

## INTRODUCCIÓN A LA BIOINFORMÁTICA

Prof. Ana Julia Velez Rueda

### UNA INTRODUCCIÓN A LA INTRODUCCIÓN

La Bioinformática es una disciplina científica destinada a la aplicación de métodos computacionales al análisis de datos biológicos, para poder contestar numerosas preguntas. Los comienzos de la bioinformática se encuentran íntimamente relacionados a los avances científicos. Entre otros proyectos, el Proyecto genoma humano marca un hito en la ciencia y muestra la necesidad de recurrir a métodos automatizados para el análisis del caudal enorme de datos que se producen con las técnicas de secuenciación. Pero no todo en la Bioinformática se trata de procesar datos, esta disciplina puede generar entre otras cosas datos de relevancia biológica derivado de simulaciones y cálculos que retomen conocimientos ya establecidos.

Hoy en día en el campo de la Bioinformática se desarrollan, por ejemplo, herramientas para la predicción de la unión de fármacos a proteínas; o la presencia de cavidades y bolsillos en las estructuras proteicas, etc. Así es como echando mano de la tecnología podemos percibir distintos procesos biológicos y poner en imágenes conceptos que pueden resultarnos de otra forma muy abstractos. Un paso importante para aplicar el pensamiento computacional (binario) a la Biología, es reconocer qué datos o qué información nos aportan un conocimiento biológico. En principio, se podría decir que toda descripción de un sistema biológico podría ser un “dato biológico”. Por ejemplo, se pueden considerar datos biológicos el número de murciélagos en una región dada, la cantidad de pacientes enfermos con gripe en una población, la cantidad de glóbulos rojos por mililitros de sangre, entre otros.

Ahora bien, para comenzar desde lo micro a lo macro podríamos pensar en las células. Dentro de cada una de las células del cuerpo hay información importante almacenada, también hay otras tantas cosas, pero enfocándonos en la información que las hace a cada “célula quienes son” podríamos destacar unas moléculas muy importantes.

**RETO I: ¿Podrías buscar un ejemplo de macromoléculas que almacenen información sobre la ‘identidad’ de un organismo dado?**

### BIOMOLÉCULAS

La **química de los organismos vivos** se organiza **alrededor del carbono**. Este elemento tiene la capacidad de formar enlaces sencillos con hidrógeno o dobles con átomos de oxígeno o/y nitrógeno; dos átomos de carbono pueden compartir dos o hasta tres enlaces con otros átomos de carbono. Los **átomos de carbono enlazados covalentemente** pueden formar cadenas lineales, con ramificaciones o estructuras circulares; formando los **esqueletos sobre los que se añaden grupos de otros átomos “grupos funcionales” que le confieren las características funcionales** específicas.

La disposición espacial de los grupos sustituyentes de una molécula orgánica determinan su configuración, dos isómeros conformacionales solo pueden ser inter convertidos rompiendo

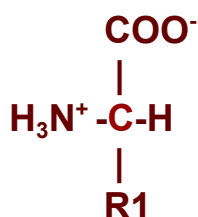
enlaces. Los dobles enlaces, alrededor de los que no existe libertad de rotación; y los centros quirales, alrededor de los cuales los sustituyentes se disponen según una secuencia específica definen la configuración de una molécula. Y ya que las interacciones moleculares entre biomoléculas son estereoespecíficas, esta característica se hace relevante para su función. La **conformación molecular** se refiere, entonces, a la disposición espacial de los grupos sustituyentes que tiene libertad de movimiento/de adoptar distintas disposiciones en el espacio, y según estas conformaciones en el espacio varíen, la función y/o funcionalidad de una dada molécula variará.

Se llaman **biomoléculas** a todas las moléculas que intervienen en la estructura y funcionamiento del organismo vivo, lo mismo sean grandes moléculas poliméricas (macromoléculas) como los **hidratos de carbono, los lípidos, las proteínas y los ácidos nucleicos** o sus monómeros: monosacáridos, ácidos grasos, aminoácidos y nucleótidos, así como sus intermediarios metabólicos. Sin tener que entrar en el detalle químico de la estructura de los monómeros: (monosacáridos, ácidos grasos, aminoácidos y nucleótidos) que se verá más adelante conviene tener una idea inicial de cómo son las grandes moléculas de los organismos vivos, los carbohidratos, los lípidos, las proteínas y los ácidos nucleicos. Cada una de estas macromoléculas se encuentra formada por el encadenamiento de los monómeros, unidos entre sí mediante enlaces característicos.

## PROTEÍNAS

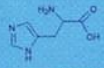
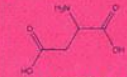

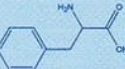



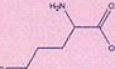
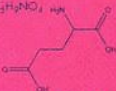
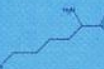
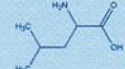



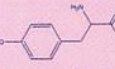
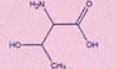
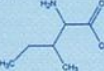
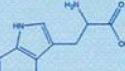



Las [proteínas](#) son polímeros de deshidratación de los aminoácidos. Las proteínas pueden estar constituidas por una o más cadenas peptídicas, que se denominan subunidades iguales o diferentes. Tienen, además, una composición aminoacídica característica y algunas pueden tener grupos químicos adicionales, no aminoacídicos (Grupos prostéticos).

Los aminoácidos son anfóteros, con una estructura básica que cuenta con un grupo amino ( $\text{NH}_2$ , básico) y un grupo carboxílico ( $\text{COOH}$ , ácido). Siempre hay al menos un átomo de carbono entre el grupo amino y el grupo carboxílico. La fórmula general de los aminoácidos se representa como sigue:



Los aminoácidos difieren entre sí por la naturaleza de sus grupos R, conformando así una lista de 22 aminoácidos que se combinan para formar a todas las proteínas presentes en los seres vivos. La unión de aminoácidos mediante un enlace peptídico, que da lugar a las cadenas peptídicas, se produce entre el grupo  $\text{COOH}$  de un aminoácido y el grupo  $\text{NH}_2$  de otro. Nuestro cuerpo utiliza solo 20 y puede sintetizar 10 de estos, y por ello reciben el nombre de aminoácidos esenciales constituyéndose en componentes indispensables de la dieta diaria de un ser humano.

Por tener un centro quiral, existen dos estereoisómeros D y L (levógiros o dextrógiros), cuyo nombre deviene de su actividad óptica de desviar la luz, aunque en los organismos vivos encontramos principalmente los L isómeros. Estos además poseen propiedades ácido-base y en algunos casos poseen grupos ionizables en sus cadenas laterales: ácido aspártico y glutámico, arginina, lisina, histidina, etc. (ver Tabla Periódica de Aminoácidos).

Periodic Chart of Amino Acids									
<div>H</div> <div>155.16 137.14 C<sub>6</sub>H<sub>7</sub>N<sub>3</sub>O<sub>2</sub></div> <div></div> <div>Histidine</div>									<div>D</div> <div>133.10 115.09 C<sub>4</sub>H<sub>7</sub>NO<sub>4</sub></div> <div></div> <div>Aspartic Acid</div>
<div>R</div> <div>174.20 156.19 C<sub>6</sub>H<sub>14</sub>N<sub>4</sub>O<sub>2</sub></div> <div></div> <div>Arginine</div>	<div>F</div> <div>165.19 147.18 C<sub>9</sub>H<sub>11</sub>NO<sub>2</sub></div> <div></div> <div>Phenylalanine</div>	<div>A</div> <div>89.09 71.08 C<sub>3</sub>H<sub>7</sub>NO<sub>2</sub></div> <div></div> <div>Alanine</div>	<div>C</div> <div>121.16 103.14 C<sub>3</sub>H<sub>7</sub>NO<sub>2</sub>S</div> <div></div> <div>Cysteine</div>	<div>G</div> <div>75.07 57.05 C<sub>2</sub>H<sub>5</sub>NO<sub>2</sub></div> <div></div> <div>Glycine</div>	<div>Q</div> <div>146.15 128.13 C<sub>5</sub>H<sub>9</sub>NO<sub>3</sub></div> <div></div> <div>Glutamine</div>	<div>E</div> <div>147.13 129.11 C<sub>5</sub>H<sub>9</sub>NO<sub>4</sub></div> <div></div> <div>Glutamic Acid</div>			
<div>K</div> <div>146.19 128.17 C<sub>6</sub>H<sub>14</sub>N<sub>2</sub>O<sub>2</sub></div> <div></div> <div>Lysine</div>	<div>L</div> <div>131.17 113.16 C<sub>6</sub>H<sub>13</sub>NO<sub>2</sub></div> <div></div> <div>Leucine</div>	<div>M</div> <div>149.21 131.20 C<sub>3</sub>H<sub>7</sub>INO<sub>2</sub>S</div> <div></div> <div>Methionine</div>	<div>N</div> <div>132.12 114.10 C<sub>4</sub>H<sub>8</sub>N<sub>2</sub>O<sub>3</sub></div> <div></div> <div>Asparagine</div>	<div>S</div> <div>105.09 87.08 C<sub>3</sub>H<sub>7</sub>NO<sub>3</sub></div> <div></div> <div>Serine</div>	<div>Y</div> <div>181.19 163.17 C<sub>9</sub>H<sub>9</sub>NO<sub>3</sub></div> <div></div> <div>Tyrosine</div>	<div>T</div> <div>119.12 101.10 C<sub>4</sub>H<sub>9</sub>NO<sub>3</sub></div> <div></div> <div>Threonine</div>			
<div>I</div> <div>131.18 113.16 C<sub>6</sub>H<sub>13</sub>NO<sub>2</sub></div> <div></div> <div>Isoleucine</div>	<div>W</div> <div>204.23 186.21 C<sub>11</sub>H<sub>12</sub>N<sub>2</sub>O<sub>2</sub></div> <div></div> <div>Tryptophan</div>	<div>P</div> <div>115.13 97.12 C<sub>5</sub>H<sub>9</sub>NO<sub>2</sub></div> <div></div> <div>Proline</div>	<div>V</div> <div>117.15 99.13 C<sub>5</sub>H<sub>11</sub>NO<sub>2</sub></div> <div></div> <div>Valine</div>	<div><div><div>Basic</div><div>Acidic</div><div>Nonpolar (hydrophobic)</div><div>Polar, uncharged</div></div><div><div>1-Letter Amino Acid Code</div><div>Molecular Weight</div><div>MW(H<sub>2</sub>O)</div><div>Molecular Formula</div></div><div><div>3-Letter Amino Acid Code</div><div>Chemical Structure</div><div>Chemical Name</div></div></div> <div><div></div><div>Serine</div></div>					
© Copyright 2003 by Bochem AG, Switzerland. Reproduction forbidden without permission.									

**Tabla periódica de Aminoácidos:** en esta tabla se pueden observar los aminoácidos con sus estructuras químicas y con el código uni-letra en el margen superior izquierdo. En rojo se muestran los aminoácidos ácidos (con grupos COOH en sus cadenas laterales), en rosa los aminoácidos polares, en azul los básicos (con grupos NH<sub>2</sub> en sus cadenas laterales) y en celeste los aminoácidos apolares.

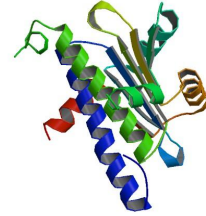
La actividad biológica de una proteína depende en gran medida de la disposición espacial de su cadena polipeptídica. Las proteínas no son un ovillo, sino que adoptan una estructura dada en el espacio. Se definen cuatro niveles distintos, conocidos como estructura primaria, secundaria, terciaria, y cuaternaria, y, cada uno de ellos se constituye a partir del anterior. Una misma cadena polipeptídica puede adquirir distintas estructuras secundarias, en distintos segmentos de la misma. La estructura tridimensional de una proteína normal en su entorno fisiológico normal está determinada por la totalidad de las interacciones interatómicas y, por lo tanto, por la secuencia de aminoácidos la composición de aminoácidos que la conforman (estructura primaria) (Anfinsen et al., 1961).

Como una forma de “acortar” la cantidad de letras que se usan para simbolizar un aminoácido, se establecieron por convención las **nomenclaturas de 3 letras y de 1 letra**. Entonces, por ejemplo, la forma más corta de simbolizar Asparagina puede ser usando “Asn” o, más simple aún, solo “N”.

La estructura secundaria que una proteína adopta se debe a la formación de puentes de hidrógeno entre los átomos que forman el enlace peptídico.

Los tipos básicos de la estructura secundaria son:

- $\alpha$ -hélice: plegamiento en espiral de la cadena polipeptídica sobre sí misma.
- Lámina plegada (beta hoja plegada): el plegamiento no origina una estructura helicoidal sino una lámina plegada en zig-zag.
- Bucles: sin una forma definida.



La estructura terciaria de una proteína corresponde al plegamiento tridimensional de las proteínas, debido a las interacciones de sus cadenas laterales. Las proteínas pueden plegarse y desplegarse repetidas veces, con la termodinámica como “fuerza impulsora”, hasta llegar a un mínimo de energía denominado estado nativo. Pauling y Mirsky, en su trabajo publicado en 1936, dan una primera definición del estado nativo de las proteínas como un plegamiento o conformación característico, que le confiere a las proteínas su función, y cuya pérdida denominaron desnaturalización (Mirsky and Pauling, 1936). Sin embargo, la actual descripción del estado nativo proteico propone que las proteínas en agua (lo que aplica a las células) presentan más de una conformación posible, que pueden interconvertirse unas en otras y explican su función (Frauenfelder et al., 1991; Wei et al., 2016). La función de una proteína y sus propiedades estarán determinadas, entonces, por la distribución de sus subestados conformacionales y las redistribuciones de las poblaciones en los diferentes entornos (Zhuravlev et al., 2009).

Desde el punto de vista menos Bio y más informático, las proteínas (y veremos luego que también los ácidos nucleicos, tanto ADN como ARN) pueden ser representadas “en unos y ceros” de múltiples formas.

**RETO II: Proponé una forma de expresar la información contenida en la estructura primaria de las proteínas usando tipos de datos de los lenguajes de programación que conocés.**

**RETO III: ¿ En qué tipo de datos podrías expresar la información de la estructura terciaria proteica?**

**RETO IV: Rosalind Franklin es una científica muy relevante, que tuvo menos reconocimiento del merecido. ¿Cuáles fueron sus contribuciones en este campo? ¿Qué nos cuenta su historia acerca del mundo de la ciencia?**

👉 **Mirá el artículo “El Caso de Rosalind Franklin” de Mujeres con Ciencia.**

Dijimos que el hecho de que una proteína adquiriera una u otra estructura depende de la composición de aminoácidos que la conforman (estructura primaria). Se ha estudiado en detalle la frecuencia de aparición de los distintos aminoácidos en una dada estructura secundaria y se observó que estos no se encuentran distribuidos de igual modo, si no que algunos aminoácidos predominan en ciertas estructuras. Es decir que conociendo la secuencia de una proteína y la preferencia de cada uno de los 20 aminoácidos para formar parte de una u otra estructura podríamos predecir qué disposición en el espacio adoptará una dada proteína.

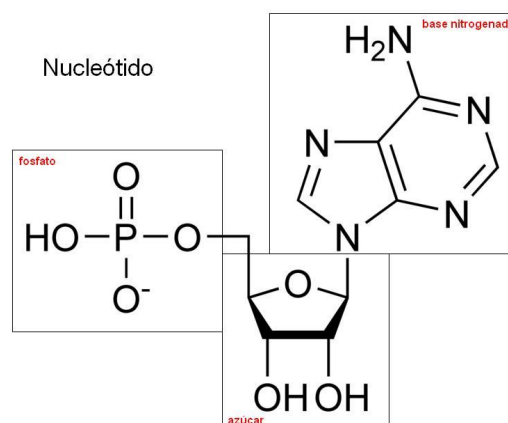
**RETO V: Proponé en pseudocódigo un programa que prediga la estructura secundaria que adoptará cada residuo (aminoácido) de la secuencia proteica dada, especificandola como H (si es una hélice), B (si es una hoja beta plegada) y L (si es un bucle o loop).**

☒ **PREGUNTAS DISPARADORAS:** ¿Qué inputs tendría tu programa? ¿De qué modo se te ocurre configurar el output?

**¡Guardate esta idea, la vamos a usar más adelante!**

## ÁCIDOS NUCLEICOS

Los ácidos nucleicos constituyen el material genético de los organismos y son necesarios para el almacenamiento y la expresión de la información genética. Existen dos tipos de ácidos nucleicos, química y estructuralmente distintos: el ácido desoxirribonucleico (ADN) y el ácido ribonucleico (ARN); ambos tienen la misma estructura general en todos los seres vivos. Desde el punto de vista químico, los ácidos nucleicos son macromoléculas formadas por polímeros lineales de nucleótidos, unidos por enlaces éster de fosfato, sin periodicidad aparente. Un nucleótido es una molécula orgánica compuesta por tres componentes: una base nitrogenada (purina o pirimidina), un azúcar (pentosa) y un grupo fosfato.



**Estructura base de un nucleótido.** Compuesto en forma genérica por tres componentes: una base nitrogenada (purina o pirimidina), un azúcar (pentosa) y un grupo fosfato.



Dentro de la célula, el ADN guarda la información necesaria para construir las proteínas. El ADN es una cadena formada por muchas combinaciones de cuatro nucleótidos A, C, G y T, que difieren entre sí en la base nitrogenada que lo compone. Las distintas combinaciones le otorgan distintas funcionalidades. Para que la información llegue del ADN a las proteínas, el mensaje genético es copiado desde el ADN a otra molécula, el ARN. El ARN es muy similar al ADN pero se diferencian por el azúcar (Pentosa) que llevan: ribosa y desoxirribosa, respectivamente.

Los nucleótidos se unen entre sí mediante el grupo fosfato, que sirve de puente de unión entre el primer nucleótido y el siguiente. Las moléculas de ADN y ARN también son capaces de adoptar una estructura en el espacio. Se definen, al igual que en el caso de las proteínas, cuatro niveles estructura: primaria, secundaria, terciaria, y cuaternaria, cada uno de los cuales se constituye a partir del anterior. La estructura primaria de los ácidos nucleicos se corresponde con la secuencia ordenada de nucleótidos desde el extremo 5' al 3' de una molécula. La información genética está contenida en el orden exacto de los nucleótidos.



Rosalind Franklin  
1920–1958

La estructura secundaria es el arreglo local de la estructura primaria estabilizada por puentes de hidrógeno, que según el modelo de Watson y Crick corresponde a una estructura en doble hélice (Watson and Crick, 1953)... ¡Bah! “de” Watson y Crick...

**PARA PENSAR: ¿Cuántas proteínas puede sintetizar un organismo? ¿De qué depende la cantidad y la característica de las proteínas que puede sintetizar un organismo? 🤔**

La estructura terciaria es un plegamiento complicado sobre la estructura secundaria adquiriendo una forma tridimensional. El ADN presenta una estructura terciaria, que consiste en que la fibra retorcida sobre sí misma, formando una especie de súper-hélice. Esta disposición se denomina ADN superenrollado y permite el empaquetamiento del ADN en las células. El ADN es una molécula muy larga en algunas especies y, sin embargo, en las células eucariotas se encuentra alojado dentro del núcleo, un compartimiento pequeño.

**RETO VI: ¿Qué hace distintos a dos individuos de una especie? Propone una forma de corroborar tu respuesta realizando un diagrama de un posible método computacional para dicho fin.**

**✓ PREGUNTAS DISPARADORAS: ¿Qué información deberías tener? ¿De qué modo deberías expresar dicha información para el análisis?**

Existen diferencias en las estructuras del ADN y ARN, el primero será una cadena doble, mientras que el ARN se encuentra por lo general como una cadena sencilla (algunos virus pueden presentar ARN doble hebra.. ¡Malditos!). Se conocen varios tipos de ARN y todos ellos participan de una u otra manera en la síntesis de las proteínas. Ellos son: ARN mensajero (ARNm); el ARN ribosomal (ARNr), que forma el armazón de los ribosomas; el ARN de transferencia (ARNt) y small nuclear RNAs (snRNAs). Otros tipos incluyen microRNAs (miRNAs), pequeños de interferencia (siRNAs), etc.

## BARRILETE CÓSMICO ¿DE QUÉ PLANETA VINISTE?

Ahora bien, no es una cuestión menor la obtención de la información biológica para su procesamiento computacional ¿De dónde viene esa información? ¿Dónde encontramos las secuencias de proteínas o su función en la célula; o inclusive información relativa a pacientes infectados con CODVID-19? 🤖

En términos generales los científicos de todo el mundo desarrollan diversos conocimientos relacionados con los seres vivos. Estos conocimientos se obtienen a base de observaciones y experimentación. Los datos y conclusiones obtenidos son compartidos entre científicos de forma organizada, ya sea por medio de publicaciones en revistas super-archi-nerds o, por ejemplo, a través de Bases de datos disponibles en internet. Una base de datos (DB por sus siglas en inglés) es una colección estructurada de datos; en particular, una base de datos biológica es una colección de información relacionada con seres vivos. Estos datos provienen de experimentos científicos, literatura publicada, análisis computacional, etc.

La información contenida en bases de datos biológicas puede incluir, por ejemplo: funciones, estructura y localización de proteínas o genes, efectos clínicos de mutaciones, así como similitudes de secuencias o distancias evolutivas, etc. Entre las bases de datos más utilizadas por científicos de todo el mundo, bioinformáticos o no, se encuentran GenBank (colección de todas las secuencias biológicas estudiadas) y PDB (que guarda la información estructural disponible acerca de ácidos nucleicos y proteínas).

👉 ¡Vamos a explorar juntos/as como es [La vida en Tres dimensiones!](#)

## Bibliography

Anfinsen, C.B., Haber, E., Sela, M., and White, F.H. (1961). The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc Natl Acad Sci USA* 47, 1309–1314.

Frauenfelder, H., Sligar, S.G., and Wolynes, P.G. (1991). The energy landscapes and motions of proteins. *Science* 254, 1598–1603.

Mirsky, A.E., and Pauling, L. (1936). On the structure of native, denatured, and coagulated proteins. *Proc Natl Acad Sci USA* 22, 439–447.

Watson, J.D., and Crick, F.H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171, 737–738.

Wei, G., Xi, W., Nussinov, R., and Ma, B. (2016). Protein ensembles: how does nature harness thermodynamic fluctuations for life? the diverse functional roles of conformational ensembles in the cell. *Chem. Rev.* 116, 6516–6551.

Zhuravlev, P.I., Materese, C.K., and Papoian, G.A. (2009). Deconstructing the native state: energy landscapes, function, and dynamics of globular proteins. *J. Phys. Chem. B* 113, 8800–8812.