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

Índice

Modelado tridimensional del ARN mediante Random Walks

Resumen:

Esto es una prueba para probar el formato del Resumen. Esto es una prueba para probar el formato del ResumenEsto es una prueba para probar el formato del ResumenEsto es una prueba para probar el formato del ResumenEsto es una prueba para probar el formato del ResumenEsto es una prueba para probar el formato del ResumenEsto es una prueba para probar el formato del ResumenEsto es una prueba para probar el formato del ResumenEsto es una prueba para probar el formato del ResumenEsto es una prueba para probar el formato del ResumenEsto es una prueba para probar el formato del ResumenEsto es una prueba para probar el formato del ResumenEsto es una prueba para probar el formato del Resumen.

Abstract:

This is a test to prove the abstract's layout.This is a test to prove the abstract's layout.This is a test to prove the abstract's layout.This is a test to prove the abstract's layout.This is a test to prove the abstract's layout.This is a test to prove the abstract's layout.This is a test to prove the abstract's layout.This is a test to prove the abstract's layout.This is a test to prove the abstract's layout.This is a test to prove the abstract's layout.This is a test to prove the abstract's layout.This is a test to prove the abstract's layout.This is a test to prove the abstract's layout.This is a test to prove the abstract's layout.This is a test to prove the abstract's layout.This is a test to prove the abstract's layout.This is a test to prove the abstract's layout.This is a test to prove the abstract's layout.This is a test to prove the abstract's layout.

Nota: el título extendido (si procede), el resumen y el abstract deben estar en una misma página y su extensión no debe superar una página. Tamaño mínimo 11pto.

Extensión máxima 20 páginas sin contar portada, contraportada y declaración responsable (sí se incluye índice, introducción, conclusiones y bibliografía

INCLUIR AQUÍ la Declaración Responsable sobre Autoría y Uso Ético de Herramientas de Inteligencia Artificial

El documento se puede descargar en la página web: https://fisicas.ucm.es/tfg-gradoim

Y en https://fisicas.ucm.es/file/declaracion-responsable-sobre-autoria-y-uso-etico-de-ia?ver

Modelado tridimensional del ARN mediante Random Walks

1. Abstract

2. Introducción

3. Fundamento Teórico

3.1. 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.

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 mediante el uso de monómeros llamados nucleótidos (A, T, C, G), complementarios dos a dos mediante enlaces de hidrógeno, y que por su estructura complementaria, la célula es capaz de replicar 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), pero diferente al ADN ya que no suelen encontrarse dos cadenas de ARN enlazadas entre sí como pasa con el ADN. 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).

3.2. 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) [3], 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.

3.3. 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 ($i < k < j < l$), 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.

3.3.1. 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 (1980) [4]. 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 $\epsilon_{ij} = 0$, 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.

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

3.3.2. 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.

3.4. 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 [1] [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 [7], así como a disciplinas ajenas al estudio del mundo natural, como puede ser el estudio de los mercados financieros[2].

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) [6].

3.4.1. 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,

$$p(n_r, N) = \binom{N}{n_r}\left(\frac{1}{2}\right)^N = \frac{N!}{n_r!(N-n_r)!}\left(\frac{1}{2}\right)^N \tag{3}$$

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

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

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 (4) 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

$$p(R, N) = \frac{1}{\sqrt{2\pi N a^2}} e^{-\frac{R^2}{2Na^2}} \tag{5}$$

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,

$$p\left(\vec{R}, N\right) = \left(\frac{3}{2\pi N a^2}\right)^{\frac{3}{2}} e^{-\frac{3R^2}{2Na^2}} \tag{6}$$

4. Materiales y Métodos

Para comparar los diferentes modelos que he expuesto en la sección 3, haré uso de programas escritos en Python.

4.1. Simulación en Python

4.2. Otros Programas

5. Resultados

6. Conclusión

7. Referencias

[1] P. G. Doyle y J. L. Snell, *Random walks and electric networks*, arXiv:math, 2000.

[2] 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.

[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] 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.

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

[6] 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.

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