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 Lalanda Delgado

Grado en Física

Curso académico 2023-2024

Convocatoria Ordinaria

UNIVERSIDAD COMPLUTENSE MADRID

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:

- [x] 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

Modelado tridimensional del ARN mediante Random Walks

1. Abstract

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 como alternativa el modelo Random Walk, una alternativa más sencilla conceptualmente, aunque en la práctica muy ineficiente. 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.

2. Fundamento Teórico

2.1. 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[1].

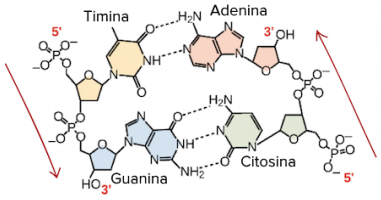

Figura 1: Cadena de ADN.

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, y por una pentosa adherida a un grupo fosfáto. La cadenas de nucleótidos, i.e., las hélices, se mantienen unidas por enlaces covalentes entre las pentosas fosfato, y por su estructura complementaria la célula es capaz de replicar las cadenas de nucleótidos con facilidad.

2. 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).

3. 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 no suelen encontrarse dos cadenas de ARN enlazadas entre sí, 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).

2.2. Estructura del ARN

Como he comentado previamente, en este estudio, me enfoco principalmente en la estructura que adopta la molécula de ARN, que al no estar enlazada a otra cadena complementaria, como pasa con el ADN, se dobla sobre si misma debido a la interacción que sufren entre sí los ácidos nucleicos que la componen. Cabe discutir, en este caso, la estructura que tiene el ARN y como tratar de predecir la conformación que dicha molécula va a tomar.

Cabe mencionar que la exposición teórica expuesta en esta sección se basa en el trabajo realizado por Paul G. Higgs (2000) [2], quién aglutinó en la revisión citada los diferentes modelos de predicción de estructura del ARN.

El ARN esta compuesto, por cuatro ácidos nucleicos (adenina, citosina, guanina, uracilo), conectados por una cadena de ribosas, unidas entre sí por fosfatos. Es esta la principal diferencia que la estructura del ARN encuentra con la estructura del ADN, cuya cadena principal esta formada por desoxiribosas, perdiendo un átomo de oxígeno respecto a la ribosa. La otra gran diferencia que encontramos entre el ADN y el ARN es en los ácidos nucleicos que usan, principalmente en que el ADN hace uso de la timina (T) y el ARN, en cambio, hace uso del uracilo (U), aunque ambas moléculas enlazan con la adenina por medio de un enlace de hidrógeno. Cabe mencionar también que el uracilo puede enlazarse también a la guanina, cosa que con la timina no pasa y habrá que tener en cuenta más adelante.

Tanto para la cadena de ADN como para la de ARN se pueden definir direcciones en función de los átomos a los que se enlazan 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 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. Al igual que el ADN, el ARN al doblarse sobre sí mismo forma estructuras en forma de hélice cuando dos segmentos complementarios de la cadena se enlazan. Principalmente, la estructura secundaria se describe encontrando las hélices que se forman y las secciones de nucleótidos desemparejadas que no forman hélices (hairpins, loops...).
- 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.

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 encontrar dicha estructura.

2.3. 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, $|i - j| \leq 4$, 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, sí son solapan o una esta contenida dentro de la otra ($i < j < k < l$ o $i < k < l < j$). Descartamos aquellas parejas que no cumplan estas condiciones ($i < k < j < l$, enlaces intercalados), 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.

2.3.1. 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) [3]. Asignamos una energía $\epsilon_{ij} = -1$ a cada par de bases enlazadas y, en caso de no estar enlazadas, asignamos una energía a la base de $\epsilon_{ij} = +\infty$, siendo ϵ_{ij} la enegía de un par de bases dado. Se pretende obtener la enegía 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,

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

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

$$E_{ij} = \text{mín}\left(E_{i,j-1}, \min_{i \leq k \leq j-4}(E_{i,k-1} + E_{k+1,j-1} + \epsilon_{kj})\right) \tag{2}$$

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.

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 ϵ_{ij}. Estos parámetros se pueden ajustar para reflejar mejor los resultados obtenidos en experimentos termodinámicos realizados con cadenas de ARN.

Cabe destacar que la energía del modelo (medida en *u.a.* del programa), E_{1N}, no es la energía libre de la cadena, sino la energía total de la misma, relacionadas mediante la entropía de la cadena de acuerdo a $G = U + TS$. Lo que nos indica este hecho realmente es que a la hora de minimizar la energía libre de la cadena, hay que tener en cuenta a que temperatura a la que hacemos la aproximación. La entropía se puede calcular fácilmente partiendo de la tabla de probabilidades de cada enlace, con la expresión de la entropía de Shannon

$$S_i = -\sum_{i,j} p_{ij} \log p_{ij} \tag{3}$$

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

2.3.2. Otros Modelos

Cabe mencionar la existencia de otros modelos de predicción de estructura del ARN, 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.

2.4. 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 [4] [5].

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 [6], así como a disciplinas ajenas al estudio del mundo natural, como puede ser el estudio de los mercados financieros[7].

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) [8].

2.4.1. Modelo Freely Jointed Chain del ADN

Supongamos que una cadena de ADN dada la podemos dividir en segmentos rígidos de una longitud dada, a, conectados entre sí por uniones flexibles. Se denomina segmentos de Kuhn a los segmentos definidos. Queremos encontrar una distribución de probabilidad $p(\vec{R}, N)$ que nos indique con que probabilidad una cadena de N nucleotidos 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,

$$p(n_r, N) = \binom{N}{n_r}\left(\frac{1}{2}\right)^N = \frac{N!}{n_r!(N-n_r)!}\left(\frac{1}{2}\right)^N \tag{4}$$

Podemos observar pues que la cantidad de pasos que tomo 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

$$p(R, N) = \frac{N!}{(\frac{N}{2} + \frac{R}{2a})!(\frac{N}{2} - \frac{R}{2a})!}\left(\frac{1}{2}\right)^N \tag{5}$$

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 (5) mediante la aproximación de Striling ($\ln n! \approx n \ln n - n + (1/2)\ln(2\pi n)$, 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

$$p(R, N) = \frac{1}{\sqrt{2\pi N a^2}} e^{-\frac{R^2}{2Na^2}} \tag{6}$$

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

$$p\Big(\vec{R}, N\Big) = \left(\frac{3}{2\pi N a^2}\right)^{\frac{3}{2}} e^{-\frac{3R^2}{2Na^2}} \tag{7}$$

3. Materiales y Métodos

Para comparar los diferentes modelos que he expuesto en la sección 2, haré uso de programas escritos en Python, subidos al repositorio Piquipato/tfgrna en GitHub [9]. 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 $n_t = 0.34\ nm$.
- La longitud de persistencia de la cadena, i.e., la longitud de cadena que se puede tomar como rígida. El valor por defecto es $\xi_p = 1.5$ `nt`, reflejando en un hairpin tiene que haber al menos tres nucleotidos libres [2].

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, (θ_i, φ_i), (con el paquete `random`).

$$\vec{s}_i(\theta_i, \varphi_i) = 2\xi_p\Big(\sin\theta_i \cos\varphi_i \hat{i} + \sin\theta_i \sin\varphi_i \hat{j} + \cos\theta_i \hat{k}\Big) \tag{8}$$

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.

$$\vec{x}_{ki} = \vec{X}_{i,0} + \frac{k}{n_{bp}} \cdot \vec{s}_i \tag{9}$$

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 esten a una distancia mayor a ξ_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 enlaces intercalados mencionados en la sección 2.3.

3.2. Promedios Modelo Freely Jointed Chain

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 estable, 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) [2], podemos determinar la estructura más probable. Gracias al trabajo de I. L. Hofacker, et.al. (1994) [10], podemos usar el paquete ViennaRNA en python para realizar este cálculo.

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:

- *SeqFold*[11]: Implementación de un algoritmo de programación dinámica, como los descritos en la sección 2.3.1, con parámetros termodinámicos ajustados.
- *ViennaRNA RNAFold*[12]: Paquete desarrollado a raíz del trabajo de I. L. Hofacker, et.al. (1994) [10], 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 `'.'`. Usaré extensivamente esta notación en la sección 4

Para comparar, calcularé el diagrama `dot_bracket` de cada uno de los modelos, así como el porcentaje de nucleotidos enlazados, que recordemos que está vinculado con la enegía libre de la conformación de la cadena.

4. Resultados

Para validar el modelo Random Walk descrito en la sección 2.4, primero monté una cohorte de cinco transcritos, que busque en la base de datos de Ensembl Genome Browser. Con los programas descritos en la sección 3.

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 ~ 156000 transcritos de genes de múltiples especies, escogí de manera aleatoria cinco genes 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	Gen	Transcrito	Especie	Longitud (*bp.*)	Biotipo	SeqFold
ENST00000662104	CEROX1-210	CEROX1	Homo Sapiens	1307	lncRNA	no
ENST00000613346	PCDH15-221	PCDH15	Homo Sapiens	580	Protein Coding	yes
ENST00000575207	VPS53-218	VPS53	Homo Sapiens	1356	Nonsense Mediated Decay	no
ENST00000506218	MACROH2A1-208	MACROH2A1	Homo Sapiens	758	Retained Intron	yes
ENST00000217043	R3HDML-201	R3HDML	Homo Sapiens	1323	Protein Coding	no

Tabla 1: Metadatos de los transcritos escogidos.

Para cada transcrito en la Tabla 1, se obtuvieron las secuencias de los genes que los codifican, así como las secuencias de cada transcrito 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.
- Simulación usando diez conformaciones aleatorias del modelo Random Walk y obteniendo los enlaces entre nucleotidos con ViennaRNA a partir de la matrix de probabilidad calculada. Con esta simulación, se obtiene la matriz de probabilidad entre nucleótidos, así como 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.
- Simulación usando el algoritmo de programación dinámica de *SeqFold*. Nótese en la Tabla 1 la columna "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.

Las gráficas resultantes del análisis se podrán encontrar en el Apéndice I 7.

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) [3], descrito en la sección 2.3.1.

Transcrito	Longitud (*bp.*)	Simulación	Nucleótidos Enlazados	E_{1N} (*u.a.*)
CEROX1-210	1307	Una Conformación	474	-237
		Diez Conformaciones	4	-2
		ViennaRNA	850	-425
PCDH15-221	580	Una Conformación	218	-109
		Diez Conformaciones	22	-11
		ViennaRNA	380	-190
		SeqFold	332	-166
VPS53-218	1356	Una Conformación	474	-237
		Diez Conformaciones	46	-23
		ViennaRNA	848	-424
MACROH2A1-208	758	Una Conformación	226	-113
		Diez Conformaciones	0	0
		ViennaRNA	478	-239
		SeqFold	418	-209
R3HDML-201	1323	Una Conformación	408	-204
		Diez Conformaciones	428	-214
		ViennaRNA	826	-413

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 nucleotidos enlazados. Esto se debe a que cada par de nucleotidos enlazados contribuye a la energía total $\epsilon_{ij} = -1$ (medido en *u.a.* del programa).

Aunque solo se trate de una medida orientativa, la energía de la cadena E_{1N} ya ofrece mucha 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.

También es aparente a partir de la tabla que calcular varias conformaciones aleatorias de una cadena dada es contraproducente. Esto puede deberse a un bajo número de iteraciones, pero aún así, exceptuando el transcrito R3HDML-201, en los demás transcritos la bajada del número de nucleotidos enlazados entre una simulación y diez simulaciones es de alrededor del 90 %.

5. Discusión

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 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.
- 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 son más que evidentes 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 de mínima energía libre tienen mucha menos enegía que las calculadas.

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 pegas del modelo es su complejidad a la hora de calcular conformaciones. Como se comentó en la sección 2.2, la estructura secundaria de las cadenas de ARN suele tener una mayor relevancia a la hora de determinar como 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 mío 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 decido en que dirección dar el siguiente paso de la cadena. Con el modelo aquí presentado, las distribuciones de (θ_i, φ_i) son uniformes: existe la misma probabilidad de coger cualquier dirección. Es evidente que esta suposición juega en contra del modelo, y que no es real. La dirección que toma la cadena depende en gran medida de la que haya tomado anteriormente. A raíz de esta idea, surgen preguntas sobre que distribución coger para decidir en que dirección ha de moverse el modelo en cada paso y con cada segmento.

- Por supuesto, también esta 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.

No puedo acabar el estudio sin mencionar posibles formas de validar estos modelos a nivel experimental. En este trabajo, se ha recurrido a modelos ya validados en trabajos anteriores con los que comparar el modelo propio. Sin embargo, resulta mucho más interesante la opción de poder validarlos a nivel experimental. En esa dirección, se me ocurren dos experimentos que serían interesantes realizar.

- Para validar los modelos, podría ser interesante realizar experimentos de fuerza extensión (force extension), ya planteados en R. Phillips, et. al. (2013) [8] o en A. Carpio, et. al. (2015) [13]. 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 mucha 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 interesante, en el campo de la biología molecular, sería el de comprobar que efectos tiene la estructura secundaria del ARN sobre su función. Si se introducen inserciones o deleciones en el transcrito y se modifica su estructura, sin modificar las regiones codificantes de proteina, ¿afectará a la traducción o al splicing del transcrito en particular? Y en caso de hacerlo, ¿cómo afecta? ¿Qué pasaría con ribozimas? ¿Se podrían llegar incluso a diseñar?

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

6. Bibliografía

[1] 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.

[2] 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.

[3] R. Nussinov y A. B. Jacobson, «Fast algorithm for predicting the secondary structure of single-stranded RNA,» *Proceedings of the National Academy of Sciencesof the United States of America*, vol. 77, págs. 6309-6313, 1980. DOI: https://doi.org/10.1073%2Fpnas.77.11.6309.

[4] 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.

[5] K. Pearson, «The Problem of the Random Walk,» *Nature*, vol. 72, pág. 294, 1905. DOI: https://doi.org/10.1038/072294b0.

[6] G. H. Weiss, *Aspects and applications of the random walk* (Random Materials and Processes). NH, 1994, ISBN: 978-0-4448-1606-1.

[7] 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.

[8] 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.

[9] P. Lalanda-Delgado. «Piquipato/tfgrna.» (2024), dirección: https://github.com/Piquipato/tfgrna.

[10] 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.

[11] 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.

[12] 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/.

[13] 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

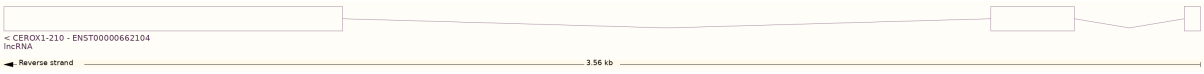

Figura 2: Diagrama del Transcrito.

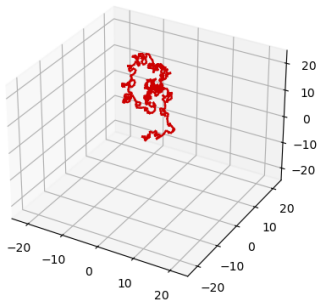

Figura 3: Conformación 3D CEROX1-210, una iteración.

Figura 4: Conformación 2D CEROX1-210, ViennaRNA.

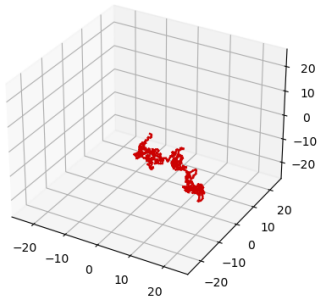

Figura 5: Conformación 3D CEROX1-210, diez iteraciones.

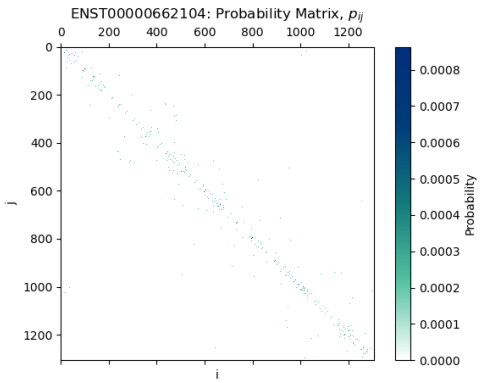

Figura 6: Matriz de probabilidades, CEROX1-210.

PCDH15-221

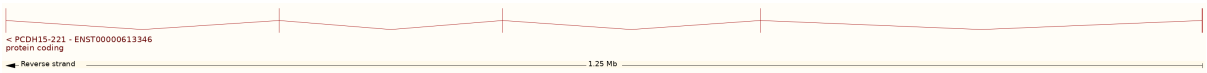

Figura 7: Diagrama del Transcrito.

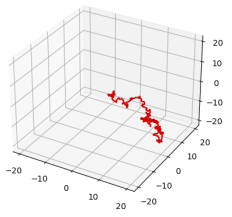

Figura 8: Conformación 3D PCDH15-221, una iteración.

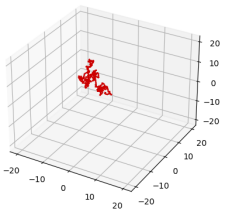

Figura 9: Conformación 3D PCDH15-221, diez iteraciones.

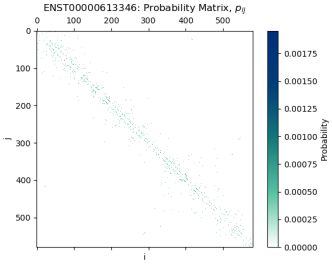

Figura 10: Matriz de probabilidades, PCDH15-221.

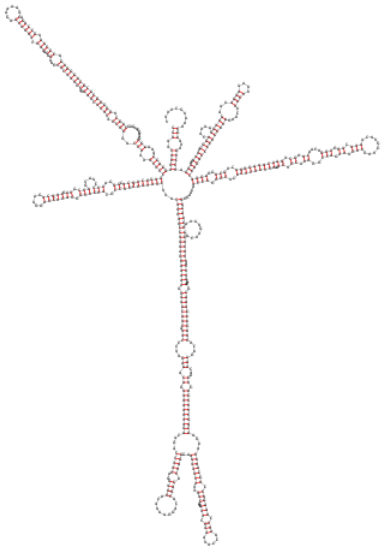

Figura 11: Conformación 2D PCDH15-221, ViennaRNA.

Figura 12: Conformación 2D PCDH15-221, SeqFold.

VPS53-218

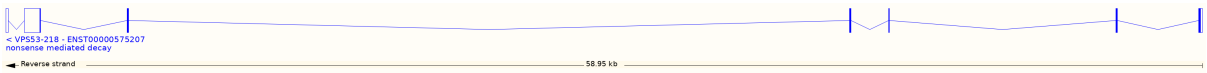

Figura 13: Diagrama del Transcrito.

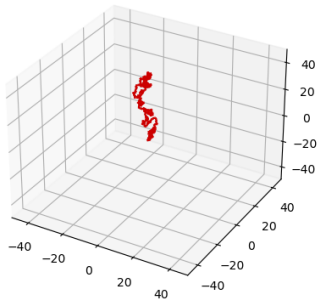

Figura 14: Conformación 3D VPS53-218, una iteración.

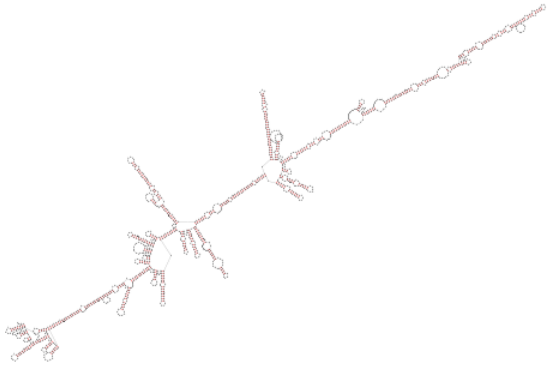

Figura 15: Conformación 2D VPS53-218, ViennaRNA.

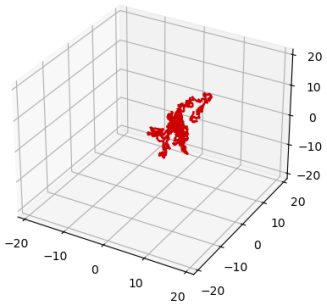

Figura 16: Conformación 3D VPS53-218, diez iteraciones.

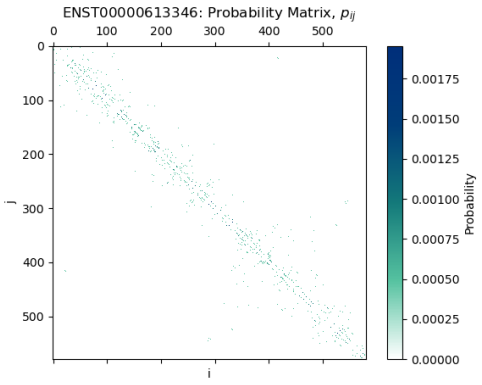

Figura 17: Matriz de probabilidades, VPS53-218.

MACROH2A1-208

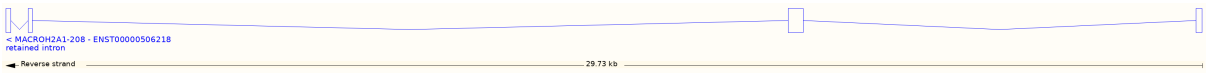

Figura 18: Diagrama del Transcrito.

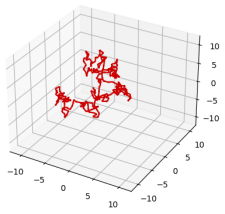

Figura 19: Conformación 3D MACROH2A1-208, una iteración.

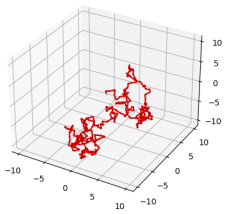

Figura 20: Conformación 3D MACROH2A1-208, diez iteraciones.

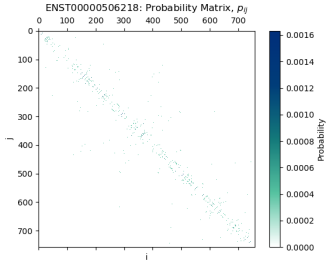

Figura 21: Matriz de probabilidades, MACROH2A1-208.

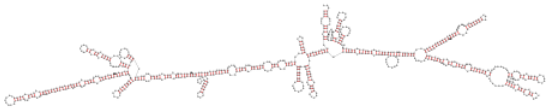

Figura 22: Conformación 2D MACROH2A1-208, ViennaRNA.

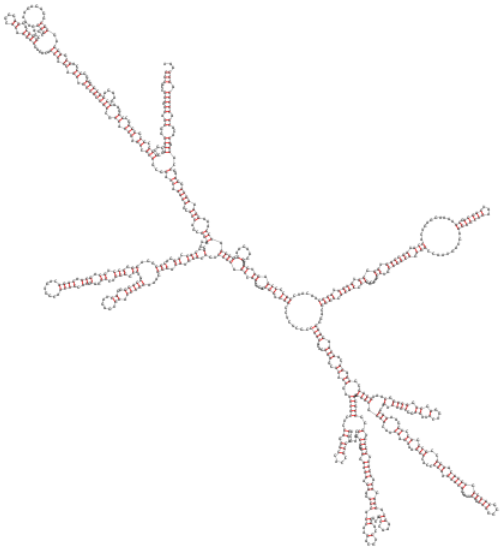

Figura 23: Conformación 2D MACROH2A1-208, SeqFold.

R3HDML-201

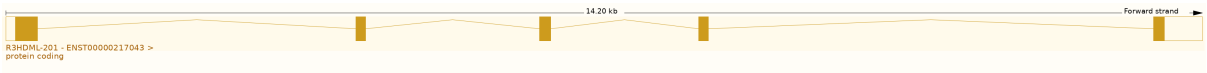

Figura 24: Diagrama del Transcrito.

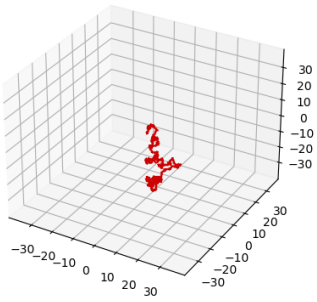

Figura 25: Conformación 3D R3HDML-201, una iteración.

Figura 26: Conformación 2D R3HDML-201, ViennaRNA.

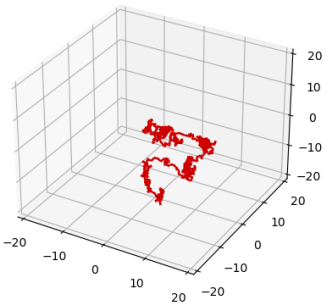

Figura 27: Conformación 3D R3HDML-201, diez iteraciones.

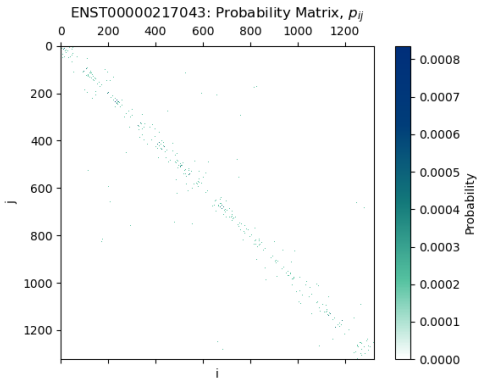

Figura 28: Matriz de probabilidades, R3HDML-201.