learning) y big data para el análisis y estudio de los sistemas biológicos, con resultados impensables hace solo veinte años y propiciando la extensión de la Bioinformática a otras áreas de investigación.

En la próxima cápsula, describiremos algunas de las áreas de aplicación más relevantes de la Bioinformática, tanto las relacionadas con las Ciencias Ómicas como las que surgen de otras disciplinas.

# Omas y Omicas



(Figura 7. Diagrama que muestra la relación entre distintas Ciencias Ómicas y sus objetos de estudio)

# MOOC MACHINE LEARNING Y BIG DATA PARA LA BIOINFORMÁTICA

## REFERENCIAS BIBLIOGRÁFICAS

- **Introduction to Bioinformatics.** A. Lesk. OUP Oxford, 2014. ISBN 0199651566, 9780199651566
- **Recurso web:** The cost of Sequencing a Human Genome.  
<https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost>
- **Recurso web:** Bioinformatics for the terrified. C. Brooksbank , A. Cowley. EMBL-EBI. doi: 10.6019/TOL.BioinfTer-c\_2016.00001.1
- **Recurso web:** Proyecto colaborativo LibreTexts: Introducción a la Biología y Biología Molecular (CK-12)  
[https://bio.libretexts.org/Bookshelves/Introductory\\_and\\_General\\_Biology/Book%3A\\_Introductory\\_Biology\\_\(CK-12\)/04%3A\\_Molecular\\_Biology](https://bio.libretexts.org/Bookshelves/Introductory_and_General_Biology/Book%3A_Introductory_Biology_(CK-12)/04%3A_Molecular_Biology)
- **Bioinformatics Curriculum Guidelines:** Toward a Definition of Core Competencies. L. Welch et al. PLoS Comput Biol 10(3): e1003496. <https://doi.org/10.1371/journal.pcbi.1003496>
- **Recurso web:** Careers in Bioinformatics. International Society for Computational Biology. <https://www.iscb.org/bioinformatics-resources-for-high-schools/careers-in-bioinformatics>
- **Searls DB (2014) A New Online Computational Biology Curriculum.** PLoS Comput Biol 10(6): e1003662. <https://doi.org/10.1371/journal.pcbi.1003662>