

# Tema 3

## Introducción a la Bioinformática

**Docente: Danilo Ceschin**

### Actividades:

#### **Biología, Bioinformática y Biología Computacional**

1. Existen controversias y similitudes entre las definiciones de Bioinformática y de Biología Computacional que puede ser evidenciada en la siguiente definición:

**“La **biología computacional** es a veces definida como sinónimo de **Bioinformática** y a veces como una disciplina emparentada, pero distinta, de esta. El NIH define a ambas disciplinas como **distintas aunque con cierto grado de solapamiento**, según esta definición la bioinformática está más relacionada con el desarrollo de herramientas computacionales con el fin de analizar y procesar datos y la biología computacional con el estudio por medios computacionales de sistemas biológicos”**


Lea cuidadosamente los textos de las siguientes páginas webs y discuta con sus compañeros si hay o no diferencia entre Bioinformática y Biología Computacional.


#### **¿Deberían fusionarse los términos y por consiguiente su definición?**

Existen dudas sobre si existe alguna diferencia entre la biología computacional y la bioinformática. Si bien como dice el NIH presentan cierto solapamiento ya que ambas usan términos como informática y computacional y de algún modo ambas palabras implican el uso de computadoras aplicadas a la Biología, son disciplinas que tienen distintos enfoques de resolución y manejo de problemas del área de la biología.


Si analizamos la biología computacional, tenemos que saber que el término computacional no implica necesariamente el uso de computadoras, sino que implica cómputo, calculaciones y el uso de técnicas numéricas en el área de la biología. Si el interés es estudiar problemas biológicos y donde la hipótesis de trabajo puede ser probada por medio de la simulación y el modelado computacional, será considerada como perteneciente al campo de la biología computacional.

La biología computacional consiste fundamentalmente en el desarrollo de algoritmos, modelos y simulaciones matemáticas para facilitar el entendimiento de problemas biológicos. La biología computacional es extremadamente

multidisciplinaria, abarca esencialmente fuertes conocimientos de: matemática, química y bioquímica, biología molecular, programación, física, estadísticas entre otras áreas del conocimiento humano. 

La bioinformática, por otra parte, se ocupa de la aplicación de la informática a la recopilación, almacenamiento, organización, análisis, procesamiento, manipulación, presentación y distribución de información relativa a los datos biológicos o médicos. Es una disciplina menos interdisciplinaria, o sea posee un menor conocimiento de matemática, estadística y programación. El objetivo es crear herramientas útiles que funcionen con datos biológicos. Además, su formación académica está formada por un fuerte componente de programación en diferentes lenguajes, desarrollo y administración de bases de datos, conocimientos profundos de Sistemas Operativos como UNIX y Linux , administración de redes y por supuesto un entendimiento de la problemática biológica, entre otros. 

En conclusión, podemos decir que ambas disciplinas impulsan a dar mejores respuestas a los problemas biológicos del mundo. Sin embargo, como se detalla anteriormente no son disciplinas semejantes sino que trabajan de forma acoplada. Funcionan como equipo y hasta podríamos decir que la bioinformática podría existir sin la biología computacional.

Podemos decir que los problemas se dividen en dos: por una parte el almacenaje, gestión de datos, el estructuramiento, comprensión de los requisitos del problema que debemos manejar para lograr acceder a toda esta información de manera eficaz, el procesamiento y modelización de los datos. Allí es cuando entra en juego la bioinformática. Por otra parte, los datos deben ser analizados, entendidos de manera lógica a través del cómputo, junto con herramientas estadísticas, matemáticas, físicas, entre otras para poder extraer información útil de ellos. Allí es cuando entra en juego la biología computacional, que juega un papel más teórico que práctico. 




**2. Explique brevemente cómo se aplican las 4V del “Big Data” a las siguientes Bases de Datos Biológicas. Para ello, indague sobre el tipo de datos, estructura de datos, curado de los datos, etc.**

Desarrolle las 4V para:

\*. GeneBank: <https://www.ncbi.nlm.nih.gov/genbank/>

### **GenBank**

- **Volumen:** almacena los datos en diferentes bases de datos dependiendo de el tipo de dato:
  - Datos de ADN (GenBank)
  - Datos de Secuenciamiento Crudo e Información de alineamiento (SRA)


- Variaciones simples de nucleotidos humanos (dbSNP)
  - Variaciones estructurales genomicas humanas (dbVar)
  - Datos de genomica funcional (GEO)
- 
- **Variedad:** GenBank solo acepta datos con información predefinida por los proveedores de información, toda la data es analizada para que ningun ejemplo tenga datos faltantes. Para ello provee herramientas de software que siguen los lineamientos de información 
  
  - **Velocidad:** El personal del NCBI ha establecido un procedimiento operativo estándar para procesar las presentaciones de HTG. Las secuencias HTG enviadas al sitio FTP se procesan a diario. Los archivos se transfieren desde el directorio SEQSUBMIT del sitio FTP a las 4 a.m., hora estándar del este (EST) y se examinan y procesan el mismo día. Los envíos exitosos se envían inmediatamente a la base de datos. Cualquier error identificado se comunica al centro de envío por correo electrónico. A las 6 pm. EST cada día, un archivo .ac4htgs y .GBFF se colocan en el directorio REPORT del sitio FTP para cada centro de envío, con una lista de las secuencias que se procesaron por completo y se depositaron en GenBank. Estas secuencias están disponibles en el sistema Entrez y las bases de datos BLASTable dentro de, en promedio, 48 horas, pero pueden recuperarse por el número de acceso inmediatamente. 
  
  - **Veracidad:** el curado de los datos en GenBank parte de 3 premisas:
    - Los laboratorios se comunican con NCBI y se informan sobre el proceso HTG de envio de datos.
    - Los datos enviados a través de HTGS van a ser publicados inmediatamente después de haber sido procesados.
    - La información enviada incluye:
      - Nombre del centro de genoma
      - Nombre de la secuencia

\*. RNACentral: <https://rnacentral.org/>

## RNACentral


- **Volumen:** Actualmente el consorcio de RNACentral está formado por 54 bases de datos de expertos, 44 las cuales ya han sido importadas a

RNAcentral, en el mismo podemos encontrar millones de datos destacando principalmente datos generales, recursos genómicos, organismos modelo, incRNA, rRNA y otros tipos de RNA.



- **Veracidad:** Los datos publicados pasan previamente por un control manual el cual evalúa si cumple o no (las secuencias menores a 10 nucleótidos están excluidas) las condiciones, el mismo previamente tiene que estar asignado a una de las bases de datos, posee los nombres de los autores y las fechas.
  - **Variedad:** secuencias, fuentes de secuencia para la predicción de estructuras secundarias, gráficos en 2D, identificadores para ncRNA e informes.
  - **Velocidad:** En la actualidad el número de datos ingresados es de millones provenientes de las 54 bases de datos disponibles, las mismas pasan por un proceso de control y se publican automáticamente en la página principal, todos los datos son actualizados cada trimestre.
- 

\*. Uniprot: <https://www.uniprot.org/>

## UNIPROT

- **Volumen:** trabaja con tres bases de datos no redundante en donde cada proteína tiene un único registro de secuencia con un único identificador (UPI)
    - UniParc: base de datos de secuencias de proteínas
    - Proteomes: base de datos de proteomas
    - UniRef: cluster de conjunto de secuencias curadas y registros únicos
  - **Variedad:** secuencia de proteínas públicamente disponible e información funcional, información sobre secuencias completas de proteomas usando fuentes como:
    - Bibliografías citadas
    - Taxonomía
    - Ubicaciones subcelulares
    - Enfermedades
    - Palabras clave y cruzamiento entre bases de datos
- 

Toda la información se suministra usando el formato UniProtKB/Swiss-Prot

- **Veracidad:** el curado de los datos es mediante dos procesos
  - Curación manual de UniProt: La curación manual consiste en una revisión crítica de los datos experimentales y previstos para cada proteína, así como la verificación manual de cada secuencia de proteínas. Los métodos de curación aplicados a UniProtKB / Swiss-Prot incluyen extracción manual y estructuración de información de la literatura, verificación manual de resultados de análisis computacionales, extracción e integración de conjuntos de datos a gran escala y actualización continua a medida que se dispone de nueva información. 
  - Anotación automática UniProt: UniProt ha desarrollado dos enfoques complementarios para anotar automáticamente secuencias de proteínas con un alto grado de precisión. UniRule es una colección de reglas de anotación seleccionadas manualmente que definen anotaciones que se pueden propagar en función de condiciones específicas, mientras que el Sistema de anotación automática estadística (SAAS) es un sistema de generación de reglas automático basado en un árbol de decisiones. Los componentes centrales de estos enfoques son reglas basadas en la clasificación InterPro y los datos seleccionados manualmente en UniProtKB / Swiss-Prot. 

\*. KEGG: <https://www.genome.jp/kegg/>

## KEGG

- **Variedad:** KEGG (Enciclopedia de genes y genomas de Kioto) es un recurso de base de datos que integra información funcional genómica, química y sistémica. En particular, los catálogos de genes de genomas completamente secuenciados están vinculados a funciones sistémicas de nivel superior de la célula, el organismo y el ecosistema.

Algunos de los datos que podemos encontrar en las distintas bases de datos son:

- genes y proteínas
- glicanos
- pequeñas moléculas
- variantes de genes humanos
- jerarquías y tablas brite

- **Volumen:** KEGG es un recurso de base de datos integrado que consta de dieciocho bases de datos (incluida la SSDB generada computacionalmente). Se clasifican ampliamente en información de sistemas, información genómica, información química e información de salud, que se distinguen por la codificación de colores de las páginas web.



\*. PubMed: <https://www.ncbi.nlm.nih.gov/pubmed>

## PUBMED

- **Volumen:** 30 millones de citas de literatura biomédica de MEDLINE PLUS, revistas de ciencias biológicas y libros en línea. Las citas pueden incluir enlaces a contenido de texto completo de PubMed Central y sitios web de editores.
- **Variedad:** Citas, textos y resúmenes de literatura biomédica. No incluye artículos de revistas de texto completo; Sin embargo, los enlaces al texto completo suelen estar presentes cuando están disponibles en otras fuentes.
- **Veracidad:** todas las citas y artículos tienen información que permite comprobar su veracidad como por ejemplo:
  - La fuente de donde proviene, año de publicación
  - Datos sobre el autor/es, profesionales involucrados, institución a cargo
  - Revisión periódica a cargo de MEDLINE PLUS



## Tema 3.2

1. Busque y desarrolle brevemente 2 o 3 (dos o tres) metodologías y técnicas utilizadas en los laboratorios, tanto a escala simple como a gran escala, para el estudio de:

\*. Ácidos Nucléicos.

\*. Proteínas

\*. Metabolitos

**Ácidos Nucléicos:**

**Electroforesis de ácidos nucleicos:**

La aplicación de la electroforesis en la separación de ácidos nucleicos permite hacer tareas simples, como verificar su síntesis o integridad, y complejas, como seguir los pasos enzimáticos de modificación durante la construcción de elaboradas colecciones de ADN. Su fundamento químico reside en que los ácidos nucleicos son polímeros de carga negativa unidos por enlaces covalentes fosfodiéster. De esta manera, el ADN y el ARN se moverán en un campo electroforético hacia el polo positivo. La matriz sólida preferida para separar los ácidos nucleicos es la agarosa pues da una capacidad de separación adecuada para la mayoría de las actividades rutinarias. Además, puede ser flexibilizada con la concentración utilizada. La electroforesis en agarosa es accesible debido a su bajo costo, a la diversidad de diseños de cámaras, y a que es fácil de preparar.



### **Secuenciación de ácidos nucleicos:**

De todos los métodos históricos de secuenciación de ácidos nucleicos diseñados por Sanger, el método por terminadores dideoxi es el más exitoso. Este método usa nucleótidos modificados con la ausencia del extremo 3'OH, grupo esencial para la polimerización por la ADN Polimerasa. De esta forma, cuando se incorpora un dideoxinucleótido, queda un trozo trunco de ADN que puede ser separado electroforéticamente. Si cada uno de los cuatro monómeros de ADN (los nucleótidos adenina, guanina, citosina y timina) se marca con una molécula fluorescente diferente, se puede determinar rápidamente el orden de los nucleótidos.



### **Clonación:**

La clonación de ADN fue posible gracias a la acumulación de conocimiento básico sobre las enzimas de restricción (er). Las ER fueron descubiertas en el estudio de los mecanismos bacterianos de defensa ante las infecciones virales (Kelly y Smith, 1970).

Las ERs cortan de forma precisa secciones de ADN y hacen compatibles los extremos de los productos escindidos con otras moléculas de ADN, como adaptadores, plásmidos, promotores y regiones codificantes, a los que se unen en una reacción *in vitro* catalizada por la enzima ligasa. Estas nuevas construcciones de ADN formadas por moléculas de orígenes genéticos distintos, se denominan quimeras. Las quimeras a su vez pueden ser escindidas para ligarse a otro fragmento de ADN, en un proceso llamado subclonación. Mediante programas de cómputo se puede simular la mejor estrategia para obtener quimeras en el laboratorio.



Las ERs sirvieron durante la segunda mitad del siglo XX para caracterizar ADN de tamaño muy grande (ej. 500,000 bases) imposible de manejar en su tamaño original. En combinación con las hibridaciones Southern Blot, las ERs fueron esenciales para descubrir los primeros genes responsables de enfermedades genéticas humanas (Gusella *et al.*, 1983) y para seleccionar características deseables en vegetales y animales (Bolívar–Zapata, 2004).


## **Proteínas:**

### **Electroforesis de proteínas:**

Las proteínas son biomoléculas cuyos monómeros son aminoácidos polimerizados por medio de enlaces peptídicos. A diferencia de los ácidos nucleicos, que poseen una carga neta negativa, la carga de las proteínas es muy variable porque en ellas puede haber aminoácidos tanto negativos como positivos. De esta manera, las primeras separaciones electroforéticas proteicas eran toscas y necesitaban un gradiente de pH que las atrapara en su punto isoelectrico (punto donde la proteína tendrá una carga neta de cero y no se podrá mover más en un campo eléctrico) (Tiselius, 1937). Laemmli (1970) implementó un tratamiento para poder separar a las proteínas con base en su peso molecular. Este consiste en la homogenización de la carga con dodecilsulfato de sodio (SDS), un detergente que simultáneamente desnaturaliza a las proteínas y las cubre de una carga neta negativa para que migren hacia el polo positivo durante la electroforesis.


### **Secuenciación de proteínas**

La secuenciación de proteínas mediante el método de Ed–man consiste en marcar el extremo N–terminal de las moléculas con el reactivo químico fenilisotiocianato (PITC). La identidad del PITC–aminoácido escindido de la proteína puede ser determinada mediante la observación de su presencia directamente en cromatografía de capa fina o de su ausencia por exclusión en un analizador de aminoácidos.



## **Metabolitos:**

Las dos plataformas tecnológicas más utilizadas para identificar y cuantificar metabolitos son: la resonancia magnética nuclear (RMN) y la espectrometría de masas (MS), esta última casi siempre acoplada a técnicas cromatográficas como la cromatografía líquida (LC-MS), la cromatografía de gases (GC-MS), o en menor medida la electroforesis capilar (CE-MS). Como consecuencia de la gran diversidad de plataformas analíticas utilizadas y la compleja naturaleza química de los metabolitos, la identificación de la estructura de estos se ha convertido en uno de los principales cuellos de botella para convertir los datos crudos de RMN y MS en conocimiento bioquímico. Sin duda, esta es la principal causa de que la





metabolómica no haya evolucionado tan rápidamente como la genómica y la proteómica. Aunque la identificación de metabolitos y proteínas se basa en la misma técnica de espectrometría de masas en tándem (o MS/MS), la diferencia principal radica en el hecho de que los espectros de fragmentación de los metabolitos permanecen impredecibles en gran medida, a diferencia de los datos de MS/MS para péptidos y proteínas.

