Primera parte
Secuencias génicas y de proteínas

Capítulo 1.

¿Cómo comenzar a trabajar?

1.1. Introducción

Todas las ciencias básicas han experimentado una revolución con el descubrimiento y desarrollo de diferentes herramientas para su estudio. La biología no se ha escapado de lo anterior. Por su parte, la introducción de la informática al estudio de las ciencias de la vida está produciendo cambios en la forma en que los biólogos percibimos la naturaleza. Por desgracia, en la mayoría de los casos nos hemos mantenido al margen del desarrollo de las principales herramientas informáticas utilizadas actualmente para el análisis de la información molecular de los organismos. Es cierto que el desarrollo de algoritmos para la elaboración de diferentes softwares en bioinformática hace uso de las matemáticas; sin embargo, esto no debería ser un impedimento para su desarrollo, uso e implementación por parte de los biólogos o de estudiantes de carreras afines.

Figura 1.1.1. Relevancia actual del manejo de datos

En este libro queremos acercar a los estudiantes de las ciencias de la vida (biología, medicina, enfermería, agronomía, etc.) a la bioinformática (informática

aplicada a la biología), ciencia que recolecta, organiza, analiza, manipula y presenta datos biológicos. De esta forma, pretendemos mostrar cómo, con un poco de dedicación y estudio, se pueden obtener resultados asombrosos de la información codificada en el ADN (ácido desoxirribonucleico), RNA (ácido ribonucleico) y las proteínas, pilares del mundo molecular en los seres vivos. En caso de que no tenga claro que es el DNA desde el punto de vista químico, visite el sitio web del National Human Genome Research Institute (2022):

Por otro lado, no existe en la actualidad un texto que cubra múltiples aspectos de la bioinformática de forma práctica o los pocos que se pueden conseguir en las librerías están desactualizados y la mayoría de sus links no están activos o migraron a otros sitios, por lo cual el estudiante se ve en problemas para encontrar la fuente de la información que el libro presenta. Algunos libros de este tipo son *Understanding Bioinformatics* de Zvelebil y Baum (2008) o *Biostatistics For Dummies* de Pezzullo (2013).

1.2. ¿Qué debe esperar el lector?

Dado su carácter aplicado en temas de bioinformática, en este libro encontrará direcciones https//: para explorar en la web. En ocasiones estas direcciones llevan a artículos científicos de acceso libre, por lo que recomendamos leerlos para dar el contexto adecuado a los temas que trataremos. En otras oportunidades estas direcciones llevan a ejemplos de los temas tratados. Por ejemplo, sitios de *software* de acceso libre o páginas de sitios donde se pueden consultar bases de datos. Algunas direcciones de enlace pueden aparecer rotos después de un tiempo ya que la internet es un sitio dinámico en permanente cambio, remodelación y actualización, por lo que hemos tratado de agregar los nombres completos de los sitios que direccionan a estos links, de tal manera que el estudiante pueda buscar el portal requerido y dar con la nueva dirección web.

1.3. Breve reseña histórica

La bioinformática es una ciencia de naturaleza interdisciplinaria que nació de los bioquímicos, matemáticos e ingenieros que desarrollaban *software* para resolver problemas en aplicaciones de manejos de datos y en el modelaje de fenómenos biológicos. Sus inicios se remontan a los comienzos de los años sesenta con los bioquímicos Margaret O. Dayhoff y Russell F. Doolittle, quienes trabajaban en el Centro de Genética Molecular de la Universidad de California en San Diego (USA) y desarrollaron algunos algoritmos para el manejo de bases de datos y fueron los primeros en utilizar computadores para hacer un atlas de estructuras y secuencias de proteínas (Hagen, 2011). En relación con la aparición de publicaciones sobre bioquímica en la web, aunque la primera vez que se reporta el uso del término bioinformática fue en 1991, su origen es incierto y es punto de debate permanente.

Una de las bases de datos más importante en la bioinformática es la base del NCBI (National Center for Biotechnology Information), fundada en 1988 por orden gubernamental y como parte del National Institutes of Health o NIH (USA). Así pues, la bioinformática tiene muy pocos años de vida y está en pleno desarrollo.

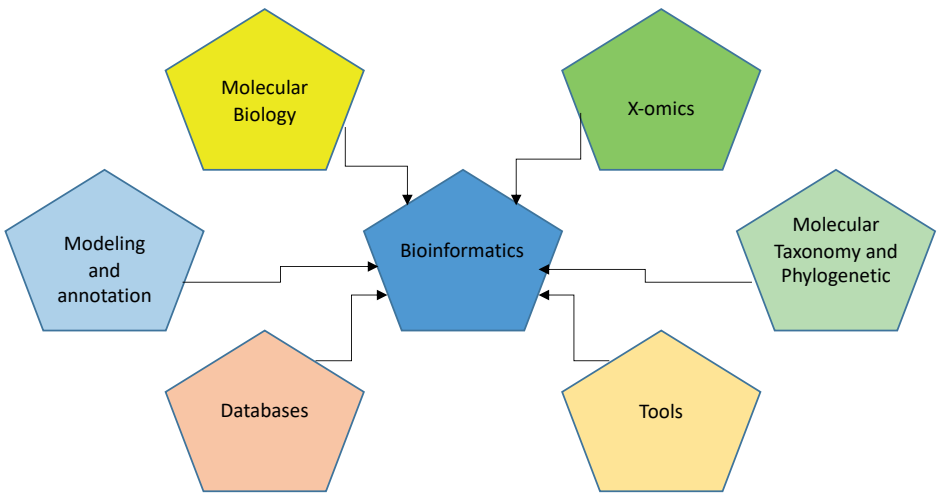
En los últimos años, la bioinformática como ciencia participa en múltiples campos de las ciencias aplicadas. Como ejemplo de lo anterior, citamos: la Medicina Molecular (Bujard y Grivell, 2009), la medicina personalizada (Agyeman y Ofori-Asenso, 2015), la terapia génica (Tipanee *et al.*, 2017), el desarrollo y diseño de nuevas drogas (Kaitin, 2010) e, inclusive, el desarrollo de aplicaciones al estudio del cambio climático en el planeta (Batley y Edwards, 2016).

1.4. La internet

La internet es el instrumento de trabajo más importante de la bioinformática. La internet, más precisamente el WWW (World Wide Web) o red informática mundial, es un sistema de manejo de hipertextos a los cuales se tiene acceso

a través de un navegador web (Explorer, Chrome, Mozilla, etc.). Fue desarrollado por los militares estadounidenses en los sesenta, pero su verdadero auge se dio entre los años 1989 y 1990 con la aparición del Instituto CERN en Suiza (French Conseil Européen pour la Recherche Nucléaire u Organización Europea para la Investigación Nuclear); allí se desarrollaron los https// o "Hypertext Transfer Protocol Secure" para la transferencia de forma segura de hipertextos. Un hipertexto es una forma de enlazar y compartir información (gráficas y textos) de diferentes fuentes por medio de redes. Para profundizar en el tema de la historia de la bioinformática puede consultar este acceso libre de Sabu Thampi (s.f.) de LBS College of Engineering en Kerala (India).

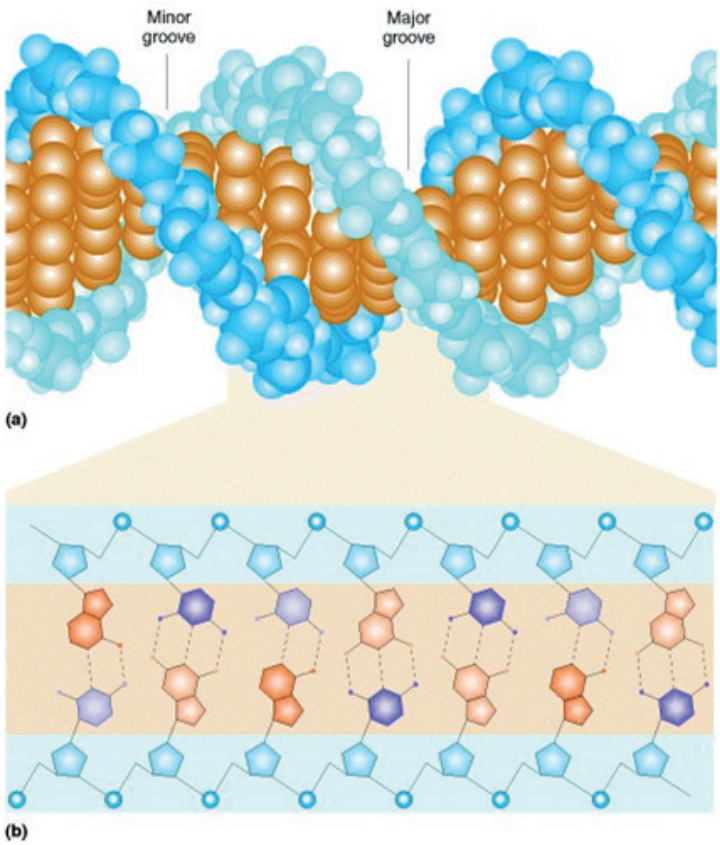
Figura 1.4.1. Áreas biológicas e informáticas que alimentan la bioinformática.



Es importante anotar aquí que este texto se ideó principalmente para usuarios de sistemas operativos (SO) Windows® o Apple Mac®, no para Linux®. Aunque la mayoría de los temas tratados aquí se pueden aplicar libremente a los SO Linux®, algunas aplicaciones mencionadas aquí no están diseñadas para correr en este SO. Para trabajar en este curso cada estudiante debe tener un e-mail (correo electrónico), ya sea personal o institucional, donde podrá, por ejemplo, descargar información que se esté procesando en una plataforma

remota. Si quiere abrir una cuenta personal de e-mail puede hacerlo en las diferentes plataformas internacionales como Gmail, Hotmail, Yahoo!, etc.). Como ejemplo y guía pueda visitar este sitio de Gmail (2021)

Figura 1.4.2. Modelo de la estructura del ADN, (Freeman, 2000).



Igualmente, se debe tener acceso a la Internet a través de un prestador de servicios (la empresa que vende el servicio al usuario), ya sea por WiFi® (sistema de interconexión inalámbrica o Wireless Fidelity), marca registrada para un sistema de interconexión o, en su defecto, por cable (Ethernet). Independientemente del tipo de conexión, usted deberá estar conectado a un servidor Web o HTTP que es un programa que relaciona y conecta máquinas (otros servidores y su máquina o computador). El equipo que utilice para

trabajar en bioinformática debe tener capacidad para procesar textos y gráficas, memoria ROM (Read Only Memory) disponible para procesar *software* que se descargue en su equipo para trabajar y suficiente memoria RAM (Random Access Memory), que es la memoria que utilizamos para procesar datos. Los teléfonos celulares o algunas *Palms* no son aptos para este trabajo por el volumen de información que se maneja.

No sobra aclarar que el idioma oficial del curso es inglés, ya que la mayoría de los sistemas internacionales de información están escritos en este idioma y son muy escasos los sistemas de información en otros idiomas, como el español. No obstante, algunas plataformas tienen la opción de escoger el idioma. Por último, para fines del manejo de este libro, asumiremos que usted desconoce algunos temas de biología molecular y de genética molecular, pero tiene un curso básico de bioquímica, de tal manera que no sabe que es un intron o un haplotipo, pero sí conoce la estructura química de los ácidos nucleicos, las proteínas, los lípidos y los carbohidratos. Por tanto, cuando nos refiramos a términos técnicos de la genética molecular, por lo general serán explicados de forma sucinta.

1.5. ¿Qué es un dato en biología?

Una de las tareas importantes en la biología ha sido por años conseguir, procesar, ordenar y guardar adecuada y ordenadamente datos. Los datos pueden provenir de un experimento científico, de la exploración de un sistema, el producto de un cálculo matemático o una estadística o de un evento particular (por ejemplo, las consecuencias de una inundación por una tormenta en plantas terrestres). Los datos pueden aparecer en informes técnicos, revistas científicas y catálogos de datos, archivados de diferente forma y con formatos diferentes.

El dato puede ser numérico, alfabético o alfanumérico, de tal forma que es susceptible de ser codificado para su compilación. Algo muy importante en el manejo de datos es la "content curation" o curación de contenidos de los datos

biológicos, es decir, que los datos que voy a trabajar o estoy produciendo deben estar curados o localizados en un lugar determinado, organizados apropiadamente, corroborados si son datos de experimentos científicos por duplicación del experimento y distribuidos de tal manera que sean útiles para propósitos de comparación. Para consultar las normas internacionales de manejo de datos acceda a Dama International. Cómo interpretar los datos obtenidos de las bases de datos ya es problema del investigador y de las herramientas que utilice para el análisis. Para dar un mayor contexto sobre un dato en biología, consulte la página del California Institute of Technology (Coltech magazin, 2019).

1.6. Algunos tópicos importantes relacionados con la bioinformática

Genómica: campo de la biología que involucra áreas diferentes (genética, bioquímica, fisiología, etc.) y que estudia la composición, la estructura y la función del genoma, conjunto de genes de un organismo (WHO, 2021). Como complemento, vea la estructura de un gen (Polyak y Meyerson, 2003). Entre las ramas de estudio de la genómica, está la genómica funcional que permite explicar la función de un gen, de su proteína, o su transcrito (RNA), que no necesariamente se traduce. La genómica estructural explica la función de un gen o una proteína desde el punto de vista de su estructura 3D, bajo el paradigma biológico una estructura / una función.

Secuencia: los ácidos nucleicos ADN o RNA son polímeros de nucleótidos (monómeros), así como las proteínas (polímeros) están formadas por aminoácidos (monómeros). La distribución ordenada de los monómeros en un polímero se denomina una secuencia. Las secuencias se obtienen por métodos bioquímicos.

Ensamblaje: método para reunir la información de secuencias parciales de una estructura génica (gen o genoma), de tal forma que se puede determinar

su estructura original. Si se hace por primera vez sin un genoma de referencia se denomina "de novo assembly" (Illumina, 2015a).

Figura 1.6.1. El ADN y el ARN son polímeros de nucleótidos y el ADN, a diferencia del ARN, no tiene la base uracilo aparte de otras diferencias químicas (Alberts, 2003).

Anotación: las anotaciones se hacen sobre los genes, los mRNAs (RNA que se traduce en proteínas) y las secuencias de las proteínas se utilizan para relacionar una secuencia con su historia biológica (composición bioquímica, genética y función). Por ejemplo, la anotación se utiliza para identificar genes de un genoma que está secuenciado *de novo* (Koonin y Galperin, 2003). A continuación, la secuencia de ADN y su correspondiente secuencia de aminoácidos de la proteína del gen de la insulina humana (AD: AH002844.2 del GenBank de NCBI):

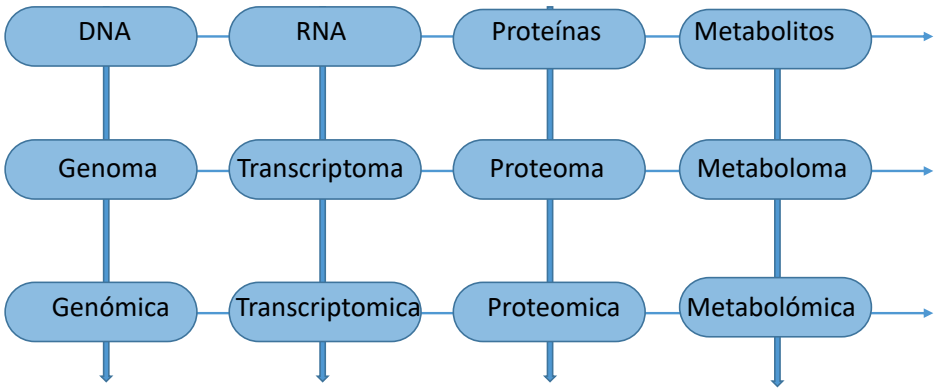
GCATTCTGAGGCATTCTCTAACAGGTTCTCGACCCTCCGC
CATGGCCCCGTGGATGCACTCCTCACCGTGCTGGCCCTG
CTGGCCCTCTGGGGACCCAACTCTGTTCAGGCCTATTCCA
GCCAGCACCTGTGCGGCTCCAACCTAGTGGAGGCACTGTA
CATGACATGTGGACGGAGTGGCTTCTATAGACCCCACGAC
CGCCGAGAGCTGGAGGACCTCCAGGTGGAGCAGGCAGAAC
TGGGTCTGGAGGCAGGCGGCCTGCAGCCTTCGGCCCTGGA
GATGATTCTGCAGAAGCGCGGCATTGTGGATCAGTGCTGT
AATAACATTTGCACATTTAACCAGCTGCAGAACTACTGCAA
TG TCCCTTA GACACCTGCCTTGGGCCTGGCCTGCTGCTCTGC
CCTGGCAACCAATAAACCCCTTGAATGAGMSKFLLQSHSANA

CLLTLLLTLASNLDISLANFEHSCNGYMRPHPRGLCGEDLHVI
ISNLCSSLGGNRRFLAKYMVKRDTENVNDKLRGILLNKKE
AFSYLTKREASGSITCECCFNQCRIFELAQYCRLPDHFFSRISR
TGRSNSGHAQLEDNFS

Epigenómica: es el estudio de las modificaciones químicas que se producen en los nucleótidos de los ácidos nucleicos y en los aminoácidos de las histonas (proteínas unidas al ADN y que dan forma a los cromosomas y regulan su función). Estas modificaciones tienen como función bioquímica activar o desactivar genes o los mecanismos de regulación de la expresión génica. Los cambios químicos (metilación, alquilación, miristilación, etc.) que se producen en las histonas y en el DNA pueden ser reversibles (NIH, s.f.)

Metagenómica: estudia la composición génica de las poblaciones. En la genómica se hace referencia a un solo individuo o al genoma de un solo individuo. Cuando se estudia la composición génica de una población se hace metagenómica.

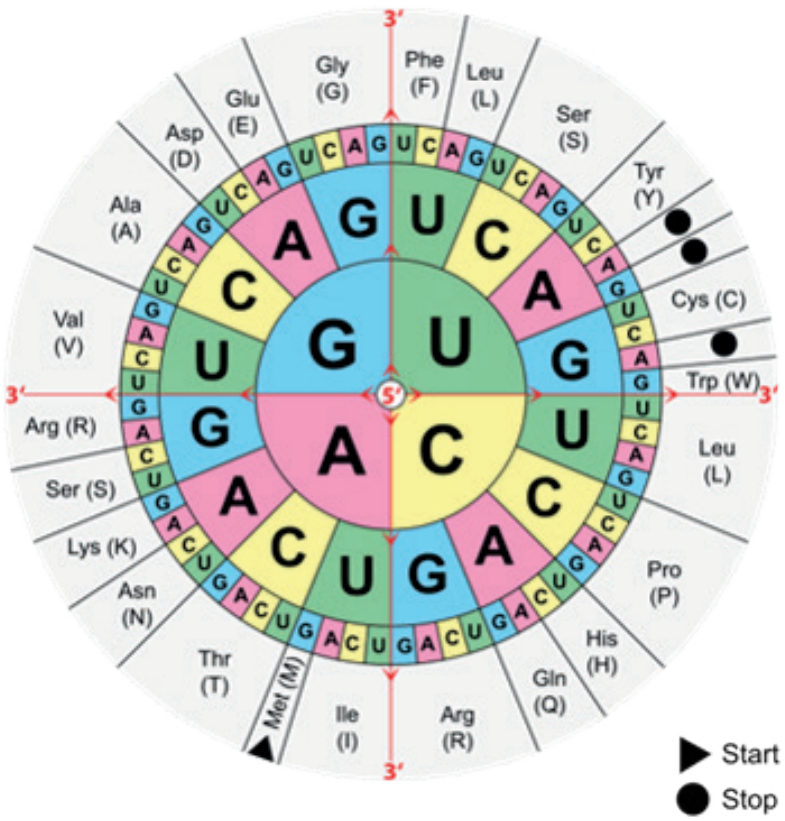
Figura 1.6.2. La relación de las especies moleculares de las células vivas (ADN, ARN, proteínas y metabolitos) con las diferentes áreas de estudio.



Ciencias ómicas: es un conjunto de ciencias que estudia la complejidad de la información biológica y se dividen en genómica (ver arriba), que estudia la estructura génica de un organismo. La transcriptómica hace referencia a todos los elementos génicos que son transcriptos en forma de ARN o Transcriptoma (ARNr, ARNt, ARNm, ARNi, miRNA), también llamado expresoma. Para profundizar en el tema, consulte Blackburn, (2017) y NaturePortafolio (2022).

La proteómica es el estudio del conjunto de proteínas que son traducidas con base en el transcriptoma de la célula. Por su parte, la metabolómica estudia los fenómenos bioquímicos que se dan en un organismo por la interacción de los compuestos químicos en la célula, también llamado metabolismo celular o metaboloma. Para saber más sobre el tema, consulte EMBL-EBI (2021).

Figura 1.6.3. El código genético (Wikimedia commons, 2017a).



Entry: es una aplicación de bioinformática, un set de datos o una unidad básica de información que se caracteriza por un tipo de contenido con un formato particular (Feagan *et al.*, 2007). Es común confundir un "Entry" con un "Query", este último es un programa de consulta de información particular o específica. Un término similar que se usa son los códigos de entrada de una secuencia (ADN, ARN o proteína) denominados ID, por sus siglas en inglés "Identification number". Lo utilizan las bases de datos para hacer referencia a una secuencia en particular y normalmente es un número.

Código genético: el ADN está construido con nucleósidos de desoxiadenosina, desoxiguanosina, desoxitimidina y desoxicitidina que se transcriben a nucleósidos de adenosina, guanosina, uridina y citidina o ARN (Nature, 2014).

Estructura 3D: la función de una proteína está determinada por su estructura tridimensional que se adquiere cuando la proteína es sintetizada y las fuerzas intermoleculares (enlaces de tipo Van Der Vaals como puentes de hidrógeno, puentes salinos o espacios hidrofóbicos) que actúan sobre los átomos de la molécula determinan su estructura tridimensional (How to: View the 3d Structure of a Protein, s.f.; Structure, s.f.). Algunas moléculas de ARN tienen estructura tridimensional y su función depende de ella; tal es el caso de las ribozimas (HERSCHLAG LAB, s.f.)

Código de aminoácidos: así como solo existe una forma de leer una secuencia de ADN o de ARN, existen dos formas diferentes de leer una secuencia de proteínas: una es el código de una letra, por ejemplo, L es leucina; y el código de tres letras, más fácil de recordar, por ejemplo, leucina es Leu.

Tabla 1.6.1. Códigos de los veinte aminoácidos más comunes en las proteínas que mantienen su derivación del nombre en inglés. Fuente: elaboración propia.

No.	Nombre completo	Código una letra	Código dos letras
1	Alanina	A	Ala
2	Arginina	R	Arg
3	Asparagina	N	Asn
4	Ácido aspártico	D	Asp
5	Cisteína	C	Cys
6	Glutamina	Q	Gln
7	Acido glutámico	E	Glu
8	Glicina	G	Gly
9	Histidina	H	His
10	Isoleucina	I	Ile
11	Leucina	L	Leu
12	Lisina	K	Lys
13	Metionina	M	Met
14	Fenilalanina	F	Phe
15	Prolina	P	Pro
16	Serina	S	Ser
17	Treonina	T	Thr
18	Triptófano	W	Trp
19	Tirosina	Y	Tyr
20	Valina	V	Val

Palíndromes: cadenas dobles de ADN o RNA que contienen la misma información al leer en dirección 3′ a 5′ que al hacerlo en dirección 5′ a 3′ (Smith, 2008). Estas secuencias se encuentran comúnmente en las bacterias que son blanco de la actividad de endonucleasas (enzimas que cortan el ADN). Un ejemplo de ello se ve a continuación:

5'- G A A T T C -3'
3'- C T T A A G -5'

Un tesaurus extendido de términos y acrónimos en bioinformática puede consultarse en Omicstutorials.

1.7. ¿Qué hacer?

En todos los capítulos de este libro el estudiante encontrará un aparte titulado "¿Qué hacer?". Allí habrá ejercicios prácticos de bioinformática que puedes utilizar, o bien con la información que te proporcionamos en el capítulo o con tu propia información de secuencias de ácidos nucleicos y proteínas.