

**UNIVERSIDAD COMPLUTENSE DE MADRID**

**FACULTAD DE CIENCIAS FÍSICAS**

DEPARTAMENTO DE ESTRUCTURA DE LA MATERIA, FÍSICA TÉRMICA Y  
ELECTRÓNICA



**TRABAJO DE FIN DE GRADO**

Código de TFG: ETE45

Física Biológica

Biological Physics

Supervisor/es: Francisco J. Cao García, Juan Pedro García Villaluenga

**Pedro Lalande Delgado**

Grado en Física

Curso académico 2023-2024

Convocatoria Ordinaria

# Modelado tridimensional del ARN mediante Random Walks

## Resumen

El estudio de la estructura secundaria del ARN ofrece la posibilidad de una mejor comprensión sobre los procesos biológicos de transcripción y traducción. La capacidad de predecir estructuras en base a una secuencia dada abriría la puerta a poder diseñar moléculas de ARN a nuestro interés, ya sea para modificar el comportamiento del genoma de un ser vivo, como para que cumpla una función catalítica en alguna reacción de interés. Con este objetivo, existen una serie de modelos de predicción estructural del ARN, con especial énfasis en la estructura secundaria de esta molécula. La gran mayoría de modelos tratan de predecir la estructura de ARN de menor energía libre, suponiendo que será la más probable en el equilibrio. En este trabajo, se estudia el modelo Random Walk como una alternativa conceptualmente más sencilla, aunque en la práctica muy inefficiente. Tras comparar con el modelo *ViennaRNA* y con el modelo *SeqFold*, se ha determinado que las cadenas calculadas por Random Walks suelen no ser las más estables, y por lo tanto no se ajustan a la realidad. Se plantea como una opción posible para un estudio posterior cambiar la distribución con la que se eligen las direcciones de los segmentos, para así tener en cuenta como contribuye un segmento a la energía de conformación.

## Abstract

The study of RNA secondary structure offers the possibility to come to a better understanding of the biological processes of transcription and translation. The ability to predict structures based on a given sequence would open the door to being able to design RNA molecules, either to modify the behaviour of the genome of a living being or to perform a catalytic function in a given reaction. To this end, there are a number of RNA structural prediction models, which focus on predicting the secondary structure of this molecule. The vast majority of models try to predict the RNA structure with minimizing the chain free energy, assuming that it will be the most likely state at equilibrium. In this paper, the Random Walk model is studied as a conceptually simpler, but in practice very inefficient, alternative. After comparison with the *ViennaRNA* model and the *SeqFold* model, it has been determined that the chains calculated by Random Walks are often not the most stable, and therefore do not fit to the biological description. A possible option for further study is to change the distribution with which the segment directions are chosen, in order to take into account how a segment contributes to the conformational energy.

---



**Declaración Responsable sobre Autoría y Uso Ético de**  
**Herramientas de Inteligencia Artificial (IA)**

Yo, Lalanda Delgado, Pedro

Con DNI/NIE/PASAPORTE: 47533870T

declaro de manera responsable que el/la presente:

- Trabajo de Fin de Grado (TFG)
- Trabajo de Fin de Máster (TFM)
- Tesis Doctoral

Titulado/a

Física Biológica

es el resultado de mi trabajo intelectual personal y creativo, y ha sido elaborado de acuerdo con los principios éticos y las normas de integridad vigentes en la comunidad académica y, más específicamente, en la Universidad Complutense de Madrid.

Soy, pues, autor del material aquí incluido y, cuando no ha sido así y he tomado el material de otra fuente, lo he citado o bien he declarado su procedencia de forma clara -incluidas, en su caso, herramientas de inteligencia artificial-. Las ideas y aportaciones principales incluidas en este trabajo, y que acreditan la adquisición de competencias, son mías y no proceden de otras fuentes o han sido reescritas usando material de otras fuentes.

Asimismo, aseguro que los datos y recursos utilizados son legítimos, verificables y han sido obtenidos de fuentes confiables y autorizadas. Además, he tomado medidas para garantizar la confidencialidad y privacidad de los datos utilizados, evitando cualquier tipo de sesgo o discriminación injusta en el tratamiento de la información.

En Madrid a 6 de mayo de 2024,



**FIRMA**

# Índice

<b>1 Introducción</b>	<b>2</b>
1.1 Polímeros en las Células . . . . .	2
1.2 Estructura del ARN . . . . .	2
1.3 Predicción de la Estructura Secundaria del ARN . . . . .	4
1.3.1 Modelos de Programación Dinámica . . . . .	4
1.3.2 Otros Modelos . . . . .	6
1.4 Modelos Random Walk . . . . .	6
1.4.1 Modelo Freely Jointed Chain del ADN . . . . .	6
<b>2 Objetivos</b>	<b>8</b>
<b>3 Metodología</b>	<b>8</b>
3.1 Simulación del ARN con el Modelo Freely Jointed Chain . . . . .	9
3.2 Modelo Freely Jointed Chain Promediado . . . . .	10
3.3 Otros Programas . . . . .	10
3.4 Conformaciones Aleatorias . . . . .	10
<b>4 Resultados</b>	<b>10</b>
<b>5 Conclusión</b>	<b>13</b>
<b>6 Bibliografía</b>	<b>15</b>
<b>7 Apéndice I: Gráficas de los Transcritos</b>	<b>16</b>

# Modelado tridimensional del ARN mediante Random Walks

## 1. Introducción

### 1.1. Polímeros en las Células

Según el dogma central de la biología molecular, la información genética dentro de cualquier célula se transmite del ADN al ARN, por medio de la transcripción, y de este último se traduce a proteína, que es la molécula que en última instancia la célula usa para realizar una acción dada [2].

Por lo tanto, para el buen funcionamiento de cualquier célula viva, es necesaria la utilización de estos tres polímeros fundamentales,

1. El ADN, molécula con dos cadenas en forma de doble hélice, que codifica las instrucciones genéticas necesarias para el buen funcionamiento de la maquinaria molecular de la célula. El ADN está compuesto por monómeros llamados nucleótidos, compuestos a su vez por bases nitrogenadas (A, T, C, G) complementarias dos a dos mediante enlaces de hidrógeno ( $A=T$ ,  $C\equiv G$ ), y por una desoxiribosa adherida a un grupo fosfato. La cadenas de nucleótidos, i.e., las hélices, se mantienen unidas por enlaces covalentes entre los grupos fosfatos adheridos a las desoxiribosas, y por su estructura complementaria la célula es capaz de replicar las cadenas de nucleótidos con facilidad.
2. El ARN, similar al ADN en su estructura en el hecho de que también esta compuesta por ácidos nucleicos (A, U, C, G), aunque en el ARN se sustituye a la timina (T) por uracilo (U). También difiere al ADN en que las hélices en lugar de estar formadas por desoxiribosas, están formadas por ribosas. Además, el ARN no suele encontrarse enlazado a otra cadena complementaria, encontrándose en su lugar en cadenas simples. Es interesante destacar que cada vez se descubren más casos para los que la célula utiliza el ARN, siendo también útil, por ejemplo, en reacciones catalíticas (ribozimas). En general, se clasifican las moléculas de ARN por función, por tamaño y por el lugar en el que se encuentran dentro de la célula (mRNA, tRNA, rRNA, miRNA).
3. Las proteínas, que son las moléculas constituidas por monómeros llamados aminoácidos, que la célula usa tanto de manera estructural como para realizar una acción dada, a modo de catalizador en reacciones químicas (enzimas).

### 1.2. Estructura del ARN

Como he comentado previamente, este estudio está enfocado a estudiar y predecir la estructura que adopta la molécula de ARN, que difiere de la del ADN al estar compuesta esta molécula por una sola hélice en lugar de dos, lo que permite al ARN doblarse sobre sí mismo, y a sus nucleótidos enlazarse entre sí.

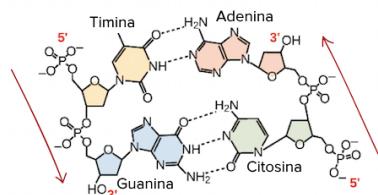


Figura 1: Cadena de ADN, Khan Academy [1].

Cabe mencionar que la exposición teórica expuesta en esta sección se basa en el trabajo realizado por Paul G. Higgs (2000) [3], quién aglutinó en la revisión citada los diferentes modelos de predicción de estructura del ARN. Los modelos que usaré de contraste con el modelo Random Walk están basados en los explicados en su trabajo.

El ARN está compuesto por cuatro ácidos nucleicos (adenina, citosina, guanina, uracilo), complementarios dos a dos ( $A=U$ ,  $C\equiv G$ ), como en el ADN. Cabe mencionar también que el uracilo puede enlazarse también a la guanina ( $U=G$ ), cosa que habrá que tener en cuenta más adelante.

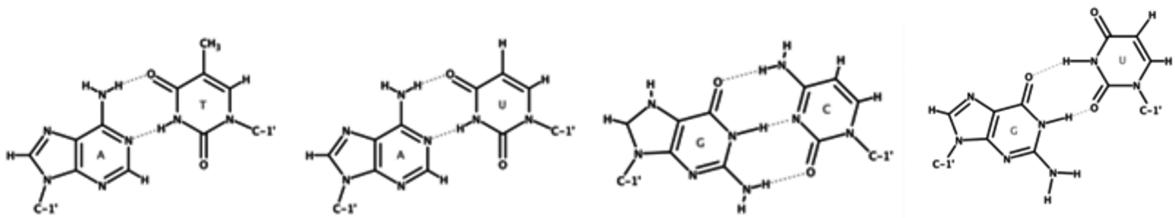


Figura 2: Pares de bases complementarias del ARN, Wikipedia [4].

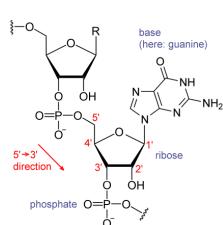


Figura 3: Estructura química del ARN, Wikipedia [5].

Tanto para la cadena de ADN como para la de ARN se pueden definir direcciones en función de los átomos a los que se enlanzan las ribosas y las desoxiribosas, también conocidas como pentosas. Se denomina al extremo de la cadena en la que el tercer carbono de las pentosas queda sin enlazar extremo 3' de la cadena. Por analogía, al extremo contrario, en el que el quinto carbono de las pentosas queda sin enlazar se le denomina extremo 5'.

En lo que a la estructura tridimensional que el ARN presenta, se puede estudiar en varias partes, a diferencia de las proteínas:

- La estructura primaria del ARN es la estructura lineal de la molécula, i.e., la secuencia de nucleótidos que componen a la molécula.
- La estructura secundaria del ARN se refiere a los enlaces que los diferentes nucleótidos de la cadena presentan entre sí, y como esos enlaces fuerzan a la cadena a doblarse sobre sí misma. El ARN al doblarse sobre sí mismo forma estructuras en forma de hélice cuando dos segmentos complementarios de la cadena se enlanzan, como las que se forman en el ADN. Principalmente, la estructura secundaria se describe prediciendo las hélices que se forman y las secciones de nucleótidos desparejadas que no forman hélices (hairpins, loops, junctions...).
- La estructura terciaria del ARN se refiere a interacciones menos relevantes entre los elementos de la cadena de ARN, como pueden ser interacciones entre tres nucleótidos, en lugar de solo dos; interacciones entre un nucleótido y la cadena de ribosas o interacciones

entre secciones complementarias de nucleótidos desemparejados.

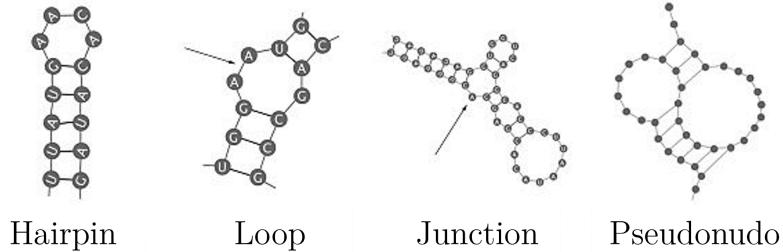


Figura 4: Motivos de la estructura secundaria del ARN, Wikipedia [6].

El principal objetivo de cualquier modelo computacional dedicado a predecir la estructura tridimensional del ARN es predecir correctamente su estructura secundaria, bajo la suposición de que la estructura terciaria se forma tras la conformación de la estructura secundaria, por ser interacciones con menos probabilidad de ocurrir y menos estables. A continuación expongo los diferentes modelos computacionales que existen para predecir dicha estructura.

### 1.3. Predicción de la Estructura Secundaria del ARN

La mayoría de modelos computacionales tratan de encontrar la estructura del ARN que minimice la energía libre de la molécula. Para ello, se parte de unas reglas básicas: suponiendo que la cadena esta formada por  $N$  nucleótidos:

- Un par de bases complementarias  $i$  y  $j$  ha de estar al menos a cuatro nucleótidos de distancia,  $|i - j| \geq 4$ , por ser la cadena demasiado rígida para permitir enlaces entre nucleótidos más próximos entre sí.
- Dadas dos parejas de bases complementarias  $i - j$  y  $k - l$ , tomamos las parejas como compatibles, i.e., que permitimos que ambas formen parte de la estructura a predecir, si solapan o una está contenida dentro de la otra ( $i < j < k < l$  ó  $i < k < l < j$ ). Descartamos aquellas parejas que no cumplen estas condiciones ( $i < k < j < l$ , pseudonudos), que corresponderían a interacciones de la estructura terciaria entre segmentos lejanos de la cadena, comentados en la sección anterior.
- La energía libre de un nucleótido libre es mayor que la energía libre de un nucleotido enlazado. Por lo tanto, suponemos que la molécula de ARN tenderá en todo caso a doblarse sobre sí misma para enlazar la mayor cantidad de pares de nucleotidos que le sea posible.

Con estas reglas básicas se desarrollan los modelos a explicados continuación.

#### 1.3.1. Modelos de Programación Dinámica

Los modelos más comunes con diferencia son los de programación dinámica. Estos modelos se basan en la recursión para optimizar la rapidez con la que encuentran estructuras candidatas a ser la más estable, en un tiempo que escala de acuerdo a  $O(N^3)$ .

El modelo más simple de esta categoría es el descrito por R. Nussinov, et. al. (1980) [7]. Asignamos

una energía  $\epsilon_{ij} = -1 \text{ u.a.}$  (unidades arbitrarias) a cada par de bases enlazadas y, en caso de no estar enlazadas, asignamos una energía a la base de  $\epsilon_{ij} = +\infty$ , siendo  $\epsilon_{ij}$  la energía de un par de bases dado. Se pretende obtener la energía mínima del segmento entre dos bases dadas  $E_{ij}$ . Supóngase que la última base,  $j$ , esta enlazada con una base  $i < k < j$ . Podemos descomponer la energía  $E_{ij}$  tal que,

$$E_{ij} = E_{i,k-1} + E_{k+1,j-1} + \epsilon_{kj} \quad (1)$$

Si la base  $j$  resulta no estar enlazada con ningún nucleotido, se puede intuir de la ecuación (1) que  $E_{ij} = E_{i,j-1}$ . Partiendo de esta ecuación, se puede encontrar la energía mínima del segmento, que corresponderá con la del enlace que lo minimice, siendo esta

$$E_{ij} = \min(E_{i,j-1}, \min_{i \leq k \leq j-4}(E_{i,k-1} + E_{k+1,j-1} + \epsilon_{kj})) \quad (2)$$

Partiendo de la ecuación (2), con un algoritmo de recursión que vaya descomponiendo el segmento a considerar en segmentos más pequeños y vaya encontrando para esos segmentos más pequeños sus energías mínimas, se puede encontrar la energía mínima de la cadena  $E_{1N}$ , así como la conformación de la cadena de ARN que la minimiza.

Así mismo, se puede calcular la función de partición del segmento  $Z_{ij}$ , así como la probabilidad de que dos nucleótidos dados estén enlazados dentro del segmento,  $p_{ij}$ , utilizando el método recursivo [8], tal que,

$$Z_{ij} = Z_{i,j-1} + \sum_{l=i}^{j-4} \left( Z_{i,l-1} Z_{l+1,j-1} \exp\left(-\frac{\epsilon_{lj}}{kT}\right) \right) \quad (3)$$

$$p_{ij} = \frac{Z_{1,i-1} Z_{j+1,N} \exp\left(-\frac{\epsilon_{ij}}{kT}\right)}{Z_{1N}} \quad (4)$$

Con la función de partición podemos obtener una expresión para la energía libre de la cadena:

$$G = -kT \log Z_{ij} \Rightarrow g = \frac{G}{kT} = -\log Z_{ij} \quad (5)$$

A partir de la matriz de probabilidad, podemos definir la *entropía posicional* de un nucleótido dado,

$$S_i = - \sum_i p_{ij} \log p_{ij} \quad (6)$$

Esta cantidad puede resultar informativa a la hora de valorar la fiabilidad de la predicción realizada con el modelo de programación dinámica.

Lo que todos los modelos de programación dinámica tienen en común es el procedimiento recursivo aquí expuesto. Las diferencias entre ellos son los parámetros termodinámicos que se tienen en cuenta en los modelos, como sería en el caso del modelo expuesto  $\epsilon_{ij}$ . Estos parámetros se pueden ajustar para reflejar mejor los resultados obtenidos en experimentos termodinámicos realizados con cadenas de ARN.

Usaré en este trabajo los modelos de programación dinámica como referencia para comparar los resultados obtenidos con el modelo Random Walk.

### 1.3.2. Otros Modelos

Cabe mencionar la existencia de otros modelos de predicción de estructura del ARN [3], aunque no haré uso de ellos en este trabajo.

- *Algoritmos Kinetic Folding*: Estos modelos parten del hecho de que las energías de enlace de las hélices encontradas en el ARN suelen ser mucho mayores a la energía de las fluctuaciones térmicas,  $kT$ , y por lo tanto una vez formadas es extremadamente difícil que estas hélices se descompongan. Por lo tanto, se pueden encontrar las estructuras de mínima energía libre montando hélices en base a la secuencia del ARN de manera secuencial.
- *Algoritmos Genéticos*: De la misma forma, se puede resolver el problema mediante un algoritmo que encuentre la estructura del ARN con un proceso que imite la evolución biológica: creando una población de conformaciones de cadenas heterogéneas, seleccionando las que minimicen la energía libre y mezclando su estructura para crear una nueva población (mutación y selección).
- *Métodos Comparativos*: Es posible deducir la estructura que tomará el ARN a partir del estudio de cadenas de ARN homólogas a la estudiada en otras especies.

## 1.4. Modelos Random Walk

Los modelos random walk son una clase de modelos basados en la idea de recorrer un espacio matemático dado mediante una sucesión de pasos tomados de manera aleatoria [9] [10].

Su aplicación se extiende a múltiples disciplinas, incluyendo entre ellas a muchas originarias de la física, como pueden ser la cristalografía o la física de fluidos [11], así como a disciplinas ajena al estudio del mundo natural, como puede ser el estudio de los mercados financieros [12].

El modelo que en este trabajo se estudia es un modelo random walk, aplicado al estudio de la estructura del ADN, conocido como el modelo freely jointed chain. En este trabajo se extiende dicho estudio al estudio del ARN. A continuación expongo el modelo freely jointed chain del ADN, que es la base de este trabajo. Se puede encontrar una explicación en detalle del modelo en R. Phillips, et. al. (2013) [13].

### 1.4.1. Modelo Freely Jointed Chain del ADN

Supongamos que una cadena de ADN se puede dividir en segmentos rígidos de una longitud dada,  $a$ , a los que denominaremos segmentos de Kuhn. Los segmentos de Kuhn están conectados entre sí por uniones flexibles. Queremos encontrar una distribución de probabilidad  $p(\vec{R}, N)$  que nos

indique con que probabilidad una cadena de  $N$  nucleótidos tenga uno de sus extremos situados en la posición  $\vec{R}$ , suponiendo que la posición del otro extremo es el origen.

Para encontrar esta distribución, vamos a limitarnos primero al caso unidimensional,  $p(x, N)$ . En el caso unidimensional, los segmentos consecutivos solo podrán estar dirigidos con respecto a los segmentos previos en dos direcciones: hacia la izquierda o hacia la derecha. Si asignamos la misma probabilidad de que esto ocurra a las dos direcciones, i.e.,  $p_l = p_r = 1/2$ , podemos obtener la probabilidad de dar un total de  $n_r$  pasos a la derecha del siguiente modo,

$$p(n_r, N) = \binom{N}{n_r} \left(\frac{1}{2}\right)^N = \frac{N!}{n_r!(N - n_r)!} \left(\frac{1}{2}\right)^N \quad (7)$$

Podemos observar pues que la cantidad de pasos que se toma en una de las dos direcciones se ajusta a la distribución binomial. Teniendo en cuenta que  $R = (n_r - n_l)a$ , siendo  $n_l$  el número de pasos dados a la izquierda, y que  $n_r + n_l = N$ , obtenemos,

$$p(R, N) = \frac{N!}{\left(\frac{N}{2} + \frac{R}{2a}\right)!\left(\frac{N}{2} - \frac{R}{2a}\right)!} \left(\frac{1}{2}\right)^N \quad (8)$$

Para el caso límite en el que el tamaño de la cadena bajo estudio es mucho menor que su longitud al estar totalmente estirada,  $R \ll Na$ , podemos simplificar la ecuación (8) mediante la aproximación de Stirling ( $\ln n! \approx n \ln n - n + (1/2) \ln(2\pi n)$ , para  $n \gg 1$ ) y la aproximación de Taylor ( $\ln(1 + x) \approx x - x^2/2$ , para  $x \ll 1$ ), para obtener la siguiente distribución gaussiana,

$$p(R, N) = \frac{1}{\sqrt{2\pi Na^2}} e^{-\frac{R^2}{2Na^2}} \quad (9)$$

Podemos extender la distribución obtenida al caso tridimensional teniendo en cuenta que  $R^2 = |\vec{R}|^2$  y recalculando los valores para normalizar la distribución,

$$p(\vec{R}, N) = \left(\frac{3}{2\pi Na^2}\right)^{\frac{3}{2}} e^{-\frac{3R^2}{2Na^2}} \quad (10)$$

Con esta distribución en mente, vamos a calcular la energía libre de la cadena. Para ello, supondremos que la cadena está siendo sometida a una fuerza  $\vec{f} = f\hat{k}$  que la deforma. La energía de la cadena al ser sometida a dicha fuerza es  $U = -\vec{f} \cdot \vec{R}_i = -fa \cos \theta_i$ . Con esta expresión para la energía de la cadena, calculamos la función de partición,

$$Z = \left[ \int_0^{2\pi} d\varphi \int_0^\pi \exp\left(\frac{-fa \cos \theta}{kT}\right) \sin \theta d\theta \right]^N = \left[ 4\pi \frac{kT}{fa} \sinh\left(\frac{fa}{kT}\right) \right]^N \quad (11)$$

De la función de partición, podemos calcular la energía libre de la cadena,

$$g = \frac{G}{kT} = -\log Z = -N \left[ \log \left( 4\pi \sinh \left( \frac{fa}{kT} \right) \right) - \log \left( \frac{kT}{fa} \right) \right] \quad (12)$$

Podemos también calcular la extensión de la cadena, derivando la expresión de la energía libre de la cadena con respecto a la fuerza empleada,

$$L = -\frac{\partial G}{\partial f} = \frac{R}{Na} = Na \left[ \coth \left( \frac{fa}{kT} \right) - \frac{kT}{fa} \right] = \mathcal{L} \left( \frac{fa}{kT} \right) \quad (13)$$

Si combinamos la ecuación (12) con la ecuación (13), obtendremos la expresión de la energía libre de la cadena, en función de su extensión,

$$g(R, N) = \frac{G}{kT} = -Nh \left( \mathcal{L}^{-1} \left( \frac{R}{Na} \right) \right) \quad (14)$$

$$h(x) = \log(4\pi \sinh x) - \log x \quad (15)$$

$$\mathcal{L}^{-1}(x) = \frac{x(3 - 1.00651x^2 - 0.962251x^4 + 1.47353x^6 - 0.46953x^8)}{(1-x)(1+1.01524x)} \quad (16)$$

donde la expresión (16) es una aproximación de la función inversa de la función de Langevin,  $\mathcal{L}(x)$  [14].

## 2. Objetivos

Partiendo de la idea de predecir la estructura secundaria del ARN a partir de la secuencia de la misma, i.e., su estructura primaria, se pretende alcanzar en este trabajo los objetivos enumerados a continuación:

- En primer lugar, se pretende comprobar si el modelo Random Walk es capaz de predecir estructuras tridimensionales del ARN que se ajusten a la descripción biológica, así como comprobar si las estructuras predichas por dicho modelo son biológicamente estables.
- También se pretende comparar el modelo Random Walk con modelos de predicción de estructura secundaria del ARN ya validados, como puede lo son *ViennaRNA* y *SeqFold*.
- Por último, se pretende comprobar si promediando varias conformaciones aleatorias del modelo Random Walk se pueden predecir estructuras más estables de una cadena de ARN dada.

## 3. Metodología

Para comparar los diferentes modelos que he expuesto en la sección 1, haré uso de programas escritos en Python, subidos al repositorio Piquipato/tfgrna en GitHub [15]. A continuación se explican los programas.

### 3.1. Simulación del ARN con el Modelo Freely Jointed Chain

La simulación que he programado en Python requiere de tres parámetros:

- La secuencia de la cadena de ARN para la que se pretende encontrar la conformación tridimensional.
- La longitud de un nucleotido en nanómetros,  $nt$ , cuyo valor por defecto es  $nt = 0.34\text{ nm}$ .
- La longitud de persistencia de la cadena, i.e., la longitud de cadena que se puede tomar como rígida. El valor por defecto es  $\xi_p = 1.5\text{ nt}$ , reflejando que en un hairpin tiene que haber al menos tres nucleotidos libres [3], i.e.,  $|i - j| \geq 4$ .

El resultado que busca el programa es una conformación aleatoria en el espacio de la cadena de ARN que se esté estudiando. Para ello, el programa sigue una serie de pasos.

Para ello, lo primero que hace el programa es calcular la longitud de cada segmento de Kuhn, en función de la longitud de persistencia,  $a = 2\xi_p$ ; la cantidad de nucleotidos en la cadena,  $n_{sgs} = \lceil N/n_{bp} \rceil$ , y el número de nucleotidos que cada segmento contiene,  $n_{bp} = \lceil 2\xi_p/n_t \rceil$ . Con estas variables se segmenta la secuencia de la cadena de ARN en estos segmentos.

A continuación se le asigna a cada segmento una dirección aleatoria en el espacio, usando coordenadas esféricas y decidiendo de manera aleatoria los ángulos,  $(\theta_i, \varphi_i)$ , (con el paquete `random`).

$$\vec{s}_i(\theta_i, \varphi_i) = 2\xi_p \left( \sin \theta_i \cos \varphi_i \hat{i} + \sin \theta_i \sin \varphi_i \hat{j} + \cos \theta_i \hat{k} \right) \quad (17)$$

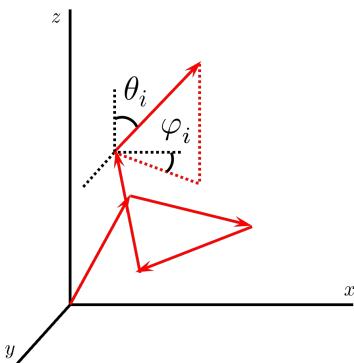


Figura 5: Direcciones aleatorias de los segmentos de ARN.

Empezando en el origen, se calcula la posición inicial de cada segmento,  $\vec{X}_{i,0}$ . Después, para cada nucleotido, se calcula la posición del mismo, usando las direcciones en las que apunta cada uno de los segmentos,  $\vec{s}_i$ , tras concatenarlos, de la siguiente manera.

$$\vec{x}_{ki} = \vec{X}_{i,0} + \frac{k}{n_{bp}} \cdot \vec{s}_i \quad (18)$$

Tras calcular la posición de cada nucleotido de la cadena,  $\vec{x}_{ki}$ , se calculan las potenciales parejas complementarias de nucleotidos en la cadena (siempre y cuando estén a una distancia mayor a  $\xi_p$ ), y si la distancia entre los nucleotidos considerados,  $d = \|\vec{x}_{k_1 i_1} - \vec{x}_{k_2 i_2}\|$  es menor a  $n_t$ , se anota la pareja de nucleotidos como enlazados.

Nótese que con este modelo se admiten los pseudonudos mencionados en la sección 1.3, al no estar restringidos los enlaces por las reglas mencionadas en la misma sección.

### 3.2. Modelo Freely Jointed Chain Promediado

El método expuesto en la anterior sección 3.1 tiene un inconveniente: solo está calculando una conformación aleatoria de las múltiples que pueden ocurrir. Sin embargo, si se repiten múltiples veces las simulaciones descritas, podemos construir una matriz de probabilidades  $p_{ij}$ , que nos de información sobre la conformación de la cadena más probable, que a efectos de este trabajo vamos a asumir que es la más estable. Esta aproximación será más fiable cuantas más simulaciones hagamos para una misma cadena.

Con la matriz de probabilidad, tal y como se describe en el trabajo de Paul G. Higgs (2000) [3], podemos determinar la estructura más probable. Gracias al trabajo de I. L. Hofacker, et.al. (1994) [16], podemos usar el paquete ViennaRNA en Python para calcular cual es la estructura que mayor probabilidad tiene de ser adoptada por una cadena dada, tras realizar múltiples simulaciones Random Walk para la misma cadena.

### 3.3. Otros Programas

Para validar la capacidad de predicción de la estructura del ARN del modelo Random Walk, voy a hacer uso de dos programas adicionales, implementados también en Python:

- *SqFold* [17]: Implementación de un algoritmo de programación dinámica, como los descritos en la sección 1.3.1, con parámetros termodinámicos ajustados.
- *ViennaRNA RNAFold* [18]: Paquete desarrollado a raíz del trabajo de I. L. Hofacker, et.al. (1994) [16], mencionado en la sección 3.2. Introduce la notación `dot_bracket`, en la que los pares de bases enlazadas se representan en una cadena de texto con '()' y las bases libres se representan con '.'. Usa también un algoritmo de programación dinámica para predecir estructuras secundarias de cadenas de ARN.

Para comparar, calcularé el porcentaje de nucleótidos enlazados, que recordemos que está vinculado con la energía libre de la conformación de la cadena, y nos dará una idea de si el modelo Random Walk realiza predicciones similares a los modelos aquí descritos, ya validados para esta tarea.

### 3.4. Conformaciones Aleatorias

Para contrastar tanto los modelos ya validados que voy a usar como los modelos Random Walk, calcularé también conformaciones aleatorias de la cadena, siguiendo las reglas de la sección 1.3. Para ello, montaré una matriz de enlaces posibles  $q_{ij}$ , asignando  $q_{ij} = 1$  si el par  $i - j$  cumple las normas y es complementario, y  $q_{ij} = 0$  si no lo hace.

De manera aleatoria tomaré elementos de la matriz, teniendo en cuenta que no se deben obtener pseudonudos, y calcularé una conformación aleatoria de la cadena dada, a modo de control.

## 4. Resultados

Para validar el modelo Random Walk descrito en la sección 1.4, primero definí una cohorte de cinco transcritos, que obtuve en la base de datos de Ensembl Genome Browser. Limité la búsqueda a transcritos de menos de 2000 pares de bases, ya que el tiempo de ejecución de estos

algoritmos es del orden de  $O(N^3)$ , un tiempo de ejecución bastante alto para un ordenador personal.

Tras obtener una base de  $\sim 156.000$  transcritos de genes de múltiples especies, escogí de manera aleatoria cinco transcritos dentro de la base de datos para montar la cohorte de validación del modelo. En la Tabla 1 se muestran los metadatos obtenidos de los transcritos:

EnsemblID	Transcrito	Gen	Especie	Longitud (bp.)	Biotipo	SeqFold
ENST00000662104	CEROX1-210	CEROX1	Homo Sapiens	1.307	lncRNA	False
ENST00000613346	PCDH15-221	PCDH15	Homo Sapiens	580	Protein Coding	True
ENST00000575207	VPS53-218	VPS53	Homo Sapiens	1.356	Nonsense Mediated Decay	False
ENST00000506218	MACROH2A1-208	MACROH2A1	Homo Sapiens	758	Retained Intron	True
ENST00000217043	R3HDML-201	R3HDML	Homo Sapiens	1.323	Protein Coding	False

Tabla 1: Metadatos de los transcritos escogidos. La columna *Longitud* (medida en pares de bases, *bp.*) indica la longitud del transcripto estudiado. La columna *Biotipo* indica la función biológica del transcripto. La columna *SeqFold* indica si se pudo ejecutar el algoritmo de *SeqFold* para dicho transcripto.

Para cada transcripto en la Tabla 1 se obtuvieron las secuencias de los genes que los codifican, así como las secuencias de cada transcripto para realizar la simulación, mediante la API de Ensembl. Además, se hicieron las siguientes simulaciones, en base a los modelos explicados:

- Simulación usando una conformación aleatoria del modelo Random Walk. Con esta simulación se obtiene una gráfica de la conformación espacial obtenida. Para las cadenas obtenidas con esta simulación, se calculó su energía libre de acuerdo a la explicación de la sección 1.4.1.
- Simulación usando diez conformaciones aleatorias del modelo Random Walk y obteniendo los enlaces entre nucleótidos con *ViennaRNA* a partir de la matriz de probabilidad calculada. Con esta simulación, se obtiene la matriz de probabilidad entre nucleótidos. Como referencia, guardé también la conformación espacial de la última simulación realizada de las diez conformaciones calculadas.
- Simulación usando el algoritmo de programación dinámica de *ViennaRNA RNAFold*. Con esta simulación, se obtiene una gráfica de la estructura secundaria obtenida en el plano. Para las cadenas obtenidas con esta simulación, se calculó su energía libre de acuerdo a la explicación de la sección 1.3.1.
- Simulación usando el algoritmo de programación dinámica de *SeqFold*. Durante la ejecución del programa para los transcritos dados *SeqFold* devolvió errores de ejecución debido a un límite de memoria por recursión. Con esta simulación, se obtiene una gráfica de la estructura secundaria obtenida en el plano. Para las cadenas obtenidas con esta simulación, se calculó su energía libre de acuerdo a la explicación de la sección 1.3.1.
- Cálculo de una conformación aleatoria de la cadena estudiada, mediante el cuál se obtiene una gráfica de la estructura secundaria de la conformación calculada, conforme a lo explicado en la sección 3.4.

Como para cada transcripto se obtienen al menos seis figuras relevantes, he incluido las gráficas resultantes del análisis en el Apéndice I.

Para comparar los resultados obtenidos de cada modelo, utilizaré el número de enlaces encontrados en una conformación dada. Así mismo calcularé, según el método de R. Nussinov, et. al. (1980) [7], descrito en la sección 1.3.1, las energías de la cadena  $E_{1N}$ .

Los resultados obtenidos se muestran en la Tabla 2:

Transcrito	Longitud (bp.)	Simulación	Nucleótidos Enlazados	$E_{1N}$ (u.a.)	$g = G/kT$
CEROX1-210	1307	Una Conformación	474	-237	-3310,16
		Diez Conformaciones	4	-2	-
		Conformación Aleatoria	578	-289	-
		ViennaRNA	850	-425	-7979,45
PCDH15-221	580	Una Conformación	218	-109	-1478,33
		Diez Conformaciones	22	-11	-
		Conformación Aleatoria	264	-132	-
		ViennaRNA	380	-190	-2498,42
		SeqFold	332	-166	-1277,09
VPS53-218	1356	Una Conformación	474	-237	-3439,11
		Diez Conformaciones	46	-23	-
		Conformación Aleatoria	608	-304	-
		ViennaRNA	848	-424	-6018,25
MACROH2A1-208	758	Una Conformación	226	-113	-1920,40
		Diez Conformaciones	0	0	-
		Conformación Aleatoria	334	-167	-
		ViennaRNA	478	-239	-4357,62
		SeqFold	418	-209	-2286,25
R3HDM1L-201	1323	Una Conformación	408	-204	-3358,15
		Diez Conformaciones	428	-214	-
		Conformación Aleatoria	602	-301	-
		ViennaRNA	826	-413	-6509,23

Tabla 2: Número de enlaces encontrados para cada cadena.

Nótese que se satisface la relación  $E_{1N} = -n_{linked}/2$ , siendo  $n_{linked}$  el número de nucleótidos enlazados. Esto se debe a que cada par de nucleótidos enlazados contribuye a la energía total  $\epsilon_{ij} = -1$  (medido en u.a.).

Aunque solo se trate de una medida orientativa, la energía de la cadena  $E_{1N}$  ya ofrece información. Nótese que las conformaciones aleatorias, en general, tienen una energía mayor, indicando que son menos estables que las conformaciones encontradas por *ViennaRNA*. Cabe esperar este resultado, dado que el algoritmo de programación dinámica está diseñado para minimizar la energía libre de la cadena.

Para los modelos para los que se ha calculado la energía libre de la cadena, que son la conformación aleatoria mediante el modelo Random Walk, el modelo *ViennaRNA* y el modelo *SeqFold*, el modelo Random Walk obtiene energía mayores, y por tanto, menos estables, que los otros dos modelos, y, en general, *ViennaRNA* siempre obtiene las cadenas más estables, con la menor energía libre.

A partir de la tabla podemos observar que el modelo promediado de diez conformaciones es el que aporta resultados menos estables. Esto puede deberse a un bajo número de iteraciones, pero aún así, exceptuando el transcrito R3HDM1L-201, en los demás transcritos la bajada del número de nucleótidos enlazados entre una simulación y diez simulaciones es de alrededor del 90 %.

## 5. Conclusión

En base a estos resultados, ¿qué se puede concluir sobre el modelo Random Walk para calcular conformaciones de cadenas de ARN? Si bien es verdad que es un modelo bastante simple, lo cual facilita su uso y aplicación, no se ajusta adecuadamente a lo que ya conocemos sobre las estructuras del ARN.

- Por un lado, no predice estructuras comunes que sabemos que en el ARN se producen de manera empírica, como las hélices, los loops o los hairpins.
- Tampoco tiene en cuenta la energía libre de la cadena. Al calcular conformaciones aleatorias, el modelo está limitado a producir, en general, conformaciones inestables o alejadas del equilibrio, porque no tiene en cuenta que son más probables los estados donde hay enlaces entre nucleótidos que los estados que no los tienen.
- Otro gran problema del modelo es que, al calcular conformaciones de manera aleatoria, en general los enlaces que encuentra suelen ser intercalados, enlaces que se dan con poca frecuencia en el ARN, y son más bien propios de su estructura terciaria, por ser más inestables.
- Cabe mencionar también que el modelo Random Walk no separa entre estructura secundaria y terciaria, haciendo aún más complejo e improbable que la conformación aleatoria que produzca se acerque a la empírica.

Estas conclusiones surgen al comparar la estructura predicha del modelo Random Walk, con las estructuras que predicen los modelos *ViennaRNA* y *SeqFold*, modelos ya ajustados a datos empíricos y validados. Queda claro que, en general, las conformaciones calculadas mediante algoritmos de programación dinámica tienen estructuras con más enlaces y por lo tanto menor energía libre, lo que las hace más estables.

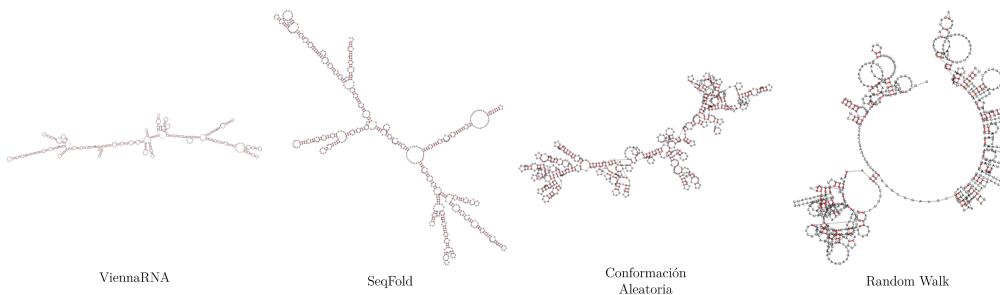


Figura 6: Ejemplo de conformaciones calculadas con los diferentes modelos del estudio.

Sin embargo, a la vista de estos resultados cabe mencionar una serie de vías de mejora del modelo, que podrían acercarlo más a predecir correctamente estructuras reales.

- Una de las limitaciones del modelo es su complejidad a la hora de calcular conformaciones. Como se comentó en la sección 1.2, la estructura secundaria de las cadenas de ARN suele tener una mayor relevancia a la hora de determinar cómo se dobla una cadena. Eso hace el cálculo de su estructura tridimensional innecesario en un principio, ya que la molécula de

ARN tenderá a formar estructuras simples, topológicamente equivalentes a un grafo planar, embebido en un plano. Es una de las características de las que otros modelos hacen uso y el modelo Random Walk no. Eso me lleva a preguntarme si restringir el modelo a dos dimensiones mejoraría su eficacia.

- Otra idea que merece atención es la de cambiar la distribución probabilística con la que se decide en qué dirección dar el siguiente paso de la cadena. Con el modelo aquí presentado, las distribuciones de  $(\theta_i, \varphi_i)$  son uniformes: existe la misma probabilidad de coger cualquier dirección. Esta suposición carece de sentido biológico. La dirección que toma la cadena depende en gran medida de la que haya tomado anteriormente. A raíz de esta idea, surgen preguntas sobre qué distribución emplear para decidir en qué dirección ha de moverse el modelo en cada paso y con cada segmento.
- Por supuesto, también está el problema de la energía libre. ¿Habrá alguna manera de tener en cuenta la energía libre de la cadena? ¿Puede esta influir en los pasos que da el modelo y, por lo tanto, en la distribución de partida que comentaba antes?

Todas estas preguntas sugieren nuevas vías de avance para seguir mejorando el modelo que en el futuro se podrían explorar.

En este trabajo, se ha recurrido a modelos ya validados en trabajos anteriores con los que comparar el modelo propio. Sin embargo, no puedo acabar el estudio sin mencionar posibles formas de validar estos modelos a nivel experimental.

- Para validar los modelos, se pueden realizar experimentos de fuerza extensión (force extension), ya planteados en R. Phillips, et. al. (2013) [13] o en A. Carpio, et. al. (2015) [19]. En estos experimentos se somete a una macromolécula, como el ARN a una fuerza, generalmente de tensión, y se mide la extensión de la molécula. Al comparar la fuerza que se ha de utilizar para llegar a una extensión dada, se pueden encontrar los enlaces de la estructura secundaria del ARN, lo cual aporta información sobre los pares de bases enlazados entre sí. Este experimento puede servir para validar los modelos aquí utilizados, así como para ajustar sus parámetros termodinámicos, en caso de ser necesario.
- Otro experimento es el de comprobar qué efectos tiene la estructura secundaria del ARN sobre su función. Si se introducen inserciones o delecciones en el transcripto y se modifica su estructura, sin modificar las regiones codificantes de proteína, ¿afectaría a la traducción o al procesamiento del transcripto en particular? Y en caso de hacerlo, ¿cómo afecta? ¿Qué pasaría con las ribozimas? ¿Cambiaría su función biológica? ¿Se podrían llegar incluso a diseñar?

Estas preguntas abren la puerta a diferentes vías de investigación en la materia de la estructura del ARN y su función biológica. Sin embargo, la complejidad de las mismas hacen que no sea posible abarcárlas en el trabajo aquí realizado, y por ello se proponen como posibilidad a futuro.

## 6. Bibliografía

- [1] Khan Academy: AP Biology. dirección: <https://es.khanacademy.org/science/ap-biology/gene-expression-and-regulation/dna-and-rna-structure/a/nucleic-acids>.
- [2] B. Alberts, R. Heald, A. Johnson y col., *Molecular Biology of the Cell*, 7th Edition. W. W. Norton & Company, 2022, cap. 1, págs. 1-48, ISBN: 978-0-3938-8482-1.
- [3] P. G. Higgs, «RNA secondary structure: physical and computational aspects,» *Quarterly Reviews of Biophysics*, vol. 33, págs. 199-253, 2000. DOI: <https://doi.org/10.1017/S0033583500003620>.
- [4] Wikipedia: Apareamiento de Bases. dirección: [https://es.wikipedia.org/wiki/Apareamiento\\_de\\_bases](https://es.wikipedia.org/wiki/Apareamiento_de_bases).
- [5] Wikipedia: RNA Chemical Structure. dirección: [https://es.m.wikipedia.org/wiki/Archivo:RNA\\_chemical\\_structure.png](https://es.m.wikipedia.org/wiki/Archivo:RNA_chemical_structure.png).
- [6] Wikipedia: Ácido Ribonucléico. dirección: [https://es.wikipedia.org/wiki/acido\\_ribonucleico](https://es.wikipedia.org/wiki/acido_ribonucleico).
- [7] R. Nussinov y A. B. Jacobson, «Fast algorithm for predicting the secondary structure of single-stranded RNA,» *Proceedings of the National Academy of Sciences of the United States of America*, vol. 77, págs. 6309-6313, 1980. DOI: <https://doi.org/10.1073/Pnas.77.11.6309>.
- [8] J. S. McCaskill, «The equilibrium partition function and base pair binding probabilities for RNA secondary structure,» *Biopolymers*, vol. 29, págs. 1105-1119, 1990. DOI: <https://doi.org/10.1002/bip.360290621>.
- [9] P. G. Doyle y J. L. Snell, *Random walks and electric networks*, arXiv:math, 2000. DOI: <https://doi.org/10.48550/arXiv.math/0001057>. dirección: <https://arxiv.org/abs/math/0001057v1>.
- [10] K. Pearson, «The Problem of the Random Walk,» *Nature*, vol. 72, pág. 294, 1905. DOI: <https://doi.org/10.1038/072294b0>.
- [11] G. H. Weiss, *Aspects and applications of the random walk* (Random Materials and Processes). NH, 1994, ISBN: 978-0-4448-1606-1.
- [12] B. Fischer y M. Scholes, «The Pricing of Options and Corporate Liabilities,» *Journal of Political Economy*, vol. 81, págs. 637-654, 1973. DOI: <https://doi.org/10.1086/260062>.
- [13] R. Phillips, J. Kondev, J. Theriot y H. G. García, *Physical Biology of the Cell*, 2nd Edition. Garland Science, 2013, cap. 8, págs. 311-354, ISBN: 978-0-8153-4450-6.
- [14] Wikipedia: Langevin Function. dirección: [https://en.wikipedia.org/wiki/Brillouin\\_and\\_Langevin\\_functions#Langevin\\_Function](https://en.wikipedia.org/wiki/Brillouin_and_Langevin_functions#Langevin_Function).
- [15] P. Lalanda-Delgado. «Piquipato/tfgrna.» (2024), dirección: <https://github.com/Piquipato/tfgrna>.
- [16] I. L. Hofacker, P. Schuster, W. Fontana y P. F. Stadler, «From sequences to shapes and back: a case study in RNA secondary structures,» *Proceedings of the Royal Society, Biological Sciences*, vol. 255, 279–284, 1994. DOI: <https://doi.org/10.1098/rspb.1994.0040>.
- [17] Z. Ouyang, M. P. Snyder y H. Y. Chang, «SeqFold: Genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data,» *Genome Research*, vol. 23, págs. 377-387, 2013. DOI: <https://doi.org/10.1101/gr.138545.112>. dirección: <https://github.com/Lattice-Automation/seqfold>.
- [18] R. Lorenz, I. L. Hofacker, P. F. Stadler y col., «ViennaRNA Package 2.0,» *Algorithms for Molecular Biology*, vol. 6, 2011. DOI: <https://doi.org/10.1186/1748-7188-6-26>. dirección: <https://viennarna.readthedocs.io/en/latest/>.
- [19] A. Carpio, L. L. Bonilla y A. Prados, «Theory of force-extension curves for modular proteins and DNA hairpins,» *Physical Review E*, vol. 91, 2015. DOI: <https://doi.org/10.1103/PhysRevE.91.052712>.

## 7. Apéndice I: Gráficas de los Transcritos

### CEROX1-210

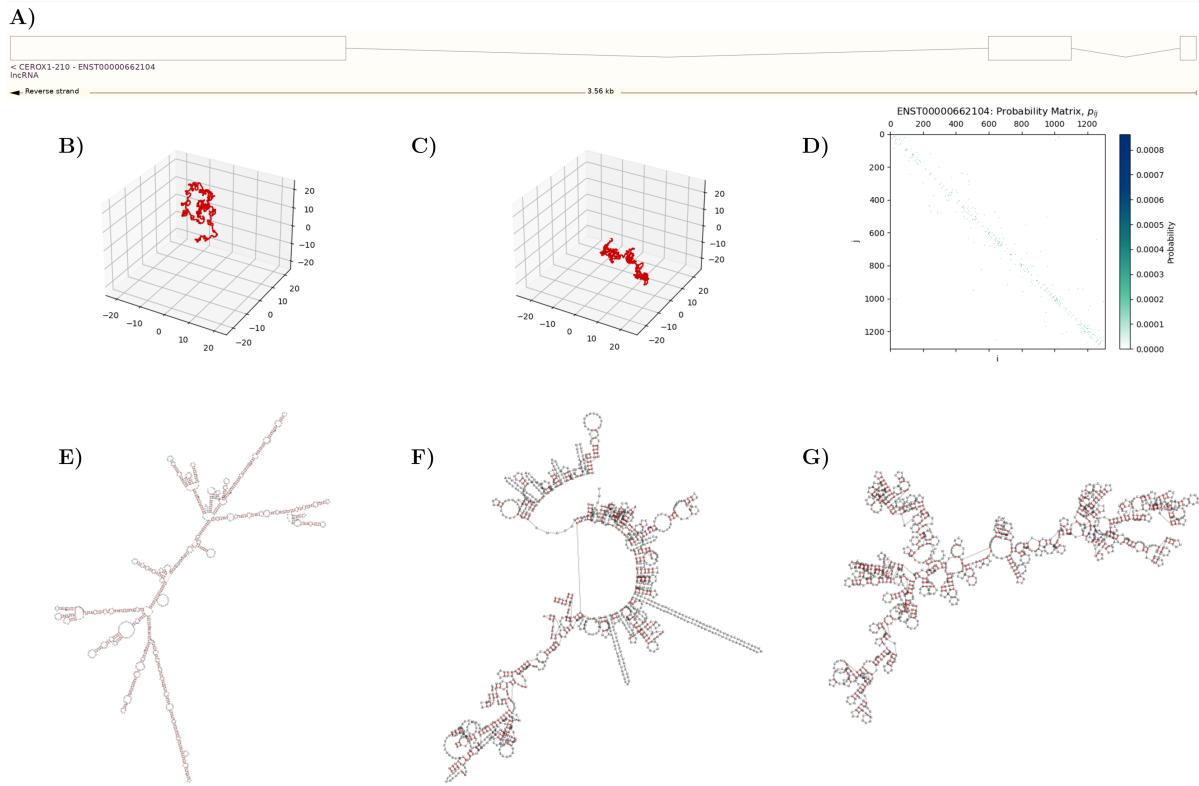


Figura 7: Resultados para el transcripto CEROX1-210: A) Diagrama del transcripto, B) Modelo Random Walk, una iteración, C) Modelo Random Walk, diez iteraciones, D) Matriz de Probabilidades, diez iteraciones, E) Conformación 2D, *ViennaRNA*, F) Conformación 2D, Modelo Random Walk, G) Conformación Aleatoria 2D.

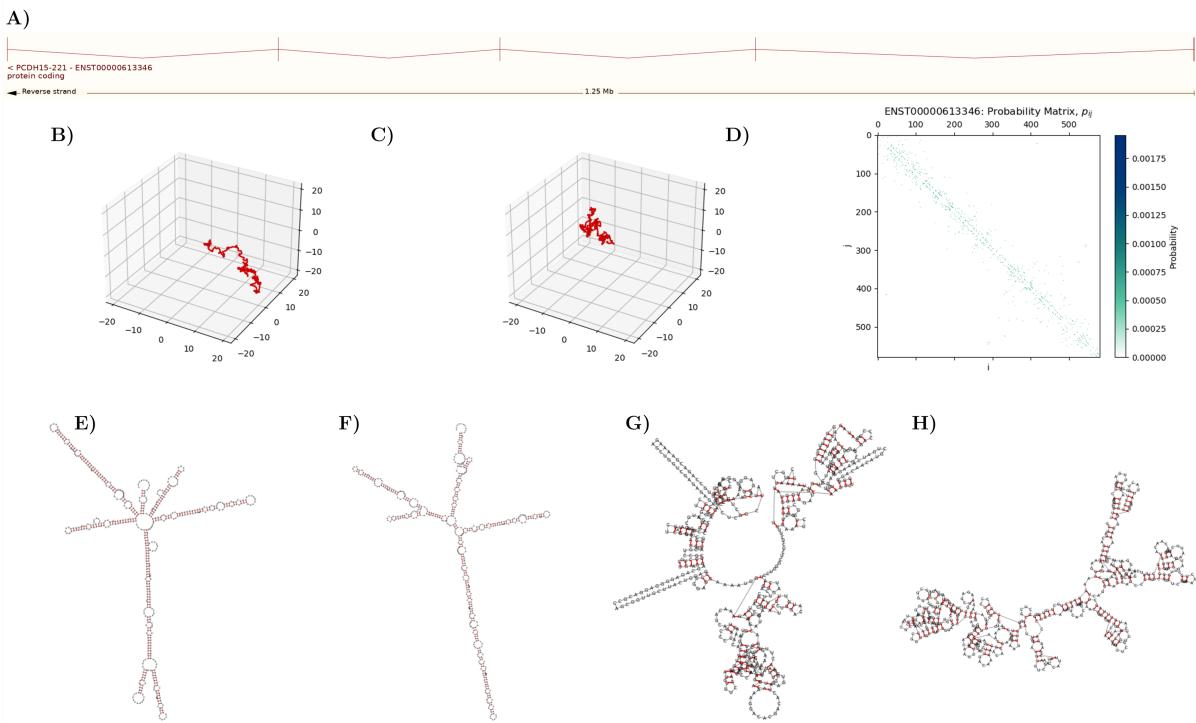
**PCDH15-221**

Figura 8: Resultados para el transcripto PCDH15-221: A) Diagrama del transcripto, B) Modelo Random Walk, una iteración, C) Modelo Random Walk, diez iteraciones, D) Matriz de Probabilidades, diez iteraciones, E) Conformación 2D, *ViennaRNA*, F) Conformación 2D, *SeqFold*, G) Conformación 2D, Modelo Random Walk, H) Conformación Aleatoria 2D.

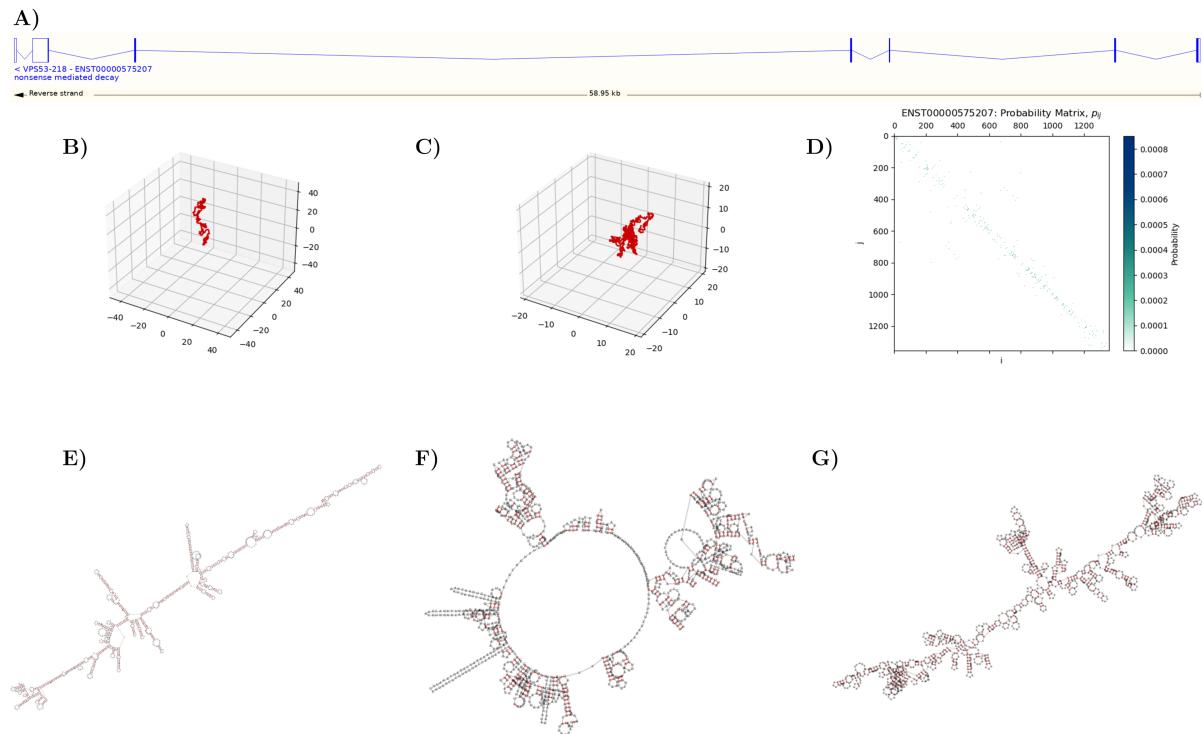
**VPS53-218**

Figura 9: Resultados para el transcripto VPS53-218: A) Diagrama del transcripto, B) Modelo Random Walk, una iteración, C) Modelo Random Walk, diez iteraciones, D) Matriz de Probabilidades, diez iteraciones, E) Conformación 2D, *ViennaRNA*, F) Conformación 2D, Modelo Random Walk, G) Conformación Aleatoria 2D.

## MACROH2A1-208

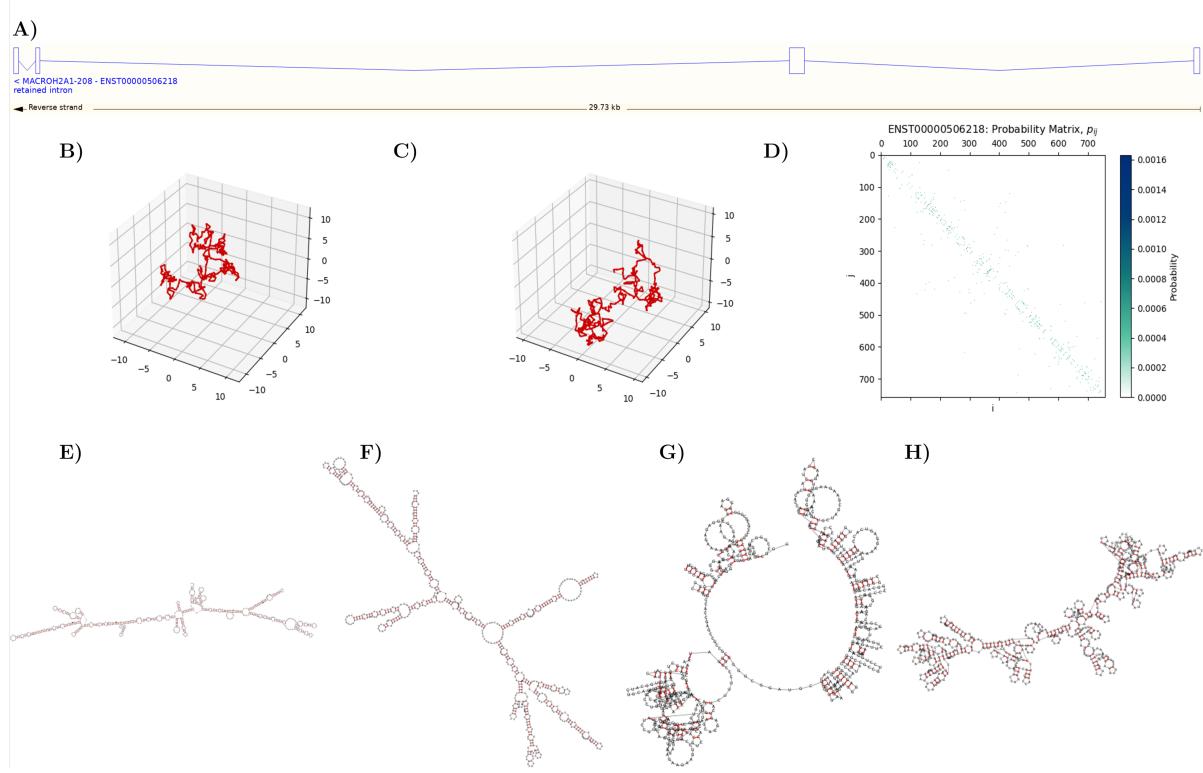


Figura 10: Resultados para el transcrito MACROH2A1-208: A) Diagrama del transcrito, B) Modelo Random Walk, una iteración, C) Modelo Random Walk, diez iteraciones, D) Matriz de Probabilidades, diez iteraciones, E) Conformación 2D, *ViennaRNA*, F) Conformación 2D, *SeqFold*, G) Conformación 2D, Modelo Random Walk, H) Conformación Aleatoria 2D.

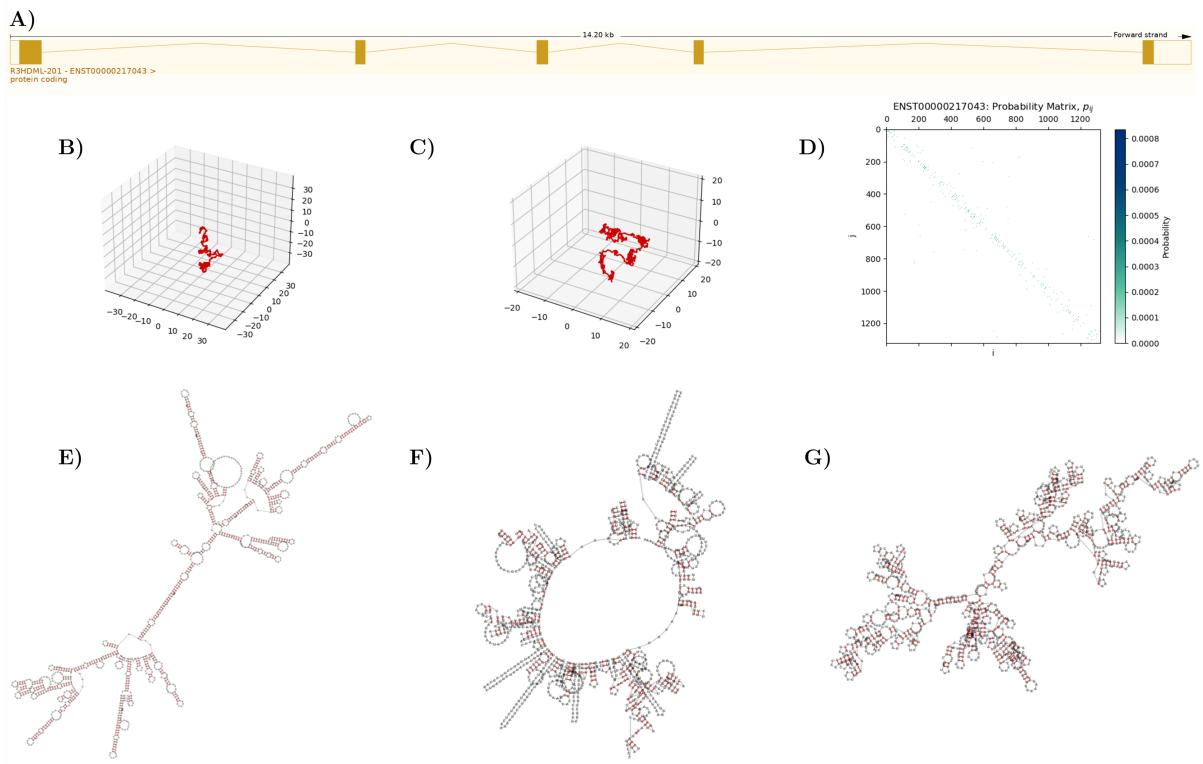
**R3HMDL-201**

Figura 11: Resultados para el transcrito R3HMDL-201: A) Diagrama del transcrito, B) Modelo Random Walk, una iteración, C) Modelo Random Walk, diez iteraciones, D) Matriz de Probabilidades, diez iteraciones, E) Conformación 2D, *ViennaRNA*, F) Conformación 2D, Modelo Random Walk, G) Conformación Aleatoria 2D.