
# Estudio de la diversidad conformacional en ARNs

Licenciatura en Biotecnología y Biología Molecular – FCE UNLP

Asignatura: Laboratorio de procesos biotecnológicos

Estudiante: Martín González Buitrón

Director del Trabajo Final: Dr. Gustavo Parisi

Tutor del Trabajo Final: Dr. Alexander Monzon

Lugar de trabajo:
Departamento de Ciencia y Tecnología.
Universidad Nacional de Quilmes.
Bernal, 2018

*Para Pame, mi compañera de vida.*

*Quien me ayudó a encontrar la motivación para terminar esta etapa.*

*Gracias por todo tu amor, alegría y sabiduría. Gracias por darme tantos momentos felices.*

*Y gracias por tu apoyo incondicional, por vos y con vos llegué hasta acá.*

*¡Te amo!*

# Índice general

# Abreviaturas

Å ……. Ångstrom

**BD** …….. Base de datos

**CoDNaS** ……. *"Conformational Diversity of the Native State"* o Base de datos de Diversidad Conformacional del Estado Nativo

**DC** …….. Diversidad Conformacional

**DDBJ** …….. "*DNA Data Bank of Japan"* o Base de datos de ADN de Japón

**DRX** …….. Difracción de rayos X

**EM** …….. *"Electron Microscopy"* o Microscopía Electrónica

**ENA** …….. "*European Nucleotide Archive"*

**IDP** …….. *"Intrinsically Disordered Proteins*" o Proteínas intrínsecamente desordenadas

**NCBI** …… *"National Center for Biotechnology Information"* o Centro Nacional para la Información Biotecnológica

**NDB** …….. "*Nucleic Acid Database*" o Base de datos de Ácidos Nucleicos

**PDB** …….. *"Protein Data Bank"* o Base de datos de Porteínas

**RMN** …….. Resonancia Magnética Nuclear

**RMSD** ……. *"Root Mean Square Deviation"* o Desviación cuadrática media

# Agradecimientos

Al estado argentino, por sostener y brindar el sistema de educación público, gratuito y laico.

A mi director, Gustavo, por haberme escuchado y alentado aquella vez que me acerqué a tu clase con la inquietud de querer trabajar con ácidos nucleicos. Aún recuerdo la primera vez que fui a la UNQ y me diste la bienvenida, mostrándome el lugar y abriéndome las puertas del grupo. Gracias por eso y por tu confianza diaria. Y gracias por no bajar los brazos a pesar de las recaídas en el camino.

A Alex, por ser mi tutor. Gracias por las ayudas para que este trabajo haya arrancado.

A Silvina, por querer continuar profundizando este trabajo junto con el de proteínas. Gracias a vos y Gustavo por las juntadas en su casa, donde siempre nos reciben muy amablemente.

A Nico, que me ayudaste cada vez que te contacté. Gracias por tus consejos justos.

Al resto del grupo de Bioinformática, por la buena onda y las charlas casuales en el almuerzo y tren. Gracias a cada uno de ustedes por los consejos y ayudas a lo largo de todo este camino. Gracias por no dejar de revivir los seminarios y el club del libro.

A mis excompañeros de militancia estudiantil, Suma, que me enseñaron a defender la educación pública y a luchar por los derechos del pueblo. Con uds. aprendí a ser estudiante. Gracias por tantos recuerdos geniales.

A mis amigos, con los que he compartido y comparto momentos hermosos. Gracias por cada juntada, risa, abrazo, salida y mate.

A mis viejos, por inculcarme sus valores y educarme desde chiquito. Gracias por ayudarme en esta etapa. Gracias má por tus charlas del corazón. A los dos, gracias por tanto amor, los amo.

A mis hermanas, Ro y Sofi, por las peleas y el amor desde chiquitos. Gracias a ustedes, por ser un punto de referencia en mi deconstrucción. Las amo.

A mi hermano, Agu, por aguantarme desde chiquito, siempre al lado molestando. Gracias por enseñarme tantas cosas. Te amo.

# 1. Introducción

## 1.1 Relación estructura-función y estado nativo

Para poder comprender el funcionamiento de las biomoléculas, componentes ubicuos y primordiales de todos los organismos, es fundamental remitirse al estudio de las estructuras que adoptan en el espacio y los movimientos que experimentan. Las proteínas son las principales biomoléculas que históricamente han sido estudiadas desde el punto de vista de su relación estructura-función. Las enzimas fueron las elegidas para tratar de comprender su actividad, función y cómo lograban catalizar las reacciones químicas.

Básicamente se conocen tres modelos que buscan explicar la relación estructura-función en proteínas. El primero de ellos fue propuesto a fines del siglo XX por Emil Fischer en 1894[1]. Este modelo es conocido como "llave y cerradura" (del inglés *lock and key*). Fischer buscaba explicar la alta especificidad de ciertas enzimas glucosídicas por sus sustratos (ligandos). Sostuvo que toda enzima debía contener una estructura única con geometría complementaria a la forma del sustrato para permitirle interaccionar solo con él y ser así altamente específica. Este concepto de relación-función se puede ejemplificar esquematizando a la enzima como una cerradura y al ligando como una llave, ambas perfectamente complementarias lo que explicaba la alta especificidad. En la **figura 1.1** se representa el modelo de Fischer:

$$P^* + L \rightleftarrows P^*L$$

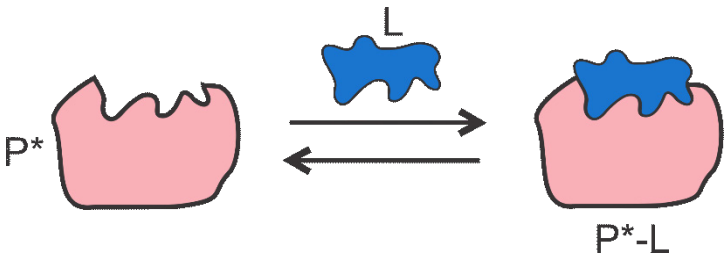



**Figura 1.1:** Modelo esquemático de "llave y cerradura". La enzima es P* y el sustrato, L.

En este esquema, $P^*$ es la proteína con su única estructura posible y $L$ es el ligando que posee una forma perfectamente complementaria al sitio de unión en la proteína. Energéticamente, esta interacción estabiliza la forma $P^*L$. Este modelo tiene como falencia la incapacidad para explicar aquellas enzimas que son multisustrato y/o poseen múltiples actividades[2,3]. El segundo modelo en explicar la relación estructura-función fue propuesto por Daniel Koshland recién en 1958[4,5]. Este modelo se conoce como de "ajuste inducido" y establece que la unión del ligando induce un cambio en la estructura de la proteína. Nuevamente, la proteína posee una única estructura en su forma libre, que sólo cambia cuando interacciona con un ligando. Así, este modelo admite una cierta flexibilidad en la proteína, como un guante de goma o latex que se adapta al momento de meter la mano. Podemos representar el modelo de Koshland de la siguiente manera (**figura 1.2**):

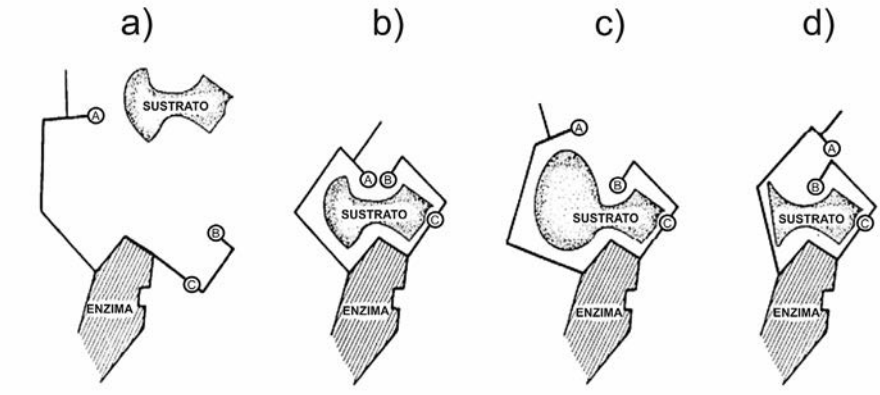



**Figura 1.2:** Modelo esquemático del "ajuste inducido". Las líneas negras indican las cadenas de la proteína que contienen grupos catalíticos A y B y un grupo de unión C. a) el sustrato y la enzima se encuentran disociados. b) el sustrato induce un cambio en las cadenas de las proteínas para llevar a A y B a la alineación apropiada y producir la unión correcta. c) un grupo voluminoso adherido al sustrato, evita la alineación correcta de A y B. d) la deleción del grupo elimina la acción de apoyo sobre la cadena que contiene A, por lo que termodinámicamente el complejo tiene una alineación incorrecta de A y B. Imagen adaptada de [5].

Koshland evidenció su modelo de ajuste-inducido utilizando métodos bioquímicos y no directamente estructurales. Observó que el número de tioles titulables de la fosfoglucomutasa cambiaba con el agregado de ligando. Formuló su modelo teniendo en consideración los conceptos derivados de experimentos claves en biología estructural como el de Mirsky y Pauling en 1936[6], donde definían los estados desnaturalizado y nativo de una proteína sobre la base de su configuración tridimensional. De forma sencilla, Mirsky y Pauling proponen la primera definición del estado nativo de una proteína, como aquella única configuración espacial que le confiere su funcionalidad biológica (configuración nativa). Es importante mencionar que Pauling amplía el concepto de configuración nativa en 1940 para poder explicar el reconocimiento estructural de los anticuerpos por sus antígenos, considerando la existencia de varias configuraciones nativas de energía similar[7]. Esta idea de múltiples configuraciones nativas fue también derivada por Karush en 1950 para explicar las diferentes afinidades de unión para distintos ligandos que presentaba la seroalbúmina bovina[8]. Karush proponía de forma temprana la existencia de un estado conformacional nativo representado por múltiples configuraciones en equilibrio termodinámico, a lo que él denominó "adaptabilidad configuracional". Desafortunadamente, su trabajo no tuvo repercusión sino hasta años recientes.

Hay que destacar que, hasta fines de la década del 50, no se conocía ninguna estructura tridimensional de proteína, siendo la mioglobina de cachalote (del inglés *sperm whale*) la primera resuelta por difracción de rayos X (en adelante DRX) y publicada en marzo de 1958[9]. Ésta fue la primera biomolécula donde se afrontó una alta complejidad para resolver su estructura utilizando la técnica de DRX debido a que tenía 1200 átomos (sin contar los átomos de hidrógeno). Anteriormente, la más compleja había sido la vitamina B12, con sólo 93 átomos.

El tercero de los modelos que da una explicación alternativa para la relación estructura-función fue presentado en 1965 por Monod, Wyman y Changeux (MWC), al que denominaron modelo del "pre-equilibrio"[10]. La hipótesis planteada por MWC se basa en la pre-existencia de al menos dos estructuras de una misma proteína que se encuentra en equilibrio termodinámico en ausencia de ligando. El conjunto de estos confórmeros

(entendiendo a un confórmero como todas aquellas estructuras tridimensionales estables que adopta una estructura primaria producto de la rotación en el espacio de sus enlaces simples y que se corresponden a un mínimo de energía potencial), define la diversidad conformacional de dicha biomolécula, esto es el conjunto de diferencias estructurales entre confórmeros. Se ampliará este concepto en la siguiente sección.

Monod, Wyman y Changeux extendieron el concepto de flexibilidad de las proteínas, considerando un equilibrio marcado entre las estructuras posibles que pudieran darse naturalmente por los cambios conformacionales (cambios producidos a nivel de interacciones entre residuos) de forma termodinámicamente estable. Claramente, este modelo logra definir un nuevo concepto de estado nativo, semejante a lo propuesto ocho años antes por Karush[8], donde éste nuevo estado funcional estaría representado por varias estructuras y no así por una única como plantean los dos modelos previos (*llave y cerradura* y *ajuste-inducido*).

Como bien dijimos, una diferencia sustancial con los otros modelos es la pre-existencia de confórmeros en ausencia de ligando. Para formular este concepto, Monod-Wyman-Changeux tuvieron en cuenta las cristalizaciones llevadas a cabo por Perutz en 1960[11] como primera evidencia experimental acerca de la existencia de dos conformaciones distintas para una proteína (la hemoglobina) en ausencia de ligando. Ahora, una vez que se tiene la presencia del ligando, aquellas conformaciones presentarán diferentes afinidades por el mismo, provocando así un desbalance en el equilibrio de las poblaciones que dará como resultado un desplazamiento hacia la población del confórmero seleccionado que aún no ha unido el ligando. Es por esto último que a este modelo de MWC también se lo conoce como modelo de “selección conformacional”. Podemos representarlo de la siguiente manera (**figura 1.3**):

$$P \rightleftarrows P^* + L \rightleftarrows P^*L$$

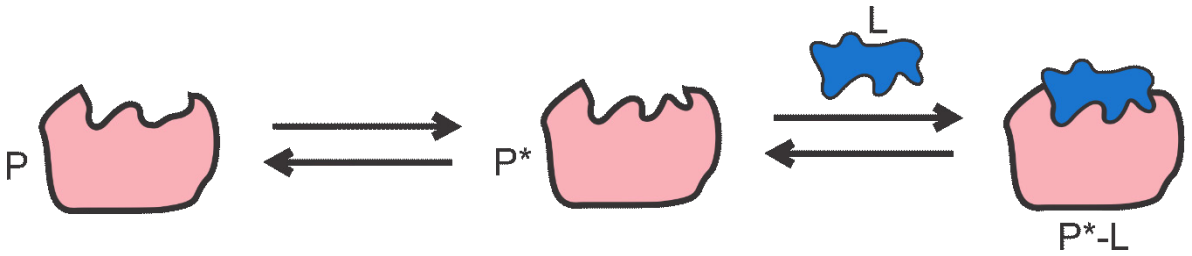



**Figura 1.3:** Modelo esquemático de "pre-equilibrio". La enzima se encuentra en dos conformaciones, P y P*. El sustrato es L. Como puede observarse, el confórmero P posee un sitio de unión diferente a P*.

Generalmente los procesos biológicos involucran fenómenos que pueden interpretarse con al menos uno de éstos dos últimos modelos, siendo, en primera instancia, mediante selección conformacional seguido de ajuste inducido. En la **figura 1.4** se muestra, de manera gráfica, cómo estos 2 modelos pueden explicar lo observado en un evento de reconocimiento para un determinado ligando.

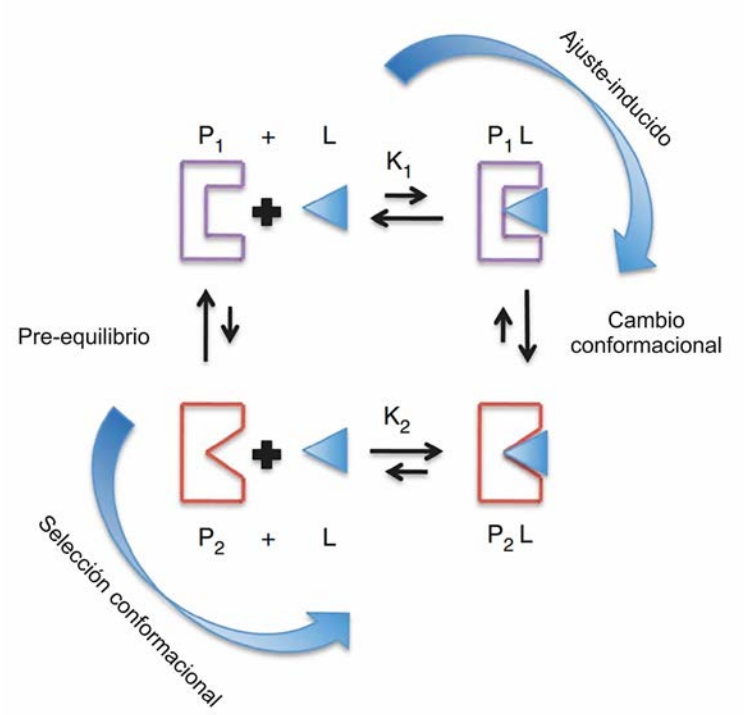



**Figura 1.4:** Representación gráfica del proceso de reconocimiento molecular. Existe un ciclo termodinámico en donde puede estar involucrado el mecanismo de ajuste inducido o selección conformacional. En la selección conformacional, la conformación roja ($P_2$) se encuentra preexistiendo en el ensamble de conformaciones y es más competente en poder unir el ligando (L). Las constantes de velocidad cinética y termodinámica pueden determinar si el proceso se produce por el mecanismo de selección conformacional o ajuste inducido[12]. Imagen adaptada de [13].

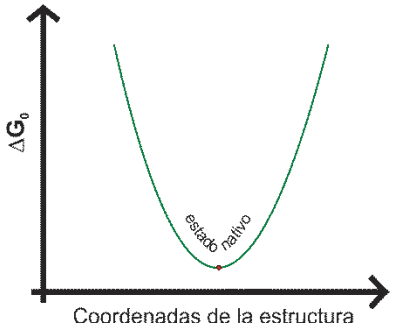
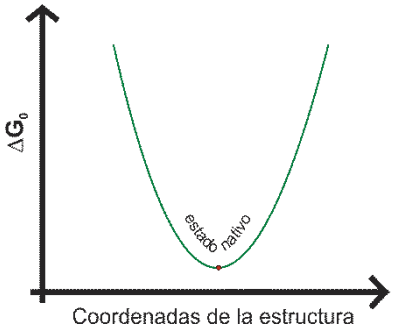
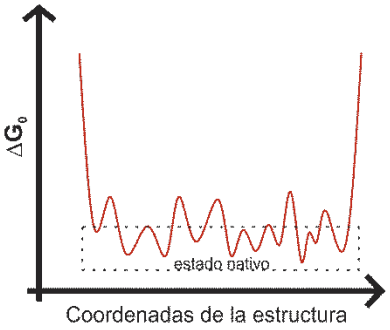
Los años posteriores a la publicación de los últimos dos modelos cambiaron el paradigma de comprender la función de las biomoléculas, en especial de las proteínas, desde la rigidez y exactitud geométrica hacia la flexibilidad (dinámica) y diversidad conformacional. Como anunció de forma metafórica en 1975[14] el argentino Gregorio Weber[15] (en contraposición a pensar a las proteínas de forma rígida y fuera de su estado en solución), las proteínas "gritan y patalean".

Traducción de cita textual en párrafo de [14]:

*…el modelo de molécula de proteína que resulta de las observaciones cristalográficas de rayos X es una proteína "platónica", bien removida de su perfección de molécula "estocástica" pataleando y gritando que inferimos debe existir en solución.*

Si observamos cada uno de los tres modelos previamente descritos, podemos decir que los dos primeros definen un único estado nativo (única conformación) y el último, otro tipo de estado nativo (múltiples conformaciones). Si lo definimos en términos de la teoría de paisajes energéticos[16], haciendo hincapié en la teoría del plegado proteico (**figura 1.5**) con una visión simplificada del modelo de embudo[17–20] (**figura 1.6**), reconoceríamos dos tipos de fondos: uno liso, con un único pico, y otro rugoso, con varios picos de energía semejante. Esto podemos representarlo de manera gráfica, entendiendo a la curva mostrada como la descripción de la energía libre de Gibbs de un único tipo de enlace simple en una única dimensión. Podemos apreciar los siguientes gráficos en la **Tabla 1.1**:

**Tabla 1.1:** En la siguiente tabla se muestran los gráficos de los tres modelos previamente explicados, donde se logra apreciar una descripción del fondo del embudo. En los estados nativos únicos, se observa un punto rojo en el fondo del embudo y se corresponde al mínimo de energía. En el estado nativo múltiple se aprecia un recuadro de línea negra entrecortada que encierra la región última del embudo con varios estados en equilibrio de energía mínima semejante.

| Modelo | Estado nativo | Gráfico |
|---|---|---|
| Llave y cerradura | Único |   |
| Ajuste inducido | Único |   |
| Pre-equilibrio | Múltiple |   |

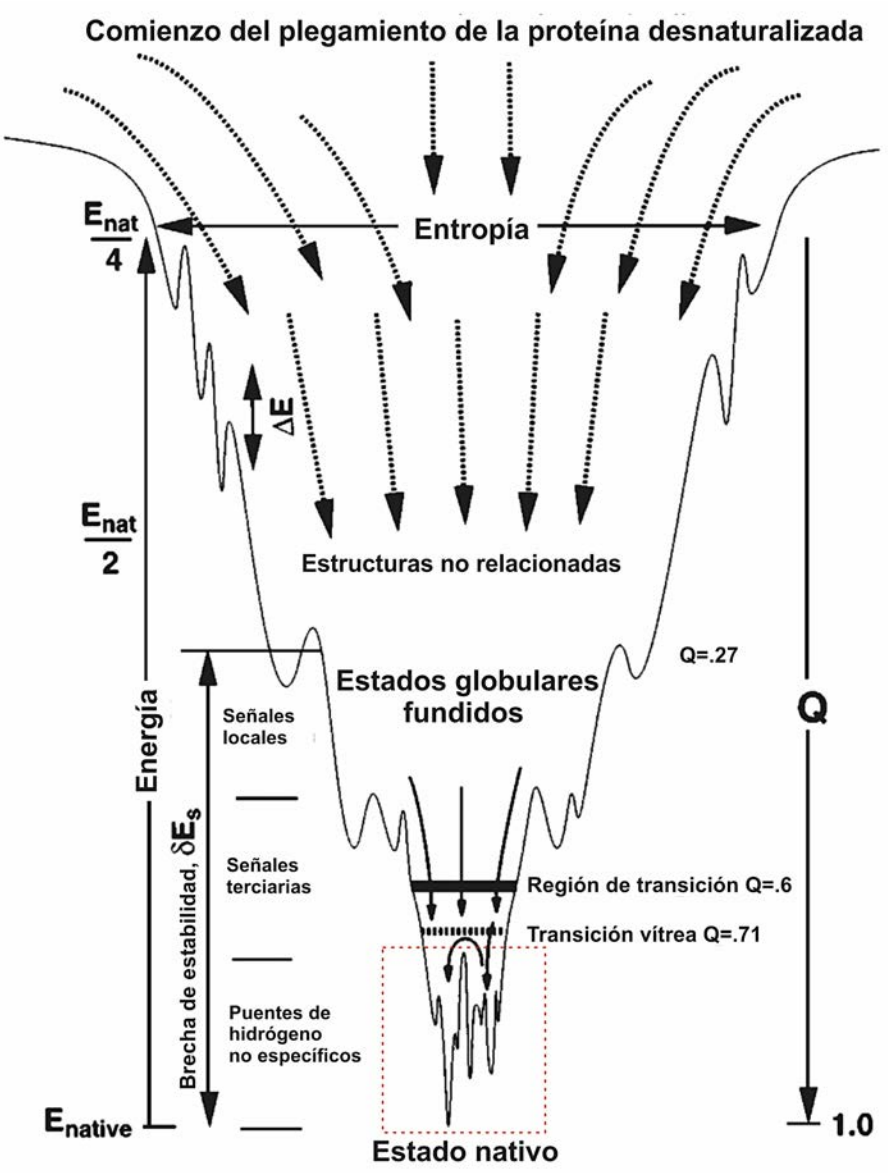



**Figura 1.5:** Esquema conceptual de la teoría del plegado proteico. El proceso de plegado de una proteína no sigue un camino único y específico, por el contrario, su plegamiento puede describirse en forma de un embudo energético en el cual la proteína puede adoptar diferentes plegamientos lo que la conlleva a seguir diferentes rutas energéticas en pos de minimizar su energía y llegar al estado nativo. El fondo del embudo, en este caso, rugoso, describe el estado nativo de la proteína formado por un conjunto de confórmeros en equilibrio, con relativamente menor energía que los estados desnaturalizados. Como puede observarse, la apertura del embudo está dominada por la enorme cantidad de estados desnaturalizados y de ahí la gran entropía en ese punto del proceso. E es la energía libre; Q es la fracción de contactos nativos. Imagen adaptada de [19].

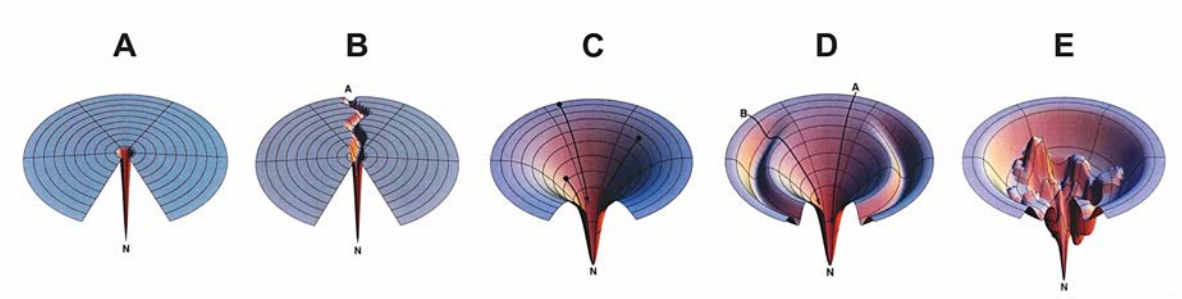



**Figura 1.6:** En la siguiente figura se muestran diferentes paisajes energéticos. A – representación del paisaje característico de la paradoja de Levinthal (paisaje tipo campo de golf) donde el plegado está dominado por la búsqueda conformacional difusional. B – Idem que la representación A aunque se asume la existencia de una ruta de plegado característica hasta llegar al estado nativo. C – Paisaje liso y de plegado rápido. En este paisaje con forma de embudo se asume que a medida que se incrementan los contactos en la cadena y disminuye su energía interna, su libertad conformacional también se ve reducida. D – Este paisaje (paisaje de embudo con fosa) presenta el plegamiento rápido (trayectoria "A") y por otro lado un plegamiento más lento debido a la presencia de una trampa cinética (trayectoria "B"). E – Paisaje rugoso. En este tipo de paisaje energético existen trampas cinéticas y barreras energéticas, siendo la representación más aceptada para biomoléculas que presentan estructuras ordenadas. Imagen adaptada de [21].

Este último concepto de estado nativo se ha ampliado a "***ensamble nativo***"[22] y representa las **distribuciones estadísticas de las poblaciones de los confórmeros de menor energía libre de Gibbs que se encuentran en equilibrio termodinámico.** Como es de suponer, esta definición es válida para todas las biomoléculas.

La utilización de paisajes energéticos ha logrado avanzar en la conceptualización del plegado de proteínas, permitiendo poder re-evaluar la paradoja de Levinthal, rutas de plegamientos y presencia de embudos lisos o rugosos[21]. Esta visión de paisaje energético en forma de embudo es ampliamente utilizada para explicar la relación termodinámica-cinética en el plegado de proteínas e interacción de unión con ligandos y biomoléculas. Básicamente, se busca comprender la/s función/es biológica/s desempeñada/s conociendo una información estadística detallada de los ensambles conformacionales junto con sus fluctuaciones termodinámicas[23]. Como se verá en la siguiente sección, la nueva visión de estos paisajes energéticos deja de considerarlos estáticos para describirlos como paisajes energéticos dinámicos[24,25].

Si nos remitimos a otras biomoléculas, como por ejemplo a los ácidos ribonucleicos (ARNs), encontraremos que también pueden desarrollarse estudios en el plegado y en la relación estructura-dinámica-función haciendo uso de paisajes energéticos[26–31].

## 1.2 Diversidad conformacional

Como se ha mencionado anteriormente, la definición de ensamble nativo aplica a todas las biomoléculas. Lo que no se ha dicho aún es el grado de heterogeneidad que puede presentar un ensamble nativo. Esto último está asociado a la diversidad de confórmeros nativos en equilibrio y puede apreciarse gráficamente como un fondo rugoso del embudo representado en un paisaje energético. Allí podríamos observar fondos más rugosos que otros, de mayor o menor extensión dependiendo de la energía relativa de los distintos confórmeros. Esta extensión heterogénea de poblaciones de confórmeros nativos coexistiendo en equilibrio termodinámico, separados por barreras energéticas semejantes, está definida como *diversidad conformacional*.

Podemos decir entonces que todas las biomoléculas presentan diversidad conformacional, lo cual es una propiedad intrínseca y, en nuestro caso, de interés biológico. Justamente, si lo pensamos detenidamente, estamos desafiando el paradigma de "una secuencia, una estructura, una función". Es por eso que cuando hablamos de función biológica, la visión actual establece "una secuencia, un ensamble, al menos una función". Para no dejar lugar a pequeños malos entendidos, cuando pensamos la diversidad conformacional debemos lograr apreciar aquellas conformaciones que presentan un vínculo directo con la función biológica. Es decir, debemos entender que, de todas las posibles conformaciones nativas que existan para una biomolécula dada, incluyendo las más mínimas diferencias energéticas o poseyendo coordenadas de conformación prácticamente iguales, nos interesará establecer la relación entre las proporciones relativas de los confórmeros y la función biológica. Poder determinar y asignar de manera acertada la función biológica (o al menos una de ellas) para una biomolécula es un reto muy complejo y para nada fácil de realizar, permaneciendo como un tema central a trabajar en este nuevo siglo.

Si nos remontamos a los inicios de las primeras evidencias experimentales de confórmeros en equilibrio, debemos mencionar el trabajo realizado por Perutz hacia fines de la década de 1950 y publicado en 1960[11]. El grupo liderado por Perutz cristalizó y

obtuvo por DRX la estructura de dos confórmeros de la hemoglobina en ausencia de ligando, el confórmero R (relajado, "*relaxed*" en inglés) y el T (tenso, "*tense*" en inglés). El calificativo del estado de rigidez es dado por la ausencia o presencia de un mayor número de interacciones débiles intermoleculares. La función biológica de la hemoglobina es la de transportar oxígeno en la sangre. Para lograrlo, la hemoglobina debe poder unir la molécula de oxígeno el tiempo suficiente para que sea transportada través del sistema circulatorio y finalmente ésta pueda ser liberada. Esta unión y liberación del oxígeno está explicada en términos de las constantes de afinidad de los confórmeros ($K_d$) y de sus constantes cinéticas ($k_{on}$ y $k_{off}$). De este trabajo surgió la teoría del pre-equilibrio (MWC) discutida anteriormente. Como vemos, la diversidad conformacional evidenciada en la hemoglobina por DRX, junto a la teoría del pre-equilibrio, han aportado herramientas para poder formular un modelo que pudiera explicar mejor la función biológica de dicha biomolécula.

Inicialmente, el concepto de diversidad conformacional fue introducido por Pauling en 1940 para explicar la unión de los anticuerpos[7] y ampliado por Karush en 1950[8], ambos trabajos fueron previos a tener conocimiento de información estructural a nivel atómico. A pesar de que a lo largo de los años se ha incrementado exponencialmente la información estructural a nivel atómico, muchas veces la dinámica conformacional de las moléculas no es tenida en cuenta para poder entender su función biológica. Esta mirada lleva consigo la idea equivocada de pensar a las estructuras de las biomoléculas de forma estática y no dinámica (en ausencia o presencia de ligandos); aisladas y no en constante interacción con el medio; sin interaccionar y no interaccionando con múltiples (bio)moléculas; sin formar complejos; entre otros.

Actualmente, la diversidad conformacional tiene un rol clave en el concepto de selección conformacional[13], inicialmente introducido en el modelo MWC. El modelo de selección conformacional considera la existencia de confórmeros en equilibrio previo a la unión de algún ligando, presentando cada uno de ellos diferentes afinidades de unión. Una vez que el ligando se ha unido preferentemente a alguna de estas conformaciones, se dice que se ha "seleccionado" el confórmero y, en consecuencia, el equilibrio químico es desplazado

hacia este mismo. Esta selección ocurre sin que el ligando haya inducido un cambio conformacional, como lo propone el modelo de Koshland.

Ahora, si nos detenemos a pensar en los movimientos que experimentan las estructuras en sus medios (extra)celulares, debemos pensar en las dinámicas de las mismas. No solo alcanza con poseer información espacial para entender cómo funciona una biomolécula: una descripción abarcativa y detallada implica tener al tiempo presente como cuarta dimensión. El estudio de la dinámica de biomoléculas es un campo con gran impacto en la relación estructura-función. Entonces, si nos enfocamos en describir estos movimientos desde la perspectiva de los paisajes energéticos, veremos paisajes rugosos con "valles y montañas". Así, para proteínas, podemos encontrar diferentes niveles dinámicos que describen las fluctuaciones de movimientos en diferentes tiempos. En la **figura 1.6** se esquematiza un corte transversal del paisaje energético de los movimientos de proteínas y las principales técnicas empleadas para medir éstos mismos.

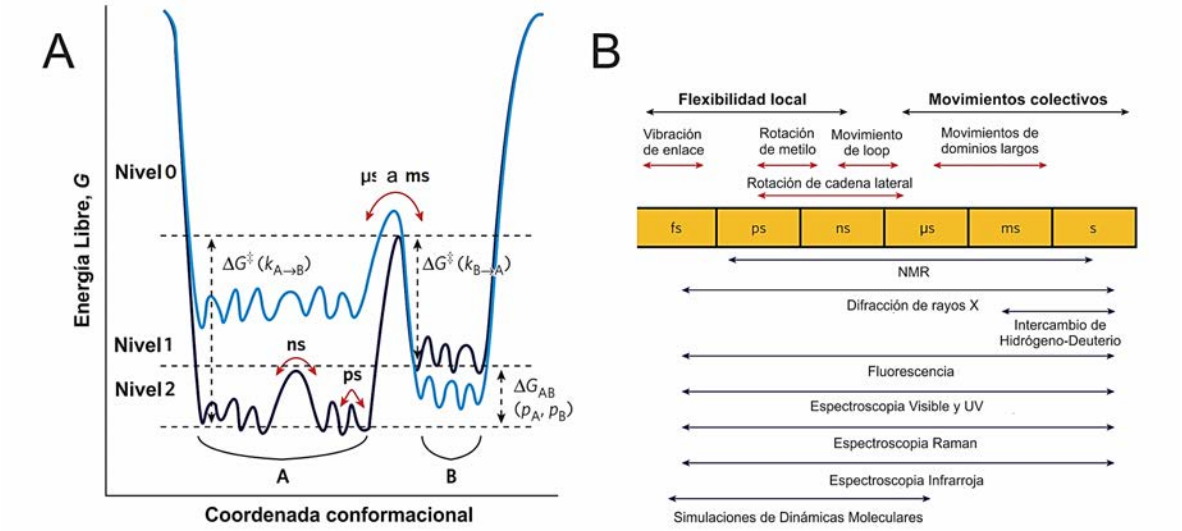



**Figura 1.6:** El paisaje energético está definido por la amplitud y escala de tiempo de los movimientos de las proteínas. A – Se observa un corte transversal unidimensional del paisaje energético multidimensional. En él se muestra la jerarquía de los diferentes niveles de movimientos definidos por Frauenfelder[16] donde se observan las barreras y escalas de tiempos en que éstos son llevados a cabo. Un estado es definido como un mínimo en la superficie de energía, donde el estado de transición es el máximo entre los pozos. Las poblaciones de los estados A y B en el nivel 0 ($\rho_A$, $\rho_B$) son definidas como distribuciones de Boltzmann basadas en su diferencia de energía libre($\Delta G_{AB}$). La barrera entre estos estados ($\Delta G^{\ddagger}$) determina la tasa de interconversión $k$. En los niveles inferiores se producen las fluctuaciones más rápidas entre muchos subestados relacionados dentro de cada estado de nivel 0. Un cambio en el sistema, por ejemplo, producto de la unión con un ligando, mutación(es) o condiciones externas, produce un cambio en el paisaje energético (de línea azul a celeste o viceversa) el cual se evidencia en un cambio en el equilibrio entre los estados conformacionales. B – Escala de tiempo de cada proceso dinámico. Se detallan los diferentes métodos experimentales empleados para la detección de las fluctuaciones según su escala temporal. Imagen adaptada de [25].

Debemos notar que, dependiendo la escala de tiempo en que se desarrollan los procesos, se pueden observar ciertos tipos de movimientos característicos. Para cada nivel de movimiento se corresponden métodos específicos, por ejemplo, las transiciones lentas entre conformaciones se suelen registrar con diferentes métodos experimentales de alta y baja resolución (ver **figura 1.6 - B**). A su vez, como mencionamos en la sección anterior, estos paisajes energéticos no son estáticos, sino que poseen un dinamismo que depende de las condiciones del entorno y/o de la unión a otras (bio)moléculas. Ésto se observa gráficamente en las líneas de azul oscuro a celeste de la **figura 1.6 – A** y en la **figura 1.7** donde el paisaje energético de la proteína E-THF cambia al unir el cofactor NADPH durante el proceso de selección conformacional.

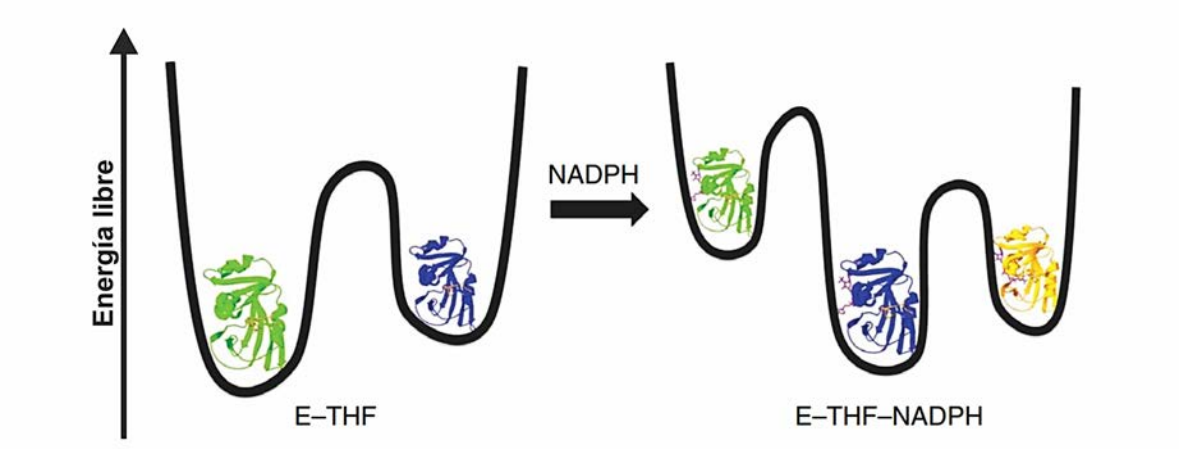



**Figura 1.7:** La dinámica conformacional de la enzima dihidrofolato reductasa es dependiente de ligando[32]. El paisaje energético del complejo binario (E-THF) cambia al unir el cofactor NADPH (paisaje dinámico). En el proceso de unión, se selecciona el confórmero azul (complejo binario) y, luego de la unión del NADPH, se ve energéticamente favorecido el complejo terciario (E-THF-NADPH, complejo también azul debido a similitud estructural). Como puede observarse, luego de la unión del NADPH y posterior redistribución del ensamble conformacional, se ve favorecida la presencia de un nuevo confórmero, en este caso, de color amarillo. Imagen adaptada de [13].

Un ejemplo paradigmático de selección conformacional es el de la ubiquitina, que a través de su ensamble nativo debe poder reconocer y unir diversas proteínas. En la **figura 1.8** se puede observar de forma gráfica cómo cambia el paisaje energético de la ubiquitina al unirse a 4 proteínas diferentes. Para concluir, podemos decir entonces que, los ensambles conformacionales facilitan la interacción entre diversos patrones de unión.

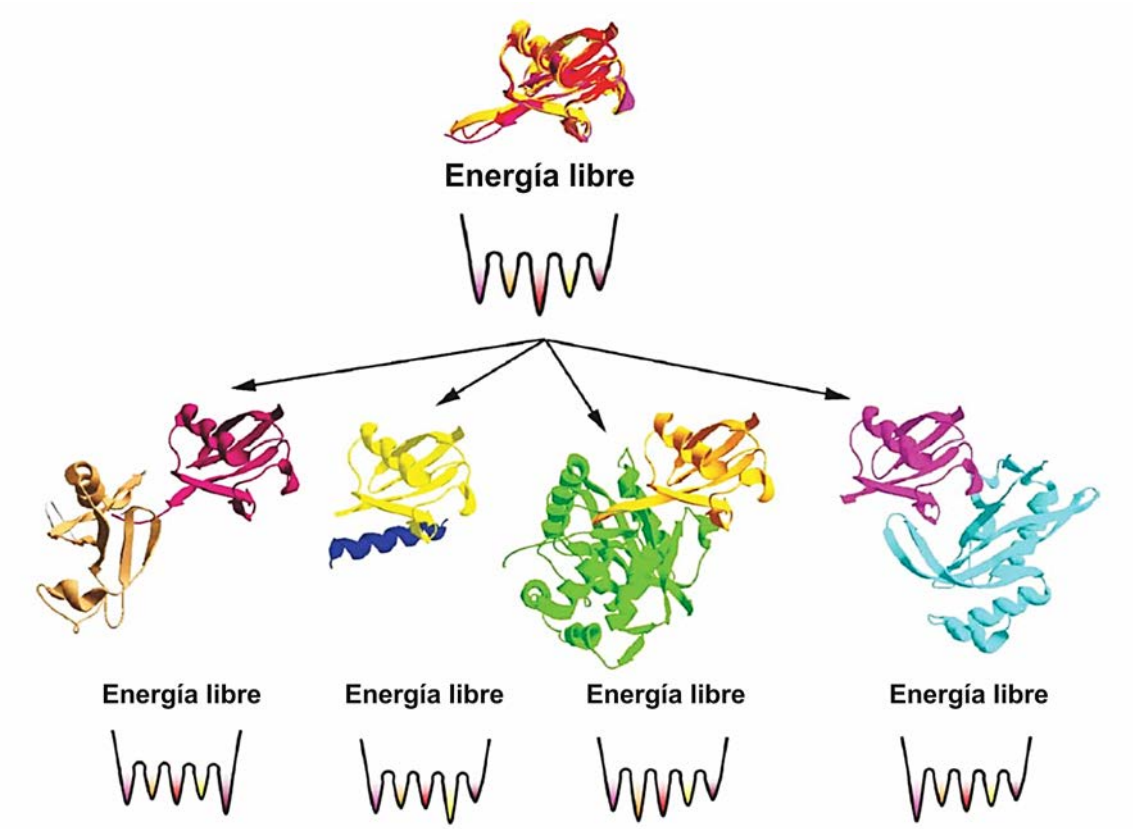



**Figura 1.8:** Cambio del paisaje energético de la ubiquitina al unirse con diferentes proteínas. Los estudios de NMR indican que todas las conformaciones que se unen a cada proteína se encuentran presentes en ausencia de ellas. Actualmente, se tienen 46 estructuras cristalizadas del ensamble conformacional de la ubiquitina, aunque en esta imagen sólo se muestran 5 de ellas (códigos PDB 1F9J, 1S1Q, 1XD3, 2D36 y 2G45). Los paisajes energéticos mostrados son hipotéticos ya que se desconocen las poblaciones relativas y las barreras energéticas que separan cada uno de los confórmeros en el ensamble. Imagen adaptada de [13].

Es importante mencionar que no todos los paisajes energéticos suelen poseer una profundidad apreciable. Existen biomoléculas, por ejemplo, las proteínas intrínsecamente desordenadas (IDP, del inglés *Intrinsically Disordered Protein*) cuyos paisajes energéticos carecen de picos pronunciados. Ésto puede observarse con notoriedad en la **figura 1.9** donde se comparan paisajes energéticos de IDP y no-IDP. Estas proteínas, que pueden ser total o parcialmente desordenadas, desafían la relación estructura-función ya que son funcionales sin adquirir un estado globular y de estabilidad característica. Desde un punto de vista evolutivo, estas proteínas demuestran que el hecho de tener una estructura 3D bien definida no es prerrequisito para ejercer su función. Claramente, la diversidad

conformacional en éste tipo de proteínas tiene un rol aún más importante que excede el objetivo de éste trabajo y está siendo arduamente investigada[33–37].

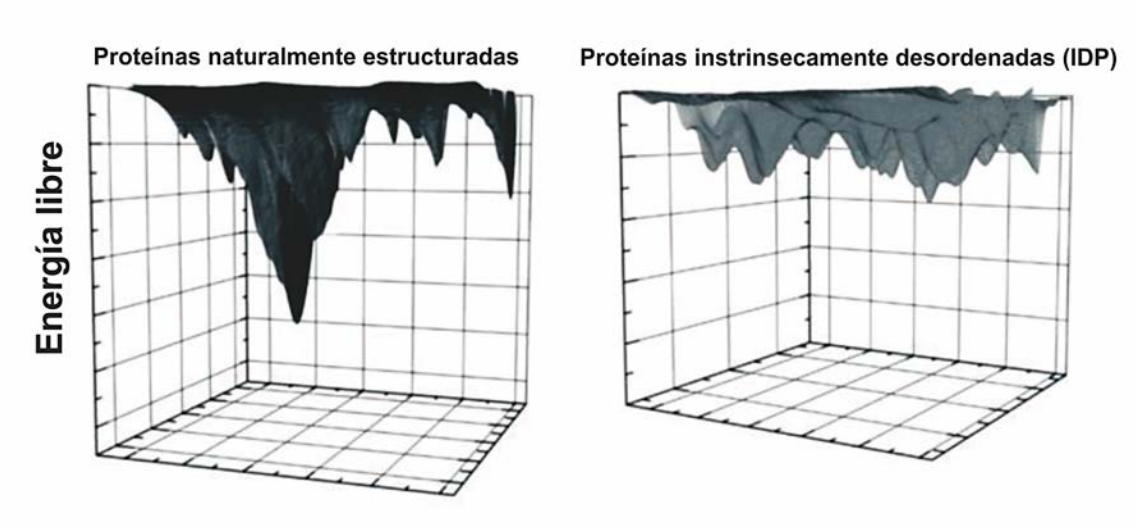



**Figura 1.9:** Clásicos paisajes energéticos para proteínas. Como puede apreciarse, en el paisaje energético que se muestra a la izquierda, se observan diferentes picos de embudos siendo uno de ellos claramente más pronunciado y profundo que el resto. Éste pico se corresponde a aquellas estructuras que poseen las menores energías libres en sus estados conformacionales. Por otro lado, en el paisaje de la derecha, se observan muchos picos con energías libres semejantes. Estos picos se corresponden a conformaciones que carecen de estructuras de estabilidad apreciable, siendo la fluctuación entre las diferentes conformaciones lo que mayormente ocurre. Además, una proteína que tenga un gran número de picos con energías semejantes, transitará un menor tiempo en cada conformación. Imagen adaptada de [38].

Por otro lado, la dinámica de los ARNs también es de suma importancia para el desempeño de su función biológica[39–41]. Muchos de los procesos celulares donde se ve involucrado el ARN, se producen en una escala de tiempo desde los picosegundos a segundos o, inclusive, minutos (ver **figura 1.10**).

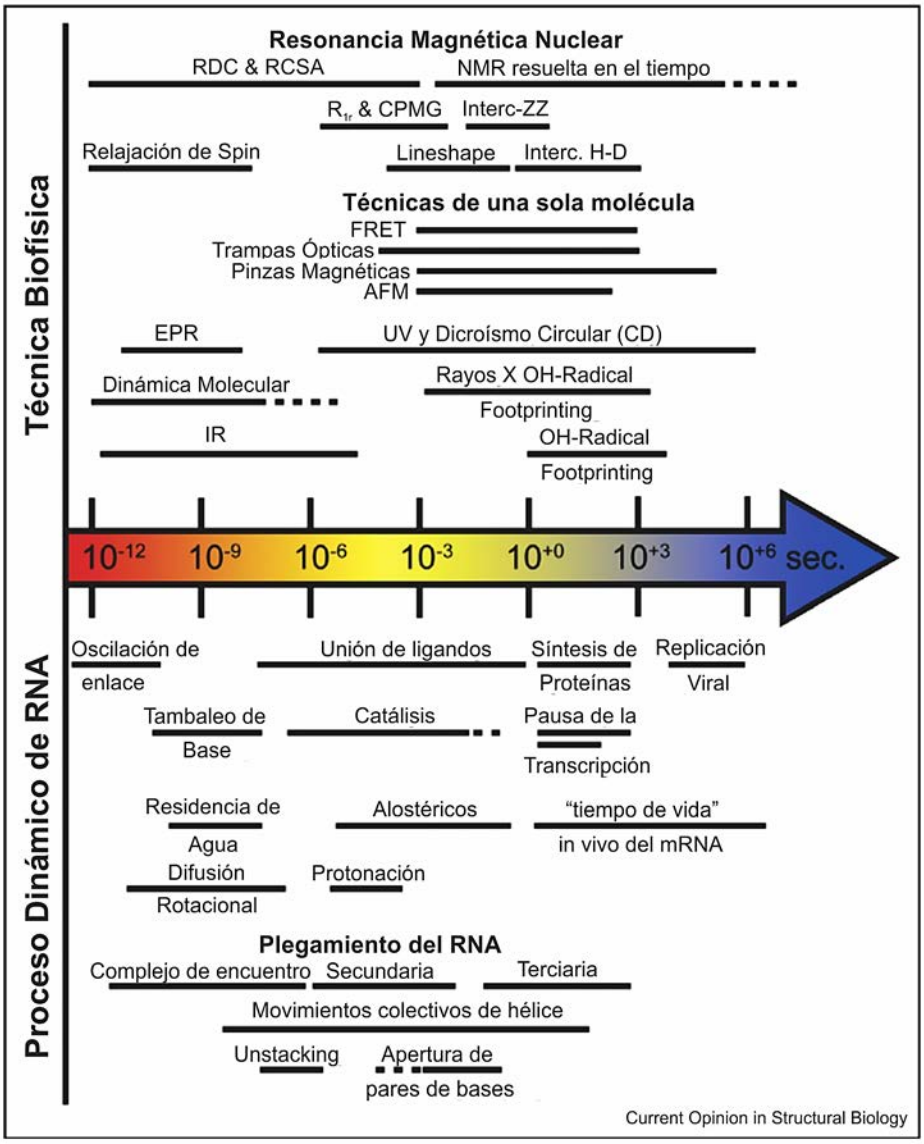



**Figura 1.10:** Diagrama temporal de los procesos dinámicos en el ARN y las correspondientes técnicas biofísicas que pueden aplicarse para su caracterización. Imagen adaptada de [40].

Antes de continuar con ejemplos biológicos característicos de la dinámica de ARNs, debemos hacer notar que recién en el año 2014 se propuso la primera clasificación detallada de niveles dinámicos para el ARN[41]; y si lo comparamos con el propuesto por Frauenfelder en 1991 para proteínas[16], nos encontramos con una separación de casi 25 años. Esta clasificación establece 3 niveles dinámicos que, descritos sobre la base de los estados conformacionales separados por barreras energéticas, se diferencian de la siguiente manera: el nivel-0 es aquel que se refiere a las conformaciones de ARN con

distinta estructura secundaria (es el nivel que implica movimientos más lentos); el nivel-1 es aquel donde las conformaciones presentan diferencias de apareamiento de bases; y por último el nivel-2 es aquel donde las conformaciones presentan estructuras secundarias y apareamiento de bases similares pero difieren en otros aspectos en su estructura, producto por ejemplo de dinámicas de regiones interhélices y dinámica de loops (es el nivel más rápido). Una representación gráfica de los niveles dinámicos en ARNs puede observarse en la **figura 1.11**.

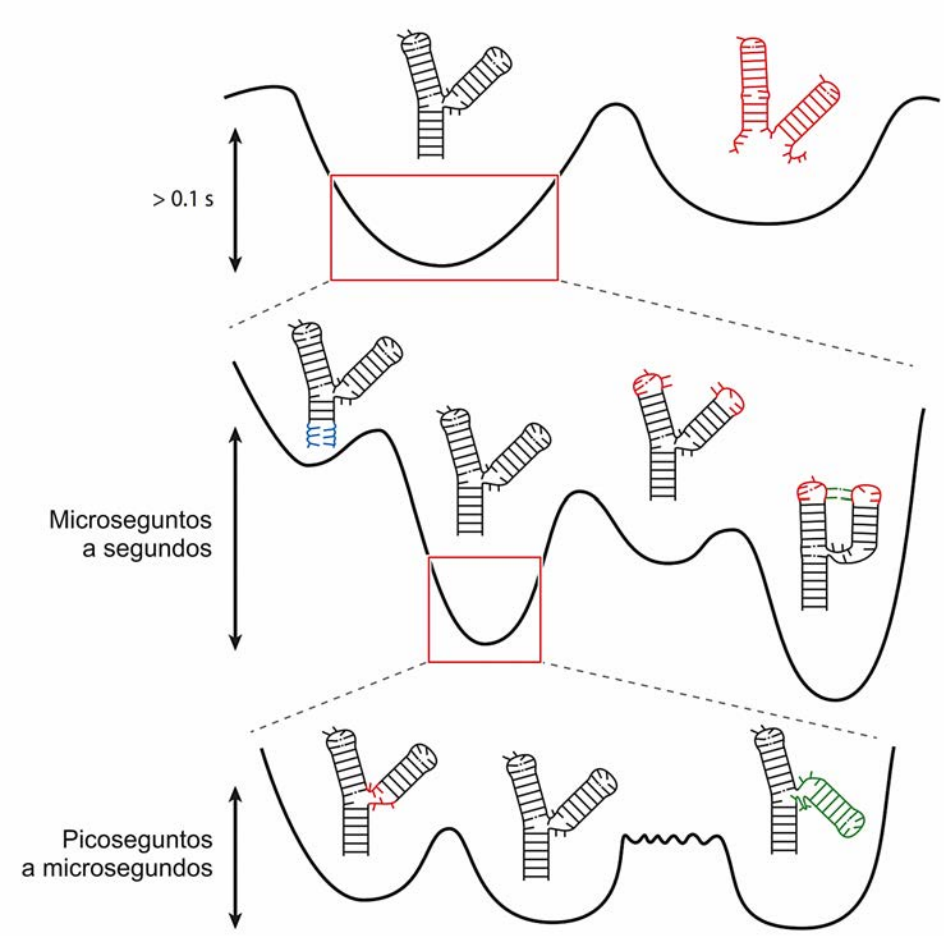



**Figura 1.11:** Representación gráfica de los diferentes niveles dinámicos en moléculas de ARNs. En el nivel más bajo, el Nivel 0, ocurre la dinámica de estructuras secundarias y las conformaciones se encuentran en fosas anchas separadas por barreras de energía libre muy bien definidas. Dentro de cada fosa de estructura secundaria, se llevan a cabo movimientos/arreglos locales alternativos de emparejamiento de bases que definen la dinámica del Nivel 1. Estos movimientos incluyen fusión de pares de bases (izquierda, azul), reorganización (extremo derecho, rojo) y emparejamiento terciario (verde). Cada fosa de emparejamiento local, a su vez, define un conjunto limitado de conformaciones tridimensionales con transiciones entre estas fosas que constituyen la dinámica del Nivel 2. Estas dinámicas incluyen dinámica de bucle (izquierda, roja) y dinámica de interhélices (derecha, verde). Aunque la dinámica de interhélices y la dinámica de bucle tienen alturas de barrera similares, debido a la mayor cantidad de coordenadas involucradas, la dinámica de interhélices suele ocurrir más lentamente (posee una barrera de separación larga y áspera). Imagen adaptada de [41].

A modo de ejemplo en cómo la dinámica de cambios conformacionales del ARN se ve reflejada en su función biológica podemos citar el caso de regulación traduccional del riboswitch de adenina[42], donde se establece un equilibrio de tres estados conformacionales dependientes de la temperatura y uno de ellos posee alta afinidad de unión por la adenina (efector). Además, en la conformación que une adenina los sitios de inicio de la traducción y unión al ribosoma se encuentran libres. Es por esto que, a temperatura ambiente y en presencia de adenina, ésta se une al confórmero con mayor afinidad (selección conformacional) permitiendo que éste se estabilice, desplace el equilibrio, aumente la población de confórmeros con sitios de unión a ribosoma e inicio de la traducción libres y se produzca la activación de la traducción. En la **figura 1.12** se puede observar dicho ejemplo.

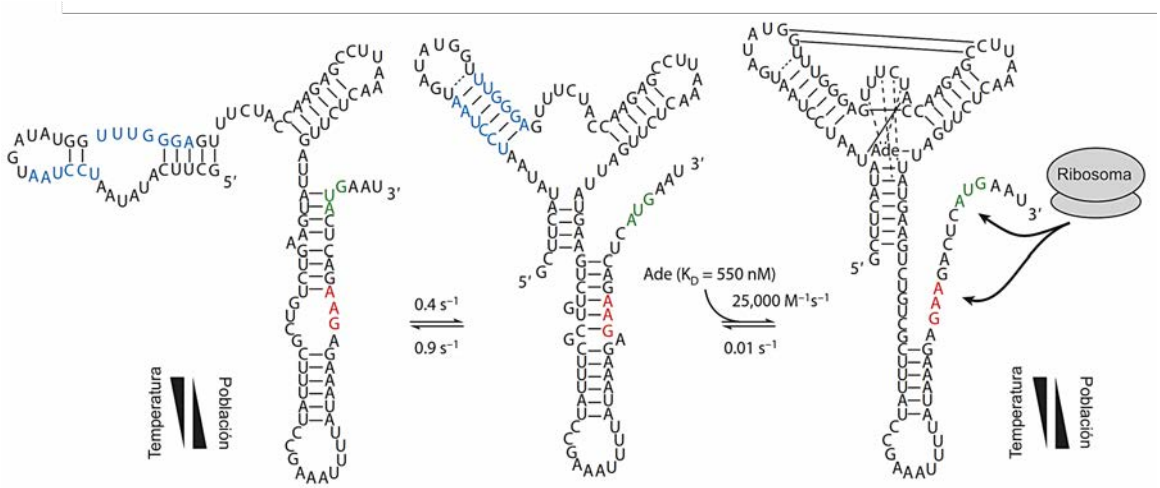



**Figura 1.12:** Equilibrio de tres estados de estructuras secundarias del riboswitch de adenina. En la conformación unida a adenina, tanto el codón de inicio (verde) como el sitio de unión al ribosoma (rojo) están expuestos, dando comienzo al proceso de traducción. La dependencia de la temperatura en el equilibrio de las estructuras secundarias de las formas apo (sin adenina unida) compensa el aumento de la afinidad por el ligando de la conformación competente a baja temperatura [42]. Las constantes de velocidad y equilibrio corresponden a las medidas a 25 ° C. Imagen adaptada de [41].

Por otro lado, también existe regulación a nivel transcripcional mediada por efectores como la temperatura[43], moléculas pequeñas[44], metales[45], pH[46], proteínas[47], otros ARNs[48], etc; todos producen la estabilización del confórmero seleccionado permitiendo así continuar con la transcripción. El intercambio entre las conformaciones con diferente estructura secundaria no se vería favorecido a ocurrir en tiempos razonables

debido a las altas barreras energéticas que las separan, a menos que algún tipo de intervención externa ocurriera. En la **figura 1.13** se pueden observar dos ejemplos gráficos de cómo los efectores pueden regular la transcripción.

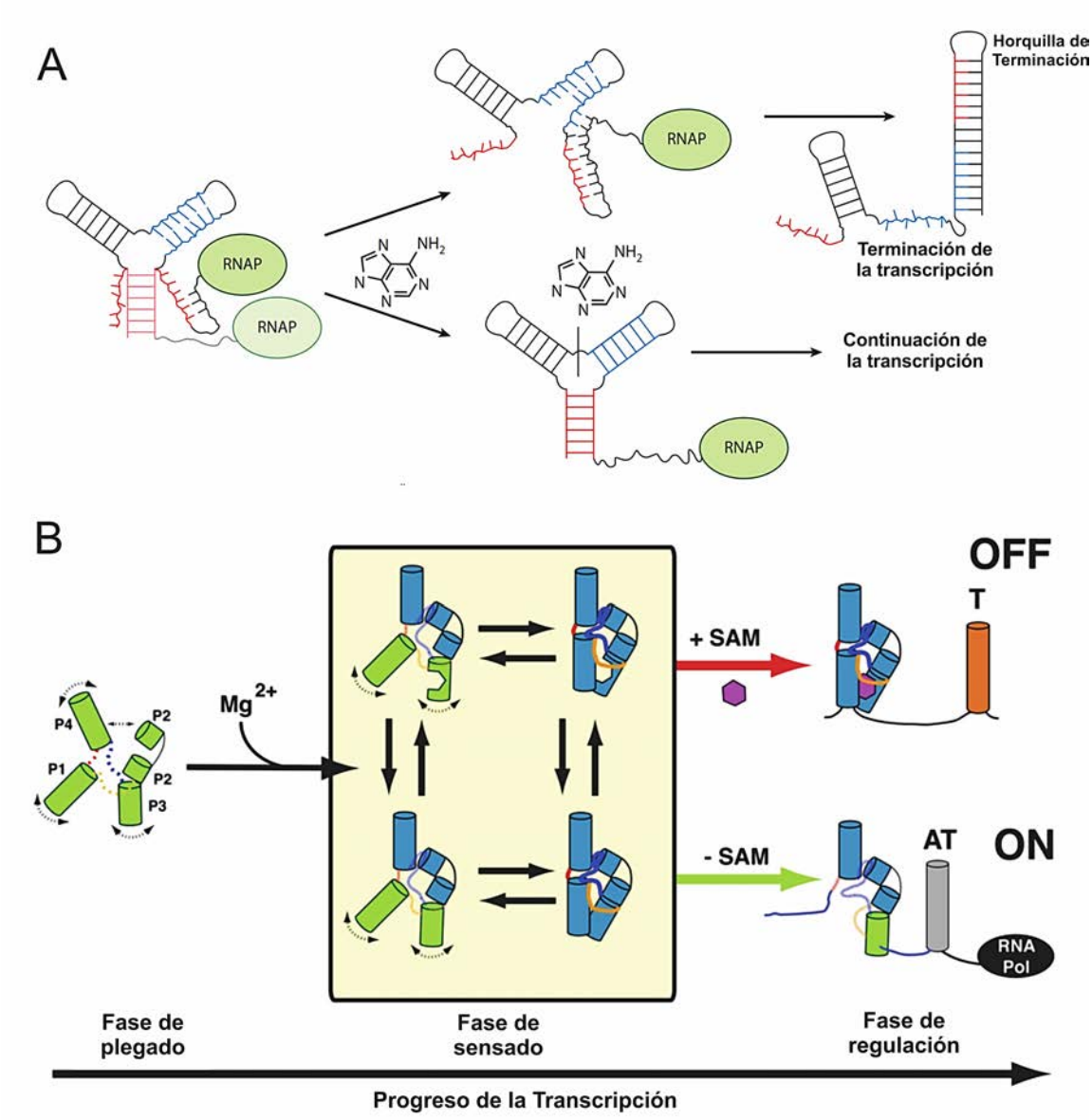



**Figura 1.13:** Ejemplos de regulación transcripcional mediada por confórmeros de ARNs. A – Regulación transcripcional de un riboswitch de adenina. La unión de la adenina (ligando) estabiliza al confórmero que la une, esto produce el secuestro de residuos que de otra manera se aparearían con secuencias transcritas corriente abajo dando lugar a la formación de la horquilla de terminación que se ve favorecida termodinámicamente. B – Regulación transcripcional de un riboswitch SAM-I. El $Mg^{+2}$ favorece la formación del ensamble conformacional estabilizados por la presencia de interacciones terciarias (tipo pseudoknots) e inicia la fase de sensado. Luego, la presencia de SAM-I selecciona un confórmero del ensamble (selección conformacional) con posteriores cambios en la conformación (ajuste-inducido) permitiendo así la formación de la horquilla de terminación. En ausencia de SAM-I, el extremo 3' del dominio P1 es libre de formar interacciones alternativas, ésto estabiliza la horquilla antiterminación y permite que el proceso de transcripción continue. Imágenes adaptadas de [41,49].

Por último, mencionaremos dos ejemplos más, uno correspondiente al nivel 1 y el otro al nivel 2. En el nivel 1 podemos encontrar 4 clases de dinámicas diferentes en apareamiento de bases, siendo (a) fusión de pares de bases (del inglés *base-pair melting*), (b) reorganización de pares de bases (*base-pair reshuffling*), (c) isomerización de pares de bases (*base-pair isomerization*) y (d) interacciones terciarias de largo alcance (*long-range tertiary interactions*). En la **figura 1.14** se esquematizan estas 4 clases y se presenta el ejemplo de la reorganización de bases en la conformación mayoritaria del sitio A ribosomal durante el proceso de reconocimiento de la minihélice codón-anticodón del complejo mRNA/tRNA[50].

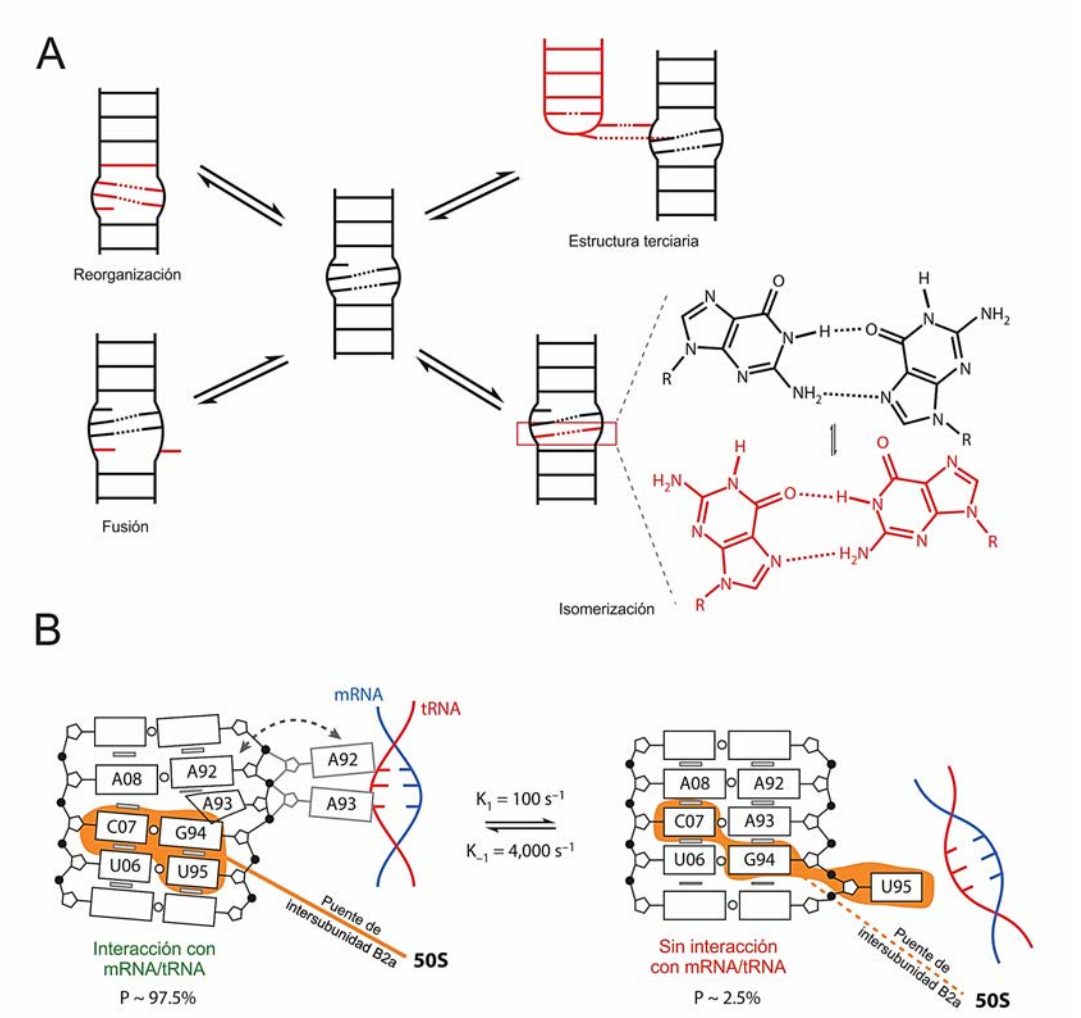



**Figura 1.14:** A – Clases de movimientos del Nivel 1. B – Dinámica de las conformaciones mayoritaria y minoritaria del sitio A del ribosoma en presencia/ausencia del complejo mRNA/tRNA. Como puede observarse en la conformación mayoritaria (P ~ 97,5%), los residuos A1492 (A92) y A1493 (A93) pueden interaccionar y estabilizar el complejo mRNA/tRNA. La conformación minoritaria secuestra estos residuos, inhibiendo la decodificación e interrumpiendo el puente de intersubunidad B2a [51]. Imágenes adaptadas de [41].

Cuando el ribosoma interacciona con la minihélice codón-anticodón del complejo mRNA/tRNA, se estabilizan pares de la hélice para prevenir la pronta disociación del tRNA. Luego, el complejo quintenario cambia su conformación para producir la activación la de hidrólisis del GTP en el factor de elongación. Debemos notar que este tipo de interacción dinámica tan precisa es una interacción terciaria que posee su propia dinámica finamente regulada durante el complejo proceso de decodificación de la información.

La diversidad conformacional de cualquier biomolécula puede relacionarse con la evolución molecular. La ubiquitina es un ejemplo clave, pues su ensamble conformacional ha evolucionado para contener un gran número de subestados conformacionales, dejando entrever que quizás haya evolucionado para permitir un gran número de diversos patrones de interacción o, dicho de otra forma, los patrones de interacción han aprovechado la diversidad conformacional preexistente de la ubiquitina. Una situación similar se produce en la relación anticuerpos-antígenos[52]. Segundo, cuando la secuencia de una biomolécula adopta diferentes conformaciones, menor tiempo atraviesa en cada una de ellas. A su vez, cuando la diversidad conformacional de una biomolécula se ve reducida, la variabilidad (capacidad de aceptar mutaciones una conformación) y la capacidad de evolucionar, es decir, la probabilidad de alcanzar mejores estados de *fitness* a través de la variabilidad, decrecen fuertemente[53]. Y considerando al evento de duplicación de genes como un mecanismo de evolución eficaz en la generación de nuevas funciones, debido a que una de las copias mantiene la función mientras las otras evolucionan para dar nuevas funciones; diremos que lo mismo se puede pensar con la presencia de diversidad conformacional en alguna biomolécula. Entonces, supongamos la evolución de una población de una cierta biomolécula. Si ésta población la representamos con un cierto número de individuos que, además, cada una de ellas sólo tiene una única conformación en su estado nativo, éstas tendrán un determinado fitness, donde algunas pueden encontrar leves mejores estados que otras, permitiéndole así a la población posicionarse sobre ciertos picos en el paisaje de *fitness* (del inglés "*fitness landscape"*). Esta condición de única conformación reduce considerablemente -a nivel poblacional- la posibilidad de alcanzar los estados más altos en el paisaje de fitness. Por el contrario, cuando éste cierto número de individuos poseé más

de una conformación en su estado nativo, se incrementa fuertemente la posibilidad de alcanzar los picos más altos del paisaje de fitness a la población. Esto último puede apreciarse gráficamente en la **figura 1.15**.

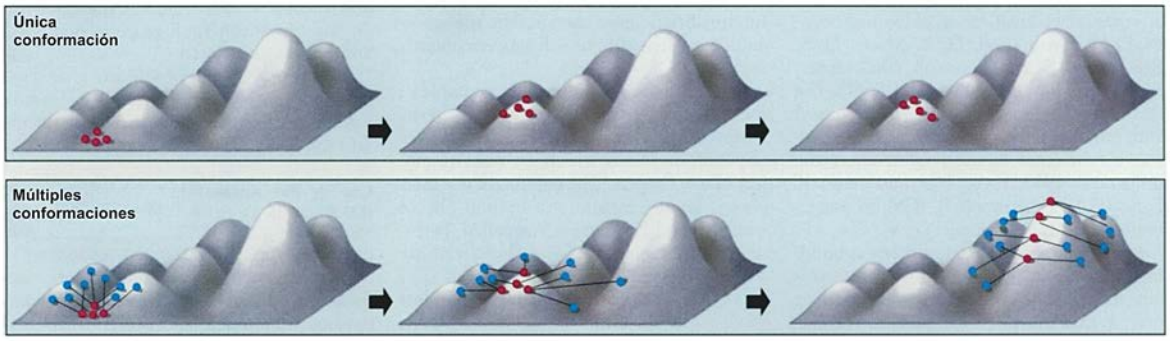



**Figura 1.15:** Representación gráfica simplificada del proceso de evolución molecular de una población en un paisaje de fitness considerando moléculas con única conformación (arriba) y múltiples conformaciones (abajo). Los picos representan estados de fitness mejorados. Imagen adaptada de [54].

# 2. **Objetivos**

Hemos presentado conceptos fundamentales sobre propiedades intrínsecas de las biomoléculas (proteínas como ácidos nucleicos) que contribuyen a nuestra comprensión de cómo llegan a desempeñar sus funciones biológicas. Como es de esperar, la mayoría de los conceptos tratados han surgido y se han basado en proteínas. Nosotros nos abocaremos a estudiar los ARNs, entendiendo que el conocimiento con el que se cuenta de ellos actualmente ha venido creciendo en forma considerable los últimos 30 años.

Con intención de poder contar qué nos ha motivado a realizar este trabajo final, enunciaremos algunos de los principales sucesos que han contribuído en la contemporaneidad de la biología molecular del ARN. Podemos decir que el primer hito en la biología molecular del ARN fue el descubrimiento de la actividad autocatalítica del intrón del grupo I de *Tetrahymena* (1982), el cual les valió el premio Nobel de química a Thomas Cech y Sidney Altman en 1989, dejando asentado la importancia de las ribozimas en la biología y fortaleciendo la teoría del "mundo del ARN". El hecho impactó fuertemente en la comunidad científica de entonces, ya que se sabía muy poco de los otros roles que desempeña el ARN, exceptuando el de ser portador de la información genética. Otro paso en la dilucidación de la biología del ARN, a pesar de que ya se tenía conocimiento de su rol regulador posttranscripcional[55,56], fue la demostración de la regulación génica por ARNs de interferencia en 1998[57]. Esto también valió otro premio novel, en este caso de Fisiología o Medicina, a Andrew Fire y Craig Mello en 2006. Más recientemente, las nuevas tecnologías de secuenciación brindaron la posibilidad de obtener con relativa sencillez la información secuencial sobre moléculas biológicas. Actualmente se acumula de forma

acelerada una gigantesca cantidad de información secuencial, cuyo volumen impide comprender por completo su participación en procesos biológicos. Teniendo esto en cuenta, es importante resaltar unos de los aportes más recientes e inquietantes que podemos mencionar al respecto, siendo éste el conocimiento que se tiene sobre la transcripción de una gran proporción del genoma humano, superando el 83% del material genético total[58,59]. Esto, comparado al material codificante traducido para construir proteínas (casi 2% del genoma), no deja de ser simplemente impresionante. ¿Qué podemos decir de todas esas moléculas? ¿Qué estructuras tienen? ¿Cuántas de ellas tienen una función biológica asociada? ¿Cuál es la relación entre su estructura y su función? Existen novedosos trabajos que persiguen estas preguntas[60–66], pero aún restan muchos interrogantes por enfrentar. Así como en 1944 la publicación de Avery y colaboradores desafió y cambió el paradigma de dónde reside la información biológica; o cuando en 1961 Monod y colaboradores dieron el primer paso en lograr entender cómo la información biológica era transportada del ADN a una proteína, es claro que la biología molecular hoy está transitando un avance en la comprensión del rol del ARN.

## 2.1 Objetivo general

El objetivo general de este trabajo se centrará en el estudio del ARNs, desde el campo de la biología estructural, haciendo uso de herramientas computacionales y diferentes bases de datos biológicas, con el fin de mejorar nuestra comprensión de su relación estructura-función.

## 2.2 Objetivos específicos:

- Crear una base de datos de diversidad conformacional de ARNs.
- Anotar dicha base con propiedades biológicas.
- Derivar información biológica desde la base de datos.

# 3. Materiales y métodos

# Biología estructural del ARN

## 3.1 El comienzo, el dogma central y las primeras evidencias

Se puede decir que la biología estructural del ARN comienza a desarrollarse luego del modelo de la doble hélice de ADN presentado por J. Watson y F. Crick el 25 de abril de 1953[67]. Nada se conocía de la estructura tridimensional del ARN, tampoco de los tipos de ARNs y menos sobre sus roles biológicos. Sí se conocía su localización celular, diversidad de tamaños y su posibilidad de formar estructuras ramificadas mediante el grupo hidroxilo (-OH) del C2'de la ribosa. Comparado al ADN, estas diferencias sumaban incertidumbre sobre todo en lo referido a su estructura. A pesar de presentar dificultad técnica en la obtención de ARN puro para la realización clara de patrones de difracción, ya para fines de la década de 1950 se había demostrado que el ARN podía formar estructuras de doble hélices ARN-ARN[68], ARN-ADN[69] y triple hélices ARN-ARN-ARN[70,71].

Fue en 1957 que F. Crick anuncia la hipótesis del adaptador molecular y el dogma central de la biología molecular[72]. Ya para 1958 se logra aislar el primer ARN que aportaría evidencia sobre la hipótesis del adaptador molecular (tRNA) involucrado en la síntesis de proteínas[73]. Aún quedaba por demostrar cómo la información contenida en el ADN era transportada/transferida a la síntesis de proteínas. Luego de años de erróneas hipótesis y equivocadas interpretaciones de resultados experimentales por diferentes grupos de investigación[74], S. Brenner, F. Jacob y M. Meselson logran aislar, hacia mediados de los

años 60, el primer mRNA. En mayo de 1961, su trabajo culminará en el artículo demostrando la existencia de la molécula encargada de transportar la información genética desde el DNA al ribosoma para que se produzca la síntesis de proteínas[75]. Por acuerdo previo entre los autores, el mismo día en que se publicó el artículo de Brenner-Jacob-Meselson, también se publica el artículo liderado por F. Gros en que demuestra la existencia de un ARN vinculado a la síntesis de proteínas, éste era diferente al ribosomal y tRNA[76]. Ese mismo mes de mayo también sale publicado el trabajo de Jacob y Monod (enviado en diciembre de 1960) donde anuncian su modelo de regulación génica en la síntesis de proteínas[77], que ya contaba con la idea de la existencia del mRNA. La primera evidencia que una molécula de ARN provenía del ADN fue publicada en 1962[78], cuando se logró demostrar que el tRNA de E.coli se hibridaba con el ADN del genoma de E. coli. En ese momento, comprobar la existencia de la formación de una doble hebra híbrida (ADN-ARN) era una fuerte evidencia para proponer un mecanismo sobre cómo la información genética era transferida.

Se estaban formando así las bases para el entendimiento de cómo la información biológica, contenida en el ADN, es transportada en forma de mRNA hasta el ribosoma y, con ayuda del adaptador de Crick (tRNA), es traducida a una proteína. Rápidamente se logró descifrar el código genético[79], sentando así las bases de la biología molecular. Paralelamente, el tRNA era objeto de estudio y recién en 1965, luego de desarrollar el primer método de secuenciación de oligonucleótidos largos[80], R. Holley y colaboradores logran evidenciar por primera vez la estructura primaria de una molécula de ARN biológica, el ARN de transferencia de alanina ($tRNA^{Ala}$) de levadura[81]. En dicho trabajo también dejan asentado, de forma especulativa, una primera hipótesis de tres posibles estructuras secundarias que ésta adoptaría en solución (**figura 3.1.1**), aclarando que, cualquiera sea la conformación real, dependería de las condiciones de la solución.

Claramente, el trabajo de Holley tiene presente la idea de que las biomoléculas presentan una única conformación en su estado nativo.

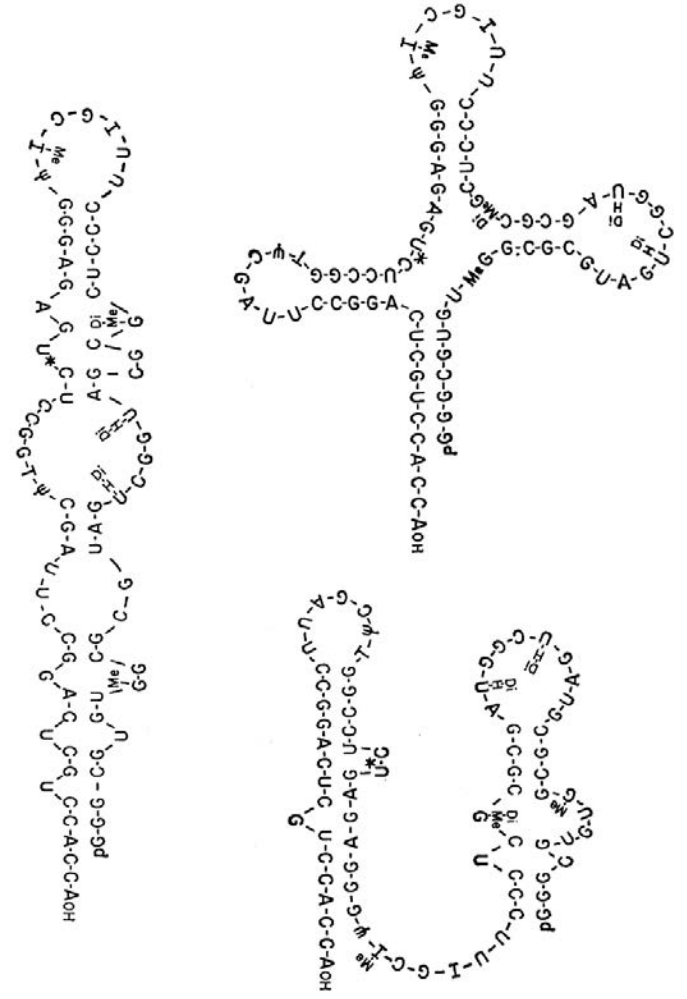


**Figura 3.1.1:** Representación esquemática realizada por Holley y colaboradores en 1965 sobre tres conformaciones posibles del $tRNA^{Ala}$. En cada una de ellas puede observarse pequeñas regiones doble hélice. Imagen adaptada de [81]

La primera molécula de ARN a la cual se le conoció su información tridimensional fue el $tRNA^{Phe}$ de E. coli. Esta evidencia experimental, luego de varios intentos fallidos y de baja resolución cristalográfica[82,83,84], fue presentada en 1974 con una resolución de 3 Å[85,86]. El largo proceso implicó una mejora en los métodos de purificación de tRNA durante la década del 60 y también en las condiciones de cristalización[82]. En la **figura 3.1.2** se puede observar el modelo propuesto.

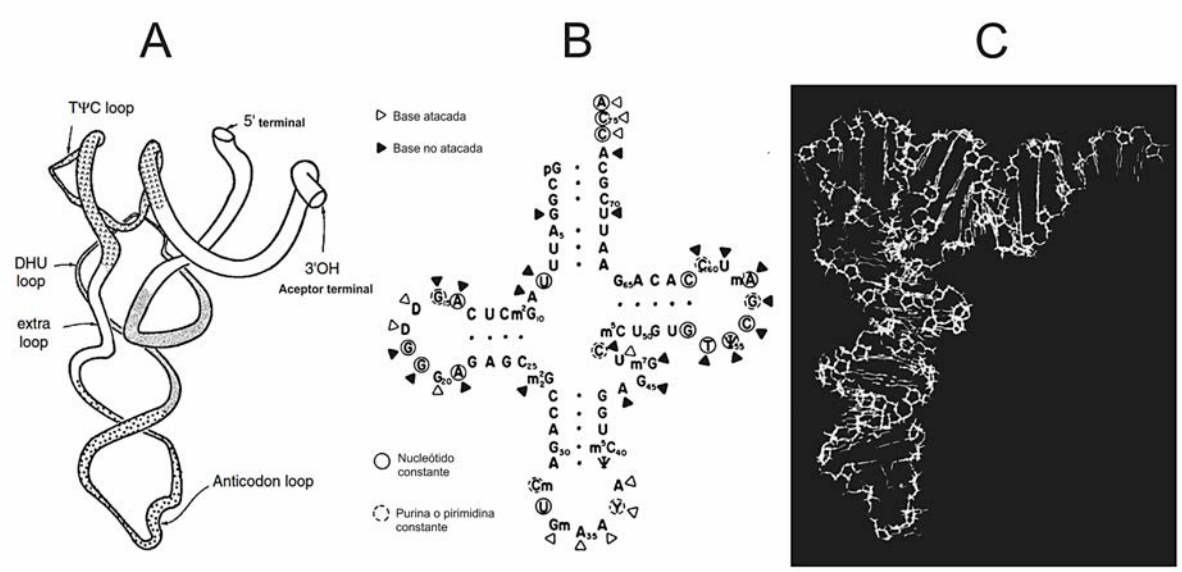



**Figura 3.1.2:** A – Trazado de toda la cadena completa del $tRNA^{Phe}$ de levadura a 4Å de resolución. La forma indica diferentes segmentos de horquilla-loop. La molécula muestra una clara forma de "L"[84]. B – Secuencia del $tRNA^{Phe}$ de levadura en su configuración on forma de trébol. Las bases encerradas por círculos son constantes en todos los tRNAs, mientras que los círculos punteados indican las posiciones que son ocupadas constantemente por purinas o pirimidinas. Los triángulos blancos y negros indican la accesibilidad de ser modificados químicamente, siendo accesibles y no accesibles, respectivamente[85]. C – Estructura tridimensional completa del $tRNA^{Phe}$ de levadura publicada en 1974[85].

Una de las principales evidencias que demostró la información estructural del $tRNA^{Phe}$ fue la observación de interacciones terciarias, fundamentalmente entre sus dos brazos. También, observar la forma de "L" del tRNA junto a la presencia de múltiples nucleótidos modificados. Todas estas evidencias dieron lugar a plantear nuevos interrogantes en la biología del ARN, especialmente debido a su capacidad de formar estructuras terciarias con variadas interacciones nunca antes previstas. Se daba así el comienzo y desarrollo al campo de la biología estructural del ARN, pero a un ritmo muchísimo menor comparado al de las proteínas. Deberían pasar unos 20 años desde la aparición de la primera estructura de ARN

para conocer la segunda estructura de otro ARN, la del Hammerhead ribozyme (HHR; ribozima cabeza de martillo)[87], aunque recién para 2006 se tendrá la primera estructura biológicamente relevante de HHR[88,89]. Otro gran paso en la obtención de estructuras de ARN se dio en 1996, cuando se logró duplicar el tamaño máximo de una molécula resuelta por cristalografía que, hasta por ese entonces, era el tRNA. Esta molécula era el dominio P4-P6 del intrón del grupo I de Tetrahymena thermophila y contaba con 160 nucleótidos (nt)[90] que comparados con los 75 nt de los tRNAs constituían un progreso importante.

Como ya se comentó previamente, el progreso en la obtención de estructuras tridimensionales de ARN, comparado al de proteínas, comenzó mucho después e incluso se dio a ritmo más lento. Esto puede verse de manera gráfica en la **figura 3.1.3**. También lo fue la creación de la primera sociedad dedicada exclusivamente al ARN (*The RNA Society* - *https://www.rnasociety.org/*) siendo conformada en 1993, produciendo luego la primera revista bajo la misma intención en 1995 (*http://rnajournal.cshlp.org/*)[91].

Desde 1974 ha habido un crecimiento sostenido en lo que respecta al conocimiento sobre la función y estructura del ARN, así como también el descubrimiento de muchos tipos nuevos de ARNs, incluyendo ribozimas, pequeños ARNs nucleares del espliceosoma (snRNAs), microRNAs, etc. Hoy en día, bajo los gigantescos avances de nuevas tecnologías de secuenciación y/o computacionales, vemos con especial interés esta biomolécula, tanto desde el enfoque evolutivo como en lo estructural-predictivo. Así, a pesar de que cada día se dan pequeños avances en materia de aplicación, descubrimiento y comprensión de la biología de los ARNs, aún nos queda mucho camino por recorrer.

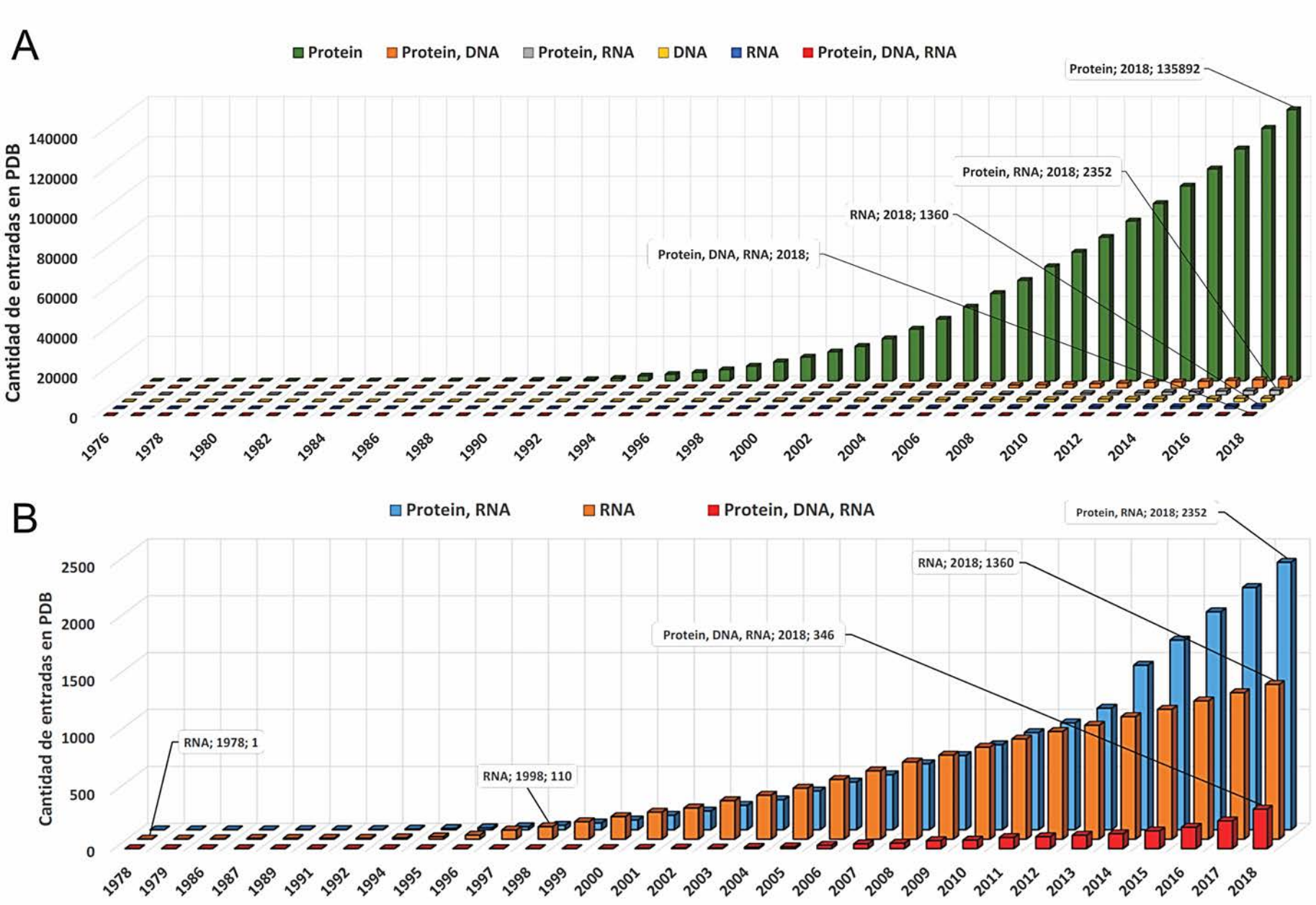



**Figura 3.1.3:** A – Evolución temporal de la deposición de estructuras en PDB hasta la actualidad. B – Evolución de deposición en PDB de las principales entradas conteniendo al menos una molécula de ARN. Se resaltan datos actuales e históricos cada 20 años desde la primer estructura de ARN depositada en la PDB.

## 3.2 Estructura química y *backbone*

Cuando nos referimos a las biomoléculas debemos tener en cuenta sus arreglos espaciales, las interacciones a las que están sometidas, ya sean interacciones intra o intermoleculares, sus dinámicas, pero también sus constituyentes. Si nos remitimos en describir al ARN, debemos, primero, saber que su estructura primaria está definida por la secuencia de nucleótidos de bases nitrogenadas púricas (A: adenina; G: guanina) y pirimidínicas (C: citosina; U: uracilo). Un ejemplo esquemático de ello puede observarse en la **figura 3.2.1**.

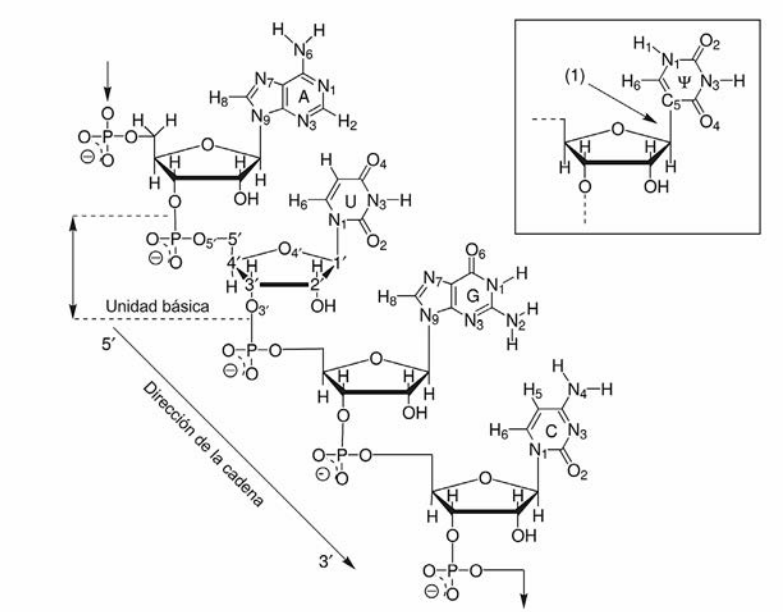



**Figura 3.2.1:** Estructura primaria del ARN "AUGC" dibujada desde el extremo 5' al 3'. Las bases de los ARNs frecuentemente son modificadas, especialmente el tRNAs, donde se suele encontrar timina (5 metil uracilo) o pseudouridina (Ψ, contiene enlace glucosídico entre el C5 de la base y el C1' de la ribosa). Imagen adaptada de [92].

Un aspecto muy importante al momento de estudiar los constituyentes del ARN es saber que existen nucleótidos modificados. Esto ocurre frecuentemente en los ARNs lo cual redunda en una diversidad en las posibilidades de interacción, impactando de manera significativa en la estructura tridimensional, su estabilidad y función biológica[93]. En la **figura 3.2.2** se pueden observar ejemplos de estas modificaciones para un tRNA. Hoy en día se conocen al menos 100 nucleótidos modificados. Existen bases de datos (por ejemplo, *http://mods.rna.albany.edu/home*) encargadas de recopilar y clasificar estos nucleótidos presentes en la naturaleza[94–96].

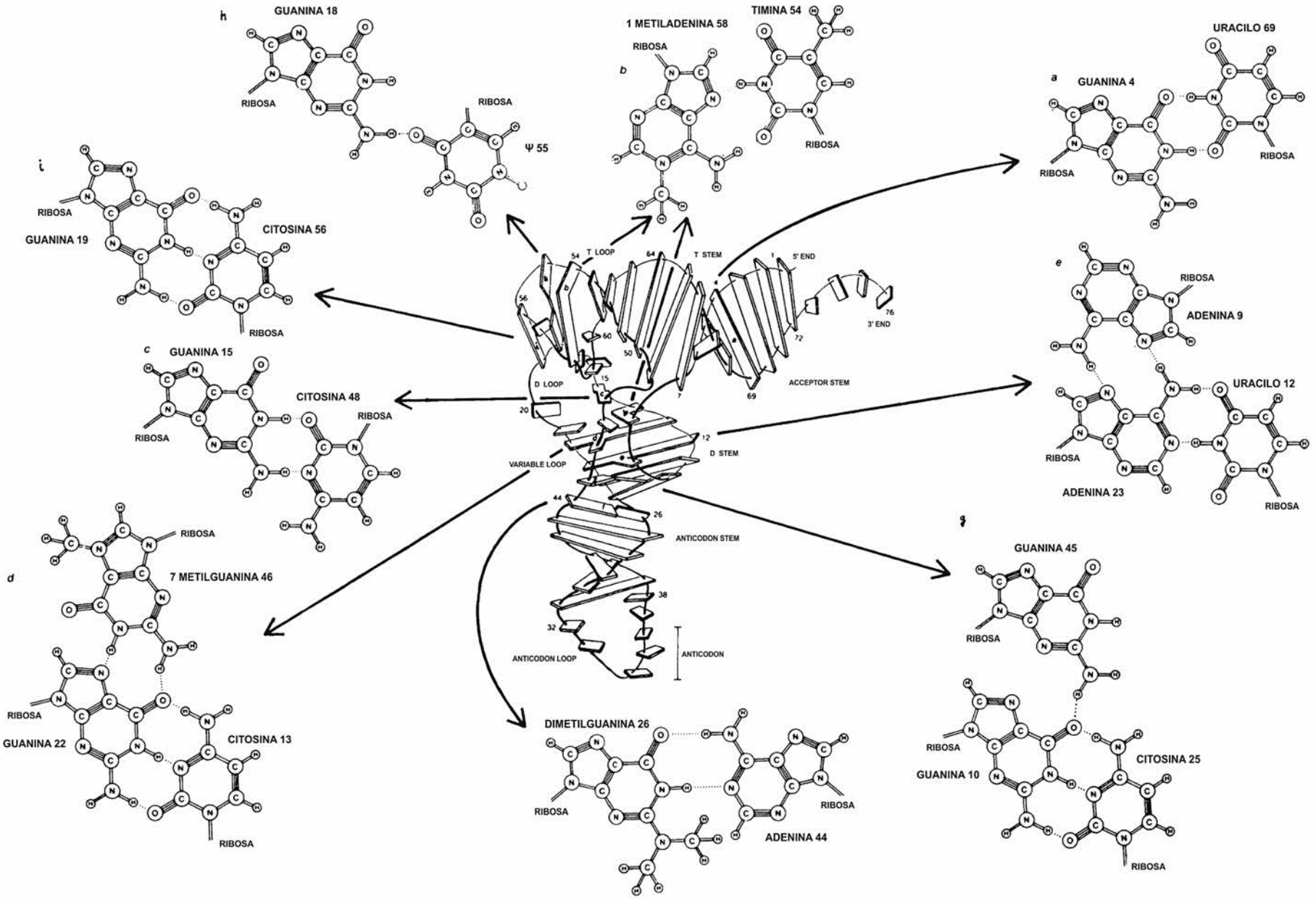



**Figura 3.2.2:** Pares de bases terciarias de la estructura del tRNA$^{Phe}$ de levadura. Se observan varias bases modificadas. Hay que notar que todas las pares de bases terciarias son diferentes a los típicos pares de bases Watson-Crick y están localizadas cerca de la esquina de la "L". Imagen adaptada de [97].

La unidad básica o monómero, que define al polímero de ARN es el nucleótido. Está conformado invariablemente por una ribosa fosforilada en el carbono 5' (C5') unida a una base nitrogenada por el C1'. Su organización secuencial, mediante la unión fosfodiéster con el oxígeno 3' (O3') del C3' de la ribosa, delimita un claro esqueleto estructural llamado esqueleto "ribosa-fosfato". Este esqueleto ribosa-fosfato (de ahora en adelante, *backbone*, del inglés de "esqueleto"), posee seis ángulos de torsión (**α**, **β**, **γ**, **δ**, **ε**, y **ζ**). Cada uno de ellos pueden moverse condicionadamente en ciertos grados (grados de libertad). A pesar de ello, el *backbone* de los ARNs presentan una gran flexibilidad. Existe, a su vez, un séptimo ángulo de torsión (**χ**) muy importante, aunque no pertenece al *backbone*, entre el C1' de la ribosa y el N (nitrógeno) de la correspondiente base nitrogenada. Por otro lado, la ribosa tiene sus propios ángulos de torsión en el anillo. Preferentemente, en la naturaleza, el anillo de ribosa de los ARNs se encuentra en su conformación C3'-endo, lo cual lleva adoptar, al *backbone* de una doble hélice, la conocida forma A-RNA. En la **figura 3.2.3** se puede observar un esquema de la unidad básica del ARN con sus ángulos correspondientes y también la distancia de separación entre los átomos de fósforo consecutivos para las dos mayores conformaciones del anillo de la ribosa en los ARNs.

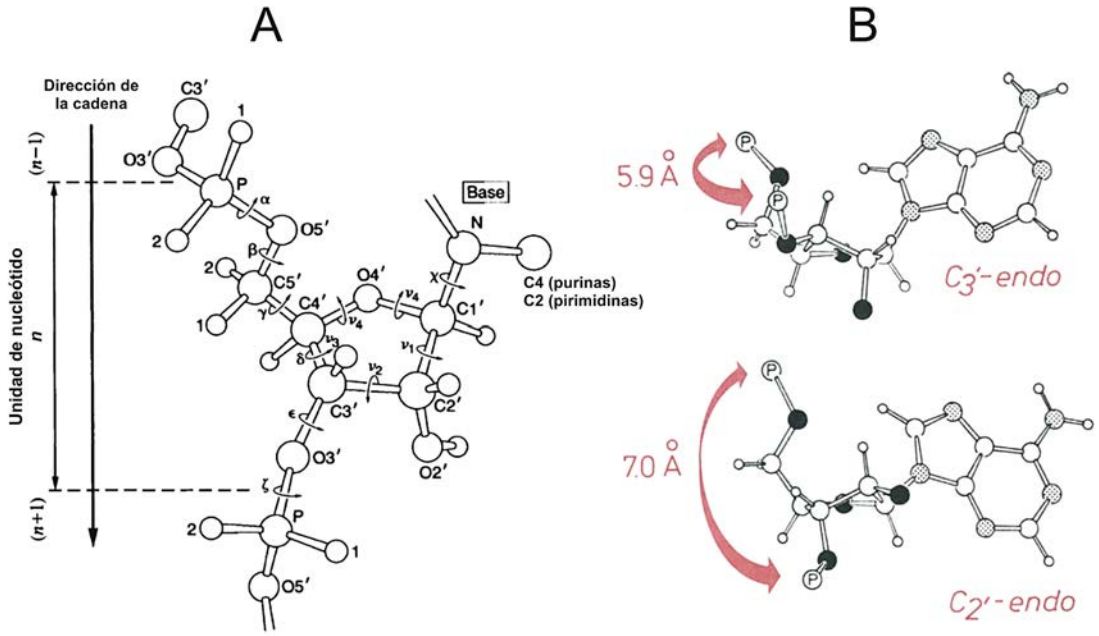



**Figura 3.2.3:** A – Representación esquemática de la unidad básica del *backbone* de ARN con sus ángulos de torsión correspondientes. Los nucleótidos se cuentan desde arriba hacia abajo, en la dirección del O5' → O3'. B – Conformaciones observadas en las formas de hélices A y B, siendo C3'-endo (arriba) y C2'-endo (abajo) respectivamente. A su vez, se indican las distancias (en angstroms) fosfato-fosfato cada cada una de ellas. Imágenes adaptadas de [98,99].

Otra forma de estudiar el *backbone* ribosa-fosfato, es definiendo una unidad de estudio diferente, llamada "suite", que involucra aquellos ángulos de torsión entre dos ribosas consecutivas, estos son $\delta_{(i-1)}$, $\varepsilon_i$, $\zeta_i$, $\alpha_i$, $\beta_i$, $\gamma_i$ y $\delta_i$. Esto permitió realizar estudios sobre los confórmeros detectables, concluyendo así su propiedad rotamérica[100]. De esta definición surgió una nomenclatura modular de dos caracteres para cualquier confórmero en una secuencia de una estructura (**Tabla 3.1**). Contar con información estructural permitió realizar estudios sobre las conformaciones del *backbone* del ARN, crear diferentes librerías, desarrollar software específico, e identificar 46 familias de confórmeros[100,101]. En la **figura 3.2.4** se puede observar el *backbone* de un dinucleótido con los ángulos que conforman la suite.

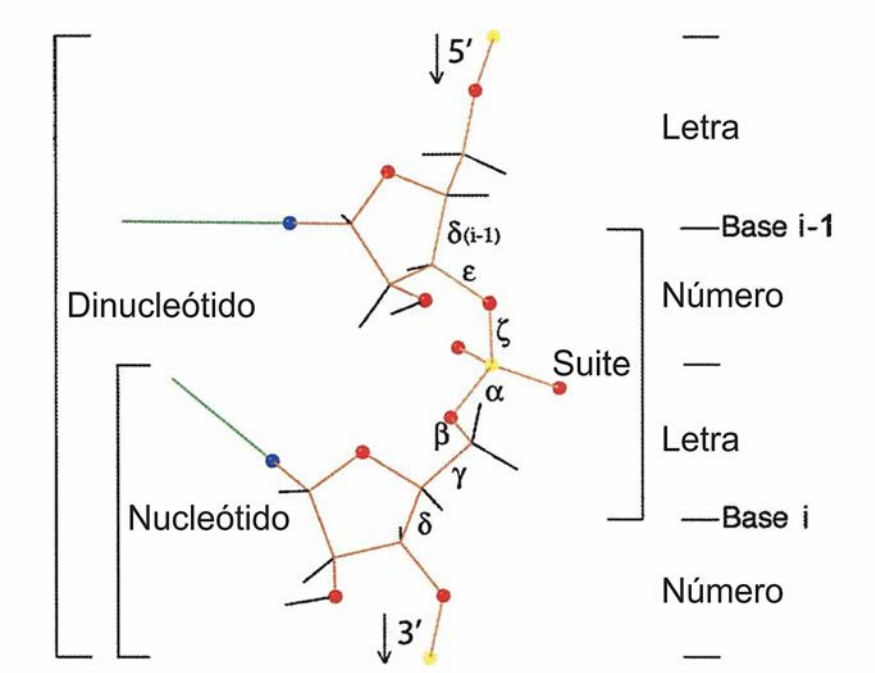



**Figura 3.2.4:** *Backbone* del ARN con ángulos de torsión y división de *Suite*, nucleótido y dinucleótido. Las unidades modulares del heminucleótido se muestran a lo largo del lado derecho, donde recibe letra o número en la nueva nomenclatura. Imagen adaptada de [101].

| | |
|---|---|
| **Heminucleótidos δεζ :** | |
| Conformaciones C3'-endo: Números impares: | Conformaciones C2'-endo: Números pares: |
| Código de ángulos δεζ | Código de ángulos δεζ |
| **1** = 3′-em | **2** = 2′-em |
| **3** = 3′-et | **4** = 2′-et |
| **5** = 3′-ep | **6** = 2′-ep |
| **7** = 3′-e-e | **8** = 2′-e-e |
| **9** = 3′-ee | **0** = 2′-ee |
| **&** = 3′t-e | **#** = 2′te |
| **Heminucleótidos αβγδ :** | |
| Código de ángulos αβγδ | Mnemotécnico |
| Conformaciones C3'-endo: | |
| **a** = mtp3′ | **1a** is A-form |
| **c** = ttt3′ | **1c** is "crankshaft" variant of A-form |
| **d** = ptp3′ | inverted "p"; see below |
| **e** = -ept3′ | **1e** is stack-shift dent; only eclipsed α |
| **f** = tet3′ | |
| **g** = ttp3′ | **1g** is suite 1–2 of GNRA tetraloop |
| **h** = mtt3′ | |
| **i** = p-et3′ | |
| **j** = pet3′ | |
| **L** = mep3′ | |
| **m** = m-ep3′ | Minor 1a shoulder |
| **n** = ptt3′ | **6n** is 2′3′ Z-form; "N" is rotated form of "Z" |
| Conformaciones C2'-endo: | |
| **b** = mtp2′ | **2b** would be B-form DNA |
| **o** = mtm2′ | **1o** and **2o** both put bases opposite each other |
| **p** = ptp2′ | Most p angles in 2′ set |
| **q** = pet2′ | |
| **r** = ptm2′ | Rare reverse order of common m t p |
| **s** = mpt2′ | **4s** is commonest suite 2–3 of S-motif |
| **t** = ttt2′ | All-*trans* |
| **z** = ttp2′ | **5z** is 3′2′ Z-form |
| **[** = m-ep2′ | **1[** is commonest intercalation conformation |

**Tabla 3.1:** En la siguiente tabla se muestra la nomenclatura modular de la suite de ARN. Para todos los heminucleótidos: (_) Suites con algún ángulo sin definir (se da en terminación de cadena o loops desordenados). Se usa "L" para mayor claridad, pero debería estar en minúscula en los estudios computacionales. (!) Conformaciones inusuales: Suites o heminucleótidos que no se encuentran en la lista, malos valores de $\varepsilon$, etc. Entonces, ! denota algo malo o interesante. Nota: en la lista de $\delta\varepsilon\zeta$, el código es un número (significando un símbolo en el rango 002-003 de Unicode) para el primer carácter del nombre del confórmero de consenso modular; en las listas de $\alpha\beta\gamma\delta$ que describen el segundo carácter de los nombres del confórmero, el cógido es una letra (un símbolo >005A en Unicode). Para los ángulos diedros, m significa -60° (minus); t , 180° (trans); p, +60° (plus); e, 120° $\pm$ 25°; -e, -120° $\pm$ 25°. Tabla adaptada de [101].

Haber logrado un consenso en la nomenclatura de los confórmeros descriptos por las suites correspondientes ha sido un avance en el estudio e interpretación de la información estructural. Sin embargo, es necesario decir que hay tres principales problemas en la identificación de suites estrechamente relacionadas[101] : las interacciones terciarias que las bases, ribosas y fosfatos experimentan entre ellas y con otras moléculas como el agua y iones ($Mg^{+2}$, $K^{+}$, $Na^{+}$, etc); la composición de bases de cada especie y más aún, la composición de bases en las estructuras secundarias; y las poblaciones de suites que estadísticamente necesitan ser significativas para ser detectadas. Posiblemente, dichos problemas irán disminuyendo al contar con una mayor cantidad de información tridimensional de buena resolución, con una previa evaluación de calidad, como también al aumentar la diversidad en la procedencia de la información biológica, buscando así generar un mejor *dataset* de trabajo.

## 3.3 Estructura secundaria, terciaria y motivos de ARNs

Las definiciones de estructura secundaria y terciaria rigen para cualquier biomolécula. Podemos decir, entonces, que una estructura secundaria de ARN será aquella en donde la estructura primaria adopte un arreglo local y que se encuentre estabilizada por puentes de hidrógeno; una estructura terciaria de ARN será aquella en donde uno o más tipos de estructuras secundarias, de una misma cadena, adopten un arreglo tridimensional estable producto de fuerzas de distintas interacciones débiles.

Como ya vimos, el *backbone* de los ARNs es muy flexible. Sumado a esto, se sabe que la información biológica para la formación de las estructuras terciarias está contenida en la estructura primaria, la cual, en este caso, solo tiene como posibilidad de variación los cuatro nucleótidos antes descritos. Si lo comparamos con las proteínas, la diversidad en monómeros es 5 veces menor, dejando entrever una limitación en la formación de estructuras terciarias. A pesar de ello, las moléculas de ARN presentan una gran cantidad

de estructuras secundarias muy estables. En la **figura 3.3.1** se pueden ver tipos de estructuras secundarias.

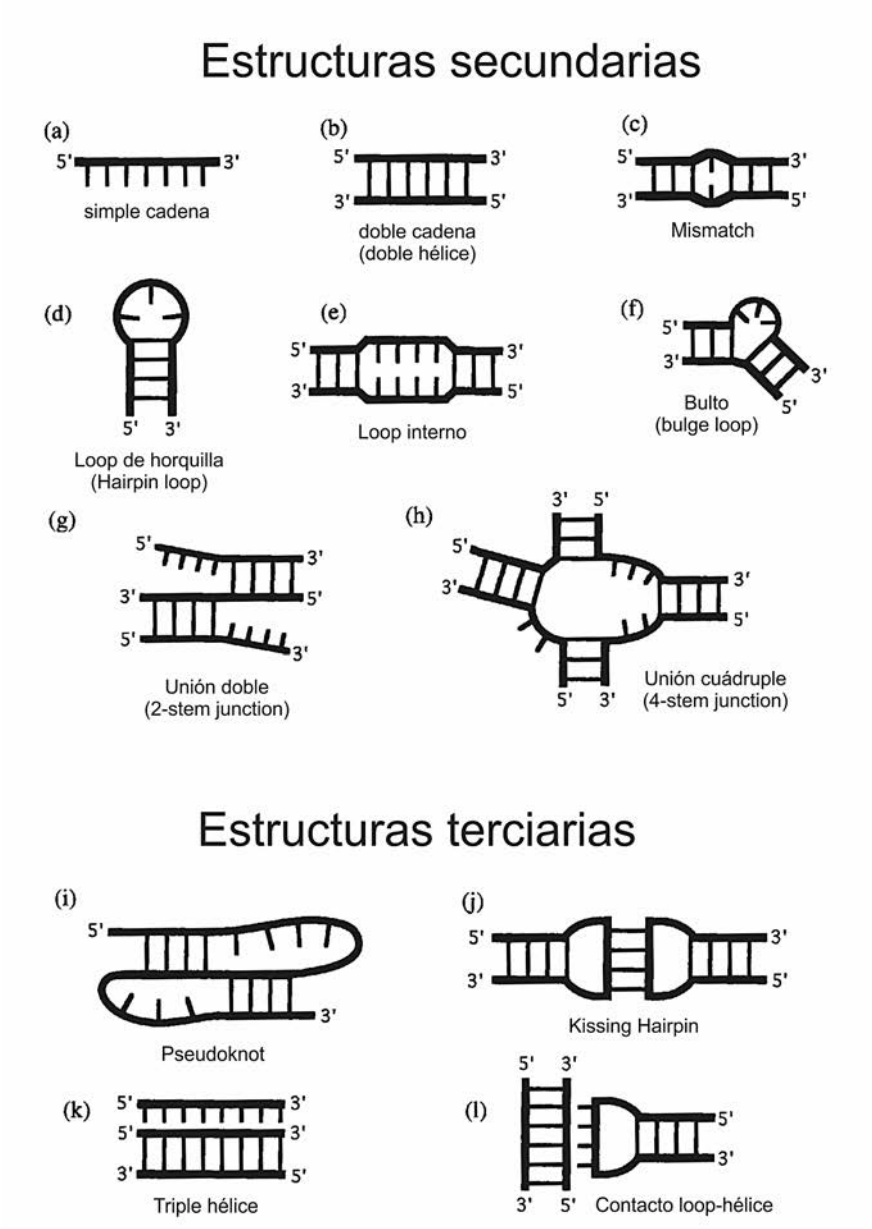



**Figura 3.3.1:** Representación simplificada de ejemplos de estructuras secundarias y terciarias en ARNs. Las líneas gruesas representan el *backbone* y las líneas finas las bases nucleotídicas. Imagen adaptada de [102].

Al estudiar las estructuras secundarias y terciarias debemos tener en cuenta las interacciones que los nucleótidos pueden presentar entre ellos[103–106]. Esto es un aspecto de gran importancia en la relación estructura-función. Ahora, si nos enfocamos en el estudio de estas interacciones, debemos tener presente una clasificación lo más representativa posible[106,107]. Una de las clasificaciones más aceptadas se basa en la abstracción geométrica de los nucleótidos en forma de triángulos, donde cada lado está simbolizado por interacciones de diferentes grupos químicos funcionales[107,108]. En la **figura 3.3.2** se resumen las interacciones de pares de nucleótidos posibles.

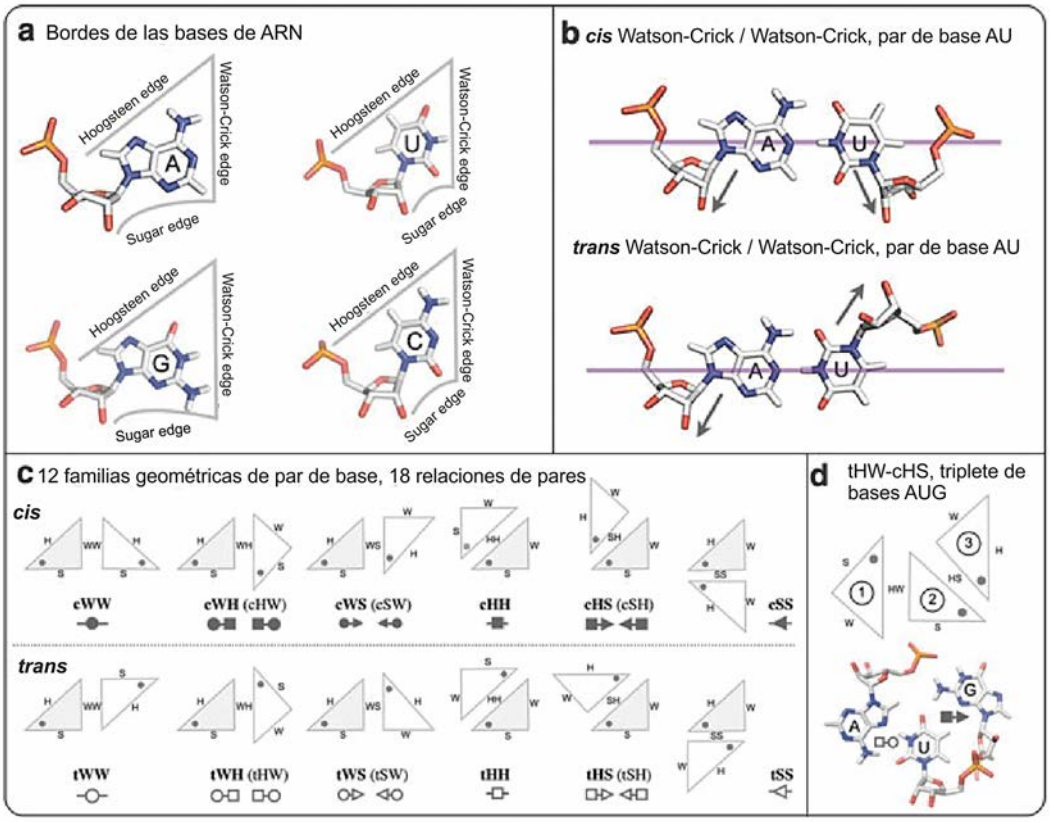



**Figura 3.3.2:** Resumen de la clasificación de Leontis/Westhof para las interacciones de bases. **a** – Cada nucleótido de ARN sin modificar puede ser representado por tres bordes de interacción, Hoogsteen (H), Sugar (S) y Watson-Crick (W). Por lo tanto, las bases nucleotídicas poder ser simplificadas y representadas por triángulos. El borde Sugar incluye al grupo 2' OH de las ribosas. **b** – Para cada par de interacción, los nucleótidos pueden poseer interacciones tipo ***cis*** o ***trans*** dependiendo de cómo se ubiquen respecto del eje imaginario de color magenta que es perpendicular a la interacción entre 2 bordes. De forma gráfica, los enlaces glucosídicos que se encuentren sobre el mismo lado del eje corresponde a la configuración ***cis***, mientras si se encuentran en lados opuestos se corresponde a la configuración ***trans***. **c** – Representación esquemática de las 12 familias geométricas utilizando la nomenclatura H, S y W, y símbolos de círculos, cuadrados y triángulos, llenos o vacíos para las 18 relaciones de pares existentes debido a la presencia de asimetría. Como se puede observar, los símbolos llenos corresponden a las configuraciones tipo ***cis***, por el contrario, los símbolos vacíos corresponden a las configuraciones tipo ***trans***. **d** – Representación esquemática de un triplete de bases AUG. La base central, nombrada base "2" (base U), interacciona con las otras dos bases mediante bordes diferentes. El triplete tHW/cHS es nombrado acorde a las relaciones que lo forman, siendo la interacción de "1" con "2" (tHW) y "2" con "3" (cHS). Imagen adaptada de [109].

Los ARNs también pueden presentar interacciones entre más de dos nucleótidos de forma simultánea, formando tripletes o cuartetos de bases. Un ejemplo esquemático de ello puede verse en la **figura 3.3.3** donde el nucleótido de guanina interacciona con otros tres nucleótidos desde sus diferentes "lados" de interacción. Esta propiedad de presentar múltiples tipos de interacción, diferentes a las comúnmente conocidas Watson-Crick (WC), le confiere a los ARNs la posibilidad de formar estructuras terciarias muy diversas que están fuertemente asociadas a su función biológica. Se conoce que casi un tercio de las interacciones, provenientes del estudio de estructuras de ARNs resueltas por métodos experimentales, presentan interacciones diferentes a las del apareamiento WC[110].

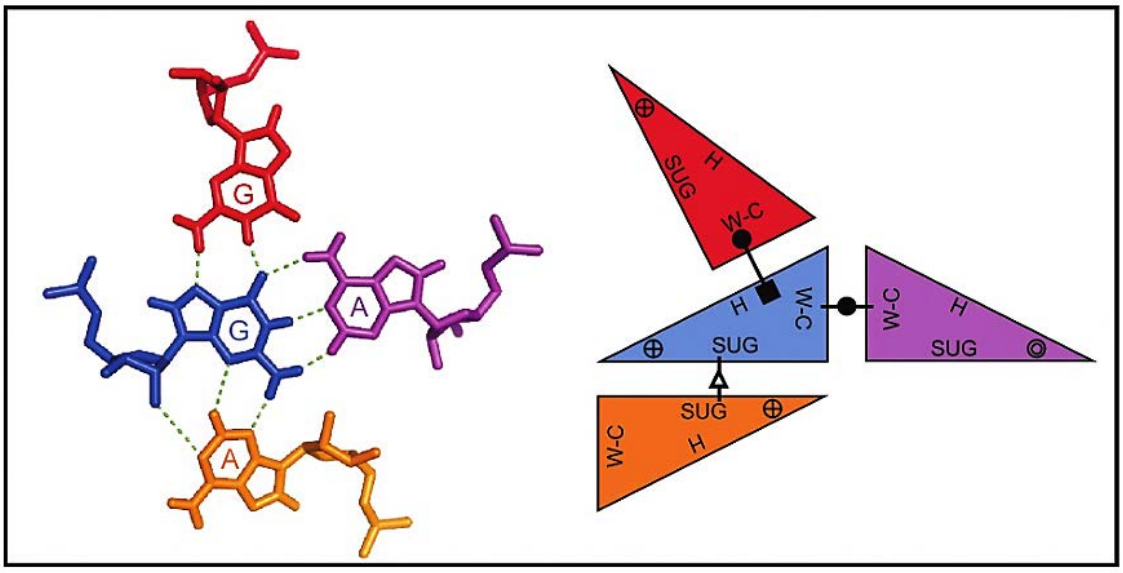



**Figura 3.3.3:** Representación gráfica de un cuarteto de bases interaccionando. En el lado derecho se representa mediante la abstracción de triángulos. Como se logra observar, cada base puede interaccionar hasta con otras tres bases utilizando los 3 bordes de la representación simplificada de Leontis/Westhof. La representación gráfica de bastones mostrada a la izquiera fue obtenida del PDB 1j5e que pertenece al ARN 16S de *T. thermophilis*. Las líneas verdes indican puentes de hidrógenos. Imagen extraída de [111].

Del estudio de estructuras tridimensionales se definen los motivos. Éstos, por definición, son puramente estructurales más que secuenciales[112–114], pues puede haber más de una secuencia con la misma estructura 3D. La definición de un motivo tiende a estar condicionada desde el punto de estudio a realizar (para más información ver el capítulo 15 del libro[115] "*RNA Structure and Folding: Biophysical Techniques and Prediction Methods*"). Un motivo es un tipo de estructura secundaria o terciaria, conformada por una región nucleotídica relativamente corta, modular, con una forma tridimensional y

estabilidad específica definida por al menos una interacción no canónica (diferente al tipo WC), acompañado de características dinámicas y que se encuentra de manera recurrente en las moléculas de ARNs. Una definición más estricta involucra una descripción tridimensional lo más completa posible, desde la conformación del *backbone*, las diferentes interacciones tipo puente de hidrógeno y apilamiento de bases, preferencia secuencial, presencia o no de co-factores tales como moléculas de agua, iones y otras moléculas. También debería incluir su rol funcional en la estructura terciaria y considerar las restricciones evolutivas[116]. Es así, por ejemplo, que en la naturaleza podemos encontrar diferentes motivos en moléculas de ARNs sin importar el organismo en cuestión. Los motivos pueden comprender la estructura específica de un loop o interacciones entre dos loops (loop-loop), entre un loop y una hélice (loop-hélice), entre hélices, etc. Generalmente estas estructuras terciarias se forman en presencia de iones metálicos y/o moléculas cargadas positivamente. Como iniciativa frente a los numerosos trabajos en búsquedas de motivos, hace más de una década que se ha propuesto su caracterización y clasificación por medio del Consorcio Ontológico del ARN (del inglés "*RNA Ontology Consortium (ROC)*", pág. web: *http://bioportal.bioontology.org/ontologies/RNAO/* o *https://github.com/BGSU-RNA/rnao*)[117,118]. En consecuencia y como veremos en la siguiente sección, existen bases de datos de motivos de ARNs[119].

Es importante mencionar que en muchos motivos se pueden observar estructuras más pequeñas que están muy conservadas y carecen de una secuencia característica. A estas estructuras se las llaman "elementos estructurales o atributos estructurales". En la **figura 3.3.4** pueden verse los diferentes elementos que se encuentran en el motivo "loop de sarcina-ricina". Por otro lado, en la **Tabla 3.2** se resumen las principales características entre elementos estructurales y los motivos.

| Características | Elemento | Motivo de loop | Motivo de interacción terciaria |
|---|---|---|---|
| **Tamaño** | Pequeño, local | Puede abarcar todo el loop | Múltiples loops o hélices involucradas |
| **Conservación de la secuencia** | Poco o nada | Frecuentemente con preferencia de secuencia/isosterico | Sitios de interacción |
| **Conservación de la estructura** | Por definición | Frecuentemente conservado | Evolutivamente conservado |
| **Rasgos particulares (ej.: apareamiento, stacking, turn)** | Usualmente de único rasgo | Multiples rasgos/elementos | Múltiple en cada motivo de interacción |
| **Ocurrencia** | Encontrado dentro de varios motivos | No anidado; suele ocurrir en motivos con interacción terciaria | Suele incluir multiples elementos y motivos |

**Tabla 3.2:** Características de elementos y motivos estructurales de ARN. Tabla adaptada de [116].

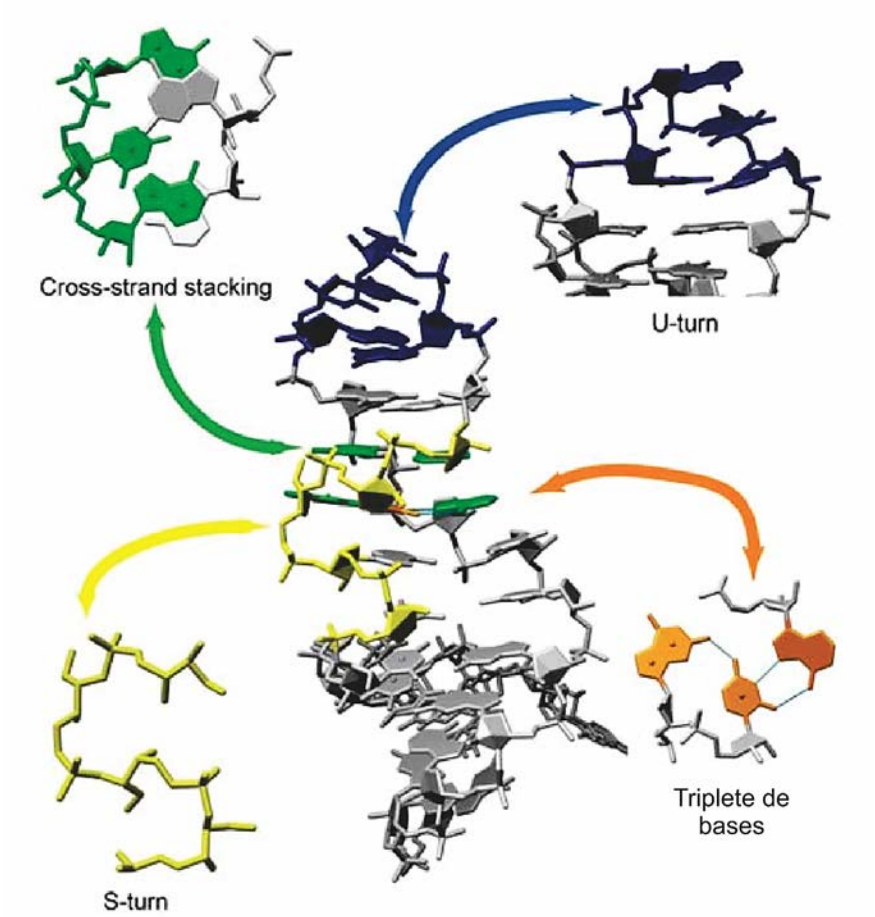



**Figura 3.3.4:** Representación por elementos del loop sarcina-ricina del ARN ribosomal 28S de Rata (código PDB 483D). El loop de sarcina-ricina es un ensamblaje de dos motivos secundarios: un loop interno (bulge G) y un hairpin loop (GNRA tetraloop). También puede considerarse estar compuesto por ciertos elementos, siendo cross-strand stacking (en verde y gris), U-turn (en azul) en el tetraloop GNRA, un triplete de bases (en naranja) y un S-turn (en amarillo). No se muestran interacciones no-canónicas entre bases ni tampoco puentes de hidrógeno entre el *backbone* y bases. Se observa un triplete de bases involucrado también en el cross-strand stacking. Imagen extraída de [116].

Al igual que las proteínas, los ARNs largos se organizan de forma jerárquica. La estructura primaria, constituida por la secuencia de nucleótidos, adopta una estructura en el espacio, la cual va cambiando en su proceso de plegamiento hasta formar estructuras secundarias de mayor estabilidad, donde luego regiones cortas de éstas interaccionan entre sí formado estructuras terciarias más estables, los motivos. Estos motivos, a su vez, presentan una complejidad menor a los dominios estructurales, los cuales son estructuras terciarias mucho más largas y complejas. Los dominios estructurales están unidos covalentemente mediante doble hélices o cadenas simples. Los dominios pueden presentar interacciones internas o entre ellos de forma no-covalente por medio de motivos. Por eso, podemos decir que los motivos tienen al menos 3 roles: (1) participan en la arquitectura de los dominios; (2) median las interacciones terciarias intra- e inter- dominios; y (3) desempeñan un rol funcional en la unión con otras moléculas, como proteínas, ARNs y también moléculas pequeñas.

El estudio de los motivos también involucra la observación y descripción de diversas conformaciones que éstos presentan, por ejemplo, al interaccionar con otra (bio)molécula. En la **figura 3.3.5** se puede observar, a nivel atómico, dos confórmeros del motivo receptor tetraloop, uno sin interaccionar y otro interaccionando con otro motivo tetraloop.

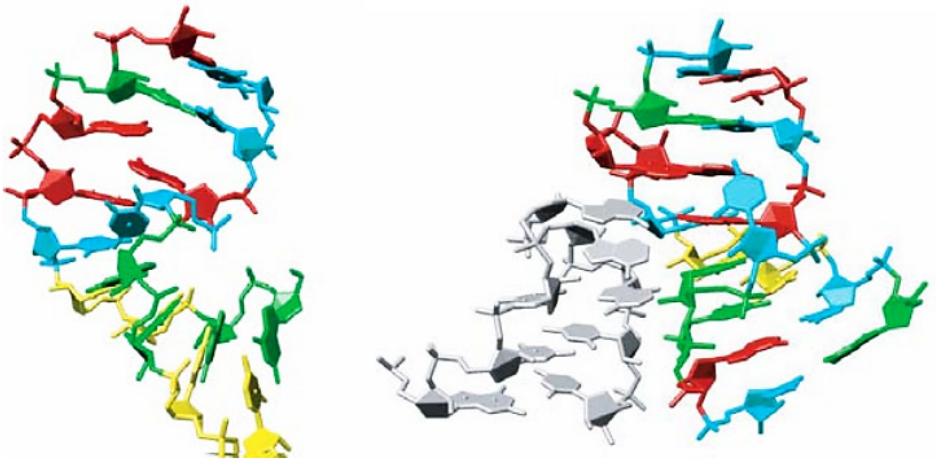


**Figura 3.3.5:** Cambio conformacional en el receptor de tetraloop unido al tetraloop GNRA. En la imagen se puede apreciar las estructuras cristalinas del motivo del receptor de tetraloop en su forma libre (PDB 1TLR) y formando una interacción terciaria con el tetraloop GNRA en el dominio P4-P6 del intrón del grupo I (PDB 1GID). La secuencia del receptor de tetraloop es variable, por lo que, en cada caso mostrado, la secuencia es diferente, siendo la de la derecha "GUCCUAAGU/AUAUGGAU" y la de la izquierda "GGCCUAAGA/UUAUGGCC". Las bases del receptor de tetraloop fueron coloreadas de la siguiente manera: roja (A), cyan (U), verde (G) y amarillo (C); mientras que el tetraloop GNRA es de color gris. Imagen extraída de [116].

## 3.4 Bases de datos estructurales

El correcto uso de la información es crucial para cualquier estudio que se quiera llevar adelante. Como sabemos, la información biológica se encuentra contenida, expresada y transmitida a nivel molecular. La forma que ésta tiene en el espacio es aún más importante a nivel biológico por la relevancia de la relación estructura-función. Por ejemplo, cuando realizamos estudios moleculares, partimos desde una estructura primaria (secuencia) y luego avanzamos hacia los demás niveles estructurales. Esto nos acerca en buscar una interpretación hacia la función desempeñada; en fin, buscar desenmascarar su sentido evolutivo. Debemos tener en cuenta que la información biológica se nos presenta de varias formas. Actualmente poseemos muchas maneras de adquirir la información, aunque las decisiones adoptadas sobre su interpretación siguen siendo muy importante y el principal desafío actual. Podemos decir, entonces, que la generación de espacios (bases de datos; "BDs") para contener toda esa información es un requisito central y necesario. Lo que queda es trabajar en identificar, organizar y validar la información. Esto último es un requisito indispensable para poder interpretarla fructíferamente. Por eso, esta sección se centrará en describir las principales bases de datos que dedican su trabajo al estudio de la información estructural del ARN.

Antes de continuar, es relevante indicar que, generalmente, las BDs suelen clasificarse por el tipo de información que contienen y la manera en que ésta es curada (se entiende por curada al proceso y modo de evaluación, validación y corrección de la información). El forma en que se obtiene la información es de suma importancia ya que, por ejemplo, puede ser información primaria derivada de descubrimientos experimentales verificados, o recolectada por procedimientos automáticos de escaneos computacionales (prestaremos especial atención a esto en el siguiente capítulo). A su vez, las BDs pueden ser muy específicas o generales. Sabiendo esto, la elección y el uso que se les dé dependerá de la pregunta a contestar por el usuario: si es una pregunta específica o más bien, general. Por último, una de las ventajas a presentar por las BDs es la interconexión con otras BDs u otros enlaces de interés referidos a dicha información. El diseño es otra ventaja que es importante

mencionar. Esto último, debe asegurar la facilidad en el uso de la misma (del inglés, ser *user-friendly*).

Existen diversas BDs que contienen estructuras primarias de ARNs. Comúnmente, se las llama BDs de secuencias de ARN[120]. Actualmente, por ejemplo, en la página *http://www.oxfordjournals.org/nar/database/c/*, desplegando la categoría correspondiente, podemos encontrar una lista de 99 bases de datos de secuencias de ARN. En los siguientes capítulos, se hará uso de la base RNAcentral (*https://rnacentral.org*)[121] la cual alberga aproximadamente 14 millones de secuencias provenientes de casi 30 bases de datos ARNs-específicas curadas por expertos y que se corresponden a secuencias de ARNs que no codifican para proteínas, denominados ARNs no codificantes (en adelante "ncRNAs"). A pesar de presentar una controversia en el nombre de la clase al cual éstos tipos de ARNs pertenecen, debido a que contienen información biológica codificada, se los ha denominado por la negativa a la clásica visión codificante del ARN mensajero (mRNA), ribosomal (rRNA) y de transferencia (tRNA) que, mediante el código genético universal, se decodifica la información biológica dando así con la síntesis de péptidos y proteínas[122,123]. Esto último denota una mirada proteinocéntrica. Más allá de su semántica, prestaremos principal interés en la información brindada por esta BD ya que la utilizaremos para nombrar la gran mayoría de los tipos de moléculas de ARNs a trabajar más adelante. Además de presentar una gran diversidad de organismos y tipos de ncRNAs, RNAcentral asigna un identificador único para cada secuencia de ARN. Estos identificadores son estables y no cambian debido a que son exclusivos para cada secuencia en cuestión. El formato del identificador es "**URS + 10-dígitos hexadecimales**", por ejemplo, **URS000040D674**. A su vez, incluye otro nivel de asignación, el de especificidad de especie. Esto lo realiza agregando una barra inclinada hacia la derecha o "guion bajo" al identificador URS, seguido del identificador taxonómico del NCBI (*"National Center for Biotechnology Information"*, web: *https://www.ncbi.nlm.nih.gov/*), por ej.: **URS00003B7674_9606** o **URS00003B7674/9606**. Es importante tener en cuenta que RNAcentral excluye la asignación de identificadores URS a secuencias menores a 10 nucleótidos, como así también, a aquellas con más de 10% de sus nucleótidos desconocidos.

Como ya se mencionó en las secciones anteriores, la biología estructural del ARN tiene sus comienzos a mediados de la década del 50' luego del modelo de doble hélice de ADN de Watson y Crick, pero no fue hasta principios de la década del 70' que comienza a recopilarse información tridimensional producto de la obtención de la estructura del $tRNA^{Phe}$ resuelta por DRX a nivel atómico. Paralelamente, se estaba consolidando la presencia de la mayor de las BDs a nivel tridimensional, la PDB ("*Protein Data Bank*" - *https://www.rcsb.org/*)[124]. Ésta se fundó en 1971[125] pero, a nuestro interés, no fue hasta 1991 que se estableció la base de datos de ácidos nucleicos (NDB, "*Nucleic Acid Database*" - *http://ndbserver.rutgers.edu/*)[126,127] con la tarea de controlar todas las estructuras tridimensionales de ácidos nucleicos depositadas en la PDB. Recién para 1996, la NDB asume total responsabilidad, siendo el principal sitio donde se depositan las estructuras 3D de ácidos nucleicos para luego ser compartidas a otras bases de datos, entre ellas, la misma PDB. Hoy, la NDB contiene cerca de 10000 estructuras de ácidos nucleicos y es el principal sitio en línea que provee, de forma detallada y con validez, la más completa información tridimensional sobre estas biomoléculas.

En estos casi 30 años desde el surgimiento de la NDB se han desarrollado numerosas BDs específicas que contienen información 3D de ácidos nucleicos, en particular ARNs. De forma semejante a lo nombrado anteriormente para secuencias, podemos encontrar una descripción accediendo a *http://www.oxfordjournals.org/nar/database/c/*. Desplegando la categoría correspondiente a bases de datos estructurales y luego en la subcategoría de estructura de ácidos nucleicos, se puede observar una lista de 22 bases de datos, mayormente de ARNs. También existen otras BDs que proveen información 3D del ARN o derivada de ella. En la **Tabla 3.3** se describen algunas de estas BDs. Es importante mencionar una BD, hoy en día obsoleta pero de gran importancia a nivel estructural, conocida como SCOR ("Structural Classification of RNA" - *http://scor.berkeley.edu/*)[128,129]; Esta base de datos organizaba manualmente los motivos desde una clasificación jerárquica, considerando la relación entre la estructura, interacciones terciaria y su función. Dicho de otra forma, **SCOR** fue similar a lo que es **SCOP**[130] (hoy **SCOP2,** *"Structural Classification of Proteins 2"* - *http://scop2.mrc-lmb.cam.ac.uk/*)[131] para proteínas, donde se detallan

las relaciones estructurales y evolutivas entre las estructuras de proteínas. También lo fue **MeRNA** (*http://merna.lbl.gov*)[132]**,** una BD curada manualmente que analizaba sitios de unión a metales de cada estructura de RNA, clasificando así diferentes motivos de unión a metales.

En la sección previa describimos la importancia de los motivos y nombramos la existencia de BDs. Una de las más importantes es *RNA 3D Motifs Atlas* (*http://rna.bgsu.edu/rna3dhub/motifs*)[119,133], ya que proporciona una clasificación detallada de grupos de motivos (loops internos y loops de horquillas) provenientes de un set no-redundante de estructuras 3D de ARNs. La extracción de los motivos se realiza mediante el uso de un software (**FR3D**[134]) comparando geométricamente todas las estructuras contra sí mismas en búsqueda de patrones de similitud geométrica. Actualmente, *RNA 3D Motifs Atlas*, cuenta con 295 entradas para motivos del tipo *hairpin-loop* y 297 para *internal-loop*. La importancia de esta BD radica en la identificación de motivos estructurales como evidencias evolutivas, logrando a su vez, probar la diversidad de secuencias asociadas a cada motivo. Esta información es de gran importancia en los estudios que relacionan la divergencia secuencial *vs* estructural. Como se sabe, el *dataset* de estructuras 3D, desde en donde se obtienen los motivos, es crucial y determinante para comprender cómo están representadas las estructuras biológicas. En el capítulo siguiente veremos porqué nuestro *dataset* de trabajo debe ser redundante.

**Tabla 3.3:** En la siguiente tabla se listan diferentes bases de datos que contienen información estructural a nivel secundario y/o terciario. Se muestra un detalle y el link correspondiente de cada BD. En la columna **Actualización** se puede observar la última fecha en que fue actualizada la BD. En caso de actualizarse periódicamente, se indica la periodicidad. *ND*: no declarado.

| Nombre | Descripción | Actualización | Link - Referencia |
|---|---|---|---|
| InterRNA | Proporciona una descripción detallada de las diferentes interacciones entre bases en los motivos de ARNs. Usa *datasets* de DRX curados. | Mensualmente | http://mfrlab.org/interrna/ - [135] |
| MetalPDB | Provee información detallada sobre los metales y sus sitios de unión presentes en las diferentes estructuras de las macromoléculas contenidas en la PDB (132972 estructuras en la última actualización). Permite hacer búsqueda por similitud (BLAST), códigos PDB o EC, nombre de molécula, tipo de metal, etc. | 2018-06-04 | http://metalweb.cerm.unifi.it/ - [136,137] |
| MINAS | Contiene información geométrica de las dos primeras capas de coordinación de metales unidos a estructuras de ácidos nucleicos presentes en PDB y/o NDB. Además contiene información de la secuencia nucleotídica próxima a la unión del ligando. | 2013-03-08 | http://www.minas.uzh.ch/ - [138] |
| PseudoBase++ | Contiene información sobre estructuras secundarias de pseudoknots obtenidas por métodos experimentales determinados por cristalografía, resonancia magnética nuclear (del inglés *nuclear magnetic resonance, "NMR"*), mutacionales, y de comparación de secuencias. | *ND* | http://pseudobaseplusplus.utep.edu/ - [139] |
| RNA Bricks 2 | Es una base de datos de motivos 3D y sus contactos consigo mismos y/o con proteínas. Provee información estructural mediante anotación de scores, y una serie de herramientas de búsqueda y comparación de estructuras 3D de ARN. Usa notación de dot-bracket para estructuras 2D. | semanalmente | http://iimcb.genesilico.pl/rnabricks2 - [140] |
| RNA CoSSMos | Posee una caracterización detallada de motivos en estructuras secundarias de ARN (loops, bulge, harpin). Permite realizar búsquedas en las diferentes estructuras 3D catalogadas en PDB. | *ND* | http://cossmos.slu.edu/ - [141] |
| RNA FRABASE 2.0 | Busca en la PDB, de forma automática, fragmentos 3D de ARN usando como input secuencias y/o estructuras 2D. Usa notación dot-bracket. Actualmente posee 2753 estructuras con al menos una molécula de ARN provenientes de la PDB. Proporciona variada información de valores de ángulos de torsión, coordenadas de átomos, parámetros conformacionales de la ribosa, tipos y clasificación de pares de bases. | 2016-03-07 | http://rnafrabase.ibch.poznan.pl/ - [142,143] |
| RNA SSTRAND | Contiene información de estructuras secundarias de ARN provenientes de diversas bases de datos de diferentes organismos. Permite analizar, buscar y subir información de estructuras secundarias de ARNs. | 2008 | http://www.rnasoft.ca/sstrand/ - [144] |
| RNAJunction | Proporciona y contiene información estructural de nivel secundario (*helical junctions, internal loops, bulges* e interacciones *loop-loop*). Utiliza software propio para extraer la información desde las estructuras depositadas en la PDB. | *ND* | https://rnajunction.ncifcrf.gov/ - [145] |
| Voronoia4RNA | Esta base de datos almacena y proporciona información de densidades de empaquetamiento atómico (del inglés *atomic packing densities*) de estructuras que | 2013-09-09 | http://proteinformatics.charite.de/voronoia4rna/ - [146] |

| | | | |
|---|---|---|---|
| | contengan moléculas de ARNs. Para estimar las fuerzas subyacentes y las interacciones de van der Waals tiene en cuenta al solvente y los espacios libres entre átomos vecinos. Actualmente incluye 1766 estructuras de PDB con al menos una molécula de ARN y una resolución menor o igual a 3.5 Angstrom. También tiene en cuenta estructuras resueltas por NMR. Para el cálculo de valores de referencia utiliza la base de datos de estructuras 3D de ARN no-redundante del grupo de Bioinformática estructural del ARN, BGSU. | | |
| RNArchitecture | Reciente base de dato de clasificación estructural de moléculas de ncRNAs. Provee una descripción detallada de los niveles de clasificación, enfocándose en las familias estructurales mediante el estudio de similitudes en sus secuencias y estructuras. RNArchitecture también proporciona información bibliográfica y links a otras bases de datos (Rfam, PDB, etc). La relación secuencial se define de acuerdo a alineamientos múltiple. Por otro lado, la relación estructural se establece en base a la superposición de estructuras. | 2017 | http://iimcb.genesilico.pl/RNArchitecture/ - [147] |
| URS database (URSDB) | La base de datos universo de estructuras de ARNs (URSDB, del inglés *Universe of RNA Structures Database*) relaciona y proporciona, de forma detallada, información de todas las estructuras de la PDB que contengan moléculas de ARNs. Se actualiza cada mes y básicamente muestra la siguiente información en 49 tablas divididas en 4 grupos:<br>1) Información procesada de los archivos de PDB;<br>2) Anotación de puentes de Hidrógeno, stems, etc. (utiliza el software DSSR);<br>3) Otra información creada por DSSR;<br>4) Anotación de loops, regiones de cadena única, pseudoknots, regiones cerradas elementales y múltiples. | semanalmente | http://server3.lpm.org.ru/urs/ - [148] |
| The RNase P Database | Base de datos disponible desde 1994. Contiene todas las estructuras de las RNasas P depositadas en la PDB. Proporciona información secuencial de las moléculas de ARNs y proteínas involucradas en cada complejo. También provee alineamientos secuenciales, estructuras secundarias, modelos 3Ds, links para información taxonómica e información secuencial. Organiza su información filogenéticamente pero particularmente mediante información estructural secundaria. Cesó sus actualizaciones en 2004-2005. | 2005 | http://www.mbio.ncsu.edu/RNaseP/ - [149] |
| HD-RNAS | HD-RNAS proporciona una clasificación funcional de las estructuras de ARNs disponibles en la PDB. Utiliza todas aquellas estructuras resueltas por DRX con una resolución mejor a 3.5 Angstrom y que contengan al menos una cadena de ARN mayor a 9 nucleótidos. Cada clase, a su vez, es clasificada taxonómicamente y también respecto al organismo en que fue cristalizada. Actualmente contiene 1060 estructuras cristalizadas con 3066 cadenas de ARNs, organizadas en 9 clases. Se lograron anotar, con su organismo correspondiente, 62 moléculas sintéticas de ARN mediante el uso de software BLAST. Esta base generó un *dataset* no-redundante de estructuras de ARNs consistiendo en 376 entradas PDB, indicando la mejor representación considerando la resolución y el factor R. | 2012-10-17 | http://www.saha.ac.in/biop/www/HD-RNAS.html - [150] |

Otra de las bases de datos más importantes o con mayor impacto en el campo de bases estructurales de ARN, es **Rfam** (*http://rfam.xfam.org/*)[151,152]. Semejante a **Pfam** para proteínas (*http://pfam.xfam.org/*)[153,154], **Rfam** tiene como objetivo estudiar, clasificar y organizar moléculas de ARNs en familias (anotación por homología). La anotación actual de Rfam (versión 14.0, agosto 2018) se basa en la colección de referencia de 14434 genomas completos, curados, suministrados por **proteomas de Uniprot** (del inglés *"Uniprot proteomes". http://www.uniprot.org/proteomes*). Hoy en día, **Rfam** contiene 2791 familias representadas en los 3 dominios principales más los virus. Su clasificación se fundamenta en similitud secuencial y estructural a nivel secundario, donde cada familia comparte un ancestro en común. Utiliza, principalmente, la estructura primaria para realizar alineamientos múltiples (MSA) y la estructura secundaria para complementar dicha información en el MSA (aplica modelos de covarianza. Software *INFERNAL*). A su vez, basándose en la conservación de estructura secundaria, **Rfam** clasifica a las familias en tres grandes grupos: genes ncRNAs, elementos estructurados cis-regulatorios y ARNs que desarrollan auto-empalme (de inglés "*self-splicing RNAs*"). Para cada familia, **Rfam** provee el alineamiento semilla, las secuencias, estructuras secundarias, especies provenientes, árboles filogenéticos, estructuras 3D (si hay disponible en **PDB**), referencias de otras BDs y archivos de curación (contienen información del modelo de covarianza, autor del alineamiento, referencias sobre las secuencias, el número de secuencias, etc).

Debido a la importancia que presenta para nuestro trabajo, ampliaremos brevemente la descripción de la información brindada por **RNArchitecture**[147]. Como se menciona en la **Tabla 3.3**, esta base de datos tiene como objetivo la clasificación estructural de las moléculas 3D de ncRNAs. La información que usa y organiza es proveniente de las BDs Rfam y PDB. Como es de suponer, sus niveles de clasificación conllevan un entendimiento evolutivo. El nivel central es el de **Familia**, el cual describe aquellas moléculas de ARNs relacionadas evolutivamente y que poseen una estructura conservada con similitud secuencial detectable. Las **Familias** que presentan variaciones estructurales se subdividen en **Subfamilias**. Por otro lado, las **Familias** que presentan divergencia secuencial, pero tienen estructuras y funciones similares, están relacionadas evolutivamente y se las clasifica

como **Superfamilias**. A su vez, a las superfamilias de estructuras de ARNs que comparten relaciones geométricas en su núcleo estructural, pero se desconoce si pueden ser derivadas de procesos evolutivos divergentes o convergentes, se las agrupa en el nivel **Arquitectura**. Finalmente, el nivel más alto de clasificación es el de **Clase**. Este último nivel organiza la información de manera amplia en estructuras complejas, estructuras simples y dominios de lncRNAs de estructura desconocida. En la **figura 3.4.1** se observa un gráfico en forma de "rayo de sol" con la clasificación de **RNArchitecture**. Es importante mencionar que **RNArchitecture** contiene 2688 **Familias** donde solo 74 de ellas (2,54%) posee un modelo estructural resuelto experimentalmente (DRX, RMN, EM). Esto quiere decir que casi la totalidad de las **Familias** están definidas por modelos 3D teóricos. La presencia de modelos teóricos, predichos computacionalmente, presentan desventajas y sesgos al estudiar e inferir propiedades biológicas en ARNs. Este inconveniente irá disminuyendo a medida que se vayan depositando un mayor número de moléculas diferentes de ARNs resueltas por métodos experimentales.

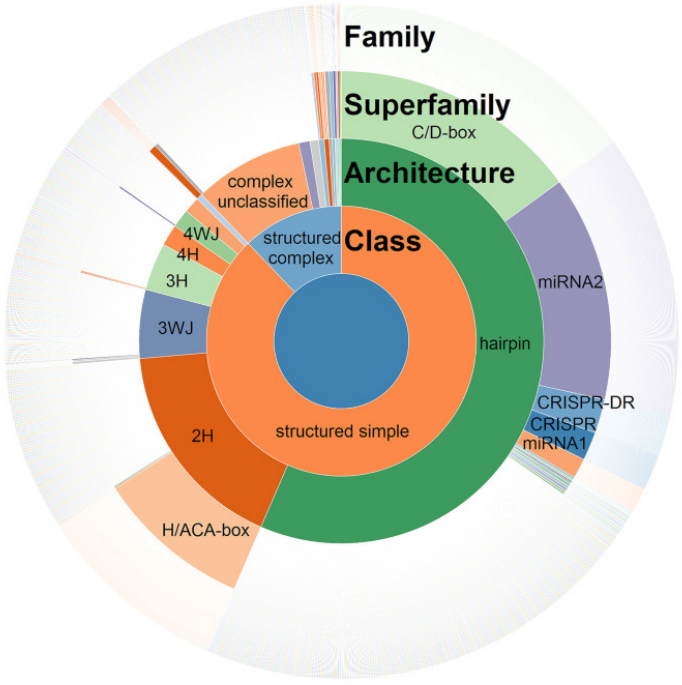



**Figura 3.4.1:** En este gráfico se ilustra la jerarquía de RNArchitecture y el contenido referido a la liberación 1.0. La capa más externa indica las 2688 Familias de ARNs. Las capas sucesivas combinan estas Familias en 1722 Superfamilias, 22 Arquitecturas y finalmente en dos Clases. Se muestran los nombres de cada clase y de las arquitecturas y superfamilias más grandes. Imagen extraída de [147].

# Desarrollo preliminar de una base de datos de Diversidad Conformacional en ARNs

## 3.5 ¿Por qué hacer una base de Diversidad Conformacional en ARNs?

Antes de comenzar con las principales razones que justifiquen la realización de una base de datos de diversidad conformacional en ARNs, no es redundante mencionar que, para cualquier biomolécula, es de suma importancia tener un vasto conocimiento de su estructura debido a que está fuertemente relacionada con su función biológica. Ya hemos mencionado en la introducción que, la comprensión del concepto de *estado nativo* es fundamental para poder realizar un estudio lo más atinado posible. Actualmente, la definición más precisa y funcional de *estado nativo* es aquella que considera a un *ensamble nativo* (conjunto de confórmeros de menor energía libre de Gibbs que coexisten en equilibrio termodinámico y sus distribuciones son estadísticamente significativas). Teniendo presente esto último, entendemos a la *diversidad conformacional* como el conjunto de diferencias estructurales entre los confórmeros del ensamble. Cualquier estudio que requiera informacional tridimensional debe tener presente los confórmeros de la biomolécula en cuestión. Hoy en día, no existe BD que represente la diversidad

conformacional en ARNs. Por eso, una de las principales razones que impulsa la realización de una base de datos de DC en ARNs, es la necesidad de contar con esta información contenida, organizada y clasificada para estimar la DC de las moléculas de ARNs. Otras razones igualmente importantes serían: poder brindar la información biológica, fisicoquímica y experimental necesaria a aquellos estudios que consideren la presencia de múltiples conformaciones; animar a la comunidad a obtener más y mejores estructuras tridimensionales de forma experimental en post de mejorar sus estudios; encontrar propiedades generales en las estructuras de los ARNs; comprender la función biológica asociada; ofrecer información relevante con la unión a ligandos; promover un cambio de enfoque en el uso de estructuras únicas para el campo de la predicción estructural, como así también en el estudio evolutivo; etc.

Es de suma importancia mencionar que la realización de este trabajo es una extensión del estudio realizado para proteínas por el Grupo de Bioinformática Estructural de la Universidad Nacional de Quilmes, SBG-UNQ (del inglés *Structural Bioinformatics Group*), donde se han desarrollado múltiples trabajos científicos sobre diversidad conformacional incluyendo la base de datos de diversidad conformacional en proteínas, *CoDNaS* (del inglés *Conformational Diversity of Native State in proteins;* pág. web: *http://ufq.unq.edu.ar/codnas/*)[155,156].

## 3.6 Estructuras redundantes y estimación de la diversidad conformacional

En la sección 3.3 del capítulo anterior vimos que el uso de *dataset* de trabajo para el estudio de la información tridimensional e implementación en bases de datos es crucial. Todos los trabajos mostrados de BD usan *datasets* no redundantes de estructuras para representar a la molécula de ARN en cuestión. Esto produce un sesgo en la información biológica ya que el ensamble nativo está representado por las distribuciones de las poblaciones de confórmeros. En la introducción detallamos la importancia de la dinámica estructural y el uso de paisajes energéticos para explicar las funciones biológicas cuando las

condiciones del entorno cambian o se evidencian uniones con otras moléculas. Existen numerosos métodos que estudian la dinámica de ARNs y hacen estimaciones de la diversidad conformacional. Algunos de ellos son experimentales (RMN, FRET) y otros por simulación computacional (dinámica molecular de grano grueso, ENM). Los métodos computacionales por dinámica molecular tienen la desventaja de explorar sólo movimientos relativamente rápidos, hasta el orden de los 10µs. También, si se considera hacer uso de todos los átomos, las simulaciones están limitadas a moléculas pequeñas para reducir las exigencias computacionales. Además, en los métodos de grano grueso, existe una pérdida de información biológica producto de aplicar una simplificación de la estructura. Otra de las formas que existe para estudiar la diversidad conformacional es a partir de diferentes estructuras experimentales de la misma biomolécula obtenidas en distintas condiciones. Cada estructura representa un confórmero estable del ensamble para la condición en la que fue obtenido.

Nuestro objetivo radica en evidenciar y estimar la diversidad conformacional para cada molécula de ARN, por lo que debemos trabajar con la mayor colección de información biológica estructural disponible, esto son *datasets* redundantes, donde se cuenta con la representación de distintas estructuras para la misma molécula (confórmeros) obtenidas bajo distintas condiciones. Para tal fin, tomaremos en consideración todas las estructuras experimentales de ARNs contenidas en la PDB, la cuales provienen mayoritariamente de estructuras resueltas por métodos de DRX y, en menor medida, por RMN y microscopía electrónica (EM, del inglés *Electron Microscopy*). Por ejemplo, la cristalización de una molécula de ARN en diferentes condiciones (presencia o no de ligando, cambio de pH, etc) puede dar una descripción de aquellos confórmeros estables para cada una de las condiciones en las que fue obtenido cada cristal. La modificación de las condiciones experimentales favorece el desplazamiento de las distribuciones de confórmeros que coexisten termodinámicamente en equilibrio. Es importante resaltar que cada método experimental posee desventajas intrínsecas, por ejemplo, para DRX, las coordenadas atómicas provienen de un mapa de densidad electrónica en fase sólida, en cambio para RMN, las coordenadas provienen de las distancias interatómicas obtenidas en fase líquida

y posteriormente ajustadas a un modelo. Las estructuras experimentales también se basan en modelos científicos, ya que surgen de la interpretación de la información empírica[157]. Una opción alternativa es la aplicación conjunta de datos experimentales junto a simulaciones computacionales[158,159] (ver **figura 3.6.1**), por ejemplo, recientemente se realizó la determinación del ensamble de un tetraloop de ARN mediante el uso de datos provenientes de RMN junto a estudios de dinámica molecular (MD, del inglés *Molecular Dynamic*)[160]. Sin embargo, el costo computacional de las simulaciones suele ser muy alto, lo que impide realizar estudios de gran escala para analizar la dinámica de cientos de moléculas globalmente. Además, las conclusiones obtenidas a partir de simulaciones computacionales deberían ser sustentadas por observaciones experimentales.

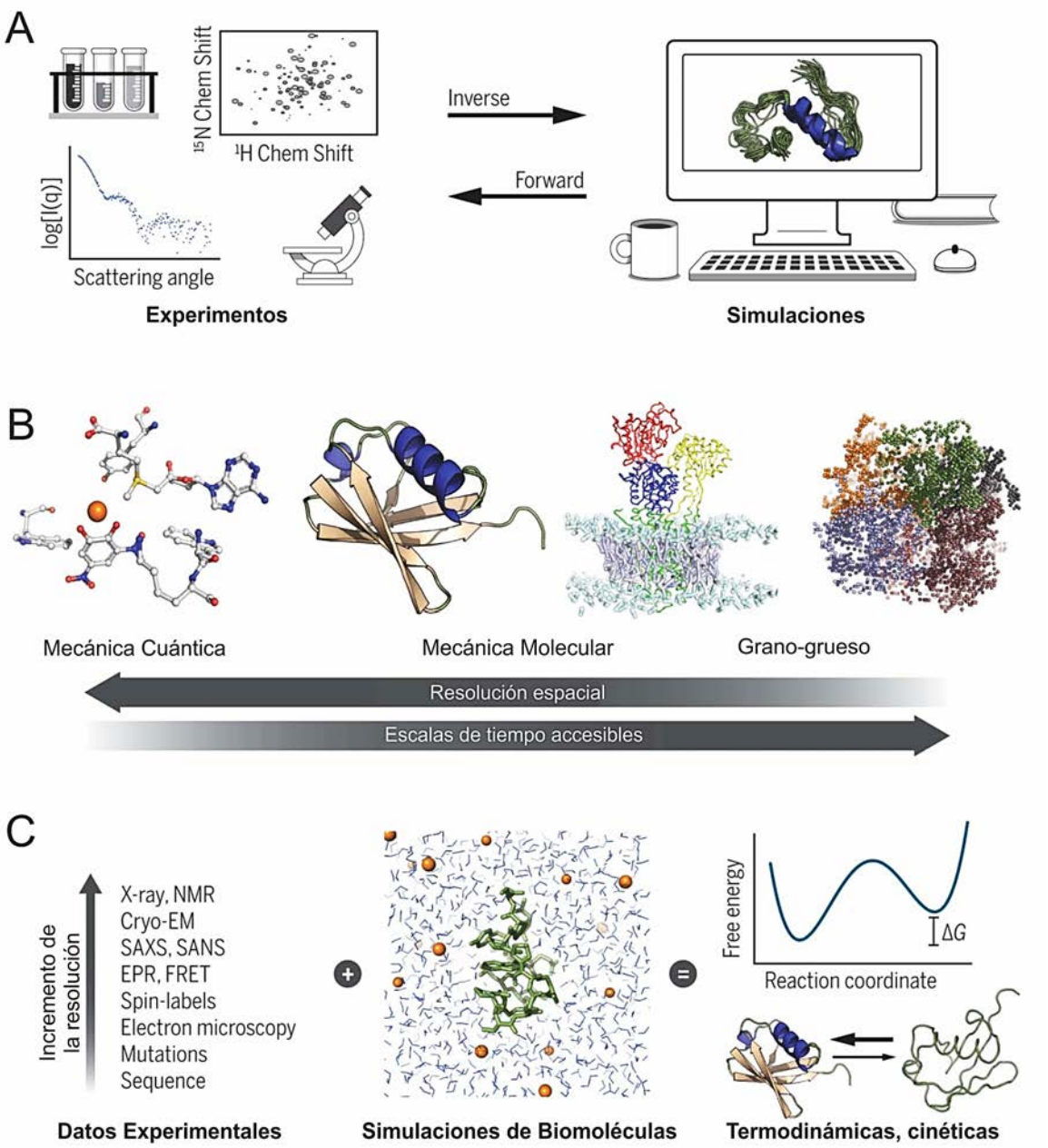



**Figura 3.6.1:** Las simulaciones y los experimentos son complementarios. A – Resolver un problema de forma inversa reside en describir factores causales que producen un set de observaciones. Las simulaciones moleculares, en cambio, pueden usarse para construir un conjunto de conformaciones moleculares microscópicas que pueden compararse con observaciones experimentales mediante el uso de un modelo avanzado. B - Los enfoques computacionales usados para estudiar biomoléculas abarcan desde detallados modelos mecánicos cuánticos a mecánicos moleculares de resolución atómica hasta modelos de grano grueso, donde se agrupan varios átomos. La disminución de la complejidad computacional otorgada por el grano-grueso progresivo hace posible el acceso a escalas de tiempo más largas y escalas de mayor longitud. C - Los datos experimentales se pueden combinar con modelos físicos para proporcionar una descripción termodinámica y cinética de un sistema. A medida que mejora la calidad del modelo, es posible describir fenómenos más complejos con menos datos experimentales. SANS, dispersión de neutrones de ángulo pequeño (del inglés *small-angle neutron scattering*); EPR, resonancia paramagnética electrónica (del inglés *electron paramagnetic resonance*); FRET, transferencia de energía de resonancia de fluorescencia (del inglés *fluorescence resonance energy transfer*); $\Delta G$, energía libre de Gibbs. Imagen adaptada de [158].

## 3.7 Cuantificación de la diversidad conformacional

La cuantificación de la diversidad conformacional de una biomolécula puede ser estimada comparando estructuralmente sus confórmeros. La comparación estructural es un procedimiento que puede realizarse superponiendo pares de confórmeros hasta lograr alinear los centros de masas de cada estructura, logrando así establecer un mismo sistema de ejes de coordenadas. Este tipo de alineación se la conoce como superposición rígida. La alineación estructural puede conseguirse rápidamente alineando, en primero instancia, la secuencia de dicha molécula (es igual para cada confórmero), seguido de una translación estructural hasta compartir el mismo centro de masa y, finalmente, realizar una rotación de alguna de las estructuras a comparar. Las rotaciones tienen como principal objetivo minimizar la distancia cuadrática media (RMSD, del inglés *Root Mean Square Deviation*) entre los átomos equivalentes de las estructuras. El valor del RMSD se corresponde a la medida euclidiana promedio entre átomos equivalentes y es ampliamente utilizado en la biología estructural ya que es fácilmente calculable cuando se cuenta con información de coordenadas espaciales. A continuación, se indica su expresión matemática:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d_i^2}$$

Donde $d_i$ es la distancia euclidiana entre el i-ésimo par de átomos, uno de cada estructura, y $N$ es el número total de pares de residuos a considerar. El RMSD se expresa en unidades de longitud, habitualmente se utiliza el $\mathring{A}ngtrom$ ($1\mathring{A} = 1^{-10}m$). En la **figura 3.7.1** se puede observar una interpretación gráfica del método de superposición rígida.

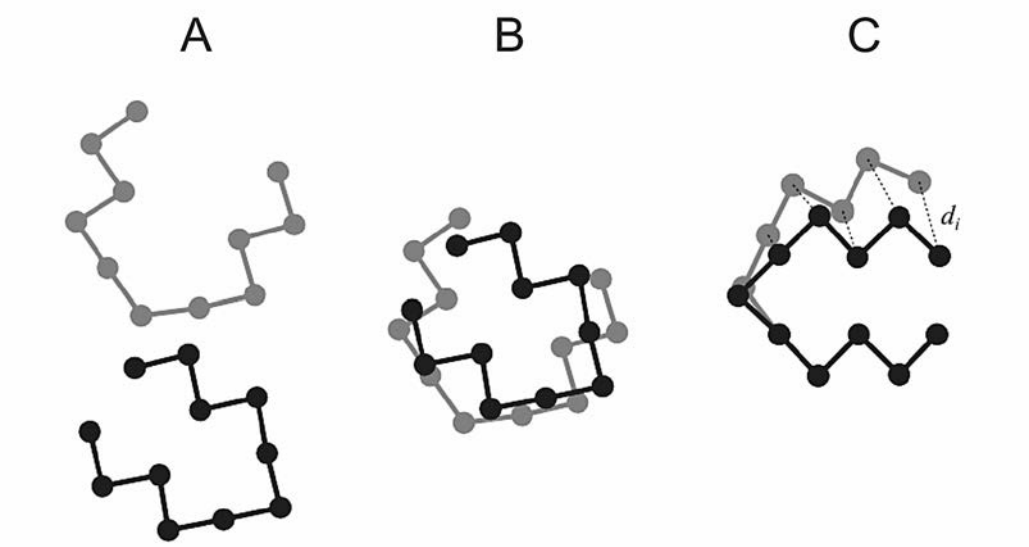



**Figura 3.7.1:** A – Representación simplificada de dos confórmeros. B – Los dos confórmeros de (A) se trasladan al mismo centro de masa y luego se rota la estructura de uno de ellos hasta minimizar completamente el valor del RMSD de los C3' de los residuos equivalentes. C – Las estructuras de los dos confórmeros se encuentran alineadas.

Al igual que en el campo de las proteínas, la superposición estructural toma como referencia las coordenadas de un átomo específico por cada residuo, siendo el carbono alfa ($C_\alpha$) frecuentemente elegido para los aminoácidos de las proteínas y el C3' de la ribosa para los ARNs (también se suele usar el átomo de fósforo – P). Generalmente, esta elección está sujeta a un átomo contenido en el *backbone* previamente definido. Este trabajo utilizará el C3' de la ribosa para cada medición que se vaya a efectuar, por lo que cuando nos remitiremos a algún valor de RMSD omitiremos su aclaración.

Por último, debemos hacer notar que cuando se quiere analizar el RMSD de dos estructuras de DRX idénticas y en igual condiciones de cristalización, éste debería tomar un valor cercano a 0 $\text{Å}$ o idealmente 0 $\text{Å}$. En proteínas, la desviación a dicho valor está asociado al error cristalográfico y se encuentra entre 0 – 0,5 $\text{Å}$[161]. Como es de suponer, a medida que el valor de RMSD aumenta, las estructuras difieren entre sí. Esta medición es un indicador confiable en la cuantificación de la variabilidad estructural entre dos confórmeros de una misma biomolécula. Esto está fuertemente asociado a los movimientos que experimenta la biomolécula en su ensamble nativo. Actualmente, no existe tal información asociada al error cristalográfico en ARNs, pero hipotetizamos que también podría encontrarse cercano a de las proteínas contenidas en la PDB ($0-0{,}5$ $\text{Å}$).

## 3.8 Construcción preliminar de la base de datos

El desarrollo preliminar de una base de datos de diversidad conformacional en ARNs fue motivado por el trabajo previo realizado en proteínas por el grupo de bioinformática estructural de la Universidad Nacional de Quilmes, el cual ha culminado en diferentes bases de datos de diversidad conformacional para dominios de proteínas [162] y proteínas[155,156]. También, tiene como motivación poder presentar a la comunidad científica la importancia de los equilibrios conformacionales en moléculas de ARNs. Como ya se mencionó, estos equilibrios están directamente relacionados con su función biológica.

Ante la necesidad de contar con esta base de datos global, abarcando todas las moléculas de ARNs conocidas que presentan diversidad conformacional, hemos decidido recopilar todas las estructuras redundantes de un ARN que estén depositadas en la PDB con el objetivo de evidenciar la diversidad conformacional en ARNs sometidos a diferentes condiciones. Estas condiciones pueden ser producto de factores fisicoquímicos (pH, temperatura, fuerza iónica, etc), factores biológicos (ligandos, mutaciones, modificaciones post-transcripcionales, etc) o mezcla de ellos. En la **figura 3.8.1** se esquematizan los pasos realizados para el armado de la base de datos de DC en ARNs.

Cada molécula de ARN que se encuentra contenida en esta base de datos está representada por un conjunto de estructuras las cuales han sido comparadas mediante un software de alineamiento estructural. Esta comparación nos permite identificar similitudes o diferencias entre cada uno de sus confórmeros. Como mencionamos en los objetivos específicos, la idea principal es identificar los diversos factores que afectan a la diversidad conformacional del ensamble nativo de cada ARN. La puesta en funcionamiento de esta base de datos permitirá reclutar ARNs con diversos grados de diversidad conformacional y estudiar sus factores asociados. También realizar diferentes estudios que abarquen aspectos dinámicos, estructurales, funcionales y evolutivos. A su vez, se buscará tener una amplia interconexión con links hacia otras bases de datos, referencias bibliográficas, publicar los identificadores correspondientes (secuenciales, estructurales, taxonómicos,

GO, etc.) y por último, desarrollar un diseño de interfaz gráfica amigable para el usuario con actualizaciones periódicas.

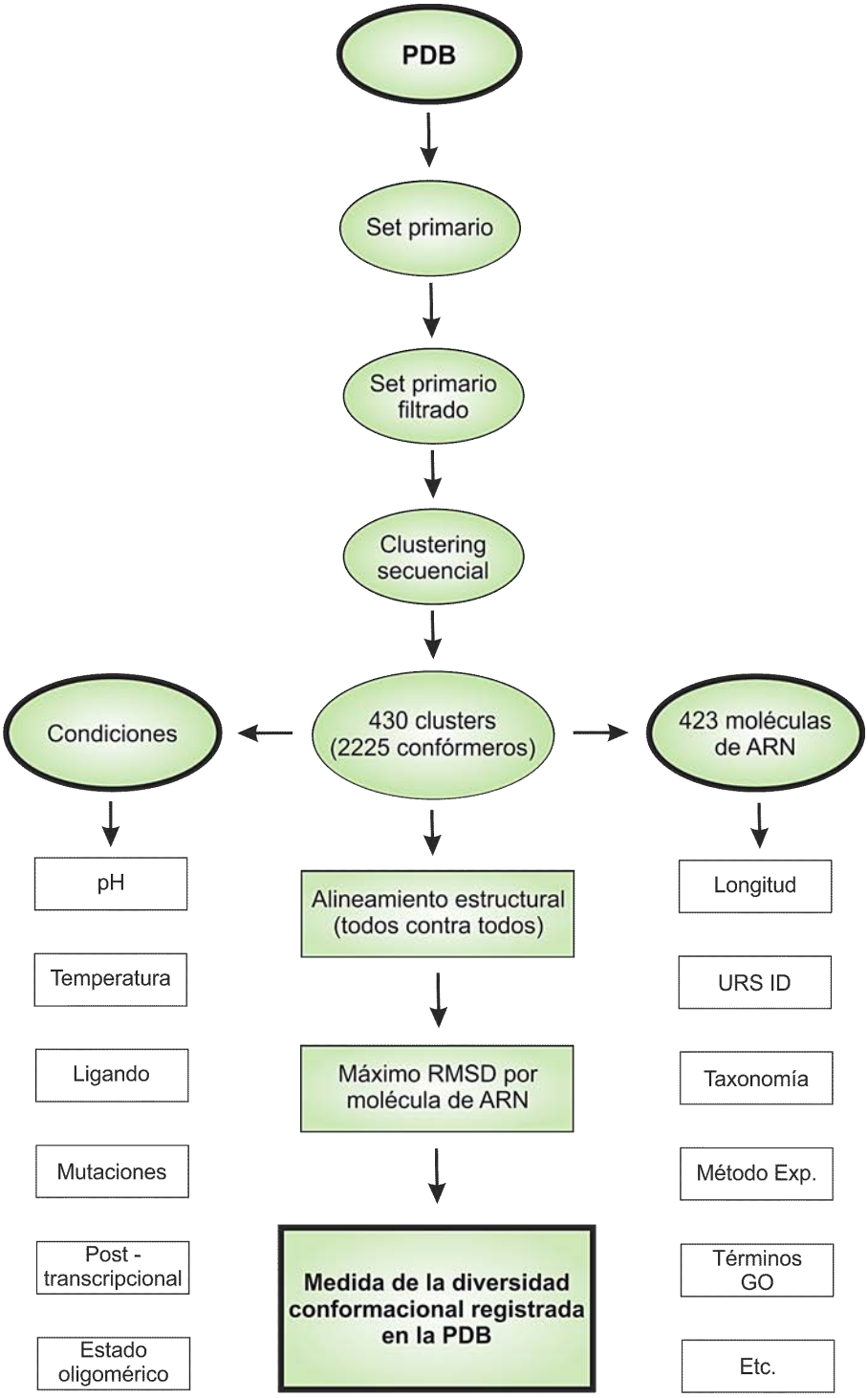



**Figura 3.8.1:** Diagrama de flujo de los pasos seguidos para la construcción de la base de datos de diversidad conformacional en ARNs.

### 3.8.1 Reclutamiento de la información estructural y posterior filtrado

Para el reclutamiento de las estructuras de ARNs se utilizó la base de datos PDB. Esta BD contiene el mayor repositorio tridimensional de biomoléculas a nivel mundial, principalmente proteínas, seguido de los ácidos nucleicos, ARN y ADN, respectivamente. Como se mencionó anteriormente en la sección 3.4, la PDB es la principal BD receptora de los depósitos de la NDB, donde se decidió no hacer uso directo de ésta última debido a la facilidad de manejo que presenta la PDB. Los archivos contenidos en la PDB tienen formato *.pdb* y/o *mmCIF* (del inglés "*macro-molecular Crystallographic Information File"*). En este trabajo, por conocimiento previo de manipulación y compatibilidad con el software de alineamiento estructural, se utilizaron todos archivos con formato *.pdb*.

Las moléculas de ARN pueden formar parte de estructuras más complejas que simplemente estar en estado monomérico, es decir, pueden formar oligómeros (homo y/o heteros), estar interaccionando con otras biomoléculas como por ejemplo proteínas, ADNs, otros ARNs, o combinación de ellas formando estructuras cuaternarias y/o quintenarias. Esto lo podemos observar en aquellos complejos macromoleculares resueltos por DRX que se encuentran depositadas en la PDB y presentan -múltiples- co-cristalizaciones de biomoléculas. Allí, el archivo *.pdb* contiene información tridimensional de cada una de las cadenas de cada tipo de biomolécula. Dentro de un archivo *.pdb* es posible encontrar estructuras de proteínas, ARNs, ADNs, híbridos ARN-ADN, diferentes ligandos como metales, moléculas orgánicas pequeñas, etc.

Para la recolección se utilizó el sistema de búsqueda avanzada de la PDB, donde se obtuvieron todos los archivos *.pdb* de aquellas estructuras depositadas que contenían al menos una molécula de ARN y también se generó un archivo único en formato *.csv* conteniendo, previa selección personalizada, información de campos descriptivos de interés, creando así una gran tabla madre (61707 filas por 20 columnas) -nuestro *set primario*- para cada confórmero de cada molécula contenida en cada uno de los archivos *.pdb* (proteínas y ácidos nucleicos). Posteriormente, se filtró el set primario a través del uso de comandos específicos desde la interfaz de línea de comandos (CLI). Se descartaron aquellos confórmeros que no fueran ARN y que no tuvieran un largo de al menos 9

nucleótidos (en adelante “nt”) y un máximo de 500 nt, es decir, 9≤nt≤500. Además, para realizar correctamente la posterior comparación estructural, se descartaron aquellas cadenas pertenecientes a archivos *.pdb_bundle* ya que las estructuras correspondientes fueron separadas en múltiples archivos *.pdb*, lo que complicó su manejo en identificación y edición. También se descartaron aquellos confórmeros que tuvieran una resolución $> 4{,}5\text{Å}$.

El paso siguiente consistió en identificar los clusters de moléculas de ARNs idénticas, esto es, moléculas que comparten 95% de identidad secuencial, con un *coverage* mínimo del 90% entre todas las secuencias del cluster al que pertenece. Este procedimiento fue necesario para reclutar las distintas instancias de cristalización de moléculas que fueran idealmente idénticas (confórmeros).

Para realizar esta tarea de clustering secuencial, se utilizó el software CD-Hit[163,164] con servicio web: *http://weizhong-lab.ucsd.edu/cdhit-web-server/cgi-bin/index.cgi*. CD-Hit permite realizar comparaciones entre secuencias y generar clusters personalizados. Para ser más específicos, dentro del paquete de programas proporcionados por el CD-Hit, utilizamos el correspondiente a secuencias nucleotídicas, *cd-hit-est*. Como *input* se proporcionó un archivo *FASTA* conteniendo todas las secuencias de moléculas de ARNs que pasaron los filtros aplicados al set primario. No se eligió 100% de identidad secuencial con la intención de incluir aquellos confórmeros que presenten variaciones mínimas, producto, por ejemplo, de mutaciones puntuales en su secuencia. El output del CD-Hit consistió en 1611 clusters identificados con 95% identidad secuencial y 90% de coverage, donde sólo 779 poseían al menos 2 confórmeros. De estos 779, sólo consideramos aquellos que contenían al menos 2 confórmeros provenientes de archivos *.pdb* diferentes, dando un número final de clusters igual a 430. Para este último paso de búsqueda y comparación de confórmeros, se realizó un script en lenguaje python, con la finalidad reclutar clusters que contengan al menos 2 instancias experimentales independientes donde se depositó la misma secuencia en la PDB. Es importante resaltar que cada cluster es único en secuencia y proporciona información de diversidad conformacional curada, haciendo a éstos futuras entradas a la base de datos de diversidad conformacional en ARNs.

### 3.8.2 Estimación de las diferencias estructurales entre confórmeros

Anteriormente nombramos la importancia de estimar la diversidad conformacional en ARNs. Esta medida puede realizarse calculando el RMSD en todos los pares de confórmeros por cluster. La medida del RMSD es muy común en el ámbito de la biología estructural, por lo que podemos encontrar varios tipos de softwares que hacen uso de ella. Para nuestro interés, elegimos aquellos destinados a comparar, por superposición estructural, moléculas de ARNs de manera local y remota, culminando en la elección final del programa SARA [165,166]. En la **tabla 3.4** se pueden observar diferentes softwares que realizan comparaciones estructurales.

El algoritmo del SARA consiste en el alineamiento de 2 estructuras de ARNs mediante el enfoque del vector unidad (del inglés *unit-vector approach*), inspirado por el programa MAMMOTH[167] usado en el alineamiento estructural de proteínas. La exactitud de su alineamiento proviene de la evaluación de 3 valores diferentes: uno para la identidad secuencial (**PID**, del inglés *percentage of sequence identity*) que es el porcentaje de nucleótidos (igual tipo) alineados, respecto del largo ($N$) de la menor de las estructuras; otro para la estructura secundaria (**PSS**, del inglés *percentage of aligned secondary structure*) que es el porcentaje de pares de bases, alineadas y definidas por el software 3DNA, que se encuentran dentro de los $4.0\text{Å}$, respecto del menor número de pares de bases en las 2 estructuras; y el último para el valor de identidad de la estructura terciaria (**PSI**, del inglés *percentage structural identity*) que es el porcentaje de átomos C3' superpuestos dentro de $4.0\text{Å}$ respecto del largo ($N$) de la menor de las estructuras. Adicionalmente, SARA calcula los valores de sus *P-values* y estima la probabilidad de obtener valores de alineamientos por azar que sean mejores o iguales a los obtenidos. A continuación, se presentan sus expresiones matemáticas:

$$PID = 100\frac{(nt\ idénticos)}{N} \qquad PSS = 100\frac{p_{al}}{NP} \qquad PSI = 100\frac{n_{al}}{N}$$

Donde $p_{al}$ es el número de pares de bases alineadas dentro de un umbral de $4.0\text{Å}$. $NP$ es el menor número de pares de bases de las 2 estructuras alineadas. Por otro lado, $n_{al}$ es

el menor número de pares de nucleótidos alineados dentro de un umbral de $4.0\text{Å}$. Y $N$ es el largo de la más corta de las 2 estructuras de ARN a comparar.

Básicamente, SARA toma como *input* 2 archivos *.pdb* y compara las cadenas que se les asigna. La comparación puede tomar como valor de referencia el átomo C3' de la ribosa o el P del fosfato en la construcción de 3 vectores unidad por átomo. SARA usa por defecto el C3' ya que está demostrado que es la distancia con menor varianza [165]. Los valores de los vectores son normalizados a unidad de distancias y llevados a origen en una unidad esférica. La comparación finaliza con la rotación y posterior cálculo de la distancia URMS (del inglés "*Unit-vector Root Mean Square"*) entre las 2 estructuras de ARNs usadas como *input*. Dicha distancia es la medida de RMSD de las 2 estructuras 3D a comparar. Una ventaja presentada por el programa SARA es la opción de utilizar información estructural secundaria derivada mediante el software **3DNA**[168]. SARA puede calcular esferas unidad usando hasta 8 átomos sucesivos y no puede comparar estructuras con un largo menor a 9 nucleótidos. La comparación consecutiva de esferas unidad permite realizar una matriz de valores de similitud de todos contra todos que luego es usada para un procedimiento de programación dinámica (DP, del inglés *"dynamic programming")* usando penalizaciones con valor cero para las terminaciones con el fin de realizar un alineamiento global de las estructuras de ARNs a comparar. El alineamiento *output* es refinado maximizando el número de átomos equivalentes o pares de bases dentro de los $3.5\text{Å}$ de RMSD. Esto lo logra usando una variante del algoritmo MaxSub[169] el cual asegura de contener el mejor alineamiento local dentro del alineamiento resultante.

| Nombre | Ref. | Año | URL | Lenguaje | Output | Algoritmo |
|---|---|---|---|---|---|---|
| *Local* | | | | | | |
| LaJolla | [170] | 2009 | http://raphaelbauer.github.io/lajolla/ | Java | Lista de archivos *.pdbs*[1] | Coincidencia de n-grama mediante una query |
| FRIES | [171] | 2013 | http://www.zbh.uni-hamburg.de/?id=436 | C & Perl | 2 archivos *.pdb*[2] | Basado en WURST[172]. Usa fragmentos *k*-mer. |
| *Local & server* | | | | | | |
| ARTS | [173,174] | 2005 | http://bioinfo3d.cs.tau.ac.il/ARTS/index.html | Compilado | 2 archivos *.pdb*[2] | Tuplas estructuralmente similares de átomos de P |
| SARA | [165,166] | 2008 | http://structure.biofold.org/sara/?aform=ali | Python | 1 archivo *.pdb*[3] | Basado en MAMMOTH[167]. Representación estructural por vector unidad |
| R3D Align | [175] | 2010 | http://rna.bgsu.edu/r3dalign/ | MATLAB[4] | 1 archivo *.pdb*[3] | Búsqueda del máximo clique |
| FRASS | [176] | 2010 | https://sourceforge.net/projects/frass/?source=directory | C & Perl | 1 archivo *.pdb*[3] | Integrales de Gauss[177] |
| SETTER | [178,179] | 2012 | http://setter.projekty.ms.mff.cuni.cz/ | Compilado | Información de Rotación/Translación | Unidades de estructura secundaria generalizadas |
| MultiSETTER | [180,181] | 2015 | http://setter.projekty.ms.mff.cuni.cz/ | Compilado | Información de Rotación/Translación | SETTER + ClustalW[182] |
| RASS | [183,184] | 2014 | http://cloud.stat.fsu.edu/RASS/ | MATLAB | *ND* | Basado en ESA[7] |
| SARA-Coffee | [185,186] | 2013 | http://www.tcoffee.org/Projects/saracoffee/index.html<br>http://tcoffee.crg.cat/apps/tcoffee/do:saracoffee | Compilado | Alineamiento secuencial | SARA + R-Coffee[187] |
| STAR3D[5] | [188] | 2015 | http://genome.ucf.edu/STAR3D/ | Java | Alineamiento estructural; 1 archivo *.pdb*[3]; imagen del alineamiento | Detecta topología 2D con Mc-Annotate[189,190], genera *k*-stacks con información 3D y detecta *k*-stacks conservados mediante el cálculo del RMSD. Luego busca el máximo clique. El último paso consiste en el alineamiento de loops por DP con información 3D. |
| SupeRNAlign | [191] | 2017 | http://genesilico.pl/supernalign/ | Python | 1 archivo *.pdb*[3] | Superposición interactiva de fragmentos estructurales |
| SupeRNAlign-Coffee | [191] | 2017 | http://genesilico.pl/supernalign/ | Python | Alineamiento secuencial | SupeRNAlign + R-Coffee[187] |
| *Server* | | | | | | |
| DIAL | [192] | 2007 | http://bioinformatics.bc.edu/clotelab/DIAL/ | Python | Alineamiento secuencial y 2 archivos con información de ángulos | Basado en PRIMOS[193]. Utiliza DP con información de: similitud nucleotídica, ángulos de torsión, pseudo-ángulos y apareamiento de bases. |
| SARSA[6] | [194] | 2008 | http://genome.cs.nthu.edu.tw/SARSA/ | *ND* | 1 archivo *.pdb*[3] | Alfabeto estructural discreto basado en ángulos de torsión **α**, **γ**, **δ** y **ζ**. |
| iPARTS | [195] | 2010 | http://genome.cs.nthu.edu.tw/iPARTS/ | *ND* | 1 archivo *.pdb*[3] | Alfabeto estructural discreto de 23 letras basado en pseudo-ángulos de torsión **η** y **θ**. |
| Rclick | [196,197] | 2015 | http://mspc.bii.a-star.edu.sg/minhn/rclick.html | *ND* | 2 archivos *.pdb*[2] | Basado en CLICK[198]. Coincidencia de cliques |
| iPARTS2 | [199] | 2016 | http://genome.cs.nthu.edu.tw/iPARTS2/ | *ND* | Alineamiento secuencial y 1 archivo *.pdb*[3] | Alfabeto estructural discreto de 92 letras basado en pseudo-ángulos de torsión **η** y **θ**. |

**Tabla 3.4:** En la siguiente tabla se listan diferentes softwares que producen alineamientos de estructuras 3D de ARNs. Se agruparon según su modo de presentación para el usuario. [1] Cada archivo *.pdb* contiene una única estructura. Se producen múltiples modelos de superposición. [2] Cada archivo *.pdb* contiene una única estructura. [3] El archivo *.pdb* contiene ambas estructuras. superpuestas. [4] Puede ejecutarse con GNU Octane. [5] Posee servicio web desde 2016 - WebStar3D[200] - URL: http://rna.ucf.edu/WebSTAR3D/. [6] Ofrece servicio de alineamiento múltiple de estructuras terciarias (MARTS). [7] Análisis de forma elástica (ESA, del inglés *"Elastic Shape Analysis"*). *ND*: no declarado.

Para cada ARN de la base de datos preliminar se identificó el par de confórmeros que maximiza el RMSD y se lo denominó “par máximo”. Este par máximo es una medida de la extensión de la diversidad conformacional, pero en principio estaría influenciada por la cantidad de información tridimensional suministrada por la PDB. Más allá de esto, utilizaremos este par máximo como prueba fehaciente de la extensión de la diversidad conformacional con el fin de desarrollar los siguientes estudios que se detallan en el capítulo 4 de este trabajo.

Siguiendo el procedimiento detallado, la BD preliminar incluye 1347 entradas diferentes de PDB, esto sería un 33% del total de entradas actuales de la PDB que contienen al menos una cadena de ARN. Y si solo se tiene en cuenta las entradas de PDB que pasaron los filtros anteriores (2475), representa un 54%. En el momento del desarrollo de este trabajo, se incluyen 430 cadenas de ARNs, 2225 confórmeros y alrededor de 10 mil alineamientos estructurales de a pares. Actualmente, un 2,25% de la PDB almacena solo ácidos nucleicos (ARN, ADN e Híbridos), un 2,5% ARNs junto a Proteínas y aproximadamente un 2,8% contiene al menos una cadena de ARN. Esto concluye que prácticamente su totalidad es destinada a proteínas y presenta una gran limitante para realizar estudios generales desde la bioinformática estructural del ARN. A pesar de esto, en los últimos 20 años la disponibilidad de estructuras de ácidos nucleicos, en especial de ARNs, ha ido incrementando fuertemente.

# 4. Resultados y discusión

# Descripción de los datos de la base de datos preliminar de diversidad conformacional en ARNs

En este capítulo describiremos algunos de los datos presentes en la BD preliminar de DC de ARNs. A través de gráficas representaremos características del conjunto de moléculas de ARNs involucradas, por ejemplo: se mostrará, con histogramas, la distribución del largo de las moléculas, los valores máximos de RMSD, las resoluciones de las estructuras; usaremos el gráfico de torta para ver la distribución taxonómica; aplicaremos gráficos de cajas (del inglés *boxplots*) para evidenciar la diversidad conformacional de tipos de ARNs, etc. También es importante comentar que, para la total realización de este trabajo, toda la extracción, manipulación, cálculo y gráfica de la información fue realizada haciendo uso del lenguaje *Bash* en entorno CLI del SO (sistema operativo) *GNU/Linux*, y en mayor medida, con el lenguaje *Python 3.5*, dentro de los softwares *Jupyter-notebook* y *Jupyter-lab* los cuales se basan principalmente de *IPython*.

# 4.1 Cantidad de confórmeros

Anteriormente dijimos que la BD preliminar de DC en ARNs contiene 2225 confórmeros. La distribución de la cantidad de confórmeros por molécula de ARN no es equivalente por eso es importante resaltar cómo se encuentran distribuidos. A fin de no entrar en una descripción minuciosa, en la **figura 4.1.1** podemos observar los gráficos que describen la cantidad de moléculas de ARNs que tienen 2, 3, 4, 5, 6 y ≥7 confórmeros que han pasado los filtros mencionados en el capítulo 3.

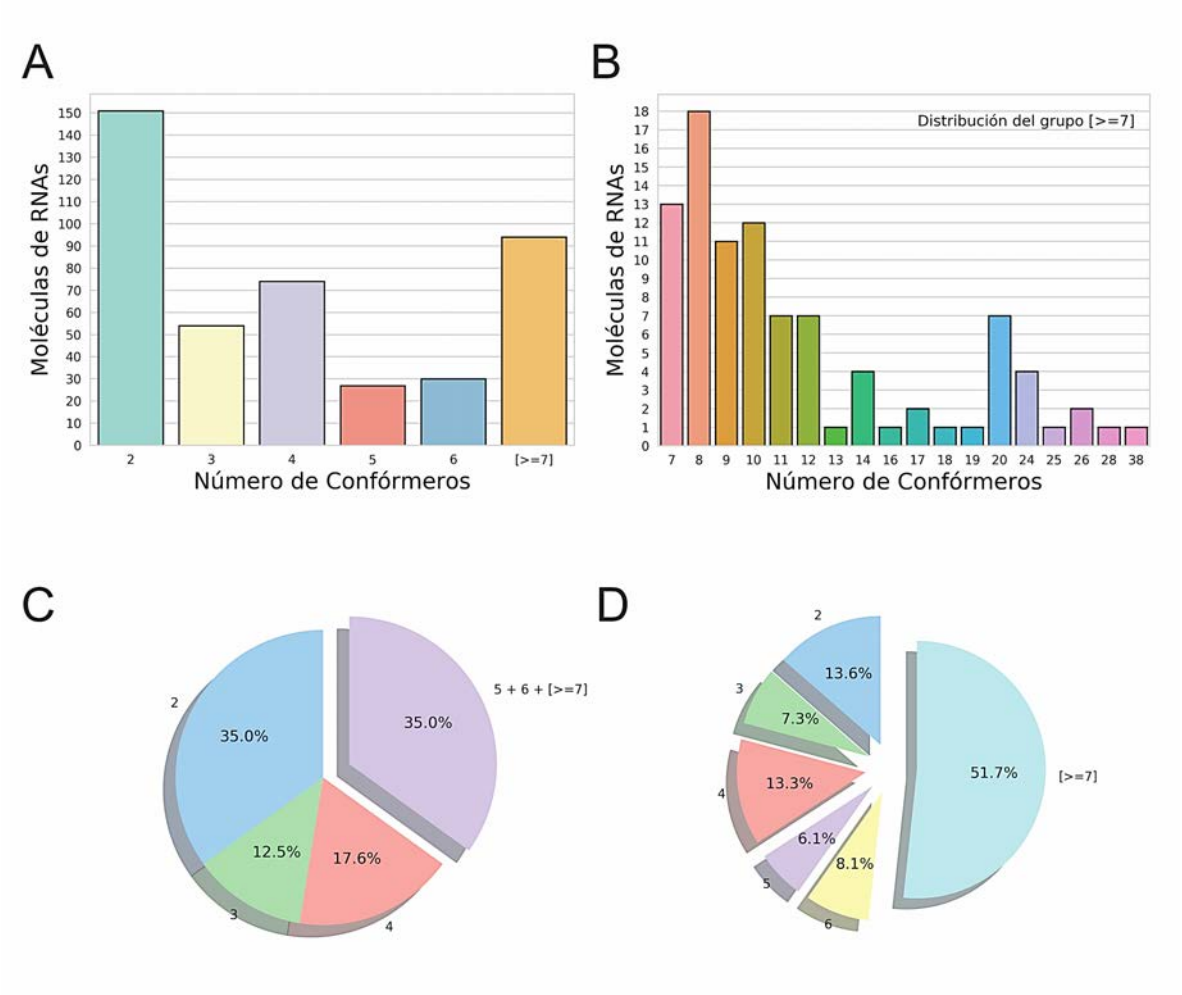



**Figura 4.1.1:** A – Distribución del número de confórmeros por cada molécula de la BD preliminar. B – Distribución del grupo [>=7] de (A). C – Gráfico de torta representando los respectivos porcentajes de la cantidad de confórmeros que contiene cada grupo respecto del total de moléculas contenidas en la BD preliminar. Como se puede observar, los grupos 5, 6 y [>=7] se sumaron como un único conjunto. D – Gráfico de torta representando los respectivos porcentajes de la cantidad de confórmeros que contiene cada grupo respecto del total de confórmeros contenidos en la BD preliminar.

La distribución de la cantidad de confórmeros por molécula de ARN tiene una mediana que es de 4 y un promedio de 6 confórmeros. Como puede verse en el gráfico de barras, la gran mayoría de las moléculas de ARNs tiene un número menor a 5 confórmeros. Expresado en términos porcentuales implica que un 65,1% de las moléculas de ARNs están representadas por menos de 5 confórmeros. Inclusive, por el gráfico de torta podemos observar que el 35% de las moléculas de ARNs contiene 2 confórmeros. Estos valores porcentuales tan elevados se deben principalmente a la limitada información tridimensional con la que actualmente cuenta la PDB comparada a la de las proteínas y a la baja redundancia en la deposición de estructuras 3D de una misma molécula de ARN resuelta por métodos experimentales. Los motivos vinculados a la baja deposición de estructuras no tienen relación biológica, sino más bien son motivos prácticos, es decir, por dificultad de cristalización o limitaciones intrínsecas al método experimental (longitud de las moléculas en RMN), interés comercial (industria), importancia médico-académica (temáticas novedosas en salud), entre otros. Es necesario resaltar que el interés biológico estructural en ARNs es un fenómeno que ha crecido recién en los últimos 20 años. Como mencionamos anteriormente, esto último puede observarse de manera gráfica en la **figura 3.1.3**.

Otra medida interesante que obtuvimos fue la relación de la cantidad de confórmeros por grupo respecto de los confórmeros totales en la BD preliminar de DC. En el gráfico de torta (**figura 4.1.1 – D**) se observa que el 51,7% de los confórmeros totales pertenecen al grupo de moléculas de ARN que tienen $\geq$7 confórmeros; el 13,6% al grupo de 2 confórmeros, etc. Ahora, en el conjunto de moléculas que tienen 5, 6 y $\geq$7 confórmeros, vemos que allí se encuentra el 65,9% de la totalidad de los confórmeros contenidos en la BD preliminar. Esto quiere decir que, a pesar de ser el 35% de la totalidad de las moléculas de ARNs, hacen uso de más de la mitad de la información tridimensional contenida en la BD preliminar. Por esto, podemos decir que del total de moléculas de ARNs contenidas en la BD preliminar, más de la mitad de ellas se ven poco representadas y más de un tercio están bien representadas.

Por último, realizamos un estudio de correlación entre la cantidad de confórmeros y la medida del RMSD del par máximo, donde hemos sometido a estas 2 variables a un estudio estadístico de correlación, precisamente el *test* de *Spearman*. En la **figura 4.1.2** se muestra el gráfico de dispersión e histogramas correspondientes a los valores en cada eje de coordenadas.

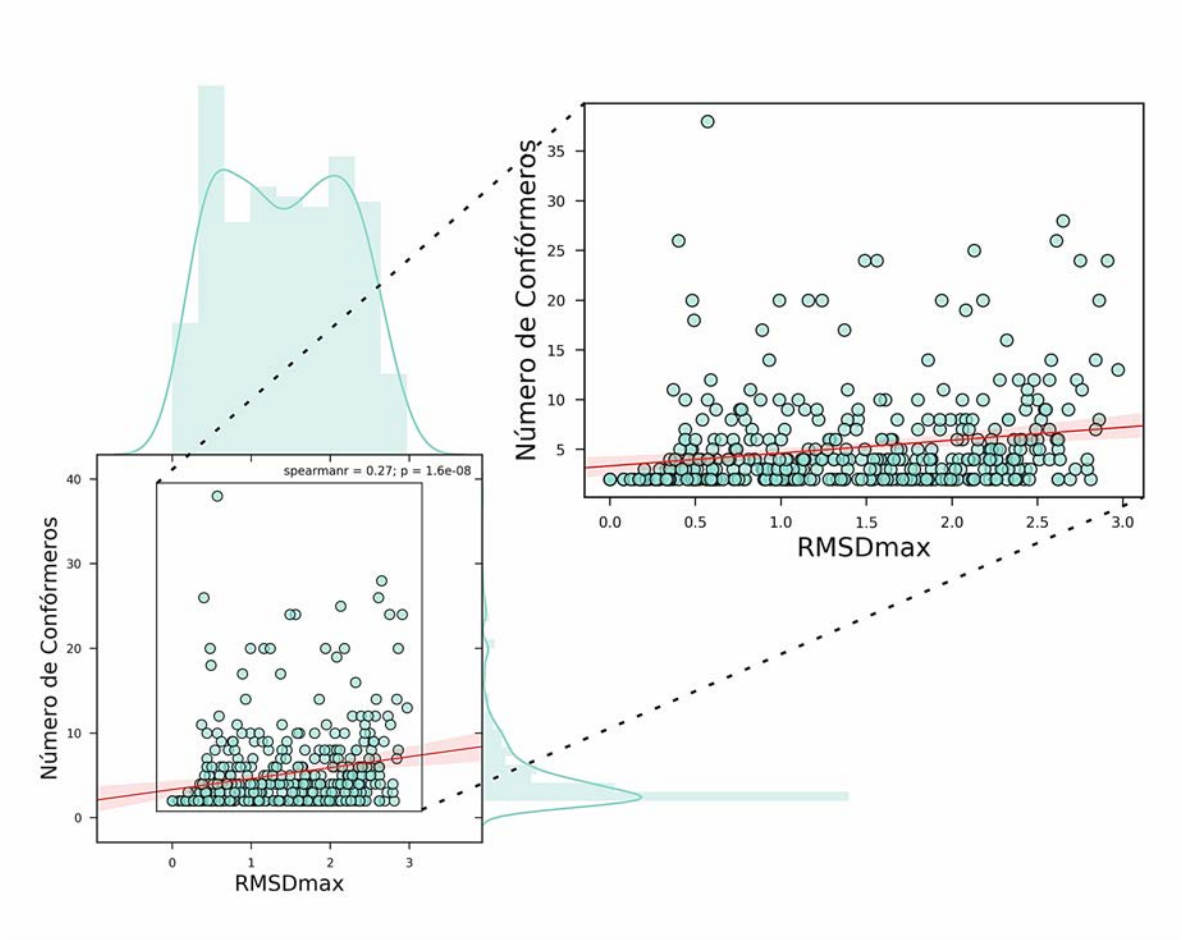



**Figura 4.1.2:** Dispersión de los valores de RMSD respecto del número de confórmeros. En cada eje se adjunta el histograma de frecuencia correspondiente junto a la curva de densidad. En la parte superior derecha se adjunta el valor de la correlación para el *test* de *Spearman*. Se amplía la zona de los puntos para mayor claridad.

El *test de Spearman* (no paramétrico) arrojó un valor de $rho = 0{,}27$ y un $p - valor = 1{,}60x10^{-08}$. Podemos decir entonces que la correlación es insignificante y asumiremos que la medida del RMSD del par máximo es robusta frente a la cantidad de confórmeros contenidos en cada cluster.

## 4.2 Longitud de los ARNs

La longitud de las moléculas de ARNs puede ser un aspecto muy relevante cuando estudiamos estructuras simples o más complejas. Actualmente, la longitud es tomada como un criterio para clasificar los ARNs que no codifican para péptidos o proteínas (ncRNAs, del inglés *non-coding RNAs*), siendo los **ARNs pequeños** aquellos con un largo $<200$ nucleótidos (sncRNAs, del inglés *small non-coding RNAs*) y **ARNs grandes** aquellos con un largo $\geq 200$ nucleótidos (lncRNAs, del inglés *long non-coding RNAs*).

Si observamos la distribución de longitudes que presenta la BD preliminar (**figura 4.2.1**) encontraremos muy poca variedad, primero porque existen puntos de corte que responden al criterio de trabajo para el reclutamiento de la información tridimensional ($9 \leq nt \leq 500$); segundo por la reducida cantidad de información con que cuenta la PDB; y tercero por motivos prácticos mencionados en el capítulo anterior. También podríamos pensar que efectivamente existe un interés biológico en moléculas tradicionalmente conocidas, por ejemplo: tRNAs, rRNAs, RNasas, ribozimas, etc. Los datos estadísticos de la distribución muestran que la mediana se encuentra en 29 nt y el promedio en 50 nt. Además, se puede observar un pico bien marcado cercano a los 25 nt mayormente perteneciente a tipos de ARNs misceláneos y un segundo pico menos marcado cerca a los 75 nt coincidente al largo de los tRNAs.

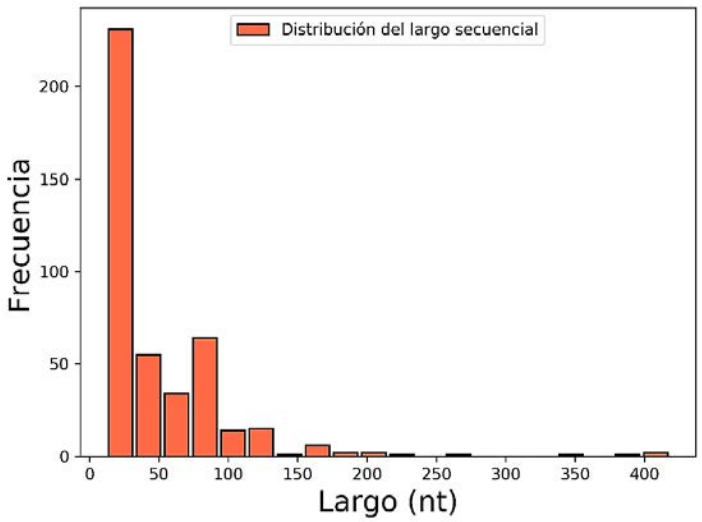



**Figura 4.2.1:** Distribución del largo secuencial para las moléculas contenidas en la BD preliminar.

También dejaremos demostrado que la longitud no es un factor que esté relacionado con la medida obtenida del RMSD del par máximo en la comparación entre par de confórmeros. Ésto puede realizarse mediante un análisis estadístico de correlación entre 2 variables. Para esto utilizamos el *test de Spearman,* obteniéndose el valor de $rho = 0{,}0633$ y un $p - valor = 0{,}1936$. Este valor de $rho$ entre el largo secuencial y el máximo RMSD, implica que no hay correlación entre las variables. En la **figura 4.2.2** se muestra una representación mixta donde aparecen puntos de dispersión e histogramas correspondientes a los valores en cada eje de coordenadas.

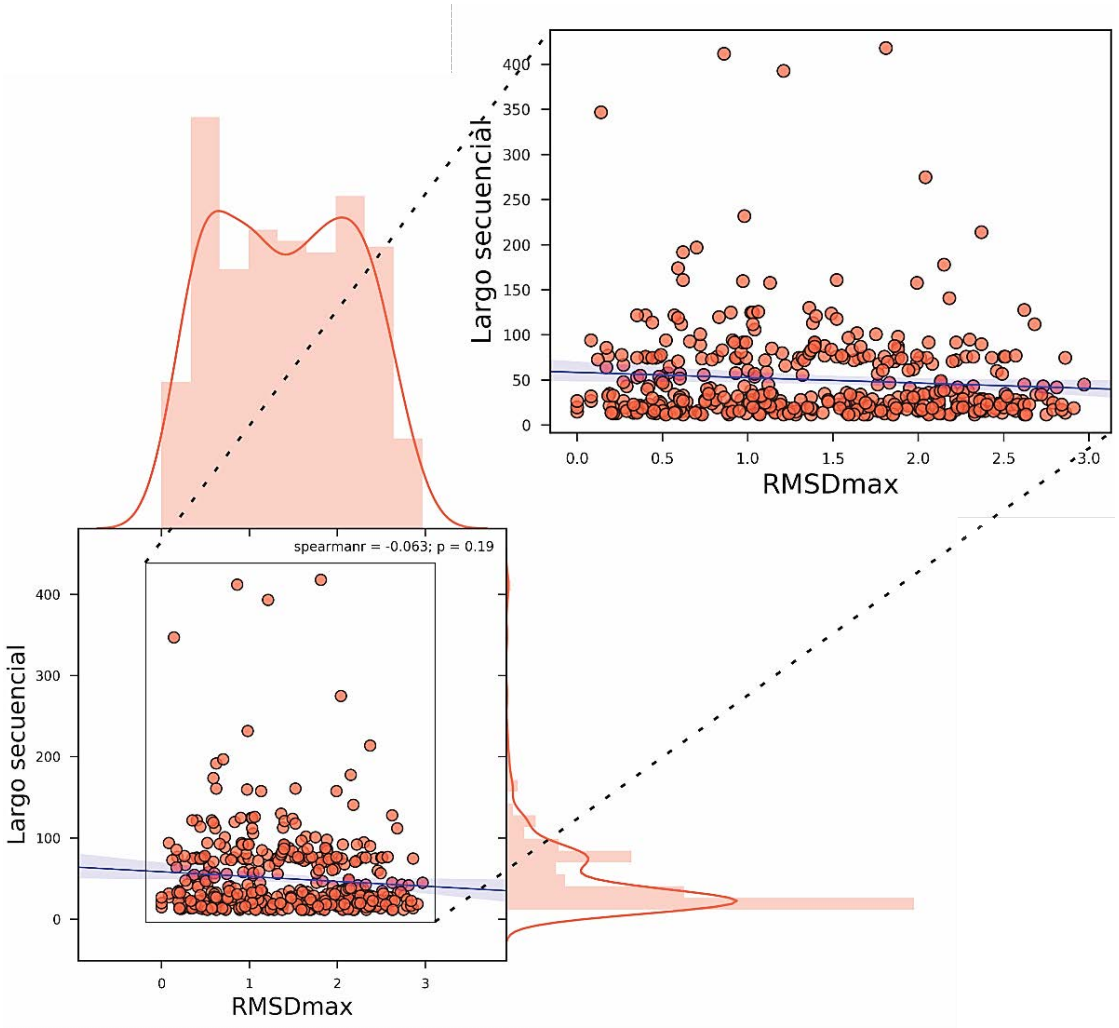



**Figura 4.2.2:** Dispersión de los valores de RMSD respecto largo secuencial. En cada eje se adjunta el histograma de frecuencia correspondiente junto a la curva de densidad. En la parte superior derecha se adjunta el valor de la correlación para el *test* de *Spearman*. Se amplía la zona de los puntos para mayor claridad.

## 4.3 Resolución de los confórmeros

La resolución de las estructuras cristalizadas es de suma importancia cuando se trabaja con información 3D proveniente de experimentos de DRX o EM. La resolución está relacionada con la capacidad de distinguir los detalles atómicos que una estructura presenta. Por esto, debemos tener en cuenta los valores que ésta toma a fin de satisfacer los requisitos mínimos necesarios de nuestro estudio. Los niveles de detalles atómicos varían en diferentes resoluciones, así, por ejemplo, para proteínas podemos encontrar la siguiente descripción (*http://proteopedia.org/wiki/index.php/Resolution*):

| Resolución | Calidad | Detalle |
|---|---|---|
| 1,2Å | Excelente | El *backbone* y las cadenas laterales se encuentran muy bien resueltas. |
| 2,5Å | Buena | El *backbone* y muchas de sus cadenas laterales están bien definidas. |
| 3,5Å | OK | El *backbone* y las cadenas laterales voluminosas están mayormente claras. |
| 5Å | Pobre | El *backbone* está mayormente claro, sin embargo las cadenas laterales no. |

Por otro lado, en los ácidos nucleicos, particularmente en los ARNs, podemos observar una clara diferencia entre las densidades de los mapas electrónicos correspondientes a resoluciones típicamente depositadas en la PDB. En la **figura 4.3.1 – A** vemos un ejemplo de cómo esto se ve representado para1.04, 1.75, 2.25, 2.75, 3.3, 3.8, 4.5 y 6.21Å.

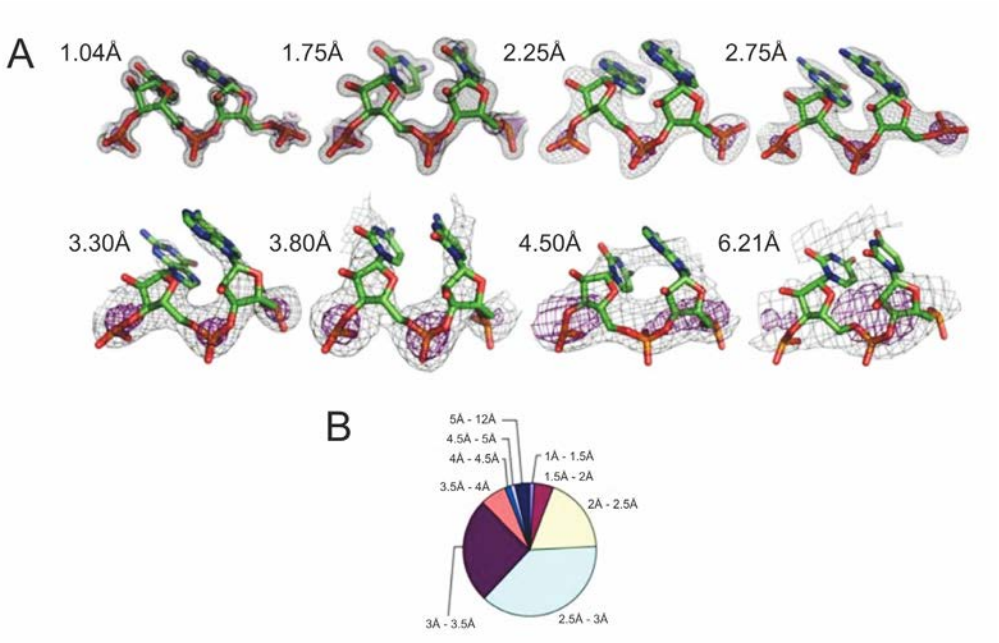


**Figura 4.3.1:** A – Representación típica de los mapas de densidades electrónicas para diferentes resoluciones de estructuras de ARNs. B – Distribución de las resoluciones de las estructuras de ARNs contenidas en PDB. Como se puede observar, casi las dos terceras partes de las estructuras de ARNs difracta en el rango 2,5 – 3,5Å. Imagen adaptada de [201].

Cuanto más alta la resolución (valores absolutos más bajos) mejor el detalle atómico, por eso aquellos valores $< 1{,}5$ Å se los llama de resolución atómica porque se logran ver los átomos de forma separada y sus distancias e interacciones son calculadas con una gran exactitud. Tener resoluciones muy buenas no asegura poseer información espacial de todos los átomos de una estructura, ya que pueden existir regiones que presenten mucha movilidad y por lo tanto la densidad electrónica alrededor de esos átomos sea difícil de definir. Generalmente se busca trabajar con resoluciones en el rango de $1{,}5 - 2{,}5$ Å. Si nos enfocamos en los ARNs, veremos (**figura 4.3.1 – B**) que la gran mayoría de las estructuras cristalizadas y depositadas en la PDB se encuentran con valores de hasta $4{,}5$ Å y la mayor proporción de ellas difracta entre $2{,}5 - 3{,}5$ Å. En esos valores las bases y fosfatos son fácilmente localizables debido a su rigidez, volumen y alta densidad electrónica, respectivamente. A resoluciones de menor calidad, el *backbone* suele no ser tan claro y presenta ciertas imprecisiones. Mejorar éste último inconveniente tiene grandes implicancias biológicas ya que la correcta conformación del *backbone* da un mejor entendimiento de la función biológica, por ejemplo, en aquellas estructuras que realizan alguna clase de catálisis o se ven involucradas en reconocimientos moleculares.

Hoy en día existen paquetes de softwares dedicados exclusivamente en corregir errores presentes en estructuras tridimensionales[202–204]. Estos errores pueden ser geométricos locales, conformacionales, estéricos, etc. Actualmente son ampliamente utilizados en los procesos de validación y refinamiento estructural. Realizar arduamente esta práctica mejoraría los análisis que se desarrollan sobre información tridimensional de origen experimental, aunque se necesitan ciertos conocimientos básicos inherentes a los métodos de obtención. Siendo esto así, existen investigaciones aplicadas únicamente a estructuras de ARNs que mejoran sustancialmente los parámetros de calidad estructural en estructuras con resoluciones entre $4 - 6$ Å[201,205–207]. Su aplicación en la información 3D utilizada para el estudio de la diversidad conformacional en ARNs sería un excelente trabajo a futuro.

Para la descripción de los datos de resolución de la BD preliminar, hemos elegido como criterio de corte tomar sólo aquellas estructuras con una resolución de hasta $4{,}5$ Å inclusive. En la **figura 4.3.2** se representa la distribución de las resoluciones de todos los confórmeros obtenidos por DRX y EM en nuestra base de datos. Se observa un pico pronunciado entre los $2{,}5 - 3{,}0$ Å coincidiendo con el estudio realizado en [201]. La mediana de la distribución es $2{,}70$ Å, el promedio $2{,}69$ Å y el 75% percentil abarca todos aquellos confórmeros con resolución hasta los $3{,}01$Å. Si consideramos sólo las estructuras que poseen alta calidad de resolución ($\leq 2.5$Å) encontramos que éstas representan un 37,74% del total de los confórmeros resueltos por DRX y EM (1982 = 1892 + 90, respectivamente), y sólo corresponden a estructuras resueltas por DRX.

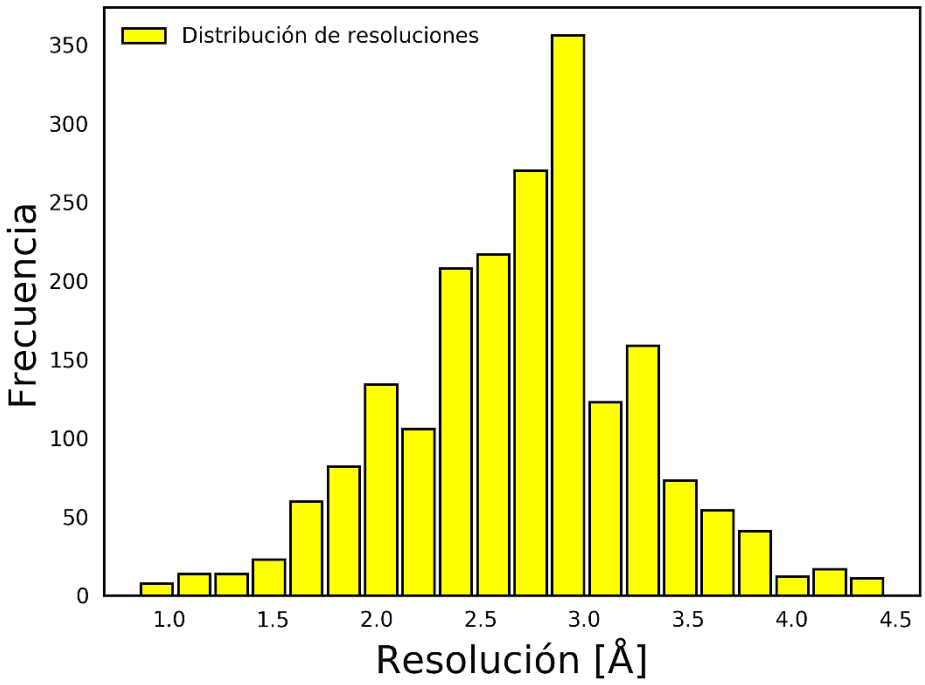



**Figura 4.3.2:** Distribución de las resoluciones para las moléculas contenidas en la BD preliminar.

Al estudiar la correlación entre la resolución promedio de los confórmeros que constituyen el par de máxima RMSD y la divergencia estructural, encontramos que no hay correlación entre las variables. Esto daría cierta robustez al estudio de la diversidad conformacional utilizando los C3' (ver **figura 4.3.3**).

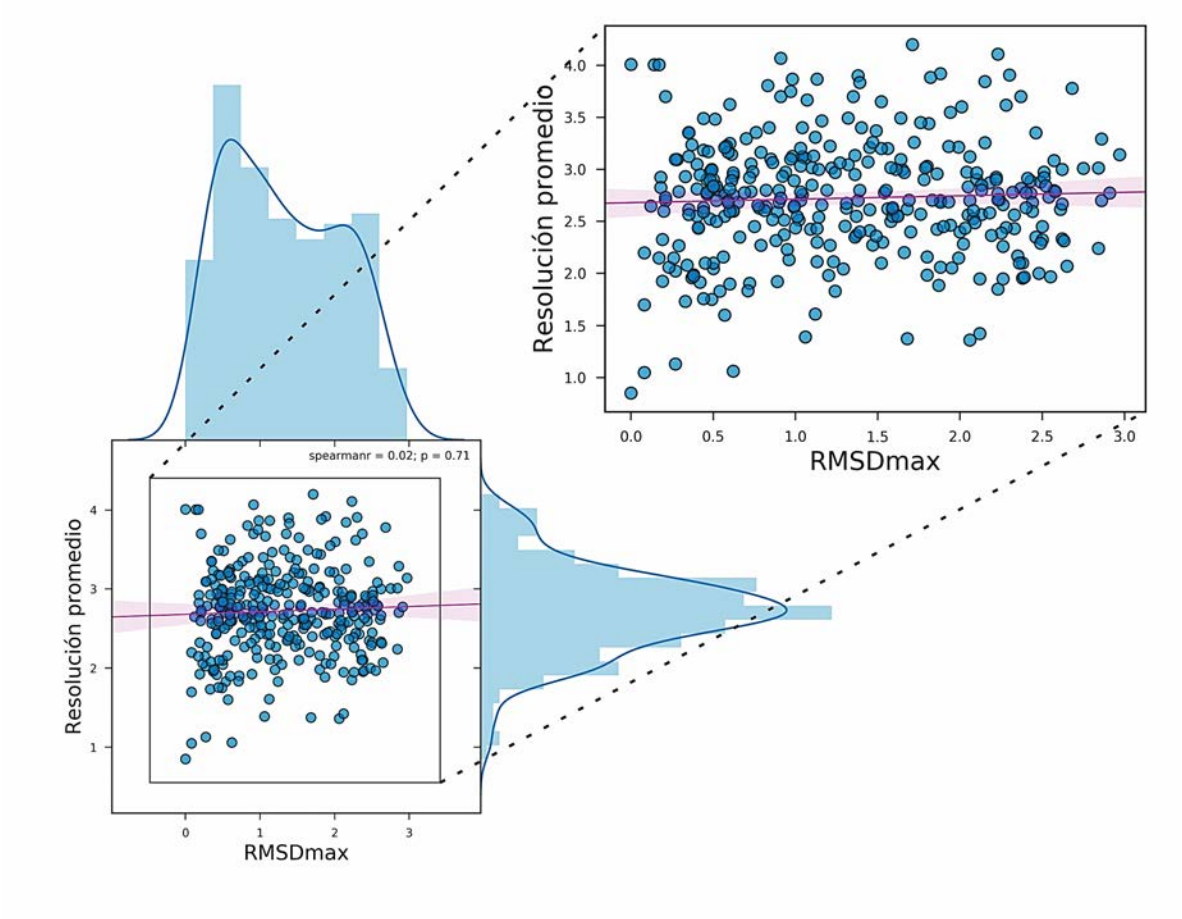



**Figura 4.3.3:** Estudio de la correlación entre la resolución promedio y el RMSD del par máximo.

## 4.4 Representación taxonómica

Es sabido que ciertos organismos, ya sea por encontrarse en nichos ecológicos diferenciales o por poseer bias composicionales en sus genomas, muestran patrones evolutivos particulares. Para estudiar si la distribución de la DC es sensible a estas particularidades, a continuación analizamos la relación entre la DC y la clasificación taxonómica provista por el NCBI. De los 2225 confórmeros contenidos en la BD preliminar de DC en ARNs, 720 confórmeros provienen de 72 organismos diferentes, pertenecientes a distintos reinos vivos. En la **figura 4.4.1** se representan los 5 primeros organismos más representados. Las moléculas de ARNs de *Escherichia coli* son las que más predominan (13,6%), seguido por las de *Saccharomyces cerevisiae* (10,1%), *Homo sapiens* (10%), *Haloarcula marismortui* (9,6%) y *Thermus thermophilus* (4,2%). En el porcentaje restante (52,5%) se encuentran agrupados todos aquellos organismos que poseen una

representación <4% del total de organismos presentes en la BD preliminar. Es importante mencionar que la gran mayoría de los confórmeros no cuenta con información taxonómica suministrada por la PDB, estos son 1228; a su vez, según lo informado por la PDB, se excluyó del análisis taxonómico aquellas cadenas provenientes de síntesis química (277).

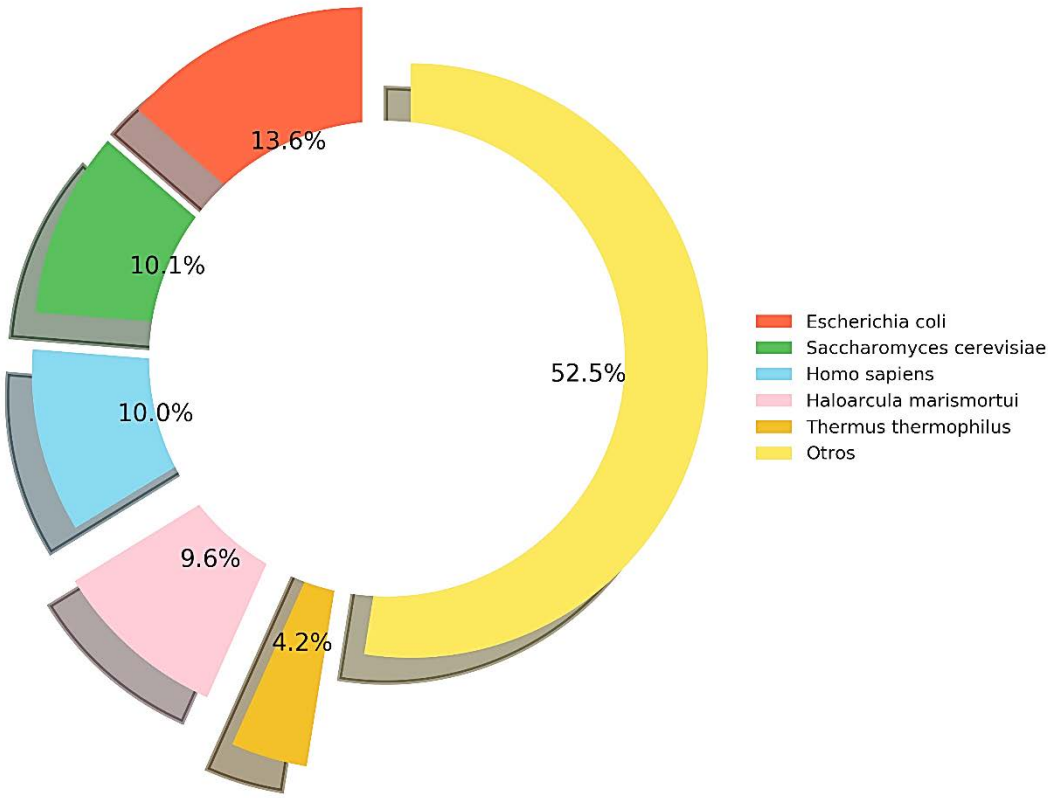



**Figura 4.4.1:** Los cinco organismos más representados en la BD preliminar.

Seguido a los análisis anteriores, realizamos la comparación de las extensiones de las diversidades conformacionales para la totalidad de las moléculas de ARN pertenecientes a cada uno de los 5 organismos más representados en la base de datos (ver **figura 4.4.2**). Si bien algunos de las distribuciones muestran diferencias significativas, es necesario realizar estudios adicionales para comprender si tales diferencias representan propiedades biológicas.

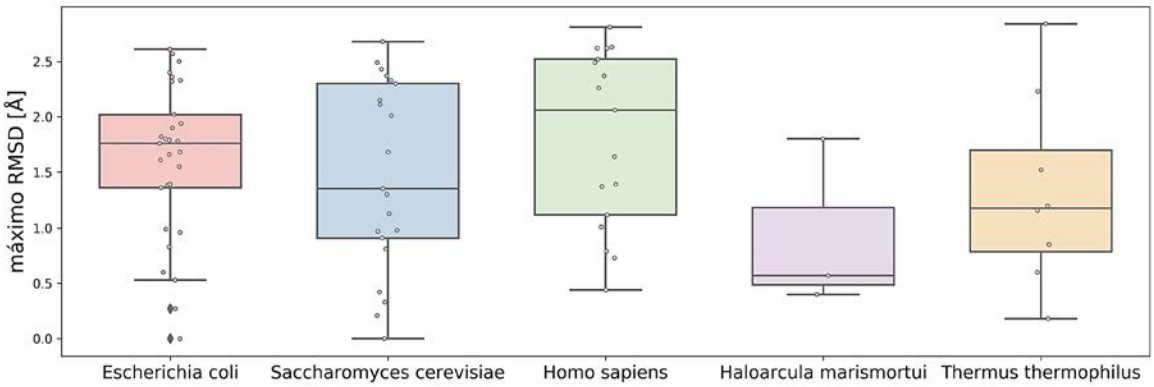



**Figura 4.4.2:** DC de los cinco organismos más representados en la BD preliminar.

## 4.5 Distribución de la diversidad conformacional

En la última sección del capítulo anterior vimos que, para cada molécula de ARN presente en la base de datos preliminar, existe un par de confórmeros que describe la máxima diversidad conformacional observada, lo que denominamos par máximo. Ésta medida proviene de la comparación estructural entre todos los confórmeros que presenta dicha molécula de ARN. Para conocer la distribución de la diversidad conformacional máxima de cada molécula de ARN realizamos un histograma de frecuencia que incluye los valores de los pares máximos de cada uno los 423 clusters. En la **figura 4.5.1** se observa dicha distribución y también se agrega la equivalente encontrada para proteínas en *CoDNaS*.

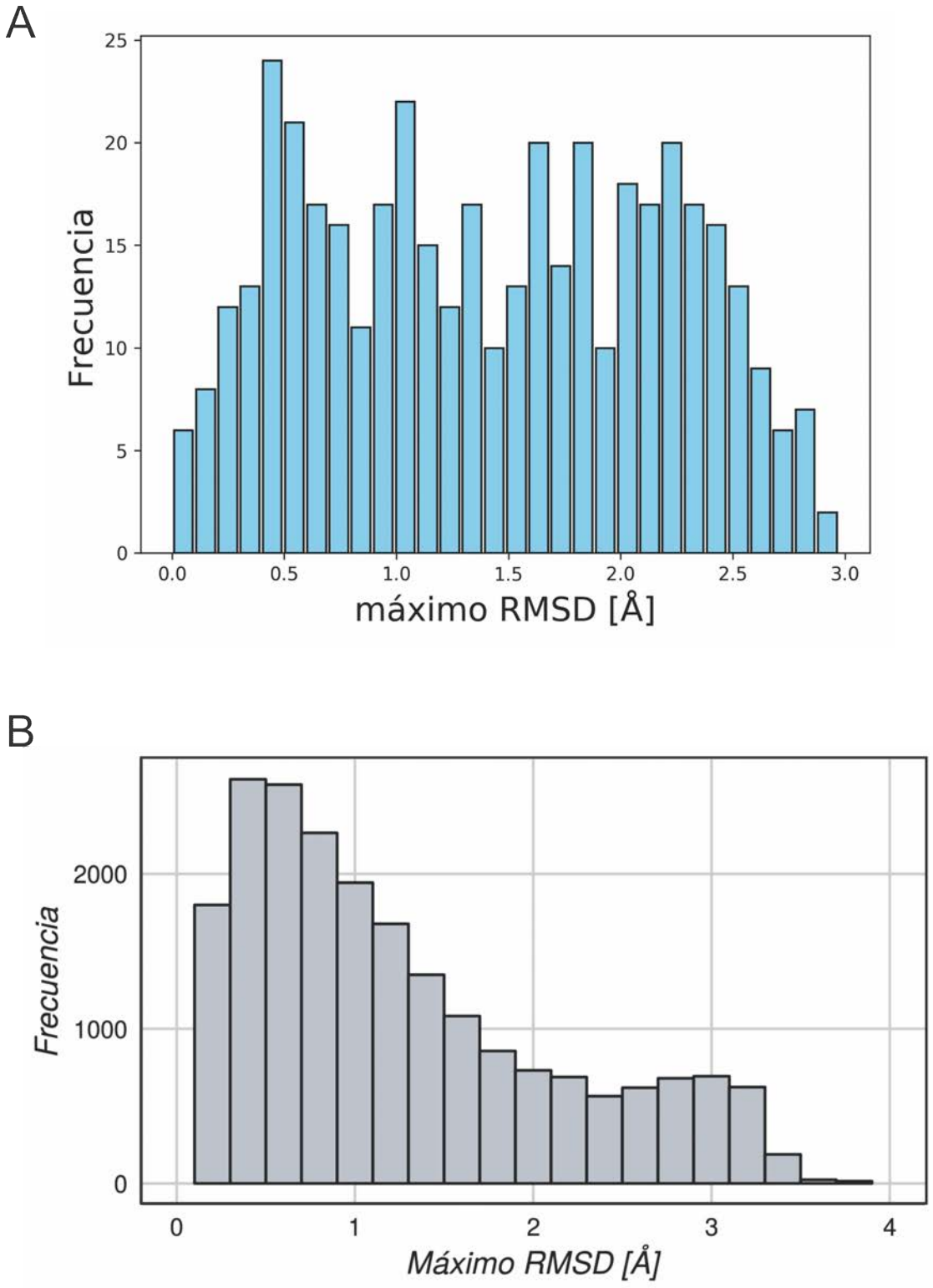



**Figura 4.5.1:** Distribución de la extensión de la diversidad conformacional de todas las moléculas. A – Moléculas de ARNs contenidas en la BD preliminar. B – Proteínas contenidas en *CoDNaS*.

De la observación de las distribuciones en la **figura 4.5.1**, podemos decir que en comparación con la de proteínas, el histograma para moléculas de ARNs carece de un pico de valor máximo de RMSD pronunciado. Lo que en principio indicaría que la movilidad de las moléculas de ARN es bastante uniforme mayormente centralizada en valores de 0,5, 1 y 2 de RMSD, donde se observa la mayor frecuencia. Esto se relaciona con la gran diferencia en la dinámica de ARN y proteínas, donde los ARN presentan gran movilidad a nivel del *backbone*. Recordemos que el *backbone* de los ARNs puede describirse considerando las ribosas y fosfatos con 6 ángulos de torsión (**α**, **β**, **γ**, **δ**, **ε**, y **ζ**).

En la distribución de los pares máximos de RMSD para ARNs encontramos que su extensión alcanza valores hasta $2{,}97\text{Å}$, que el valor de su mediana es $1{,}40\text{Å}$ y que su promedio se encuentra en $1{,}42\text{Å}$. Además, como se demostró en las secciones anteriores, la medida del RMSD del par máximo es robusta frente al largo secuencial y número de confórmeros por cluster.

Sumado a la descripción de la medida del RMSD, es importante mencionar que el 85% de los confórmeros totales fueron obtenidos por DRX, 11% por RMN y 4% por EM. A su vez, al realizar las comparaciones estructurales, encontramos que 299 clusters contenían confórmeros resueltos por DRX en su par máximo, 86 por RMN y 11 por DRX y RMN (en adelante DRX/RMN). A su vez, debido a la baja cantidad de confórmeros resueltos por EM (90), sólo 21 clusters contenían confórmeros resueltos por EM en su par máximo, 6 por EM/DRX y ninguno por EM/RMN. Entonces, de los siguientes 3 grupos (DRX, RMN y DRX/RMN), se realizaron *violinplots* para evidenciar la densidad de la distribución de los valores que toma la medición del máximo RMSD. También se resaltaron los clusters (puntos azules y naranjas) diferenciándolos en su redundancia, esto es la cantidad de cadenas presentes en el cluster en cuestión. Como puede observarse en la **figura 4.5.2**, los valores correspondientes al grupo RMN son significativamente mayores a los del grupo DRX (*test* de *Mann-Whitney* dió p-val=$2.83\text{x}10^{-10}$; *Kolmogorov-Smirnov* dió D=0.41, p-val=$1.74\text{x}10^{-10}$). Este resultado conduce a asumir que la medida del máximo RMSD se ve afectada significativamente por el método de obtención de sus confórmeros.

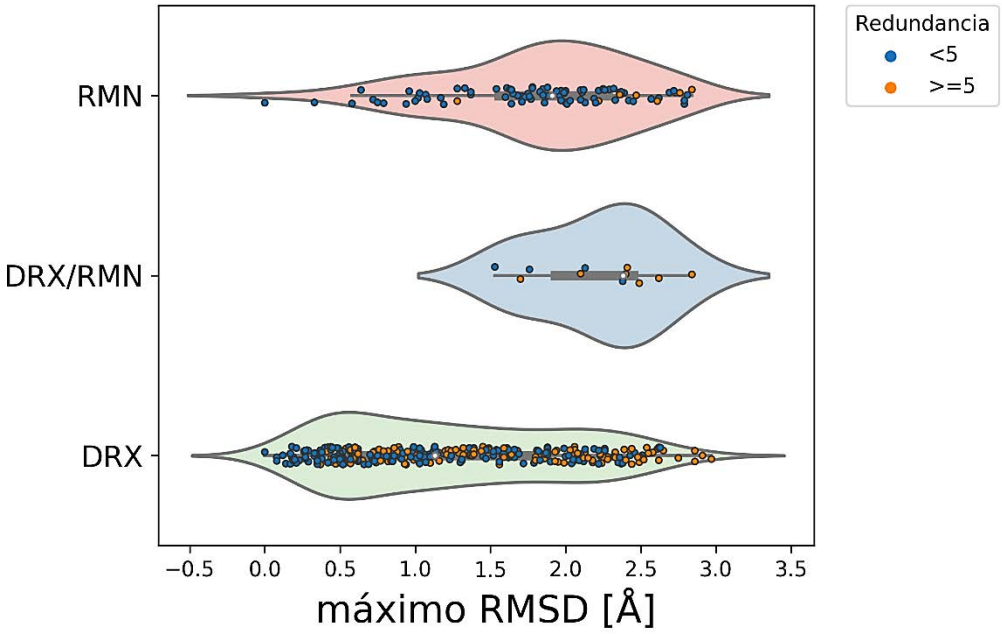



**Figura 4.5.2:** Distribución de la diversidad conformacional según el método utilizado para la obtención de los confórmeros. En rosado observamos los valores de máximo RMSD de aquellas moléculas de ARNs que tienen todos sus confórmeros obtenidos por RMN. En celeste, moléculas de ARNs con confórmeros RMN y DRX. En verde claro, moléculas de ARNs con todos sus confórmeros obtenidos por DRX.

Por lo antes dicho, no es recomendable utilizar confórmeros provenientes de diferentes métodos en la medida del máximo RMSD. Convenientemente, la BD preliminar contiene un 85% de sus confórmeros resueltos por DRX.

## 4.6 Clasificación de tipos de ARNs usando RNAcentral

La clasificación de las cadenas de ARNs contenidas en la BD preliminar se realizó con el objetivo de comparar la diversidad conformacional que presentan diferentes tipos de ARNs. Así, nos preguntamos si los tRNAs tienen mayor, igual o menor DC respecto a las ribozimas. Como mencionamos en la sección 3.4 del capítulo 3, para lograr la clasificación utilizamos la información suministrada por la base de datos **RNAcentral** debido a que es la BD secuencial no-redundante más grande que existe ya que asigna a cada cadena de ARN un único código hexadecimal y dicha información proviene de 28 bases de datos de ARNs específicas y curadas. **RNAcentral** basa su clasificación de los tipos de ncRNAs acorde a la

clasificación de la organización “Colaboración internacional de bases de datos de secuencias de nucleótidos” (**INSDC**, del inglés “*International Nucleotide Sequence Database Collaboration”*. URL: *http://www.insdc.org/*) conformada por el **NCBI**, **DDBJ** (del inglés “*DNA Data Bank of Japan”*) y **ENA** (del inglés “*European Nucleotide Archive”*).

Al ser un recurso público y abierto, **RNAcentral** permite descargar variada información para cada una de sus liberaciones históricas. En nuestro trabajo utilizamos la liberación número 9. Mediante diferentes *scripts* parseamos y cruzamos la información para lograr la anotación de nuestros clusters. Recordemos, los clusters son únicos en secuencia. De la totalidad de clusters (430) logramos anotar el 97,67% (420). Entonces, cada cluster anotado tiene asignado un único código URS y el nombre del tipo de ARN al que pertenece.

El paso siguiente fue agrupar todos aquellos clusters que pertenecen al mismo tipo de ARN y, mediante una gráfica múltiple del tipo *boxplot* comparamos sus valores máximos de RMSD. A su vez, de los 420 clusters anotados, solo 413 poseen medida del máximo RMSD. En la **figura 4.6.1** podemos observar la dispersión alcanzada de las DCs por cada tipo de ARN. Como se observa, los tipos de ARNs presentes son: *ribosima hammerhead*, *rRNA*, *tRNA*, *SRP_RNA*, *snRNA* y *ribozima*.

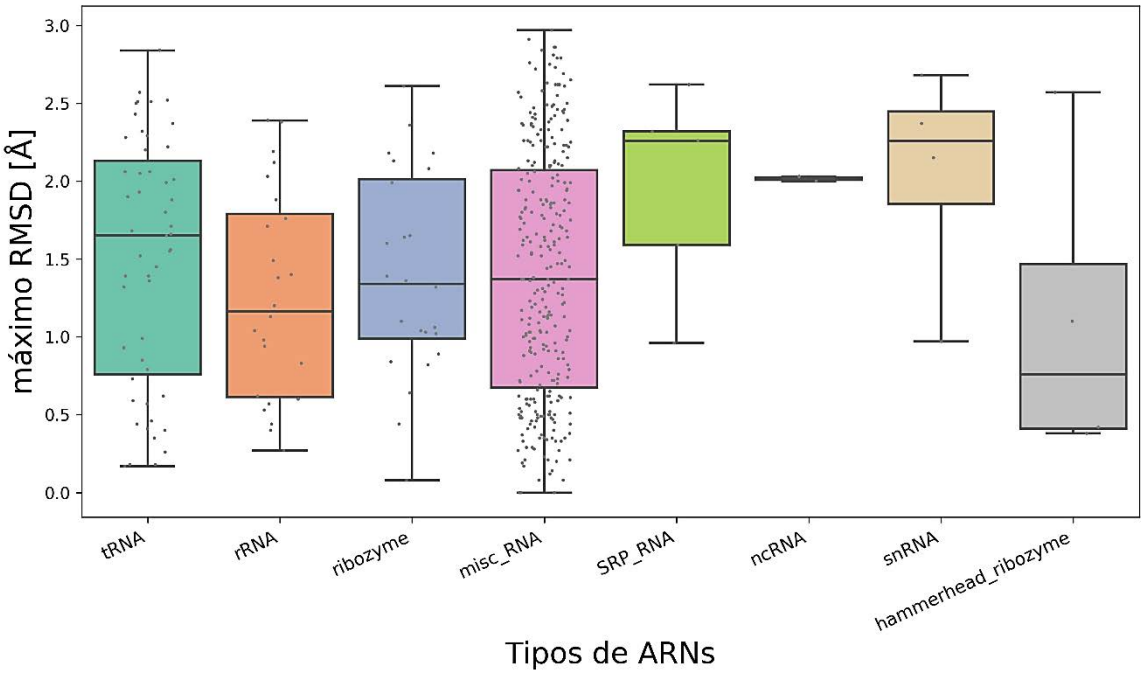



**Figura 4.6.1:** Distribución de las DCs para los diferentes tipos de ARNs.

Es importante aclarar que *misc_RNA* no es un tipo de ARN, sino que, según el **INSDC**, son todas aquellas secuencias de ARN misceláneas que no están clasificadas en alguno de los otros tipos de ARNs posibles. Lamentablemente en este grupo se encuentra la gran mayoría de los clusters anotados. Esto lo podemos observar en mayor detalle en la **tabla 4.1**.

| Tipo de ARN | Número de clusters | % total |
|---|---|---|
| *misc_RNA* | 299 | 72,40% |
| *tRNA* | 51 | 12,35% |
| *rRNA* | 24 | 5,81% |
| *Ribozima* | 24 | 5,81% |
| *SRP_RNA* | 5 | 1,21% |
| *Ribosima hammerhead* | 4 | 0,97% |
| *snRNA* | 4 | 0,97% |
| *ncRNA* | 2 | 0,48% |

**Tabla 4.1:** Número total de clusters en la BD preliminar por tipo de ARN anotado. En la columna de la derecha se representan los porcentajes respecto al total de clusters anotados.

En la **tabla 4.2** podemos observar un resumen de los valores estadísticos típicos para cada *boxplot*. No se muestran aquellos datos de los *boxplot* con menos de 24 clusters ya que carecen de significado estadístico para nuestro estudio comparativo. Adicionalmente, realizamos el *test* de *Shapiro* para corroborar si dichos datos describen una distribución normal.

| Tipo de ARN | Mediana | Media | Test de Shapiro | Distribución Normal |
|---|---|---|---|---|
| *tRNA* | 1,65 | 1,50 | W=0.934 ; p-val=0.00685 | No |
| *rRNA* | 1,16 | 1,26 | W=0.941 ; p-val=0.174 | Si |
| *Ribozima* | 1,34 | 1,39 | W=0.971 ; p-val=0.691 | SI |

**Tabla 4.2:** Número total de clusters en la BD preliminar por tipo de ARN anotado. En la columna de la derecha se representan los porcentajes respecto al total de clusters anotados.

Por otro lado, para todos aquellos tipos de ARNs con al menos 24 clusters, realizamos *tests* estadísticos para corroborar si provenían de iguales distribuciones (*test* de *Kolmogorov-Smirnov* para 2 muestras) y si las diferencias encontradas en sus medianas (*test* de *Mann-Whitney*) y medias (*test T*) eran significativamente diferentes. En la **tabla 4.3** se pueden observar los resultados de cada test realizado. Desafortunadamente no logramos resultados favorables y creemos que se deba a que la muestra poblacional de cada tipo de ARN (nuestra evidencia experimental) no alcanza a ser relevante para lograr resultados

significativos en los *tests*. Esto último se podría solucionar a futuro cuando haya un mayor número de estructuras de ARNs depositadas en la PDB.

| Tests | **KS 2_samples** | **Mann-Whitney** | **Bartlett*** | **T-Test** |
|---|---|---|---|---|
| **tRNA-Ribozima** | D=0.198 ; p-val=0.494 | U=557.5 ; p-val=0.270 | stat=1.003 ; p-val=0.316 | stat=0.562 ; p-val=0.575 |
| **tRNA-rRNA** | D=0.235 ; p-val=0.286 | U=504.0 ; p-val=0.111 | stat=0.780 ; p-val=0.378 | stat=-1.268 ; p-val=0.209 |
| **rRNA-Ribozima** | D=0.208 ; p-val=0.622 | U=255.0 ; p-val=0.251 | stat=0.0113 ; p-val=0.915 | stat=-0.690 ; p-val=0.493 |

**Tabla 4.3:** Valores de los tests estadísticos realizados a los tres tipos de ARNs más poblados con excepción de *misc_RNA*. *Se realizó este test de igualdad de varianza para poder cumplir las condiciones del T-Test donde sólo puede ser utilizado en muestras con igual varianza.

Haber realizado estas comparaciones tiene como principal interés biológico buscar entender si la DC está relacionada con el tipo de ARN y cómo ésta influye en los mecanismos de los procesos biológicos involucrados. Debemos tener en cuenta que realizar este tipo de comparación engloba muchas variables ya que nos estamos refiriendo al comportamiento de la DC en poblaciones muy grandes como lo puede ser pertenecer a un tipo de ARN particular. Dicho esto, esperamos encontrar mejores resultados situándonos en la comparación dentro de cada tipo de ARN, como por ejemplo, comparar tRNAs de aminoácidos diferentes para todos los organismos presentes en la BD preliminar o entre ribozimas. Con ésto entendemos que hemos dado pequeños pasos a grandes preguntas.

# 5. Conclusiones

El presente trabajo final planteó muchos interrogantes desde su concepción. Se partió desde la experiencia acuñada por el grupo – SBG - UNQ – en proteínas. Se esbozaron las primeras y más básicas preguntas sobre la biología del ARN, muchísimas otras se encontraron en el camino. Se aprendió a buscar, descartar y seleccionar bibliografía específica que ayudó a conocer la historia, la química y biología estructural del ARN, yendo así un poco más allá de los libros de textos comúnmente usados. Se conceptualizó el significado y se hizo uso de múltiples bases de datos específicas de ARNs. Se logró dimensionar la complejidad que ha alcanzado el campo de los ARNs en los últimos 20-30 años, aún en explosión.

En cuanto al campo de la bioinformática, se utilizaron varios lenguajes de programación (*Bash, Python* y *R)* bajo el sistema operativo Ubuntu, una de las distribuciones basadas en GNU/Linux. Mediante el uso de estos lenguajes, especialmente *Bash* y *Python*, se alcanzaron muchas respuestas a las preguntas planteadas. Es importante mencionar también que, en el transcurso del trabajo final, se concurrió a varias charlas, simposios, congresos, talleres y cursos relacionados a la temática desarrollada. Algunos de ellos fueron los siguientes: *2nd LA Student Council Symposium* - IIB ("Instituto de investigaciones Biotecnológicas") Universidad Nacional de San Martín, Campus Miguelete, Buenos Aires – Argentina; *Primera Jornada Argentina de Biología de ARNs no codificantes* - Universidad Nacional de Quilmes, Bernal – Argentina; 3er Simposio Argentino de Jóvenes Investigadores en Bioinformática - Fundación Instituto Leloir, CABA – Argentina; *Escuela de Bioinformática* - "R, Python y Linux" - Fundación Instituto Leloir, CABA – Argentina; *Workshop* “Filtrado y ensamblado NGS” – 3SAJIB - Fundación Instituto Leloir, CABA – Argentina; *II Reunión*

*Argentina de Biología de ARNs no codificantes* - Universidad Nacional de Quilmes, Bernal – Argentina.

A través de éste trabajo final, hemos estudiado y analizado la diversidad conformacional de las moléculas de ARN. A partir de la información recolectada en la PDB, y luego de reiteradas etapas de curado, se generó un valioso *dataset* redundante asegurando la confiabilidad de los datos enunciados y mostrados en las diferentes gráficas. Hemos creado la primera base de datos preliminar de DC en ARNs conteniendo la mayor cantidad de información tridimensional provistas a nivel mundial. Sin lugar a duda, el principal objetivo de ésta base es el de brindar un repositorio donde se encuentren, clasifiquen, organicen y estudien los diferentes confórmeros de ARNs para cada ensamble conocido. Asimismo, debido a la baja información 3D, también busca alentar a la comunidad científica en la obtención de nuevos confórmeros tridimensionales, en diferentes condiciones y de mejor resolución, etc. A su vez, hemos demostrado que la medida de la DC en ARNs es robusta frente al largo secuencial y al número de confórmeros con que se cuenta de cada molécula. Se encontró que la extensión de la distribución de DC para ARNs no supera los $3\,\text{Å}$ y presenta una media de $1{,}39\,\text{Å}$. Además, hemos demostrado que la estimación de la DC, mediante el uso de confórmeros resueltos por métodos experimentales diferentes, se ve afectada de forma significativa. Por otro lado, a pesar de no haber conseguido resultados estadísticamente significativos, hemos vinculado estos resultados obtenidos con los organismos más representados en la base de datos, como así también con los 3 tipos de moléculas de ARNs mejores representadas en la base de datos. Creemos que esto se deba a la limitada información con la que hemos desarrollado éstas correlaciones, la cual se irá reduciendo con la eventual deposición de nuevas y mejores estructuras de ARNs. Finalmente, creemos que los ARNs no presentan rigidez estructural, sino más bien una alta movilidad característica de los ácidos nucleicos y que la misma está fuertemente relacionada con su función biológica.

## 5.1 Trabajo a Futuro

Los resultados y conclusiones anteriores, sumado a lo conocido en proteínas, nos plantéan nuevos interrogantes, por ejemplo, estudiar: los movimientos conformacionales en complejos proteína-ARN, ARN-ARN, con las funciones biológicas; la conservación secuencial y estructural de las interacciones de éstos complejos; la divergencia estructural y secuencial en ARNs (modelos estructurales por homología); la DC en complejos ARN-Proteína y cómo éstos interaccionan; el impacto de la información estructural en enfermedades conocidas.

Sobre la concepción de la BD de DC:

- Mejorar los criterios de los filtros de la BD preliminar.

- Comparar resultados de RMSD entre diferentes softwares.

- Caracterizar y anotar, mediante otros métodos, aquellos clusters que fueron asignados como misceláneos (misc_RNA).

- Estudiar los aspectos vinculados a la diversidad conformacional, desde las condiciones de obtención de la estructura, presencia de ligandos y/o otras biomoléculas.

- Desarrollar una interfaz gráfica amigable de la BD de DC en ARNs.

Bibliografía


1. Fischer E (1894) Einfluss der Configuration auf die Wirkung der Enzyme. Ber Dtsch Chem Ges 27: 2985–2993. doi:10.1002/cber.18940270364.

2. O'Brien PJ, Herschlag D (1999) Catalytic promiscuity and the evolution of new enzymatic activities. Chem Biol 6: R91–R105. doi:10.1016/S1074-5521(99)80033-7.

3. Nobeli I, Favia AD, Thornton JM (2009) Protein promiscuity and its implications for biotechnology. Nat Biotechnol 27: 157–167. doi:10.1038/nbt1519.

4. Koshland DE, Ray WJ, Erwin MJ (1958) Protein structure and enzyme action. Fed Proc 17: 1145–1150.

5. Koshland DE (1995) The key–lock theory and the induced fit theory. Angew Chem Int Ed Engl 33: 2375–2378. doi:10.1002/anie.199423751.

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

7. Pauling L (1940) A theory of the structure and process of formation of antibodies*. J Am Chem Soc 62: 2643–2657. doi:10.1021/ja01867a018.

8. Karush F (1950) Heterogeneity of the Binding Sites of Bovine Serum Albumin. J Am Chem Soc 72: 2705–2713. doi:10.1021/ja01162a099.

9. Kendrew J, Bodo G, Dintzis H, Parrish R, Wyckoff H, et al. (1958) A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. Nature 181: 662–666. doi:10.1038/181662a0.

10. Monod J, Wyman J, Changeux JP (1965) ON THE NATURE OF ALLOSTERIC TRANSITIONS: A PLAUSIBLE MODEL. J Mol Biol 12: 88–118. doi:10.1016/S0022-2836(65)80285-6.

11. Perutz MF, Rossmann MG, Cullis AF, Muirhead H, Will G, et al. (1960) Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. Nature 185: 416–422. doi:10.1038/185416a0.

12. Weikl TR, von Deuster C (2009) Selected-fit versus induced-fit protein binding: kinetic differences and mutational analysis. Proteins 75: 104–110. doi:10.1002/prot.22223.

13. Boehr DD, Nussinov R, Wright PE (2009) The role of dynamic conformational ensembles in biomolecular recognition. Nat Chem Biol 5: 789–796. doi:10.1038/nchembio.232.

14. Weber G (1975) Energetics of ligand binding to proteins. Advances in protein chemistry volume 29. Advances in protein chemistry. Elsevier, Vol. 29. pp. 1–83. doi:10.1016/S0065-3233(08)60410-6.

15. Jameson DM (1998) Gregorio Weber, 1916-1997: a fluorescent lifetime. Biophys J 75: 419–421. doi:10.1016/S0006-3495(98)77528-9.

16. Frauenfelder H, Sligar SG, Wolynes PG (1991) The energy landscapes and motions of proteins. Science 254: 1598–1603. doi:10.1126/science.1749933.

17. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG (1995) Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins 21: 167–195. doi:10.1002/prot.340210302.

18. Onuchic JN, Wolynes PG, Luthey-Schulten Z, Socci ND (1995) Toward an outline of the topography of a realistic protein-folding funnel. Proc Natl Acad Sci USA 92: 3626–3630.

19. Onuchic JN, Luthey-Schulten Z, Wolynes PG (1997) Theory of protein folding: the energy landscape perspective. Annu Rev Phys Chem 48: 545–600. doi:10.1146/annurev.physchem.48.1.545.

20. Dill KA, Ozkan SB, Shell MS, Weikl TR (2008) The protein folding problem. Annu Rev Biophys 37: 289–316. doi:10.1146/annurev.biophys.37.092707.153558.

21. Dill KA, Chan HS (1997) From Levinthal to pathways to funnels. Nat Struct Biol 4: 10–19. doi:10.1038/nsb0197-10.

22. James LC, Tawfik DS (2003) Conformational diversity and protein evolution--a 60-year-old hypothesis revisited. Trends Biochem Sci 28: 361–368. doi:10.1016/S0968-0004(03)00135-X.

23. Wei G, Xi W, Nussinov R, 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. doi:10.1021/acs.chemrev.5b00562.

24. Kumar S, Ma B, Tsai CJ, Sinha N, Nussinov R (2000) Folding and binding cascades: dynamic landscapes and population shifts. Protein Sci 9: 10–19. doi:10.1110/ps.9.1.10.

25. Henzler-Wildman K, Kern D (2007) Dynamic personalities of proteins. Nature 450: 964–972. doi:10.1038/nature06522.

26. Treiber DK, Williamson JR (1999) Exposing the kinetic traps in RNA folding. Curr Opin Struct Biol 9: 339–345. doi:10.1016/S0959-440X(99)80045-1.

27. Woodson SA (2000) Recent insights on RNA folding mechanisms from catalytic RNA. Cell Mol Life Sci 57: 796–808. doi:10.1007/s000180050042.

28. Russell R, Zhuang X, Babcock HP, Millett IS, Doniach S, et al. (2002) Exploring the folding landscape of a structured RNA. Proc Natl Acad Sci USA 99: 155–160. doi:10.1073/pnas.221593598.

29. Ditzler MA, Rueda D, Mo J, Håkansson K, Walter NG (2008) A rugged free energy landscape separates multiple functional RNA folds throughout denaturation. Nucleic Acids Res 36: 7088–7099. doi:10.1093/nar/gkn871.

30. Solomatin SV, Greenfeld M, Chu S, Herschlag D (2010) Multiple native states reveal persistent ruggedness of an RNA folding landscape. Nature 463: 681–684. doi:10.1038/nature08717.

31. Woodson SA (2010) Taming free energy landscapes with RNA chaperones. RNA Biol 7: 677–686.

32. Boehr DD, McElheny D, Dyson HJ, Wright PE (2006) The dynamic energy landscape of dihydrofolate reductase catalysis. Science 313: 1638–1642. doi:10.1126/science.1130258.

33. Flock T, Weatheritt RJ, Latysheva NS, Babu MM (2014) Controlling entropy to tune the functions of intrinsically disordered regions. Curr Opin Struct Biol 26: 62–72. doi:10.1016/j.sbi.2014.05.007.

34. Walsh I, Giollo M, Di Domenico T, Ferrari C, Zimmermann O, et al. (2015) Comprehensive large-scale assessment of intrinsic protein disorder. Bioinformatics 31: 201–208. doi:10.1093/bioinformatics/btu625.

35. Potenza E, Di Domenico T, Walsh I, Tosatto SCE (2015) MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. Nucleic Acids Res 43: D315–20. doi:10.1093/nar/gku982.

36. Meng F, Uversky VN, Kurgan L (2017) Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions. Cell Mol Life Sci 74: 3069–3090. doi:10.1007/s00018-017-2555-4.

37. Piovesan D, Tabaro F, Paladin L, Necci M, Micetic I, et al. (2018) MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins. Nucleic Acids Res 46: D471–D476. doi:10.1093/nar/gkx1071.

38. Howton TC, Zhan YA, Sun Y (n.d.) Intrinsically disordered proteins: controlled chaos or random walk. pagepress.org.

39. Brion P, Westhof E (1997) Hierarchy and dynamics of RNA folding. Annu Rev Biophys Biomol Struct 26: 113–137. doi:10.1146/annurev.biophys.26.1.113.

40. Al-Hashimi HM, Walter NG (2008) RNA dynamics: it is about time. Curr Opin Struct Biol 18: 321–329. doi:10.1016/j.sbi.2008.04.004.

41. Mustoe AM, Brooks CL, Al-Hashimi HM (2014) Hierarchy of RNA functional dynamics. Annu Rev Biochem 83: 441–466. doi:10.1146/annurev-biochem-060713-035524.

42. Reining A, Nozinovic S, Schlepckow K, Buhr F, Fürtig B, et al. (2013) Three-state mechanism couples ligand and temperature sensing in riboswitches. Nature 499: 355–359. doi:10.1038/nature12378.

43. Giuliodori AM, Pietro FD, Marzi S, Masquida B (n.d.) The cspA mRNA is a thermosensor that modulates translation of the cold-shock protein CspA. Elsevier.

44. Haller A, Souliere MF, Micura R (n.d.) The dynamic nature of RNA as key to understanding riboswitch mechanisms. ACS Publications.

45. Cromie MJ, Shi Y, Latifi T, Groisman EA (n.d.) An RNA sensor for intracellular Mg2+. Elsevier.

46. Elgrably-Weiss M, Sheaffer A (n.d.) A pH-responsive riboregulator. genesdev.cshlp.org.

47. Babitzke P, Yanofsky C (n.d.) Reconstitution of Bacillus subtilis trp attenuation in vitro with TRAP, the trp RNA-binding attenuation protein. National Acad Sciences.

48. Grundy FJ, Winkler WC (n.d.) tRNA-mediated transcription antitermination in vitro: codon–anticodon pairing independent of the ribosome. National Acad Sciences.

49. Stoddard CD, Montange RK, Hennelly SP, Rambo RP, Sanbonmatsu KY, et al. (2010) Free state conformational sampling of the SAM-I riboswitch aptamer domain. Structure 18: 787–797. doi:10.1016/j.str.2010.04.006.

50. Yoshizawa S, Fourmy D, Puglisi JD (1999) Recognition of the codon-anticodon helix by ribosomal RNA. Science 285: 1722–1725.

51. Dethoff EA, Petzold K, Chugh J, Casiano-Negroni A, Al-Hashimi HM (2012) Visualizing transient low-populated structures of RNA. Nature 491: 724–728. doi:10.1038/nature11498.

52. Pauling L, Delbrück M (1940) The nature of the intermolecular forces operative in biological processes. Science 92: 77–79. doi:10.1126/science.92.2378.77.

53. Ancel LW, Fontana W (2000) Plasticity, evolvability, and modularity in RNA. J Exp Zool 288: 242–283. doi:10.1002/1097-010X(20001015)288:3<242::AID-JEZ5>3.0.CO;2-O.

54. Joyce GF (1997) Evolutionary chemistry: getting there from here. Science 276: 1658–1659.

55. Lee RC, Feinbaum RL, Ambros V (n.d.) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Elsevier.

56. Wightman B, Ha I, Ruvkun G (n.d.) Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. Elsevier.

57. Fire A, Xu SQ, Montgomery MK, Kostas SA, Driver SE (n.d.) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. nature.com.

58. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, et al. (2012) Landscape of transcription in human cells. Nature 489: 101–108. doi:10.1038/nature11233.

59. Hangauer MJ, Vaughn IW, McManus MT (2013) Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. PLoS Genet 9: e1003569. doi:10.1371/journal.pgen.1003569.

60. Wan Y, Kertesz M, Spitale RC, Segal E, Chang HY (2011) Understanding the transcriptome through RNA structure. Nat Rev Genet 12: 641–655. doi:10.1038/nrg3049.

61. Ding Y, Tang Y, Kwok CK, Zhang Y, Bevilacqua PC, et al. (2014) In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature 505: 696–700. doi:10.1038/nature12756.

62. Rouskin S, Zubradt M, Washietl S, Kellis M, Weissman JS (2014) Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. Nature 505: 701–705. doi:10.1038/nature12894.

63. Wan Y, Qu K, Zhang QC, Flynn RA, Manor O, et al. (2014) Landscape and variation of RNA secondary structure across the human transcriptome. Nature 505: 706–709. doi:10.1038/nature12946.

64. Ziv O, Gabryelska MM, Lun ATL, Gebert LFR, Sheu-Gruttadauria J, et al. (2018) COMRADES determines in vivo RNA structures and interactions. Nat Methods 15: 785–788. doi:10.1038/s41592-018-0121-0.

65. Lai W-JC, Kayedkhordeh M, Cornell EV, Farah E, Bellaousov S, et al. (2018) mRNAs and lncRNAs intrinsically form secondary structures with short end-to-end distances. Nat Commun 9: 4328. doi:10.1038/s41467-018-06792-z.

66. Yang SY, Lejault P, Chevrier S, Boidot R, Robertson AG, et al. (2018) Transcriptome-wide identification of transient RNA G-quadruplexes in human cells. Nat Commun 9: 4730. doi:10.1038/s41467-018-07224-8.

67. Watson JD, Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171: 737–738. doi:10.1038/171737a0.

68. Rich A, Davies DR (1956) A new two stranded helical structure: polyadenylic acid and polyuridylic acid. J Am Chem Soc 78: 3548–3549. doi:10.1021/ja01595a086.

69. Rich A (1960) A hybrid helix containing both deoxyribose and ribose polynucleotides and its relation to the transfer of information between the nucleic acids. Proc Natl Acad Sci USA 46: 1044–1053.

70. Felsenfeld G, Davies DR, Rich A (1957) FORMATION OF A THREE-STRANDED POLYNUCLEOTIDE MOLECULE. J Am Chem Soc 79: 2023–2024. doi:10.1021/ja01565a074.

71. Felsenfeld G, Rich A (1957) Studies on the formation of two- and three-stranded polyribonucleotides. Biochim Biophys Acta 26: 457–468.

72. Crick FH (1958) On protein synthesis. Symp Soc Exp Biol 12: 138–163.

73. Hoagland MB, Stephenson ML, Scott JF, Hecht LI, Zamecnik PC (1958) A soluble ribonucleic acid intermediate in protein synthesis. J Biol Chem 231: 241–257.

74. Cobb M (2015) Who discovered messenger RNA? Curr Biol 25: R526–32. doi:10.1016/j.cub.2015.05.032.

75. Brenner S, Jacob F, Meselson M (1961) An Unstable Intermediate Carrying Information from Genes to Ribosomes for Protein Synthesis. Nature 190: 576–581. doi:10.1038/190576a0.

76. Gros F, Hiatt H, Gilbert W, Kurland CG, Risebrough RW, et al. (1961) Unstable ribonucleic acid revealed by pulse labelling of escherichia coli. Nature 190: 581–585. doi:10.1038/190581a0.

77. Jacob F, Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol 3: 318–356. doi:10.1016/S0022-2836(61)80072-7.

78. Goodman HM, Rich A (1962) Formation of a DNA-soluble RNA hybrid and its relation to the origin, evolution, and degeneracy of soluble RNA. Proc Natl Acad Sci USA 48: 2101–2109.

79. Nirenberg M (2004) Historical review: Deciphering the genetic code--a personal account. Trends Biochem Sci 29: 46–54. doi:10.1016/j.tibs.2003.11.009.

80. Holley RW, Madison JT, Zamir A (1964) A new method for sequence determination of large oligonucleotides. Biochem Biophys Res Commun 17: 389–394. doi:10.1016/0006-291X(64)90017-8.

81. Holley RW, Apgar J, Everett GA, Madison JT, Marquisee M, et al. (1965) Structure of a ribonucleic acid. Science 147: 1462–1465.

82. Rich A (2009) The era of RNA awakening: structural biology of RNA in the early years. Q Rev Biophys 42: 117–137. doi:10.1017/S0033583509004776.

83. Kim SH, Quigley G, Suddath FL, Rich A (1971) High-resolution x-ray diffraction patterns of crystalline transfer RNA that show helical regions. Proc Natl Acad Sci USA 68: 841–845.

84. Kim SH, Quigley GJ, Suddath FL, McPherson A, Sneden D, et al. (1973) Three-dimensional structure of yeast phenylalanine transfer RNA: folding of the polynucleotide chain. Science 179: 285–288. doi:10.1126/science.179.4070.285.

85. Kim SH, Suddath FL, Quigley GJ, McPherson A, Sussman JL, et al. (1974) Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. Science 185: 435–440. doi:10.1126/science.185.4149.435.

86. Robertus JD, Ladner JE, Finch JT, Rhodes D, Brown RS, et al. (1974) Structure of yeast phenylalanine tRNA at 3 A resolution. Nature 250: 546–551. doi:10.1038/250546a0.

87. Pley HW, Flaherty KM, McKay DB (1994) Three-dimensional structure of a hammerhead ribozyme. Nature 372: 68–74. doi:10.1038/372068a0.

88. Martick M, Scott WG (2006) Tertiary contacts distant from the active site prime a ribozyme for catalysis. Cell 126: 309–320. doi:10.1016/j.cell.2006.06.036.

89. de la Peña M, García-Robles I, Cervera A (2017) The hammerhead ribozyme: A long history for a short RNA. Molecules 22. doi:10.3390/molecules22010078.

90. Cate JH, Gooding AR, Podell E, Zhou K, Golden BL, et al. (1996) Crystal structure of a group I ribozyme domain: principles of RNA packing. Science 273: 1678–1685.

91. Westhof E (2015) Twenty years of RNA crystallography. RNA 21: 486–487. doi:10.1261/rna.049726.115.

92. Westhof E, Auffinger P (2000) RNA Tertiary Structure. In: Meyers RA, editor. Encyclopedia of analytical chemistry: applications, theory and instrumentation. Chichester, UK: John Wiley & Sons, Ltd. doi:10.1002/9780470027318.a1428.

93. Gilbert WV, Bell TA, Schaening C (2016) Messenger RNA modifications: Form, distribution, and function. Science 352: 1408–1412. doi:10.1126/science.aad8711.

94. Boccaletto P, Machnicka MA, Purta E, Piatkowski P, Baginski B, et al. (2018) MODOMICS: a database of RNA modification pathways. 2017 update. Nucleic Acids Res 46: D303–D307. doi:10.1093/nar/gkx1030.

95. Xuan J-J, Sun W-J, Lin P-H, Zhou K-R, Liu S, et al. (2018) RMBase v2.0: deciphering the map of RNA modifications from epitranscriptome sequencing data. Nucleic Acids Res 46: D327–D334. doi:10.1093/nar/gkx934.

96. Cantara WA, Crain PF, Rozenski J, McCloskey JA, Harris KA, et al. (2011) The RNA Modification Database, RNAMDB: 2011 update. Nucleic Acids Res 39: D195–201. doi:10.1093/nar/gkq1028.

97. Kim SH (n.d.) Transfer RNA: crystal structures. Springer.

98. Bloomfield VA, Crothers DM, Tinoco I (2000) Nucleic acids : structures, properties, and functions. Sausalito, Calif. : University Science Books.

99. Saenger W (1984) Polymorphism of DNA versus Structural Conservatism of RNA: Classification of A-, B-, and Z-TYPe Double Helices. Principles of nucleic acid structure. Springer advanced texts in chemistry. New York, NY: Springer New York. pp. 220–241. doi:10.1007/978-1-4612-5190-3_9.

100. Murray LJW, Arendall WB, Richardson DC, Richardson JS (2003) RNA backbone is rotameric. Proc Natl Acad Sci USA 100: 13904–13909. doi:10.1073/pnas.1835769100.

101. Richardson JS, Schneider B, Murray LW, Kapral GJ, Immormino RM, et al. (2008) RNA backbone: consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). RNA 14: 465–481. doi:10.1261/rna.657708.

102. Neidle S (1999) Oxford Handbook of Nucleic Acid Structure. Oxford University Press.

103. Sweeney BA, Roy P, Leontis NB (2015) An introduction to recurrent nucleotide interactions in RNA. Wiley Interdiscip Rev RNA 6: 17–45. doi:10.1002/wrna.1258.

104. Zirbel CL, Sponer JE, Sponer J, Stombaugh J, Leontis NB (2009) Classification and energetics of the base-phosphate interactions in RNA. Nucleic Acids Res 37: 4898–4918. doi:10.1093/nar/gkp468.

105. Abu Almakarem AS, Petrov AI, Stombaugh J, Zirbel CL, Leontis NB (2012) Comprehensive survey and geometric classification of base triples in RNA structures. Nucleic Acids Res 40: 1407–1423. doi:10.1093/nar/gkr810.

106. Lee JC, Gutell RR (2004) Diversity of base-pair conformations and their occurrence in rRNA structure and RNA structural motifs. J Mol Biol 344: 1225–1249. doi:10.1016/j.jmb.2004.09.072.

107. Leontis NB, Westhof E (2001) Geometric nomenclature and classification of RNA base pairs. RNA 7: 499–512. doi:10.1017/S1355838201002515.

108. Leontis NB, Stombaugh J, Westhof E (2002) The non-Watson-Crick base pairs and their associated isostericity matrices. Nucleic Acids Res 30: 3497–3531. doi:10.1093/nar/gkf481.

109. Leontis NB, Westhof E (2012) RNA 3D structure analysis and prediction. Heidelberg: Springer.

110. Stombaugh J, Zirbel CL, Westhof E, Leontis NB (2009) Frequency and isostericity of RNA base pairs. Nucleic Acids Res 37: 2294–2312. doi:10.1093/nar/gkp011.

111. Nasalean L, Stombaugh J, Zirbel CL, Leontis NB (2009) RNA 3D structural motifs: definition, identification, annotation, and database searching. In: Walter NG, Woodson SA, Batey RT, editors. Non-Protein Coding RNAs. Berlin, Heidelberg: Springer Berlin Heidelberg. pp. 1–26. doi:10.1007/978-3-540-70840-7_1.

112. Moore PB (1999) Structural motifs in RNA. Annu Rev Biochem 68: 287–300. doi:10.1146/annurev.biochem.68.1.287.

113. Leontis NB, Westhof E (2003) Analysis of RNA motifs. Curr Opin Struct Biol 13: 300–308. doi:10.1016/S0959-440X(03)00076-9.

114. Leontis NB, Lescoute A, Westhof E (2006) The building blocks and motifs of RNA architecture. Curr Opin Struct Biol 16: 279–287. doi:10.1016/j.sbi.2006.05.009.

115. Gubaev A, Klostermeier D (2013) RNA Structure and Folding.

116. Hendrix DK, Brenner SE, Holbrook SR (2005) RNA structural motifs: building blocks of a modular biomolecule. Q Rev Biophys 38: 221–243. doi:10.1017/S0033583506004215.

117. Leontis NB, Altman RB, Berman HM, Brenner SE, Brown JW, et al. (2006) The RNA Ontology Consortium: an open invitation to the RNA community. RNA 12: 533–541. doi:10.1261/rna.2343206.

118. Hoehndorf R, Batchelor C, Bittner T, Dumontier M, Eilbeck K, et al. (2011) The RNA Ontology (RNAO): An ontology for integrating RNA sequence and structure data. Appl Ontol 6: 53–89.

119. Petrov AI, Zirbel CL, Leontis NB (2013) Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas. RNA 19: 1327–1340. doi:10.1261/rna.039438.113.

120. Hoeppner MP, Barquist LE, Gardner PP (2014) An introduction to RNA databases. Methods Mol Biol 1097: 107–123. doi:10.1007/978-1-62703-709-9_6.

121. The RNAcentral Consortium (2017) RNAcentral: a comprehensive database of non-coding RNA sequences. Nucleic Acids Res 45: D128–D134. doi:10.1093/nar/gkw1008.

122. Raabe CA, Brosius J (2015) Does every transcript originate from a gene? Ann N Y Acad Sci 1341: 136–148. doi:10.1111/nyas.12741.

123. Brosius J, Raabe CA (2016) What is an RNA? A top layer for RNA classification. RNA Biol 13: 140–144. doi:10.1080/15476286.2015.1128064.

124. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The protein data bank. Nucleic Acids Res 28: 235–242. doi:10.1093/nar/28.1.235.

125. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Brice MD, et al. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 112: 535–542.

126. Berman HM, Olson WK, Beveridge DL, Westbrook J, Gelbin A, et al. (1992) The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. Biophys J 63: 751–759. doi:10.1016/S0006-3495(92)81649-1.

127. Coimbatore Narayanan B, Westbrook J, Ghosh S, Petrov AI, Sweeney B, et al. (2014) The Nucleic Acid Database: new features and capabilities. Nucleic Acids Res 42: D114–22. doi:10.1093/nar/gkt980.

128. Klosterman PS, Tamura M, Holbrook SR, Brenner SE (2002) SCOR: a Structural Classification of RNA database. Nucleic Acids Res 30: 392–394.

129. Tamura M, Hendrix DK, Klosterman PS, Schimmelman NRB, Brenner SE, et al. (2004) SCOR: Structural Classification of RNA, version 2.0. Nucleic Acids Res 32: D182–4. doi:10.1093/nar/gkh080.

130. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536–540. doi:10.1006/jmbi.1995.0159.

131. Andreeva A, Howorth D, Chothia C, Kulesha E, Murzin AG (2014) SCOP2 prototype: a new approach to protein structure mining. Nucleic Acids Res 42: D310–4. doi:10.1093/nar/gkt1242.

132. Stefan LR, Zhang R, Levitan AG, Hendrix DK, Brenner SE, et al. (2006) MeRNA: a database of metal ion binding sites in RNA structures. Nucleic Acids Res 34: D131–4. doi:10.1093/nar/gkj058.

133. Parlea LG, Sweeney BA, Hosseini-Asanjan M, Zirbel CL, Leontis NB (2016) The RNA 3D Motif Atlas: Computational methods for extraction, organization and evaluation of RNA motifs. Methods 103: 99–119. doi:10.1016/j.ymeth.2016.04.025.

134. Sarver M, Zirbel CL, Stombaugh J, Mokdad A, Leontis NB (2008) FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. J Math Biol 56: 215–252. doi:10.1007/s00285-007-0110-x.

135. Appasamy SD, Hamdani HY, Ramlan EI, Firdaus-Raih M (2016) InterRNA: a database of base interactions in RNA structures. Nucleic Acids Res 44: D266–71. doi:10.1093/nar/gkv1186.

136. Andreini C, Cavallaro G, Lorenzini S, Rosato A (2013) MetalPDB: a database of metal sites in biological macromolecular structures. Nucleic Acids Res 41: D312–9. doi:10.1093/nar/gks1063.

137. Putignano V, Rosato A, Banci L, Andreini C (2018) MetalPDB in 2018: a database of metal sites in biological macromolecular structures. Nucleic Acids Res 46: D459–D464. doi:10.1093/nar/gkx989.

138. Schnabl J, Suter P, Sigel RKO (2012) MINAS--a database of Metal Ions in Nucleic AcidS. Nucleic Acids Res 40: D434–8. doi:10.1093/nar/gkr920.

139. Taufer M, Licon A, Araiza R, Mireles D, van Batenburg FHD, et al. (2009) PseudoBase++: an extension of PseudoBase for easy searching, formatting and visualization of pseudoknots. Nucleic Acids Res 37: D127–35. doi:10.1093/nar/gkn806.

140. Chojnowski G, Walen T, Bujnicki JM (2014) RNA Bricks--a database of RNA 3D motifs and their interactions. Nucleic Acids Res 42: D123–31. doi:10.1093/nar/gkt1084.

141. Vanegas PL, Hudson GA, Davis AR, Kelly SC, Kirkpatrick CC, et al. (2012) RNA CoSSMos: Characterization of Secondary Structure Motifs--a searchable database of secondary structure motifs in RNA three-dimensional structures. Nucleic Acids Res 40: D439–44. doi:10.1093/nar/gkr943.

142. Popenda M, Blazewicz M, Szachniuk M, Adamiak RW (2008) RNA FRABASE version 1.0: an engine with a database to search for the three-dimensional fragments within RNA structures. Nucleic Acids Res 36: D386–91. doi:10.1093/nar/gkm786.

143. Popenda M, Szachniuk M, Blazewicz M, Wasik S, Burke EK, et al. (2010) RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures. BMC Bioinformatics 11: 231. doi:10.1186/1471-2105-11-231.

144. Andronescu M, Bereg V, Hoos HH, Condon A (2008) RNA STRAND: the RNA secondary structure and statistical analysis database. BMC Bioinformatics 9: 340. doi:10.1186/1471-2105-9-340.

145. Bindewald E, Hayes R, Yingling YG, Kasprzak W, Shapiro BA (2008) RNAJunction: a database of RNA junctions and kissing loops for three-dimensional structural analysis and nanodesign. Nucleic Acids Res 36: D392–7. doi:10.1093/nar/gkm842.

146. Ismer J, Rose AS, Tiemann JKS, Goede A, Rother K, et al. (2013) Voronoia4RNA--a database of atomic packing densities of RNA structures and their complexes. Nucleic Acids Res 41: D280–4. doi:10.1093/nar/gks1061.

147. Boccaletto P, Magnus M, Almeida C, Zyla A, Astha A, et al. (2018) RNArchitecture: a database and a classification system of RNA families, with a focus on structural information. Nucleic Acids Res 46: D202–D205. doi:10.1093/nar/gkx966.

148. Baulin E, Yacovlev V, Khachko D, Spirin S, Roytberg M (2016) URS DataBase: universe of RNA structures and their motifs. Database (Oxford) 2016. doi:10.1093/database/baw085.

149. Brown JW (1999) The ribonuclease P database. Nucleic Acids Res 27: 314.

150. Ray SS, Halder S, Kaypee S, Bhattacharyya D (2012) HD-RNAS: An Automated Hierarchical Database of RNA Structures. Front Genet 3: 59. doi:10.3389/fgene.2012.00059.

151. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR (2003) Rfam: an RNA family database. Nucleic Acids Res 31: 439–441. doi:10.1093/nar/gkg006.

152. Kalvari I, Argasinska J, Quinones-Olvera N, Nawrocki EP, Rivas E, et al. (2018) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res 46: D335–D342. doi:10.1093/nar/gkx1038.

153. Sonnhammer EL, Eddy SR, Durbin R (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 28: 405–420. doi:10.1002/(SICI)1097-0134(199707)28:3<405::AID-PROT10>3.0.CO;2-L.

154. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, et al. (2016) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44: D279–85. doi:10.1093/nar/gkv1344.

155. Monzon AM, Juritz E, Fornasari MS, Parisi G (2013) CoDNaS: a database of conformational diversity in the native state of proteins. Bioinformatics 29: 2512–2514. doi:10.1093/bioinformatics/btt405.

156. Monzon AM, Rohr CO, Fornasari MS, Parisi G (2016) CoDNaS 2.0: a comprehensive database of protein conformational diversity in the native state. Database (Oxford) 2016. doi:10.1093/database/baw038.

157. Mackay JP, Landsberg MJ, Whitten AE, Bond CS (2017) Whaddaya know: A guide to uncertainty and subjectivity in structural biology. Trends Biochem Sci 42: 155–167. doi:10.1016/j.tibs.2016.11.002.

158. Bottaro S, Lindorff-Larsen K (2018) Biophysical experiments and biomolecular simulations: A perfect match? Science 361: 355–360. doi:10.1126/science.aat4010.

159. Dans PD, Gallego D, Balaceanu A, Darré L, Gómez H, et al. (2018) Modeling, simulations, and bioinformatics at the service of RNA structure. Chem. doi:10.1016/j.chempr.2018.09.015.

160. Bottaro S, Bussi G, Kennedy SD, Turner DH, Lindorff-Larsen K (2018) Conformational ensembles of RNA oligonucleotides from integrating NMR and molecular simulations. Sci Adv 4: eaar8521. doi:10.1126/sciadv.aar8521.

161. Burra PV, Zhang Y, Godzik A, Stec B (2009) Global distribution of conformational states derived from redundant models in the PDB points to non-uniqueness of the protein structure. Proc Natl Acad Sci USA 106: 10505–10510. doi:10.1073/pnas.0812152106.

162. Juritz EI, Alberti SF, Parisi GD (2011) PCDB: a database of protein conformational diversity. Nucleic Acids Res 39: D475–9. doi:10.1093/nar/gkq1181.

163. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658–1659. doi:10.1093/bioinformatics/btl158.

164. Huang Y, Niu B, Gao Y, Fu L, Li W (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 26: 680–682. doi:10.1093/bioinformatics/btq003.

165. Capriotti E, Marti-Renom MA (2008) RNA structure alignment by a unit-vector approach. Bioinformatics 24: i112–8. doi:10.1093/bioinformatics/btn288.

166. Capriotti E, Marti-Renom MA (2009) SARA: a server for function annotation of RNA structures. Nucleic Acids Res 37: W260–5. doi:10.1093/nar/gkp433.

167. Ortiz AR, Strauss CEM, Olmea O (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci 11: 2606–2621. doi:10.1110/ps.0215902.

168. Lu X-J, Olson WK (2008) 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. Nat Protoc 3: 1213–1227. doi:10.1038/nprot.2008.104.

169. Siew N, Elofsson A, Rychlewski L, Fischer D (2000) MaxSub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics 16: 776–785.

170. Bauer RA, Rother K, Moor P, Reinert K, Steinke T, et al. (2009) Fast Structural Alignment of Biomolecules Using a Hash Table, N-Grams and String Descriptors. Algorithms 2: 692–709. doi:10.3390/a2020692.

171. Wiegels T, Bienert S, Torda AE (2013) Fast alignment and comparison of RNA structures. Bioinformatics 29: 588–596. doi:10.1093/bioinformatics/btt006.

172. Torda AE, Procter JB, Huber T (2004) Wurst: a protein threading server with a structural scoring function, sequence profiles and optimized substitution matrices. Nucleic Acids Res 32: W532–5. doi:10.1093/nar/gkh357.

173. Dror O, Nussinov R, Wolfson H (2005) ARTS: alignment of RNA tertiary structures. Bioinformatics 21 Suppl 2: ii47–53. doi:10.1093/bioinformatics/bti1108.

174. Dror O, Nussinov R, Wolfson HJ (2006) The ARTS web server for aligning RNA tertiary structures. Nucleic Acids Res 34: W412–5. doi:10.1093/nar/gkl312.

175. Rahrig RR, Leontis NB, Zirbel CL (2010) R3D Align: global pairwise alignment of RNA 3D structures using local superpositions. Bioinformatics 26: 2689–2697. doi:10.1093/bioinformatics/btq506.

176. Kirillova S, Tosatto SCE, Carugo O (2010) FRASS: the web-server for RNA structural comparison. BMC Bioinformatics 11: 327. doi:10.1186/1471-2105-11-327.

177. Rogen P, Fain B (2003) Automatic classification of protein structure by using Gauss integrals. Proc Natl Acad Sci USA 100: 119–124. doi:10.1073/pnas.2636460100.

178. Hoksza D, Svozil D (2012) Efficient RNA pairwise structure comparison by SETTER method. Bioinformatics 28: 1858–1864. doi:10.1093/bioinformatics/bts301.

179. Cech P, Svozil D, Hoksza D (2012) SETTER: web server for RNA structure comparison. Nucleic Acids Res 40: W42–8. doi:10.1093/nar/gks560.

180. Hoksza D, Svozil D (2015) Multiple 3D RNA structure superposition using neighbor joining. IEEE/ACM Trans Comput Biol Bioinform 12: 520–530. doi:10.1109/TCBB.2014.2351810.

181. Čech P, Hoksza D, Svozil D (2015) MultiSETTER: web server for multiple RNA structure comparison. BMC Bioinformatics 16: 253. doi:10.1186/s12859-015-0696-8.

182. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680. doi:10.1093/nar/22.22.4673.

183. Laborde J, Robinson D, Srivastava A, Klassen E, Zhang J (2013) RNA global alignment in the joint sequence-structure space using elastic shape analysis. Nucleic Acids Res 41: e114. doi:10.1093/nar/gkt187.

184. He G, Steppi A, Laborde J, Srivastava A, Zhao P, et al. (2014) RASS: a web server for RNA alignment in the joint sequence-structure space. Nucleic Acids Res 42: W377–81. doi:10.1093/nar/gku429.

185. Kemena C, Bussotti G, Capriotti E, Marti-Renom MA, Notredame C (2013) Using tertiary structure for the computation of highly accurate multiple RNA alignments with the SARA-Coffee package. Bioinformatics 29: 1112–1119. doi:10.1093/bioinformatics/btt096.

186. Di Tommaso P, Bussotti G, Kemena C, Capriotti E, Chatzou M, et al. (2014) SARA-Coffee web server, a tool for the computation of RNA sequence and structure multiple alignments. Nucleic Acids Res 42: W356–60. doi:10.1093/nar/gku459.

187. Wilm A, Higgins DG, Notredame C (2008) R-Coffee: a method for multiple alignment of non-coding RNA. Nucleic Acids Res 36: e52. doi:10.1093/nar/gkn174.

188. Ge P, Zhang S (2015) STAR3D: a stack-based RNA 3D structural alignment tool. Nucleic Acids Res 43: e137. doi:10.1093/nar/gkv697.

189. Gendron P, Lemieux S, Major F (2001) Quantitative analysis of nucleic acid three-dimensional structures. J Mol Biol 308: 919–936. doi:10.1006/jmbi.2001.4626.

190. Lemieux S, Major F (2002) RNA canonical and non-canonical base pairing types: a recognition method and complete repertoire. Nucleic Acids Res 30: 4250–4263. doi:10.1093/nar/gkf540.

191. Piatkowski P, Jablonska J, Zyla A, Niedzialek D, Matelska D, et al. (2017) SupeRNAlign: a new tool for flexible superposition of homologous RNA structures and inference of accurate structure-based sequence alignments. Nucleic Acids Res 45: e150. doi:10.1093/nar/gkx631.

192. Ferrè F, Ponty Y, Lorenz WA, Clote P (2007) DIAL: a web server for the pairwise alignment of two RNA three-dimensional structures using nucleotide, dihedral

angle and base-pairing similarities. Nucleic Acids Res 35: W659–68. doi:10.1093/nar/gkm334.

193. Duarte CM, Wadley LM, Pyle AM (2003) RNA structure comparison, motif search and discovery using a reduced representation of RNA conformational space. Nucleic Acids Res 31: 4755–4761. doi:10.1093/nar/gkg682.

194. Chang Y-F, Huang Y-L, Lu CL (2008) SARSA: a web tool for structural alignment of RNA using a structural alphabet. Nucleic Acids Res 36: W19–24. doi:10.1093/nar/gkn327.

195. Wang C-W, Chen K-T, Lu CL (2010) iPARTS: an improved tool of pairwise alignment of RNA tertiary structures. Nucleic Acids Res 38: W340–7. doi:10.1093/nar/gkq483.

196. Nguyen MN, Verma C (2015) Rclick: a web server for comparison of RNA 3D structures. Bioinformatics 31: 966–968. doi:10.1093/bioinformatics/btu752.

197. Nguyen MN, Sim AYL, Wan Y, Madhusudhan MS, Verma C (2017) Topology independent comparison of RNA 3D structures using the CLICK algorithm. Nucleic Acids Res 45: e5. doi:10.1093/nar/gkw819.

198. Nguyen MN, Tan KP, Madhusudhan MS (2011) CLICK--topology-independent comparison of biomolecular 3D structures. Nucleic Acids Res 39: W24–8. doi:10.1093/nar/gkr393.

199. Yang C-H, Shih C-T, Chen K-T, Lee P-H, Tsai P-H, et al. (2016) iPARTS2: an improved tool for pairwise alignment of RNA tertiary structures, version 2. Nucleic Acids Res 44: W328–32. doi:10.1093/nar/gkw412.

200. Holzhauser E, Ge P, Zhang S (2016) WebSTAR3D: a web server for RNA 3D structural alignment. Bioinformatics 32: 3673–3675. doi:10.1093/bioinformatics/btw502.

201. Keating KS, Pyle AM (2010) Semiautomated model building for RNA crystallography using a directed rotameric approach. Proc Natl Acad Sci USA 107: 8177–8182. doi:10.1073/pnas.0911888107.

202. Touw WG, Joosten RP, Vriend G (2016) New Biological Insights from Better Structure Models. J Mol Biol 428: 1375–1393. doi:10.1016/j.jmb.2016.02.002.

203. Richardson JS, Williams CJ, Hintze BJ, Chen VB, Prisant MG, et al. (2018) Model validation: local diagnosis, correction and when to quit. Acta Crystallogr D Struct Biol 74: 132–142. doi:10.1107/S2059798317009834.

204. Smart OS, Horský V, Gore S, Svobodová Vařeková R, Bendová V, et al. (2018) Worldwide Protein Data Bank validation information: usage and trends. Acta Crystallogr D Struct Biol 74: 237–244. doi:10.1107/S2059798318003303.

205. Chou F-C, Sripakdeevong P, Dibrov SM, Hermann T, Das R (2013) Correcting pervasive errors in RNA crystallography through enumerative structure prediction. Nat Methods 10: 74–76. doi:10.1038/nmeth.2262.

206. Chou FC, Echols N, Terwilliger TC, Das R (2016) RNA structure refinement using the ERRASER-phenix pipeline. Nucleic Acid Crystallography: 269.

207. Jain S, Richardson DC, Richardson JS (2015) Computational methods for RNA structure validation and improvement. Meth Enzymol 558: 181–212. doi:10.1016/bs.mie.2015.01.007.