Una nueva parametrización para un modelo mesoscópico de interacción proteina-ADN

Autor: Luis Valiño Borau
Directores: Fernando Falo Forniés
Rafael Tapia Rojo
Juan José Mazo Torres

1 Introducción

"It is notoriously difficult to describe the word living"
- Francis Crick, frase inicial en "Of Molecules and Men"

La definición de la palabra "vida" ha supuesto un problema para muchas grandes mentes de diversas disciplinas a lo largo de la historia. Uno de los primeros en abordar el problema fue Aristóteles, quien le atribuyó una serie de rasgos que se pueden resumir en: nacer, crecer y morir. Más adelante, se añadió la última de las cuatro características que normalmente se utilizan para describir la vida de forma general: la reproducción; pudiéndose definir un ser vivo por tanto como aquel que nace, crece, se reproduce y muere. Aunque suficiente en muchos casos, esta definición pasa a ser incompleta con la aparición de la teoría de la evolución. De acuerdo a esta, ha de haber algun substrato dentro de los seres vivos que permita la conservación y transmisión de información. Existe en este sentido una definición alternativa de la vida, basada en la posesión de información genética. En 1944, una serie de experimentos realizados por Oswald Avery, Colin MacLeod y Maclyn McCarty [1] probaron que los ácidos nucleicos son los encargados de llevar esta información y en 1953 Crick y Watson desvelaron la estructura en doble hélice del ADN (ácido desoxirribonucleico) [2]. Se ha descubierto una sustancia que han de poseer todos los organismos vivos y que ejerce un rol central en multitud de procesos: en numerosos virus y viroides es el ARN (ácido ribonucleico) ; en las células, el ADN. Aunque una definición exacta de la vida sigue siendo dificil, se conoce ya una máxima que han de cumplir todos los seres vivos en la Tierra: la posesión de información en forma de ácidos nucleicos.

Queda por tanto fuera de toda duda la importancia fundamental de los ácidos nucleicos y en concreto la del ADN, que es el ácido nucleico que ejerce la función de contenedor de información en los organismos complejos. Esto implica que el ADN determina la estructura, función y reproducción de las células de todos los protistas, bacterias, hongos, plantas y animales, entre los que se encuentra el ser humano. Para desempeñar su función, el ADN requiere de un entramado formado por enzimas, ARN y en menor medida otras moléculas, que se encargan de ejecutar las "órdenes" contenidas en el genoma a través de los procesos de transcripción, traducción y replicación. Estos procesos, aunque bastante complejos, pueden resumirse en unas pocas líneas: La transcripción es el proceso por el cual se transmite la información contenida en una porción de ADN a una hebra de ARN; para su realización una enzima abre la doble hélice de ADN y otra lee y copia la información creando una cadena complementaria[1] de ARN. La transcripción se realiza exclusivamente en el núcleo de la célula (cuando lo hay) y dependiendo del tipo de ARN producido[2], a la transcripción le seguirá inmediatamene la traducción. En la traducción se utiliza una hebra de ARN procedente de la transcripción para sintetizar una proteína; este proceso (posiblemente el más importante en la célula) se realiza fuera del núcleo con la participación de ribosomas y ARN adicional. Por último, en la replicación se copia la totalidad del ADN de la célula. Este proceso se realiza durante la mitosis y como la transcripción, requiere de una enzima que abra la cadena y otra que lea y copie la cadena, solo que en este caso la copia también es ADN.

Todos los procesos descritos en el párrafo anterior se conocen casi a la perfección desde un punto de vista biológico, sin embargo son todavía en parte un misterio desde el punto de vista físico o químico. ¿Por qúe deben estudiarse desde estos otros puntos de vista? La razón se encuentra

[1] En el ARN la base complementaria de la adenina es el uracilo y no la timina.

[2] Existen tres tipos de ARN con diferentes funciones: ARNt (de transferencia), ARNm (mensajero) y ARNr (ribosomal), sólo el ARNm contiene información para la síntesis de proteínas.

en la capacidad de los métodos de las ciencias más básicas (física y química) de predecir el comportamiento futuro de los sistemas que estudian. Aunque esta faceta también forma parte de la biología, esta ciencia pone el foco principal en la descripción cualitativa de los sistemas, con la finalidad de entender su funcionamiento; en cambio la física y la química se centran más en la descripción cuantitativa de los sistemas, con el fin de entender su funcionamiento y predecir su comportamiento. Nuestra motivación es, por tanto, alcanzar una comprensión más detallada del ADN que la proporcionada hasta ahora por la biología.

El estudio del ADN desde el punto de vista físico lleva varias décadas dando sus frutos, gracias a él sabemos que el ADN no es una molécula rígida, sino una cadena altamente dinámica, capaz de adoptar diferentes configuraciones, así como de abrirse espontáneamente debido a fluctuaciones térmicas. Es precisamente esta capacidad de abrirse, es decir, de un par de bases de romper los puentes de hidrógeno que las unen, lo que permite los procesos de replicación y transcripción, que ya se ha mencionado requieren de una enzima que abra la cadena de ADN. La apertura del ADN (denominada normalmente "desnaturalización" o *melting*) puede producirse por tanto de manera espontánea (más acusada cuanto mayor es la temperatura) o inducida, además puede afectar tan solo a un par de bases, a una serie de bases adyacentes (formando una "burbuja") o a la totalidad de la cadena.

La importancia de las burbujas ha llevado a la proposición de varios modelos físicos para el ADN capaces de describirlas, de entre los que destaca el modelo de Peyrard-Bishop-Dauxois [3] (en adelante PBD) por su sencillez. Este modelo mesoscópico unidimensional[3] es capaz de describir correctamente el comportamiento del ADN tanto en condiciones estáticas como dinámicas, así como la formación de burbujas y la desnaturalización.

A este modelo pueden añadírsele modificaciones que permitan una representación más fiel de la realidad. Una de ellas es la inclusión de un término en el hamiltoniano del sistema que simule la interacción con el disolvente (que es generalmente agua). Este término, en forma de barrera de potencial, permite una mejor descripción de las burbujas y la desnaturalización del ADN [4]. Otra posible mejora es la adición de una partícula con movimiento browniano que simule el comportamiento de una enzima y su interacción con el ADN [5]. Esta partícula permite una determinación bastante acertada de los sitios de unión de las enzimas en el ADN y el cambio en el comportamiento del propio ADN debido a estas enzimas. También muy interesante es la modificación de uno de los términos del hamiltoniano, el llamado *stacking*, que describe la interacción entre pares de bases adyacentes, haciéndolo dependiente de la secuencia [6]. Esto permite una reproducción muy precisa de las temperaturas de melting para cualquier tipo de cadena de ADN, incluso las muy ricas en un solo tipo de par de bases.

En este trabajo se presenta primero una modificación del modelo de Peyrard-Bishop-Dauxois que incluye tanto la barrera de potencial como el término de *stacking* modificado. El empleo de este modelo es exclusivo de este trabajo, ya que se basa en lo realizado por Tapia et al. en [5], añadiendo el término de *stacking* dependiente de la secuencia. A continuación, se pasará a la determinación de los parámetros empíricos del modelo comparando las temperaturas de desnaturalización obtenidas con las reales. Después se aplicará el modelo con los parámetros empíricos y la partícula a cadenas biológicas reales: simulando una cadena de un solo gen con distintas mutaciones se intentará reproducir el efecto de la mutación en la expresión de la proteína correspondiente. Para ello se estudiará el tiempo de permanencia de la partícula en las zonas de relevancia biológica de la cadena, así como la apertura media y la probabilidad de apertura de la cadena en dichas zonas.

[3]Para cada par de bases, es decir, a cada par de bases le corresponde únicamente una dimensión espacial.

2 Sistema biológico

En biología se suele considerar que existen cuatro tipos principales de moléculas orgánicas: ácidos nucleicos, proteínas, lípidos y glúcidos.

La función principal de los glúcidos es energética, la mayor parte de la energía que obtienen las células proviene del procesamiento de este tipo de moléculas, comúnmente llamadas hidratos de carbono. Los lípidos, en cambio, realizan una función más notable en el ámbito estructural. Estas moléculas forman las membranas de la célula y los orgánulos y además desempeñan funciones en el transporte de determinadas moléculas dentro de la célula. Por otro lado las proteínas son las que realizan la mayor parte de las tareas en un organismo vivo; se encargan de catalizar reacciones, transportar otras moléculas, comunicarse con el exterior de la célula, formar estructuras como el citoesqueleto... El "mapa" con la información para construir las proteínas está contenido en un ácido nucleico, el ADN (ácido desoxirribonucleico). Puesto que las proteínas se encargan de la gran mayoría de las funciones de la célula, así como de la organización y transporte de las otras moléculas orgánicas, la práctica totalidad de los procesos que ocurren en la célula están controlados, directa o indirectamente, por el ADN.

El ADN de la célula es un conjunto de dos polímeros entrelazados (Ver figura 2.2), cada uno de los cuales está compuesto por una serie de monómeros denominados "nucleótidos". A su vez estos monómeros se componen de tres elementos distintos: un grupo fosfato PO_4^-, un anillo azucarado y una base nitrogenada, que es un grupo orgánico complejo de uno o dos ciclos. La espina dorsal del ADN está formada por tanto por una sucesión de grupos fosfato y azúcares, que además está orientada, ya que cada grupo fosfato está unido por un lado a un carbono directamente en el anillo del azúcar (sentido 3') y por el otro a un carbono fuera del anillo en si (sentido 5'). Esto es importante, ya que en condiciones naturales cada una de las dos hebras que forman el ADN lleva un sentido distinto.

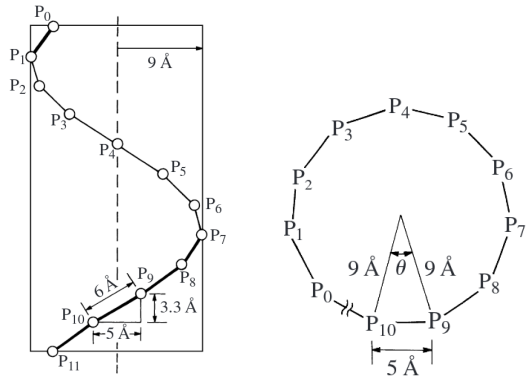

Figura 2.1: Configuración espacial esquemática del ADN. A la izquierda alzado esquemático, las líneas negras representan el azúcar, los puntos blancos los fosfatos. Pueden contarse 11 azúcares (por tanto 11 bases) en un giro. A la derecha planta esquemática, el ángulo θ corresponde a unos 33°.

La diferencia entre unos nucleótidos y otros radica entonces en el tipo de base nitrogenada que posean; existen cuatro posibilidades divididas en dos grupos: adenina (A) y guanina (G) que poseen dos anillos cíclicos y pertenecen al grupo de las purinas, y citosina (C) y timina (T) que poseen solo uno y pertenecen al de las pirimidinas.

El ADN puede adoptar diferentes configuraciones en función de la tensión a la que esté sometido o de la polaridad del disolvente en el que se encuentre. Esta última dependencia se hace evidente cuando se tiene en cuenta el carácter hidrofóbico (apolar) de las bases nitrogenadas y el hidrofílico (polar) de la cadena de azúcar-fosfato. La configuración natural en un medio acuoso (por tanto polar) como es el interior de la célula, será una en la cual las bases nitrogenadas entren en contacto con el agua en la menor medida posible y la cadena azúcar-fosfato quede expuesta en la mayor medida posible. Esto implica una estructura en la que las bases nitrogenadas quedan lo más solapadas posible unas con otras y situadas en el interior de la molécula, con la cadena azúcar-fosfato rodeándolas en el exterior. Si se tienen en cuenta todas las rigideces asociadas a los enlaces covalentes y la repulsión electrostática entre los átomos a poca distancia, se encuentra que esta estructura es una doble hélice con un giro de unos 33° y una distancia de unos 5Å entre una base y su contigua (Ver figura 2.1).

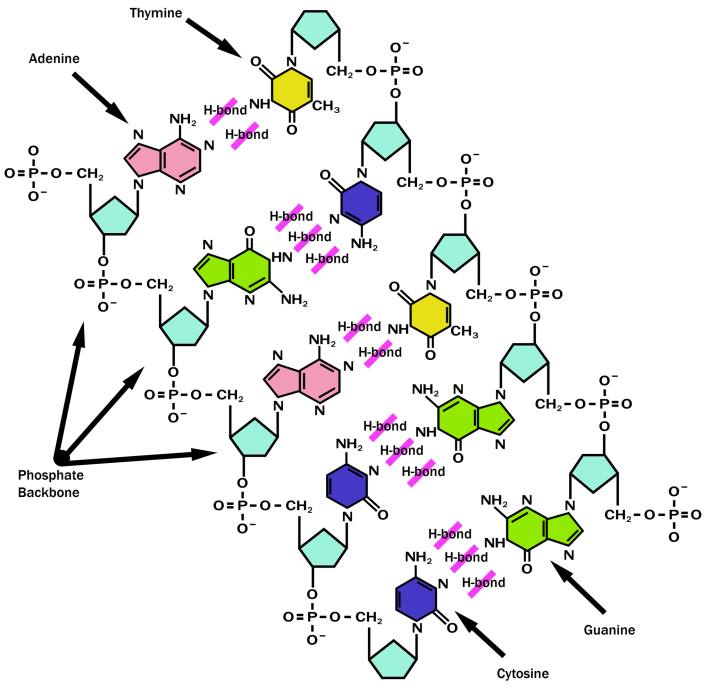

Figura 2.2: Representación esquemática del ADN. Pueden observarse las composiciones de las diferentes bases nitrogenadas, así como la de los azúcares. Solo los pares A-T y C-G son posibles, con 2 y 3 puentes de hidrógeno respectivamente. La parte señalada como *Phosphate Backbone* es la parte hidrofílica de la cadena, mientras que las bases nitrogenadas componen la parte hidrofóbica.

Las principales interacciones que ocurren dentro del propio ADN pueden agruparse en dos tipos:

- Puentes de hidrógeno: se forman entre cada par de bases opuestas A-T y G-C (el resto de combinaciones no se dan de forma natural). Si se trata del par de bases A-T se forman dos enlaces de hidrógeno y si se trata en cambio del par G-C se forman tres. Esto conlleva que en función de la secuencia haya zonas más difíciles de abrir que otras, lo que tiene una importancia biológica esencial. Además, puesto que los puentes de hidrógeno son mucho más débiles que los enlaces covalentes, la cadena se abre ante fluctuaciones térmicas sin que se modifique la estructura de cada polímero.

- Interacción de *stacking*: es la interacción entre pares de bases adyacentes. Se debe parcialmente al solapamiento de orbitales π de los electrones de las bases y parcialmente debido a

interacciones hidrofóbicas. Es una interacción muy compleja y que depende de la secuencia, lo que justifica la modificación del modelo PBD que haremos más adelante.

Este breve capítulo no es sino una introducción a la física y el funcionamiento del ADN. Para un estudio más exahustivo se recomienda consultar [7] y en especial el apéndice 2, que hacen referencia al *stacking*.

3 Modelo físico

¿Cómo podemos establecer un modelo matemático que sea capaz de reproducir las propiedades y el comportamiento del ADN observados en los experimentos?

Para responder a esta pregunta primero debemos encontrar la escala que mejor represente las características del ADN que queremos ver reflejadas en nuestro modelo. Si solo queremos un modelo que describa correctamente las curvas de fuerza-elongación del ADN para fuerzas bajas, podemos ignorar por completo la estructura interna de la molécula y representarla únicamente como una sucesión de monómeros unidos por muelles. Sin embargo, con este modelo, llamado Worm Like Chain (WLC) [8], no estaríamos diferenciando realmente entre un polímero cualquiera y el ADN y apenas podría obtenerse información de su función biológica.

Podría entonces pensarse que para obtener la mayor cantidad de información relativa al funcionamiento del ADN habría que estudiarlo a escala atómica. Efectivamente, un estudio a esta escala revelaría gran cantidad de información, aunque a un coste computacional enorme e impidiendo cualquier estudio de tipo analítico, dada la complejidad de este método. Pero la función del ADN no depende de la trayectoría exacta de cada átomo de carbono, hidrógeno, nitrógeno, oxígeno y fósforo que lo componen; tampoco de las fuerzas y distancias exactas entre ellos. La función del ADN depende de las bases nitrogenadas: cuales son y como y cuando se abre cada par de ellas. Por tanto, a pesar de que un modelo a escala atómica describe también las trayectorias de las bases y es más preciso, produce además gran cantidad de datos que no son de utilidad y requiere de capacidades de computación tan grandes que impiden observar efectos colectivos, de gran importancia en la función biológica.

La escala adecuada ha de ser entonces una intermedia, que utilice la base nitrogenada como elemento fundamental. Aunque a priori más sencilla que la escala atómica, a diferencia de esta no está basada en interacciones de valor conocido (como los potenciales entre distintos átomos) y requiere por tanto de ajuste experimental.

A continuación se presentan una serie de modelos, cada uno basado en el anterior, que desembocan en el que se utiliza en los estudios realizados en este trabajo.

3.1 Modelo de dos estados

La forma más sencilla posible de modelizar el ADN a la escala de la base nitrogenada es mediante un sistema de dos estados: si un par de bases se encuentra abierto valdrá 1, si se encuentra cerrado valdrá 0. De esta forma ponemos en el centro de todo la verdadera magnitud que nos interesa estudiar, la apertura de las bases.

El principal problema de este modelo es que, al perder todos los valores intermedios de apertura, no se pueden reproducir varios fenómenos no lineales esenciales observados en el ADN, ya que el modelo de Ising unidimensional no tiene transición de fase. Para solucionar este problema podemos modificar ligeramente el modelo, haciéndolo más complejo, pero entonces se requiere de multitud de parámetros que han de ser ajustados a partir de experimentos. El modelo es, en cualquier caso, demasiado sencillo para reproducir de manera fiel los resultados experimentales.

3.2 Modelo de Peyrard-Bishop

Este modelo se puede ver como el siguiente paso a uno de tipo dos estados; cada par de bases ya no puede estar únicamente abierta o cerrada, ahora están descritas por un grado de libertad continuo

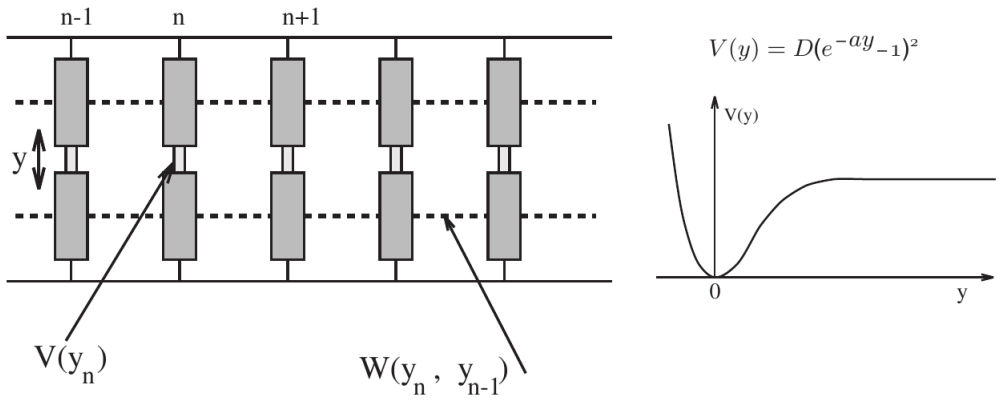

Figura 3.1: A la izquierda esquematización del modelo de Peyrard-Bishop. A la derecha representación gráfica del potencial de Morse.

que representa la distancia entre las bases del par, y_n. Un par de bases se halla completamente cerrado si $y_n = 0$ y se considera abierto cuando y_n alcanza cierto valor umbral; el parámetro y_n puede alcanzar incluso valores negativos, si las bases se acercan más allá de su distancia de equilibrio.

El hamiltoniano que describe el sistema es el siguiente (Ver figura 3.1):

$$H = \sum_n \frac{p_n^2}{2m} + W\left(y_n, y_{n-1}\right) + V(y_n) \qquad con\, p_n = \frac{dy_n}{dt} \tag{3.1}$$

con m la masa reducida del par de bases.

El potencial $V\left(y\right)$ describe la interacción entre las bases del mismo par. Su origen físico se encuentra principalmente en las fuerzas de unión que suponen los puentes de hidrógeno que se forman entre dichas bases. Es un potencial tipo Morse, un ejemplo típico de potencial de enlace químico y viene dado por la expresión

$$V\left(y\right) = D\left(e^{-ay_n} - 1\right)^2 \tag{3.2}$$

con D la energía de disociación del par y a un parámetro de dimensiones de inversa de distancia, que determina la anchura del pozo. Las características principales de este potencial son las siguientes:

- Para $y_n < 0$ presenta una interacción repulsiva muy fuerte, lo que se corresponde con la química real del enlace.
- Para $y_n = 0$ tiene una posición de equilibrio
- Para $y_n \to \infty$ se vuelve plano, dando por tanto una fuerza entre las bases que tiende a cero conforme estas se alejan. A consecuencia de este rasgo, la disociación completa del par de bases se hace posible.

Por su parte, el potencial $W\left(y_n, y_{n-1}\right)$ describe la interacción entre bases adyacentes, este término hace referencia al conjunto de interacciones denominadas *stacking*. Ya se ha mencionado que su origen es complejo, parcialmente debido a interacciones hidrofóbicas y parcialmente a la rigidez de la cadena, así como al solapamiento de los orbitales electrónicos de las bases contiguas. Dada su complejidad, una forma de abordarlo es mediante un desarrollo en serie, de manera que la expresión estable más simple que podemos obtener es un potencial armónico:

$$W\left(y_n, y_{n-1}\right) = \frac{1}{2}K\left(y_n - y_{n-1}\right)^2 \tag{3.3}$$

Con una elección adecuada de parámetros, este modelo permite reproducir las temperaturas de desnaturalización (*melting*) del ADN con suficiente precisión. Sin embargo, la transición que se predice es bastante suave, lo que contrasta con las curvas experimentales de desnaturalización térmica, que son muy abruptas.

3.3 Modelo Peyrard-Bishop-Dauxois (PBD)

Para reproducir mejor las curvas de desnaturalización, Peyrard, Bishop y Dauxois introdujeron una pequeña modificación en el potencial de *stacking*. Se dieron cuenta de que un potencial de *stacking* de tipo armónico implicaba una fuerza muy grande en las bases adyacentes cuando un par de bases se abría; debido a esto la cadena de ADN se iba abriendo poco a poco hasta hacerlo completamente cuando se formaba una burbuja. Si en cambio se utilizan dos constantes de acoplo, una cuando ambos pares de bases (y_n e y_{n-1}) están cerrados y otra cuando al menos uno está abierto, el resultado es bien distinto. Añadiendo esto al potencial, el término de *stacking* queda:

$$W(y_n, y_{n-1}) = \frac{1}{2}K\left(1 + \rho \cdot e^{(-\alpha(y_n + y_{n-1}))}\right)(y_n - y_{n-1})^2 \tag{3.4}$$

Por tanto habrá una constante $K' = K(1+\rho)$ cuando ambos pares estén cerrados y otra constante $K' = K$ cuando al menos uno esté abierto. Esto además se corresponde con lo esperado desde un punto de vista físico-químico: los electrones de los orbitalesπ de las bases adyacentes se encuentran solapados cuando los pares están cerrados; pero cuando uno de ellos se abre, sus bases rotan respecto al eje marcado por la columna de azúcar-fosfato, y el solapamiento con los orbitales π de las bases adyacentes disminuye. Ya se ha mencionado que la interacción de *stacking* depende en parte de este solapamiento, luego el cambio en el modelo tiene sentido desde este punto de vista. Otro de los elementos que influyen en el término de *stacking* es la rigidez de la cadena. Con el nuevo modelo, cuando un par de bases se abre, se reduce la rigidez de la cadena, que pasa a tener más libertad de movimiento y por tanto más entropía. Conforme se van formando burbujas cada vez más grandes, el incremento en la entropía de cada par de bases es cada vez mayor y la apertura de la cadena se acelera. Es por eso (cualitativamente) que con el nuevo modelo la curva de desnaturalización es mucho más abrupta.

3.4 Barrera

Al igual que el potencial $W(y_n, y_{n-1})$ armónico imposibilitaba reproducir correctamente las curvas de desnaturalización observadas en los experimentos, el potencial $V(y_n)$ tampoco predice adecuadamente algunos rasgos observados en los mismos: el tiempo que un par de bases permanece abierto entre dos fluctuaciones que lo abren y el tiempo que permanece cerrado después de la segunda fluctuación, son mucho menores en las simulaciones que en los experimentos [9]. El siguiente paso fue por tanto la modificación del término $V(y_n)$ para corregir estos errores [10], modificación que de nuevo puede justificarse con argumentos físicos.

El mismo efecto entrópico que influye en el potencial de *stacking* debe tenerse en cuenta también en la interacción dentro de cada par. La ganancia de grados de libertad y por tanto de entropía, al abrirse la cadena, dificulta que el par se cierre de nuevo; este efecto ha de ocurrir independientemente de lo que pase en los pares de bases adyacentes, así que debe añadirse también en el potencial $V(y_n)$. Por otro lado se debe tener en cuenta la interacción de las bases nitrogenadas con el disolvente. La disociación de un par significa esencialmente la rotura de los puentes de hidrógeno que mantenían ambas bases unidas. Puesto que las cadenas de ADN se encuentran en el medio biológico rodeadas de agua, las bases pueden volver a formar estos puentes de hidrógeno para reducir su energía, solo que esta vez con las moléculas de agua. El cerrado de la cadena implicará ahora la superación de una barrera de energía, ya que para volver a formar el enlace entre bases, primero habrán de romperse los puentes de hidrógeno con el agua.

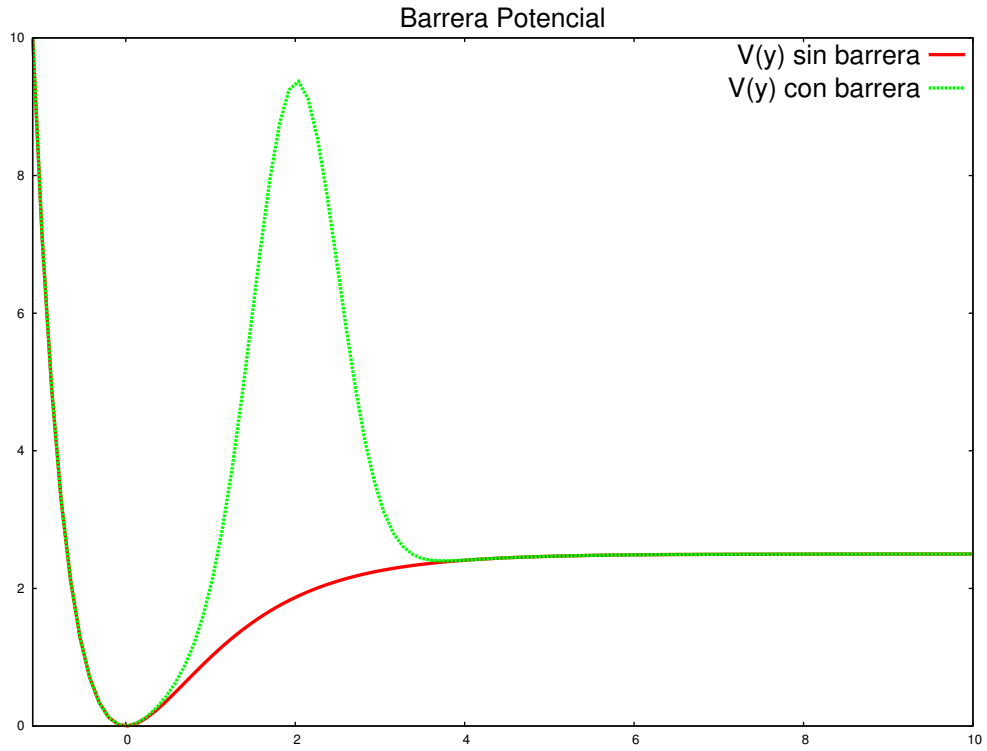

Figura 3.2: Potencial $V(y_n)$ con y sin barrera. La barrera está centrada en la posición que consideramos que marca la apertura del par.

Existen diversas modificaciones que se pueden utilizar para incluir estos fenómenos [11], en nuestro caso el potencial será de la forma

$$V(y_n) = D\left(e^{-ay_n} - 1\right)^2 + G \cdot e^{-(y_n - y_o)^2/b} \tag{3.5}$$

es decir, se utiliza una barrera gaussiana de amplitud G y anchura b, centrada en y_o (Ver figura 3.2). El efecto es claro, se dificulta tanto la apertura como el cerrado de la cadena, lo que dará tiempos de vida para las burbujas más acordes con los experimentos.

Hasta ahora no nos hemos preocupado de los valores que deben tener los parámetros para corresponderse con la realidad. Tampoco se ha hecho ninguna distinción entre los parámetros que afectan al par de bases A-T y los que afectan al par G-C. Aunque ya se ha mencionado que en los modelos mesoscópicos los valores de los parámetros deben obtenerse empíricamente, ajustándolos hasta reproducir los experimentos, existen ciertas relaciones entre ellos que se deben cumplir de acuerdo con el modelo.

La primera relación que debe cumplirse es que la unión entre el par G-C, que tiene 3 puentes de hidrógeno, ha de ser una vez y media más fuerte que la unión entre el par A-T, por tanto: $D_{G-C}/D_{A-T} = 3/2$ y $a_{C-G}/a_{A-T} = 3/2$. Esta proporción es la usada en la mayoría de trabajos relacionados con el tema. Por otro lado, de acuerdo a lo visto en los artículos que incluyen barrera gaussiana [4], un conjunto de valores razonables para ella es: $G = 3D$, $y_o = 2/a$ y $b = 1/2a^2$.

3.5 *Stacking* dependiente de la secuencia

La dependencia en la secuencia del término de *stacking* es la principal modificación que se realiza en este trabajo. El objetivo es conseguir resultados aún más cercanos a los observados experimentalmente y de nuevo, como en las modificaciones anteriores, el cambio que se introduce puede justificarse con argumentos cualitativos.

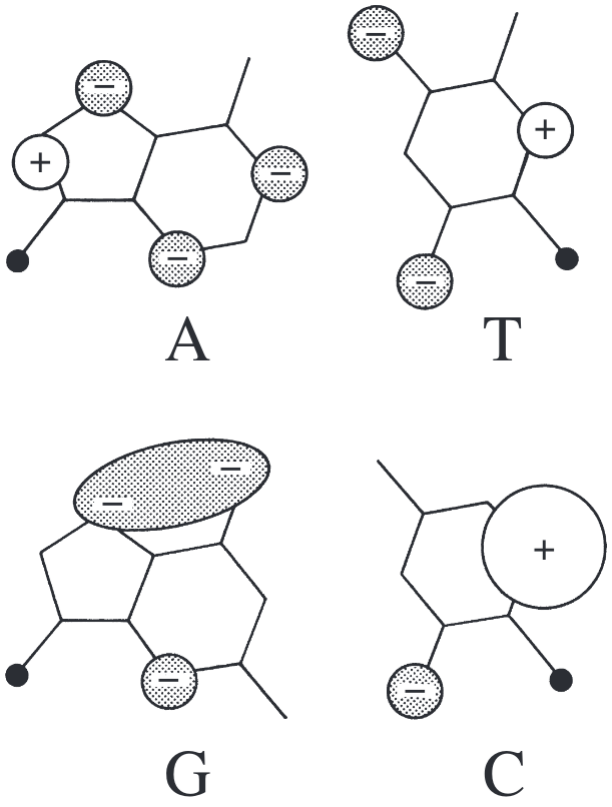

Figura 3.3: Representación esquemática de las regiones con carga parcial de las bases nitrogenadas. Se aprecia que en la guanina y en la citosina se encuentran las regiones más cargadas.

Ya se han explicado en apartados anteriores las principales interacciones que intervienen en el *stacking*; dada la importancia central de este término en nuestro estudio, hacemos un compendio con ánimo clarificador:

- Carácter hidrofóbico de las bases nitrogenadas: sabemos que en el agua se forman y rompen puentes de hidrógeno entre los átomos de H y los de O constantemente. Cuando le añadimos ADN al agua, dado que los pares cerrados de bases nitrogenadas no son capaces de formar puentes de hidrógeno (ya están formados todos los posibles en la unión entre las bases), las bases tienden a solaparse lo máximo posible para minimizar la entrada en contacto con los átomos del agua, que por tanto estarán junto a otros átomos de agua en mayor medida, lo que permitirá que se formen más puentes de hidrógeno y disminuya la energía total.

- Solapamiento de electrones: las bases nitrogenadas pueden considerarse, en primera aproximación y desde un punto de vista meramente eléctrico, como "sandwiches" en los que el pan está hecho de cargas negativas y el contenido es una enorme carga positiva. Esta analogía funciona realmente para cualquier molécula plana, ya que los electrones siempre van a estar en el exterior, rodeando el núcleo de los átomos. El principal efecto es repulsivo, pero es de menor orden que la interacción hidrofóbica en condiciones normales, de ahí que se favorezca normalmente un mayour solapamiento.

- Rigidez de la cadena y entropía: la cadena azúcar-fosfato es bastante rígida y no permite ciertos grados de libertad en el movimiento de las bases, lo que les fuerza a adoptar ciertas posiciones. Aún así cada par de bases tiene 4 grados de libertad y son precisamente estos grados de libertad los que complican tanto la obtención de un potencial de *stacking* simple y que refleje bien las interacciones reales entre bases. Además estos grados de libertad

aumentan cuando se rompen los enlaces entre pares de bases, produciendo una ganancia en entropía.

Hasta aquí no hay ninguna interacción que dependa del tipo de bases de las que estemos hablando, para encontrarla hay que hilar un poco más fino, dentro de la distribución de cargas de cada base.

- Distribución de cargas: fijándonos un poco más en la distribución de cargas en las bases nitrogenadas, observamos que no es homogénea, ya que cada base contiene varios átomos de electronegatividad elevada (oxígeno y nitrógeno) y otros de electronegatividad baja (hidrógeno). El efecto es que hay más cargas negativas rodeando los átomos de O y N y más carga positiva cerca de los de H[1]. Desde el punto de vista del "sandwich" habrían aparecido ahora algunos agujeros en el pan, mientras que en otras zonas se habría hecho más grueso. La consecuencia de esta distribución de cargas es que las bases tenderán a orientarse y colocarse de tal forma que las zonas con carga negativa de una base coincidan con las cargadas positivamente en las bases adyacentes. Puesto que diferentes bases tienen diferentes distribuciones de carga, esta interacción depende de la secuencia (Ver figura 3.3).

Este último efecto sí que es dependiente de la secuencia, de forma que queda justificada desde un punto de vista físico la introducción de esta característica en nuestro modelo.

Por sencillez, el único parámetro que haremos dependiente de la secuencia dentro del término de *stacking* será K. Teniendo todo en cuenta obtenemos 10 posibles valores, según los pares de bases adyacentes sean: AA/TT, AT/TA, TA/AT, GG/CC, GC/CG, CG/GC, CA/GT, GA/CT, AG/TC y AC/TG. Puede parecer extraño que aparezcan tanto AT/TA como TA/AT, pero hay que tomar en consideración que la direccionalidad de la cadena va a influir en la interacción entre bases; si una base está unida al carbono del anillo de azúcar, es capaz de adoptar unas posiciones distintas a si lo está al carbono fuera del anillo, además no se produce la misma ganancia en estabilidad en una base que en otra cuando están situadas la una junto a la otra. Lo mismo vale para GC/CG - CG/GC, AG/TC - GA/CT y AC/TG - CA/GT.

[1] Esto es una simplificación, ya que la distribución espacial de los electrones depende de la molécula concreta que estén formando.

4 Métodos

Para el cálculo de la dinámica de nuestro sistema nos decantamos por un método de dinámica molecular, en el que calculamos las trayectorias de cada uno de los pares de bases del sistema. El baño térmico, por su parte, estará representado por una fuerza aleatoria con distribución gaussiana, lo que resulta en una expresión para la dinámica de tipo ecuación de Langevin. Aplicada a nuestro caso este tipo de dinámica conlleva una ecuación diferencial de la forma:

$$m_b \frac{dy_n^2}{dt^2} + m_b \gamma \frac{dy_n}{dt} + \frac{\partial \left[V\left(y_n\right) + W\left(y_n, y_{n-1}\right) + W\left(y_n, y_{n+1}\right) \right]}{\partial y_n} = \xi_n\left(t\right) \tag{4.1}$$

donde m_b es la masa reducida del par de bases, γ es el *damping* del sistema y $\xi\left(t\right)$ es el ruido térmico, que cumple: $\left\langle \xi_n\left(t\right) \right\rangle = 0$ y $\left\langle \xi_n\left(t\right) \xi_k\left(t'\right) \right\rangle = 2 m_b \gamma k_B T \delta_{nk}\left(t - t'\right)$ con T la temperatura del sistema. El método de integración numérica utilizado para resolver la ecuación de Langevin es el de Runge-Kutta estocástico [12, 13] con condiciones de contorno periódicas.

También es de interés introducir la definición de dos magnitudes que van a ser importantes en nuestro estudio del sistema; la probabilidad de apertura $p_n\left(y_{o_n}\right)$ y la apertura media de un par de bases $\left\langle y_n \right\rangle$:

$$\left\langle y_n \right\rangle = \frac{1}{t_s} \sum_{t}^{ts} y_n \tag{4.2}$$

$$p_n\left(y_{o_n}\right) = \frac{1}{t_s} \sum_{t}^{ts} \Theta\left(y_n\left(t\right) - y_{o_n}\right) \tag{4.3}$$

donde t_s es el tiempo de simulación, y_{o_n} es la posición a partir de la cual consideramos un par de bases abierto y $\Theta\left(x\right)$ es la función escalón, que está definida como: $\Theta\left(x\right) = 0 \Leftrightarrow x < 0$ y $\Theta\left(x\right) = 1 \Leftrightarrow x \geq 0$.

Para reducir el coste computacional de nuestras simulaciones realizamos una adimensionalización de las ecuaciones del sistema, lo que da como resultado la definición de una serie de variables adimensionales que se presentan a continuación. La apertura de las bases (y) pasa a definirse en unidades adimensionales como $\tilde{y} = y \cdot a_{A-T}$; el tiempo t, por su parte, se expresa en unidades adimensionales como $\tilde{t} = t \cdot \left(\frac{D_{A-T} \cdot a_{A-T}^2}{m_b} \right)^{\frac{1}{2}}$ y la unidad adimensional de energía viene dada por la relación $\tilde{E} = E / D_{A-T}$. A partir de estas definiciones podemos obtener la expresión en unidades adimensionales de cualquiera de las variables del sistema: la masa adimensional se obtiene como $\tilde{m} = m / m_b$, las constantes del término de *stacking* se obtienen a partir de $\tilde{K} = \frac{K}{D_{A-T} \cdot a_{A-T}^2}$, el *damping* a través de $\tilde{\gamma} = \frac{\gamma}{a_{A-T}} \sqrt{\frac{m_b}{D_{A-T}}}$, el parámetro α como $\tilde{\alpha} = \alpha / a_{A-T}$ y la temperatura como $\tilde{T} = \frac{k_B T}{D_{A-T}}$.

5 Ajuste de los parámetros

Hasta el momento, solamente hemos expresado relaciones entre los valores de los parámetros a utilizar; esto es así porque para la optización del programa se utilizan parámetros adimensionales y para operar con ellos solo necesitamos conocer estas relaciones. Sin embargo, para que nuestro sistema pueda identificarse con una cadena real de ADN, debemos encontrar una serie de valores con significado físico para nuestros parámetros. Por otro lado, ni siquiera se han presentado relaciones entre los valores de los parámetros de *stacking*. En este apartado vamos a abordar ambas cuestiones.

Para encontrar los valores adecuados para los parámetros de *stacking* nos fijamos en los valores propuestos en artículos relacionados con el tema, en concreto en los propuestos por Alexandrov et al. [6], Weber et al. [10] y Singh et al. [11] . Todos estos datos se presentan en la tabla 5.1 junto a los de Tapia et al. [5].

Puesto que queremos añadir el término de *stacking* dependiente de la secuencia al modelo ya existente con barrera de Tapia et al., nuestra intención es mantener los valores que estos utilizan y cambiar tan solo el valor del parámetro k. Merece la pena hacer notar que los experimentos que intentan determinar la constante de *stacking* dependiente de la secuencia arrojan resultados que tienen una desviación estandar que puede llegar a ser del 50% de la diferencia entre el parámetro k de un par de bases y el de otro [14]. Por ello nos decantaremos por los valores obtenidos por Alexandrov et al., ya que son el resultado de un ajuste, de un modelo PBD muy similar al nuestro, a las temperaturas de melting del ADN (que se conocen con precisión).

Previo paso a la utilización de los nuevos valores para k, nos aseguramos de si el núcleo de nuestro programa funciona correctamente y/o de si los resultados de Alexandrov et al. son correctos. Para ello introducimos en nuestro programa sus valores para todos los parámetros y realizamos una serie de simulaciones sin barrera, es decir, con su modelo. Nuestro objetivo es comprobar si las temperaturas de *melting* (desnaturalización) experimentales para ciertas cadenas coinciden con las que obtenemos en nuestras simulaciones; estas temperaturas están recogidas en la tabla 1 de [6], aunque no se utiliza la nombrada como poly(AT), ante la imposiblidad de identificar de que secuencia se trata. El resultado es positivo, los valores coinciden perfectamente (resultados no mostrados en este trabajo).

Dado que hemos añadido una modificación al modelo y hemos introducido nuevos valores para 10 de los parámetros, es necesario comprobar si este todavía es capaz de predecir resultados experimentales, así como si los valores dimensionales asignados a sus parámetros se mantienen o han de ser modificados. El procedimiento será el mismo que el utilizado para comprobar el funcionamiento del programa, solo que esta vez se incluirá la barrera de potencial. Las cadenas a simular son cadenas muy sencillas y cortas, de tan solo 36 pares de bases y de secuencias repetidas de 2 ó 3 pares de bases; para identificarlas utilizaremos una notación en la que nombraremos tan solo el fragmento de secuencia que se repite en la hebra 5'-3' y el número de veces que lo hace. Con esta notación las cadenas simuladas son A_{36}, AG_{18}, AC_{18}, G_{36}, GC_{18} y AGC_{12}.

El primer paso en la búsqueda de la temperatura de melting para cualquier cadena es encontrar el rango en la temperatura adimensional en el cual tiene sentido que se encuentre. Si pensamos que los valores dimensionales de nuestros parámetros son correctos, entonces $D_{A-T} = 0.052\,eV$ y $D_{G-T} = 0.076\,eV$ y con estos valores obtenemos el valor dimensional de nuestra temperatura a través de la relación $T = \tilde{T} \cdot D_{A-T}/k_B$; de forma que una temperatura adimensional $\tilde{T} = 1$,

	Alexandrov[6]	Weber[10]	Singh[11]$^{(1)}$	Tapia[5]
$D_{A-T}\,(eV)$	0,05	0,031	0,064	0,052
$D_{G-C}\,(eV)$	0,075	0,073	0,096	0,076
$a_{A-T}\,(\text{Å}^{-1})$	4,2	0,029	4,2	4
$a_{G-C}\,(\text{Å}^{-1})$	6,9	0,1	6,3	6
$k_{A-A}\,(eV/\text{Å}^2)$	0,0236	0,025	0,0216	0,03
$k_{A-T}\,(eV/\text{Å}^2)$	0,0238	0,019	0,022	0,03
$k_{A-C}\,(eV/\text{Å}^2)$	0,0224	0,026	0,0238	0,03
$k_{A-G}\,(eV/\text{Å}^2)$	0,0236	0,024	0,0203	0,03
$k_{T-A}\,(eV/\text{Å}^2)$	0,0195	0,025	0,0223	0,03
$k_{C-A}\,(eV/\text{Å}^2)$	0,0255	0,033	0,0206	0,03
$k_{C-G}\,(eV/\text{Å}^2)$	0,0274	0,026	0,0246	0,03
$k_{G-A}\,(eV/\text{Å}^2)$	0,0188	0,028	0,0213	0,03
$k_{G-C}\,(eV/\text{Å}^2)$	0,025	0,034	0,028	0,03
$k_{G-G}\,(eV/\text{Å}^2)$	0,0194	0,02	0,0163	0,03
ρ	2	0	1	3
$\alpha(\text{Å}^{-1})$	0,35	0	0,35	0,8

Tabla 5.1: Compendio de diferentes valores para los parámetros del modelo PBD utilizados en distintos artículos. $^{(1)}$ Los valores del parámetro k en este artículo se obtienen de una media de los obtenidos en [6, 14, 15]. En nuestro caso haremos uso de los parámetros de Tapia et al. excepto en el caso del parámetro k, que obtendremos del artículo de Alexandrov et al.

que se corresponde con una energía $\tilde{E} = 1$ en unidades adimensionales, equivale a una temperatura dimensional $T = D_{A-T}/k_B = 0.052\,eV/8 \cdot 10^{-5}\,eV \cdot K^{-1} \simeq 603\,K$. Puesto que todas las temperaturas de *melting* de las cadenas a estudiar se encuentran entre los 300 y 400 K, lo lógico sería que se encontrasen en el intervalo $0.5 - 0.65$ adimensional, sin embargo a temperatura adimensional 0.5 se aprecia en las simulaciones que algunas cadenas están ya completamente desnaturalizadas. Esto implica que hemos de ajustar el valor dimensional de nuestros parámetros para que reproduzcan los resultados experimentales.

Lo que haremos será encontrar las temperaturas adimensionales de *melting* de todas las cadenas anteriores y representarlas frente a las reales; si la relación que observamos entre ellas es lineal, entonces la ecuación de la recta nos dirá la equivalencia dimensional de nuestra temperatura adimensional. Para la identificación de las temperaturas de desnaturalización nos centramos en el intervalo adimensional $0.4 - 0.6$, de acuerdo a lo visto para temperatura 0.5. En la figura 5.1 podemos ver como evoluciona la probabilidad media de apertura de la cadena en función de la temperatura para las distintas secuencias; se observa como a bajas temperaturas la probabilidad es baja y a cierta temperatura, cerca de la de *melting*, se dispara hasta llegar al máximo rápidamente. Según nuestro criterio, la temperatura de desnaturalización es aquella para la cual la probabilidad de apertura está justo a la mitad entre la máxima y la temperatura a la cual se produce el cambio brusco de pendiente. Echando un rápido vistazo a las figuras 5.1 y 5.2, vemos como el orden en el que se encuentran las temperaturas de *melting* adimensionales coincide con el real; de acuerdo a un modelo sin *stacking* dependiente de la secuencia, sería imposible que la temperatura de desnaturalización de la secuencia GC_{18} fuera distinta a la de la secuencia G_{36}, ya que ambas cadenas se componen de pares $G-C$ con tres puentes de hidrógeno. Sin embargo, dada la fuerte interacción de *stacking* entre las bases G y C, vemos como la secuencia GC_{18} requiere de una temperatura mayor para desnaturalizarse.

Una vez obtenemos las distintas temperaturas de desnaturalización realizamos un ajuste por

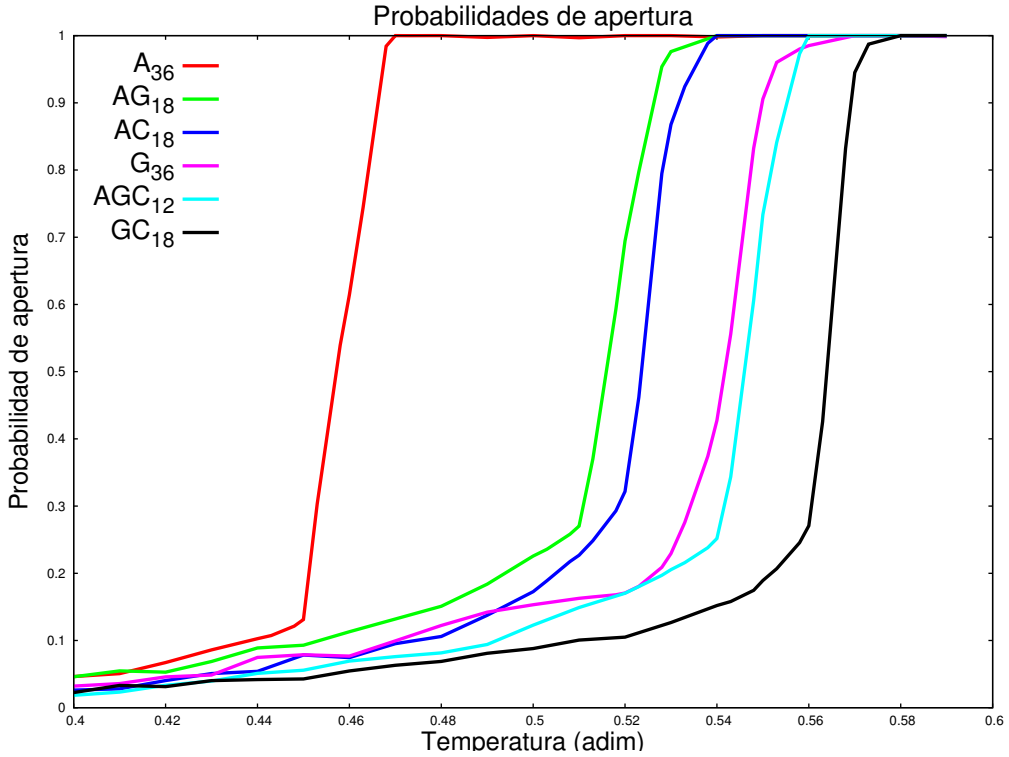

Figura 5.1: Probabilidad media de apertura de la cadena en función de la temperatura para las distintas secuencias estudiadas.

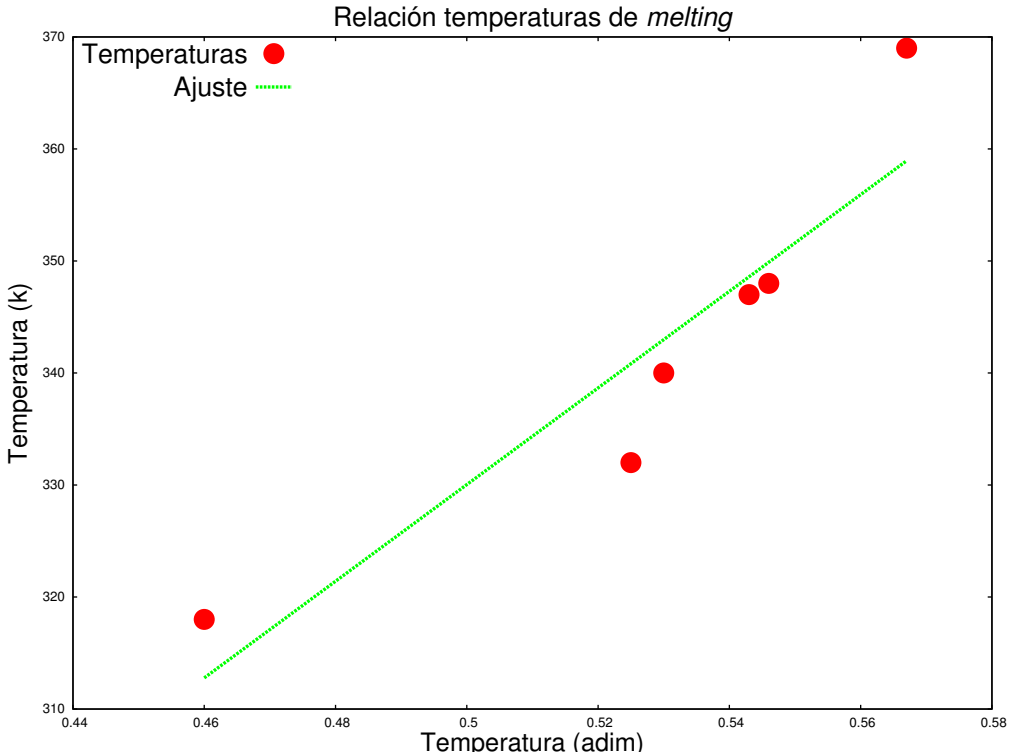

Figura 5.2: Relación entre las temperaturas adimensionales de desnaturalización obtenidas en las simulaciones y su equivalente dimensional obtenido experimentalmente. La temperatura adimensional de *melting* se obtiene como aquella para la cual la probabilidad de apertura se encuentra justo a la mitad entre la máxima y la temperatura a la cual se produce el cambio brusco de pendiente.

Valores de los parámetros utilizados							
$D_{A-T}\,(eV)$	0,05	$k_{A-A}\,(eV/\text{Å}^2)$	0,0236	$k_{T-A}\,(eV/\text{Å}^2)$	0,0195	$k_{G-C}\,(eV/\text{Å}^2)$	0,025
$D_{G-C}\,(eV)$	0,075	$k_{A-T}\,(eV/\text{Å}^2)$	0,0238	$k_{C-A}\,(eV/\text{Å}^2)$	0,0255	$k_{G-G}\,(eV/\text{Å}^2)$	0,0194
$a_{A-T}\,(\text{Å}^{-1})$	4	$k_{A-C}\,(eV/\text{Å}^2)$	0,0224	$k_{C-G}\,(eV/\text{Å}^2)$	0,0274	ρ	3
$a_{G-C}\,(\text{Å}^{-1})$	6	$k_{A-G}\,(eV/\text{Å}^2)$	0,0236	$k_{G-A}\,(eV/\text{Å}^2)$	0,0188	$\alpha(\text{Å}^{-1})$	0,8

Tabla 5.2: Valores de los parámetros que utilizaremos en nuestro estudio. Son una combinación de los de [5], con los valores para k de [6] y posteriormente ajustados a través de las temperaturas de *melting* experimentales de una serie de secuencias (Tabla 1 de [6]).

mínimos cuadrados a una recta (Ver figura 5.2) y obtenemos la siguiente ecuación:

$$T\,(K) = 431(\pm 91) \cdot \tilde{T}\,(adim) + 114(\pm 48) \tag{5.1}$$

De acuerdo a este resultado, una temperatura adimensional $\tilde{T} = 1$ equivale a una temperatura de $546 \pm 139\,K$, donde cabe destacar la gran desviación estandar observada. Con este resultado reajustamos los valores de D_{A-T} y D_{G-T} de acuerdo a la expresión $D_{A-T} = k_B T/\tilde{T}$ y obtenemos $D_{A-T} = 0.047 \pm 0.008\,eV$ y $D_{G-T} = 0.071 \pm 0.008\,eV$. Los valores para los parámetros que utilizaremos serán entonces los recogidos en 5.2.

6 Modelo interacción ADN-proteina

Como ya se ha dicho en el apartado 1, los principales procesos en los que partcipa el ADN requieren de una serie de proteínas que lo abran y lo lean. De entre todas ellas destaca la ARN-polimerasa, que es la encargada de crear la cadena de ARN complementaria a la de ADN durante la transcripción. Junto a esta proteína trabajan otras, llamadas TBP's (*transcription binding proteins*), cuya función es ayudar a la ARN-polimerasa a encontrar la posición adecuada para iniciar el proceso. También están implicadas en el proceso otra serie de proteínas (represores, activadores, correpresores y coactivadores) que regulan la intensidad de la actividad génica.

Lo increíble es que la célula es capaz de elegir qué gen, de entre las decenas de miles (en el ser humano) que contiene, es el que se ha de expresar. Este trabajo lo despempeñan las TBP's, que a través de un movimiento de difusión acaban encontrando una porción de ADN para la que presentan una geometría complementaria. Esta porción de ADN, llamada región específica, forma parte del gen que se desea expresar.

La principal cuestión es dilucidar de qué tipo es la difusión de las TBP's (el transporte activo es demasiado costoso como para ser una opción), si unidimensional o tridimensional. La estrategia óptima, de acuerdo a [16], es una combinación de ambas; la proteína se mueve a lo largo de una porción de ADN buscando su región específica, si no la encuentra pasa a soltarse y a continuación se une a otra porción cercana desde un punto de vista tridimensional, pero lejana en cuanto a la secuencia se refiere. El proceso implica por tanto en última instancia la difusión unidimensional, que es la que modelizaremos en este trabajo, basándonos en lo desarrollado en ese sentido por [5].

En su artículo, Tapia et al. representan la proteína como una partícula que se mueve a lo largo de la cadena, influida por el campo que esta genera en función de la separación entre sus bases. Para cuantificar el desplazamiento de la partícula añaden una nueva coordenada x al modelo, además incluyen su masa con el parámetro M_p. La ecuación diferencial que govierna la dinámica de la partícula es:

$$M_p \frac{d^2x}{dt^2} + M_p \gamma_p \frac{dx}{dt} + \frac{\partial V_{int}(x, y_n)}{\partial x} = \xi_p(t) \tag{6.1}$$

con γ_p es el *damping* de la partícula y $\xi_p(t)$ el ruido térmico, que cumple: $\langle \xi_p(t) \rangle = 0$ y $\langle \xi_p(t)\, \xi_p(t') \rangle = 2M_p \gamma_p k_B T \delta(t - t')$. Por otro lado V_{int} es el potencial de interacción, que está definido como:

$$V_{int}(x, \{y_i\}) = \sum_n V_n(x, y_n) = -\frac{B}{\sqrt{\pi\sigma^2}} \sum_n tanh(a y_n) \cdot e^{-(x-n)^2/\sigma^2} \tag{6.2}$$

donde B marca la amplitud de la interacción, a el rango de interacción con la separación del par de bases y σ el rango espacial de interacción con el ADN (Ver figura 6.1). En esta expresión, n es el índice que representa cada par de bases y x está normalizada de tal forma que sus valores enteros coinciden con las posiciones de estos pares, es decir, $x = 1$ es la posición en la que se encuentra el primer par de bases, que están separadas una distancia y_1; $x = 2$ la posición del segundo par de bases, separadas una distancia y_2 y así sucesivamente. Una vez queda claro el papel de los parámetros, esta expresión es sencilla de entender. En un determinado instante de tiempo la partícula se encuentra en la posición x y las bases se encuentran separadas unas distancias $\{y_n\}$; La partícula interaccionará con todos los pares de bases con una intensidad gobernada por la parte gaussiana de la expresión 6.2, por tanto interaccionará más fuertemente con los pares más cercanos y apenas interaccionará con los pares a distancia mayor de 2σ. Además, la interacción

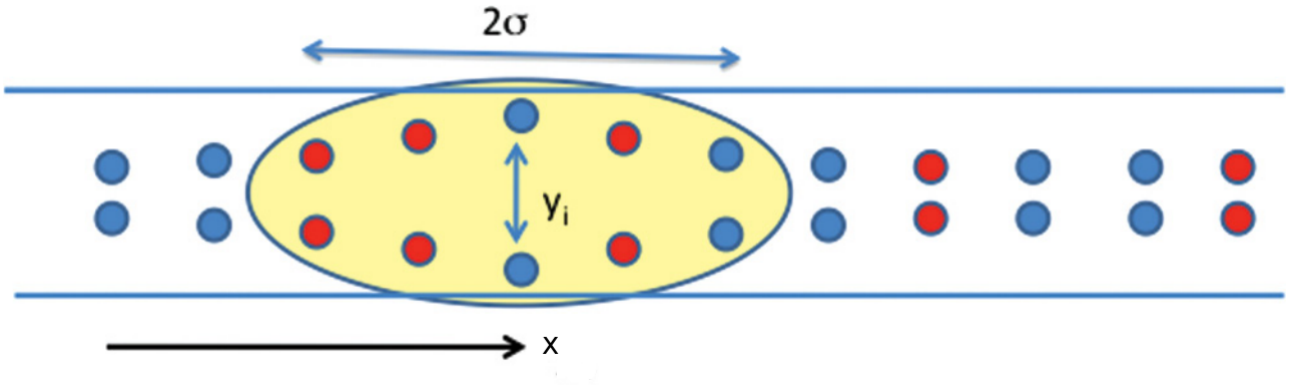

Figura 6.1: Ilustración esquemática del modelo de interacción ADN-proteína. Las bases nitrogenadas están representadas por los círculos azules y rojos, mientras que la elipse amarilla representa la partícula.

con cada par de bases viene dada por una tangente hiperbólica, por lo tanto será pequeña y lineal para valores bajos de y_n y crecerá rápidamente hasta saturarse.

Por otro lado, la partícula también tendrá influencia sobre el comportamiento de la cadena, por lo que la dinámica de esta última pasará a ser de la forma:

$$m_b \frac{dy_n^2}{dt^2} + m_b \gamma \frac{dy_n}{dt} + \frac{\partial \left[V\left(y_n\right) + W\left(y_n, y_{n-1}\right) + W\left(y_n, y_{n+1}\right)\right]}{\partial y_n} + \frac{\partial V_{int}\left(x, y_n\right)}{\partial y_n} = \xi_n\left(t\right) \quad (6.3)$$

La información principal que interesa obrener de la partícula es su trayectoria $x\left(t\right)$; a partir de ella es posible calcular las probabilidades de estancia $P\left(x\right)$ de la partícula en función de la posición que ocupa. Lo que nos permite dibujar el paisaje de energía libre:

$$F = -k_B T \cdot log\left(P\left(x\right)\right) \quad (6.4)$$

La simulación del movimiento de la partícula se realiza también con un algoritmo de Runge-Kutta estocástico.

7 Aplicación a una cadena real: R40 en *Synechocystis*

Una vez los parámetros del modelo están correctamente ajustados y la interacción ADN-proteína definida, pasamos a la aplicación del mismo a un caso real, una serie de promotores[1] alterados genéticamente para la bacteria *Synechocystis*. De acuerdo a [17], un pequeño cambio de tan solo 2 pares de bases en la secuencia de un promotor obtenido del "$\lambda P_L - derived\ BBa_R0040$" (en adelante $R40$) (Ver figura 7.1) produce grandes cambios en la expresión del gen al que precede, tanto en estado inducido como en estado reprimido[2]. En su artículo, Lindblad y Huang incluyen un estudio computacional de la secuencia a través del modelo PBD con resultados bastante limitados: aunque reproducen la mayor apertura de la cadena en ciertas secciones de relevancia biológica, no obtienen correlación alguna entre la probabilidad de apertura de la cadena en dichas secciones y la expresión del gen.

En este apartado intentamos mejorar los resultados computacionales de Lindblad y Huang utilizando nuestro modelo con *stacking* dependiente de la secuencia, barrera y partícula. Gracias a la inclusión del nuevo término de *stacking* somos capaces de distinguir entre más casos que [17]; donde ellos distinguen entre 4 casos nosotros distinguimos entre 16[3]. Además, la inclusión de la partícula permite que la influencia del pequeño cambio de los dos pares de bases se extienda con más fuerza al resto de la cadena: si la partícula pasa más o menos tiempo en la zona debido al cambio de las bases, pasará por fuerza menos o más tiempo en el resto de la cadena; sin embargo, sin partícula, la probabilidad de apertura de una zona de la cadena lejana a la de la modificación se verá afectada muy poco, ya que el efecto se ha de transmitir a través de todos los pares de bases entre ambas zonas.

En la figura 7.1 se muestran las secuencias de todos los promotores utilizados en nuestro estudio, esto incluye tanto el $R40$ como los 16 promotores $L01 - L16$. Además se incluye una tabla con los valores experimentales para la expresión del gen en estado reprimido e inducido cuando se utilizan estos promotores en la bacteria *Synechocystis*. Como ya se ha dicho, la cadena $R40$ es la que se toma como base sobre la que se realizan las mutaciones; primero se modifica hasta obtener la cadena $L31$ y sobre esta se realizan las pequeñas mutaciones de tan solo dos pares de bases que llevan a las secuencias $L01 - L16$. Para nuestras simulaciones no tomaremos el total de la cadena $L31$ tal y como aparece en la figura 7.1, excluiremos algunos elementos lejanos a las zonas relevantes biológicamente. Nuestra $L31$ contendrá los elementos: *UP-element, -35, spacer, -10* y la secuencia de 6 bases en sentido *downstream* al elemento *-10* como aparece en la figura 7.1; además se incluirán en nuestra cadena los elementos: *8-bp BioBrick scar, 10-bp ribosome binding site RBS*, 6-bp BioBrick scar* y el codon inicial ATG, tal y como se explica en la figura 1 de [17]; así como una cola de 20 pares de bases CG para aislar la secuencia (recordemos que se aplican condiciones de contorno periódicas), lo que da un total de 101 pares de bases.

El primer paso que daremos será encontrar la temperatura adecuada a la que simular nuestras cadenas de promotores. En cada caso nos interesa que la cadena se abra con facilidad, pero que no esté en proceso de desnaturalización, es decir, nos interesa que se encuentre a una temperatura a la que pueda realizar su función biológica de forma normal. Tras un pequeño estudio elegimos

[1] Región del genoma que precede a un gen y cuya función es regular la expresión de dicho gen.

[2] El estado inducido implica una alta concentración de enzimas que favorecen la expresión del gen. Mientras que el estado reprimido implica una concentración baja

[3] Dado su término de *stacking* uniforme, para ellos no hay diferencias entre la secciones de secuencia AT, TA, AA y TT; así como entre GC, CG, CC y GG.

Promoter	its downstream 8 bases[a]	Induced	Repressed	Induction[c]	Pattern[a]
L15	TATAATGGACACTA	20.1 ± 0.1	0.243 ± 0.003	79 ± 1	A ⟨17.9 ± 0.2⟩
L07	TATAATGGTCACTA	19.8 ± 0.1	0.289 ± 0.003	65 ± 1	
L05	TATAATGCTCACTA	16.3 ± 0.1	0.198 ± 0.003	78 ± 1	
L13	TATAATGCACACTA	15.5 ± 0.1	0.219 ± 0.003	67 ± 1	
L03	TATAATGGCCACTA	19.2 ± 0.1	0.220 ± 0.003	83 ± 1	B ⟨16.3 ± 0.2⟩
L11	TATAATGGGCACTA	19.1 ± 0.1	0.236 ± 0.003	77 ± 1	
L09	TATAATGCGCACTA	15.6 ± 0.1	2.88 ± 0.01	5.15 ± 0.01	
noLVA_L09[e]	TATAATGCGCACTA	11.53 ± 0.05	0.545 ± 0.004	20.1 ± 0.1	
L01	TATAATGCCCACTA	11.30 ± 0.05	0.235 ± 0.003	46 ± 1	
L02	TATAATGTCCACTA	17.7 ± 0.1	0.214 ± 0.003	79 ± 1	C ⟨15.9 ± 0.2⟩
L10	TATAATGTGCACTA	17.6 ± 0.1	0.235 ± 0.003	71 ± 1	
L04	TATAATGACCACTA	12.3 ± 0.1	0.201 ± 0.003	58 ± 1	
L12	TATAATGAGCACTA	0.043 ± 0.003	0.022 ± 0.003	1.9 ± 0.3	
L16	TATAATGAACACTA	17.5 ± 0.1	0.240 ± 0.003	69 ± 1	D ⟨15.1 ± 0.2⟩
L06	TATAATGTTCACTA	15.8 ± 0.1	0.219 ± 0.03	69 ± 1	
L08	TATAATGATCACTA	14.9 ± 0.1	0.255 ± 0.003	56 ± 1	
L14	TATAATGTACACTA	12.1 ± 0.1	0.304 ± 0.003	38.0 ± 0.4	
L21	TATAAT*GGGA*GCTA	41.6 ± 0.2	40.1 ± 0.2	0.986 ± 0.003	E
L22	GATACT*GGGA*GCTA	0.378 ± 0.003	0.023 ± 0.004	16 ± 2	F
P_{trc1O}	TATAATGTGTGGAA	46.4 ± 0.2	43.9 ± 0.2	1.007 ± 0.003	G
L31	TATAATGTGTGGTA	3.08 ± 0.01	0.169 ± 0.003	17.3 ± 0.3	H
R40	GATACTGAGCACTA	0.272 ± 0.003	0.082 ± 0.003	3.2 ± 0.1	I
J23101	TATTATGCTAGCTA	4.57 ± 0.02	4.461 ± 0.02	0.974 ± 0.003	n.s.
rnpB	CACACTAGAAAAAT	1.00 ± 0.01 (427 ± 2)	1.00 ± 0.01 (448 ± 2)	0.95 ± 0.01	

[a], The sequences of listed promoters, excluding P_{trc1O}, J23101, and *rnpB* promoters, are identical except the region shown in this table. The consensus −10 element TATAAT is underlined. The transcription start site (TSS) is boxed. The promoter sequences are detailed (Figure 1). [b], The induced and repressed conditions are *Synechocystis* sp. strain ATCC27184 (i.e. glucose-tolerant *Synechocystis* sp. strain PCC 6803) cells in LAHG growth condition treated with and without 100 ng/mL aTc, respectively for 24 hours. The mean ± standard error of mean (s.e.m.) is relative to the strength of the *rnpB* promoter in the respective regulation condition. The value in parentheses is the experimental mean ± s.e.m. of EYFP emission per cell after subtracting the auto-fluorescence of *Synechocystis* cells containing pPMQAK1 vector only. The measurement is done by a flow cytometer to collect 50,000 events for each biological sample. [c], The induction of a promoter is the ratio of its measured strength in induced compared to in repressed condition. [d], The simulated thermal opening probability patterns (A-I) at 303 K are shown (Figure 4); n.s., not simulated. The value in a bracket is the mean ± s.e.m. under the induced condition of strengths of the promoters in a given pattern. The noLVA_L09 and L12 constructs are excluded in averaging. [e], The only difference of noLVA_L09 to L09 is the introduction of a double stop codon in the end of the *tetR* gene to cease translation of a protease LVA tag tailing in C-terminal of TetR repressor. The L09 promoter regions of both are identical.

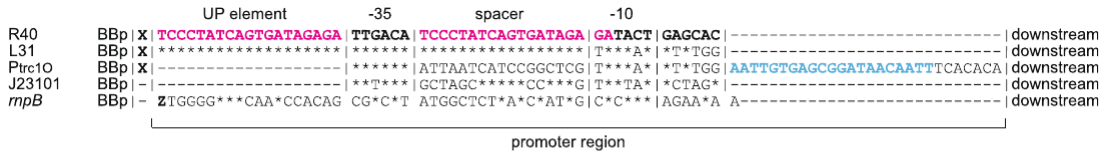

Figure 1 Alignment of selected promoters and their flanking sequences. The reporter construct of R40 promoter (i.e. BBa_R0040) is used as a reference sequence; gap (hyphen '-') and identical base (asterisk '*') are indicated for the each promoter. 'BBp' is the BioBrick prefix (GAATTCGCGGCCGCTTCTAGAG); 'X' is the TetR repressor expressing device [BBa_J23101]-[BBa_P0440] and the 8-bp BioBrick scar (TACTAGAG); 'downstream' sequence to the initial codon ATG of the reporter EYFP gene are the 8-bp BioBrick scar, the 10-bp ribosome binding site RBS*, the 6-bp BioBrick scar, and the 3-bp initial codon (TACTAGAG|TAGTGGAGGT|TACTAG|ATG); the *tetO2* operator of the Tn*10* resistance operon was placed upstream of the −35 element and in the spacer region for the two TetR-binding sites (bases shown in magenta). Core bases of each operator for the TetR binding are underlined. The R40-derived promoters L01 to L16, L21 and L22 are compared in Table 1. The *lacO1* operator of the P_{trc1O} promoter contains the bases shown in cyan. The *rnpB* promoter is aligned with the −10 element CACACT [25] and Z is the remaining 153 bases.

Figura 7.1: Arriba: tabla obtenida directamente de [17] con los resultados experimentales de la expresión del gen anexo a los distintos promotores diseñados. Debajo: varias secuencias de promotores, entre ellas la $R40$ y la $L31$, obtenidas también directamente de [17]; nuestras cadenas contienen todas las bases del promotor $L31$ (salvo los dos cambios efectuados en las posiciones -5 y -6) más los elementos *8-bp BioBrick scar*, *10-bp ribosome binding site RBS**, *6-bp BioBrick scar* y el codon inicial *ATG*.

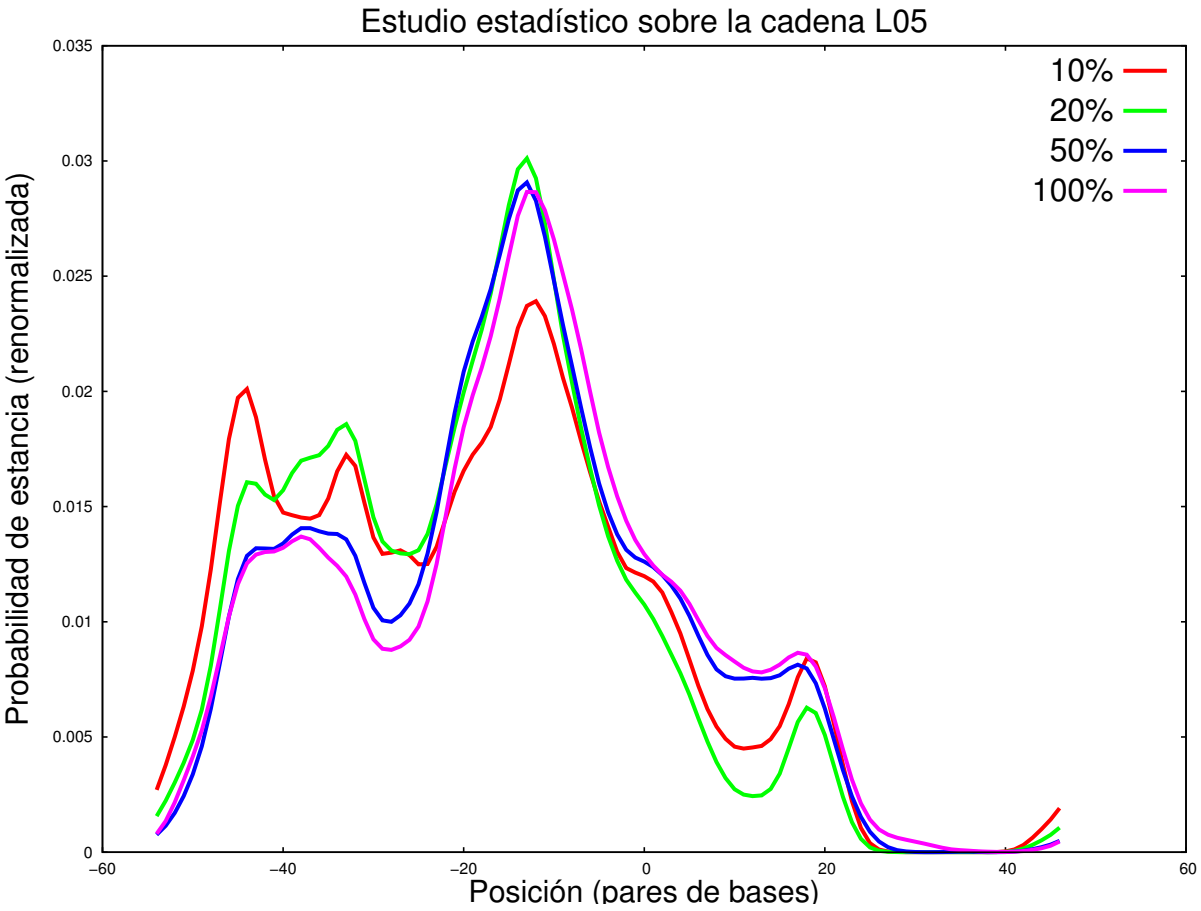

Figura 7.2: Estudio de la convergencia de nuestros cálculos para diferentes tiempos de simulación. El tiempo equivalente al 100% es $t \sim 10^{-6}\, s$. Se observa como el sistema converge para el orden de t entre el 50% y el 100%, especialmente en la zona del pico central, la de mayor importancia en nuestro estudio.

la temperatura a la cual ningún par de bases tiene una probabilidad de apertura mayor de 0.5, esta temperatura resulta ser, en unidades adimensionales, $\tilde{T} = 0.45$, que equivale a $T = 308\, K$ con dimensiones, aproximadamente temperatura ambiente. Este estudio se ha realizado sobre la cadena $L05$, dada la gran similitud entre las cadenas $L01 - L16$, los resultados se consideran válidos para todas ellas.

Una vez fijada la temperatura, queremos determinar el tiempo de simulación necesario para que el comportamiento del sistema converja. Para ello se realizan simulaciones con diferentes tiempos en la cadena $L05$ y se comparan los resultados (ver figura 7.2). Se observa como para $t = 50\%$ y $t = 100\%$ del tiempo total $t \sim 10^{-6}\, s$, las gráficas son muy similares, por lo que puede considerarse que a ese orden de t el sistema converge. Para nuestros cálculos utilizaremos el mayor de ambos, $t =$, para asegurarnos de que el sistema alcanza la convergencia en todos los casos, ya que esto puede conllevar tiempos ligeramente diferentes en otras secuencias.

En la primera imagen de la figura 7.3 se presentan una trayectoria típica de la partícula a lo largo de la cadena. Puede observarse claramente como la partícula va alternando periodos de estancia en una sección de la cadena con transiciones a otras zonas de la misma, ya sean estas rápidas o lentas. Cuando comparamos esta trayectoria con cualquier otra con condiciones iniciales distintas para la partícula, las diferencias son notables; debido a esto, el tiempo de simulación que hemos escogido se reparte entre 20 condiciones iniciales distintas colocadas uniformemente a lo largo de la cadena (de hecho esto ya está aplicado en la figura 7.2).

En la segunda imagen de la figura 7.3 se muestra el histograma calculado a partir de las 20 trayectorias estudiadas de la cadena $L05$. Este histograma representa la fracción de tiempo que la partícula permanece en cada posición de la cadena, expresada dicha posición con el índice del

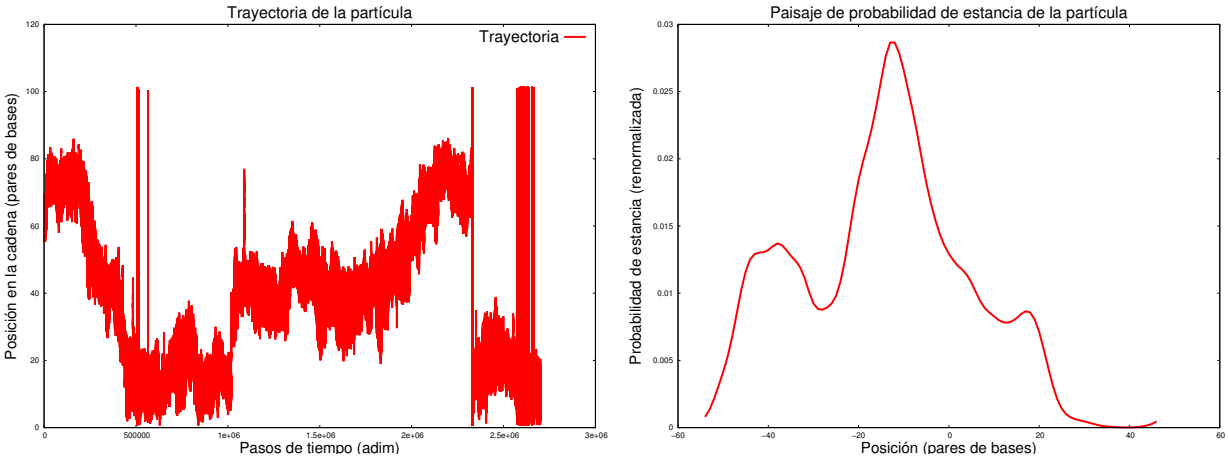

Figura 7.3: A la izquierda: trayectoria típica de la partícula a lo largo de una cadena, pueden observarse transiciones rápidas y lentas entre las zonas más abiertas de la misma. A la derecha: histograma a partir de las trayectorias de la partícula por la cadena $L05$; se observan tres picos, destacando en especial el central, el cual coincide justamente con una zona de importancia biológica, la *TATA box*.

par de bases más cercano a ella. En él se puede observar como la partícula pasa claramente más tiempo en tres secciones de la cadena, las que contienen los picos en el histograma, y especialmente en la sección que contiene el pico central, entre los pares de bases -19 y 1. Es precisamente en esta sección en la que se encuentra la *TATA box*, posiblemente la región más importante en la regulación de la expresión génica que se conoce. El modelo predice correctamente, por tanto, la importancia de la región de la *TATA box*, al permanecer más tiempo la partícula sobre ella.

Hasta este punto los resultados no difieren mucho de los obtenidos por Lindblad y Huang, tan solo en que en nuestro caso los picos están mejor diferenciados, gracias a la presencia de la partícula; pero lo que nos interesa es encontrar correlaciones entre las probabilidades de estancia de la partícula y la expresión del gen que sigue al promotor. Para ello, primero se presentan los histogramas obtenidos para las cadenas $L01 - L16$ en grupos de cuatro, tal y como están agrupados en [17], junto al histograma para la cadena $R40$ (Ver figura 7.4, cadena $R40$ en azul claro en la primera gráfica). El hecho de que se observen diferencias dentro de cada grupo ya es algo nuevo respecto a lo visto en el artículo de Lindblad y Huang, pero en la figura 7.4 no se observa correlación alguna a simple vista entre la expresión del gen y la probabilidad de estancia de la partícula , que era lo que buscábamos. Sí que se observa, sin embargo, como en el histograma de la cadena $R40$, que no contiene la secuencia de la *TATA box*, el pico central es mucho menor, tanto en altura como en área.

Para realizar un estudio más exahustivo de los datos obtenidos, buscamos encontrar variables cuantitativas que nos ayuden a reflejar mejor las probabilidades de estancia de la partícula en las diferentes regiones. La opción más obvia pasa por integrar la probabilidad de estancia de la partícula en aquellas zonas donde pueda ser relevante desde un punto de vista biológico, lo que implica el estudio, principalmente, del elemento *-10,* el cual contiene la *TATA box*. En la tabla 7.1 se presentan los valores de las integrales de la probabilidad de estancia para varias regiones distintas: el elemento *-10* y los tres picos. Aunque no se muestren en este trabajo se han realizado numerosas pruebas intentando encontrar alguna correlación entre estos números y los valores de expresión génica del artículo [17]. Ninguna de las combinaciones de operaciones entre los valores de las integrales de los picos que se han probado arroja resultados de los que pueda extraerse una dependencia cuando se comparan con los valores experimentales. Tampoco el estudio de otros elementos de menor importancia como el *-35, UP element* o el *spacer*, así como las regiones *tet02-operator* y *tet01-operator*[4] (no se muestran en el trabajo) han arrojado resultados diferentes.

[4]Todos estos elementos se hallan definidos en el artículo de Lindblad y Huang.

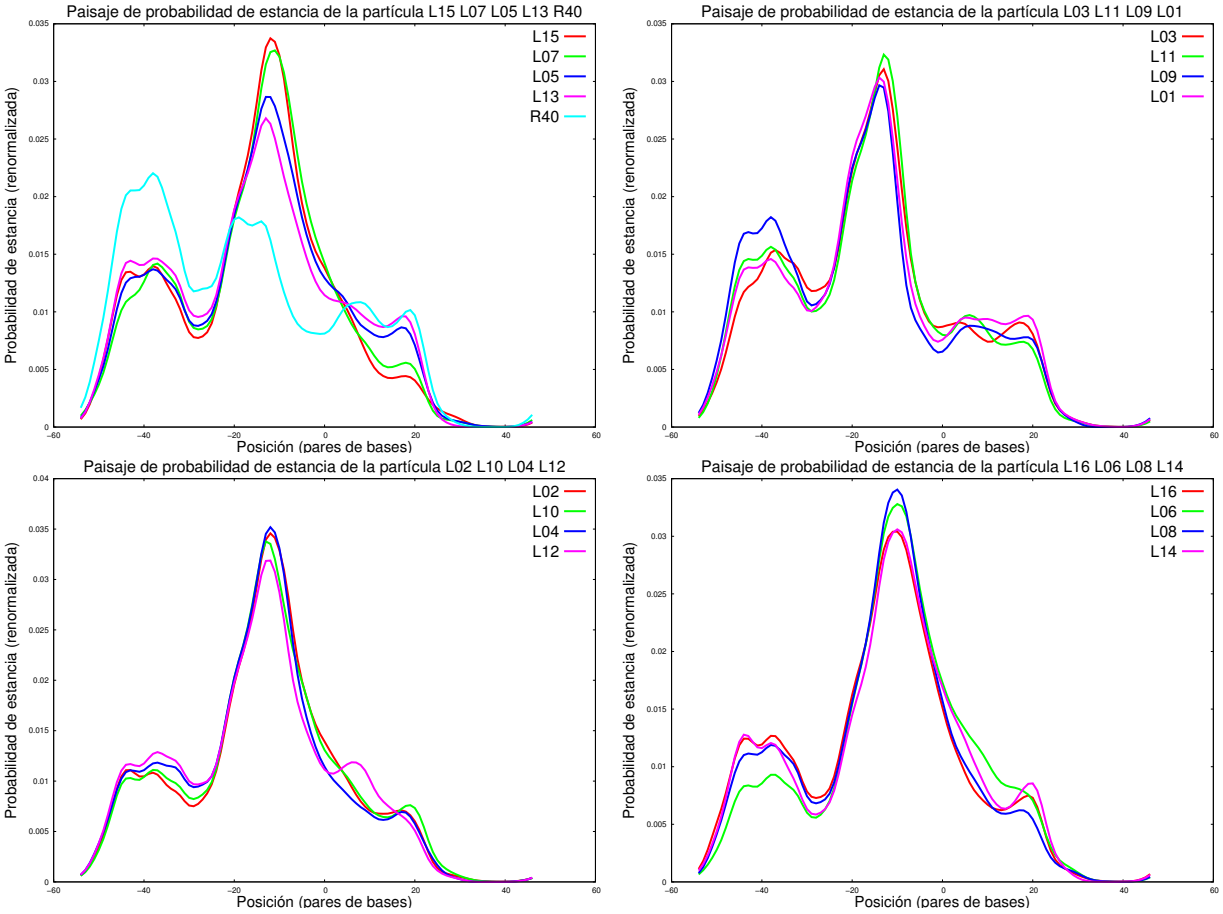

Figura 7.4: Comparativa de todos los histogramas de las cadenas estudiadas. Se observa que no existe relación entre la altura de los picos y los valores de la tabla de la figura 7.1, salvo para el primer caso. Para los promotores $L15$, $L07$, $L05$ y $L13$ sí que parece haber una relación directa entre la altura del pico central y la expresión del gen en estado inducido; también se observa una relación inversa entre la altura del primer pico y la expresión del gen en estado inducido en este grupo; pero dado que en el resto no se aprecian las mismas relaciones, no podemos descartar que sea un resultado fortuito.

Promotor	Elemento *-10*	Pico 1	Pico central	Pico 3
*L*15	0.190	0.218	0.496	0.037
*L*07	0.189	0.210	0.498	0.045
*L*05	0.160	0.217	0.446	0.067
*L*13	0.144	0.238	0.408	0.075
*L*03	0.151	0.233	0.404	0.074
*L*11	0.163	0.244	0.413	0.061
*L*09	0.128	0.279	0.357	0.067
*L*01	0.137	0.230	0.383	0.080
*L*02	0.195	0.176	0.511	0.057
*L*10	0.182	0.177	0.496	0.062
*L*04	0.196	0.194	0.497	0.054
*L*12	0.172	0.207	0.454	0.053
*L*16	0.178	0.202	0.490	0.061
*L*06	0.192	0.144	0.524	0.067
*L*08	0.198	0.188	0.523	0.050
*L*14	0.178	0.190	0.489	0.067
*R*40	0.084	0.335	0.267	0.081

Tabla 7.1: Resultados de integrar las curvas de los histogramas en los tres picos observados en todos los promotores, así como en la región *-10*. La región *-10* comprende las bases -13 a -8, el primer pico -47 a -30, el pico central -19 a +1 y el tercer pico +14 a +22. Puede verse como los valores no parecen guardar la relación esperada con la expresión génica observada experimentalmente en la figura 7.1, salvo en el caso de los promotores *L*15, *L*07, *L*05 y *L*13, para los que el valor de la integral en el elemento *-10* sigue el mismo orden que en los resultados experimentales en condiciones inducidas.

El hecho de que en uno de los grupos, el compuesto por los promotores $L15$, $L07$, $L05$ y $L13$, sí que se observen las relaciones esperadas, puede deberse a la casualidad. Sin embargo, siempre cabe la posibilidad de que con algo más de estadística o una mejor elección de los parámetros los resultados fuesen satisfactorios. Es importante decir, en cualquier caso, que las diferencias entre los valores de la integral del elemento *-10* en la tabla 7.1 son muy pequeñas como para describir las diferencias en los resultados experimentales de [17].

8 Conclusiones

El objetivo principal de este trabajo era estudiar el comportamiento de la molécula de ADN utilizando el modelo PBD con interacción con solvente (barrera), partícula browniana y término de *stacking* dependiente de la secuencia. Para ello primero se ha realizado una selección entre diversos valores para el término de *stacking* encontrados en la literatura, entre los que se han escogido los de Alexandrov et al. [6], debido principalmente al modo en que han sido obtenidos. Una vez elegidos, se ha realizado un ajuste de los parámetros del sistema a través de la reproducción de las temperaturas de *melting* de una serie de cadenas sencillas de ADN. El modelo ha sido capaz de reproducir estas temperaturas, aunque de forma poco precisa. Por último se ha aplicado el modelo a un sistema más complejo, una serie de promotores diseñados para la bacteria *Synechocystis*. Nuestro objetivo era encontrar alguna correlación entre las probabilidades de estancia de la partícula en determinadas zonas de importancia biológica y la expresión del gen anexo a los promotores, recogida en [17]. Primero se ha intentado obtener una relación de tipo cualitativo, fruto tan solo de la observación de los histogramas de la probabilidad de estancia de la partícula; ante la falta de éxito se ha utilizado un método algo más sofisticado, se ha introducido el valor de la integral en distintas zonas del histograma como variable a comparar con los valores experimentales, tampoco se han obtenido resultados. Pese a todo, se han obtenido algunos resultados positivos: hemos conseguido mejorar los resultados obtenidos por Lindblad y Huang, ya que en nuestros histogramas quedan mejor definidas las zonas de la cadena que presentan un comportamiento más relevante; por otro lado hemos sido capaces de diferenciar entre más secuencias que ellos, algo esencial si se busca reproducir lo observado experimentalmente.

Las posibles causas por las que no hemos encontrado las relaciones esperadas entre nuestros resultados y los experimentales son muy numerosas, pero pueden resumirse, básicamente, en que el modelo y en especial la parte de interacción partícula-ADN, es demasiado sencillo como para describir un proceso tan complejo como la transcripción o incluso la iniciación de la misma. Un ejemplo muy ilustrativo lo encontramos en la comparación entre las secuencias de los promotores $L12$ y $L10$; la única diferencia entre ambas secuencias es la base nitrogenada situada en la posición -6, en el caso del promotor $L12$ es una adenina, mientras que en la cadena $L10$ es una timina. Este pequeño cambio, que en nuestro modelo apenas supone la variación de los valores de dos parámetros (los términos de *stacking* de la interacción entre las bases -7 y -6, así como entre las -6 y -5), implica, de acuerdo a las observaciones experimentales, un sorprendente cambio en la expresión en estado inducido, el gen con el promotor $L10$ se expresa cientos de veces más que aquel con el promotor $L12$. Nuestro modelo no está preparado para reproducir semejante comportamiento. Dadas, sin embargo, las diferencias menos acusadas entre los valores observados del resto de promotores no parece disparatado que pudiera existir alguna correlación del tipo que hemos buscado. El problema radica en la complejidad de la interacción ADN-proteína: esta interacción puede incluir más de una proteína al mismo tiempo; puede incluir también interacciones de mayor rango que el reflejado en nuestro modelo, incluso puede incluir interacciones en tres dimensiones, ya que el ADN puede doblarse y pueden entrar en juego simultáneamente regiones del ADN muy separadas desde el punto de vista de la secuencia; también puede incluir interacciones reguladas de forma más compleja que por la simple apertura de la cadena.

Nuestro modelo es en cualquier caso el siguiente paso lógico en la mejora del modelo PBD con barrera y partícula; quizá un modelo un paso más allá sea capaz de obtener mejores resultados en el caso de los promotores, abordando los problemas aquí expuestos con alguna modificación en el término de interacción partícula-ADN.

Bibliography

[1] O.T. Avery, C.M. MacLeod, and M. McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. *The Journal of Experimental Medicine*, 79:137–156, 1944.

[2] J.D. Watson and F.H.C. Crick. Molecular structure of nucleic acids. *Nature*, 171:737–738, 1953.

[3] M. Peyrard. Nonlinear dynamics and statistical physics of dna. *Nonlinearity*, 17:R1–R40, 2004.

[4] R. Tapia-Rojo, J. J. Mazo, and F. Falo. Thermal and mechanical properties of a dna model with solvation barrier. *Phys. Rev. E*, 82(031916), 2010.

[5] R. Tapia-Rojo, J. J. Mazo, and F. Falo. Mesoscopic model for free-energy-landscape analysis of dna sequences. *Phys. Rev. E*, 86(021908), 2012.

[6] B. S. Alexandrov, V. Gelev, Y. Monisova, L. B. Alexandrov, A. R. Bishop, K. Ø. Rasmussen, and A. Usheva. A nonlinear dynamic model of dna with a sequence-dependent stacking term. *Nucleic Acids Research*, 37(7):2405–2410, 2009.

[7] C. R. Calladine, H. Drew, B. Luisi, and A. Travers. *Understanding DNA: The molecule and how it works*. Academic Press, 1992 (1 edition) 2004 (3 edition).

[8] M. Doi and S. F. Edwards. *The Theory of Polymer Dynamics*. Oxford University Press, 1988.

[9] J.L. Leroy, M. Kochoyan, T. Huynh-Dinh, and M. Guéron. Characterization of base-pair opening in deoxynucleotide duplexes using catalyzed exchange of the imino proton. *Journal of Molecular Biology*, 200:223–238, 1988.

[10] G. Weber. Sharp dna denaturation due to solvent interaction. *Europhysics Letters*, 73:806–811, 2006.

[11] A. Singh and N. Singh. Effect of salt concentration on the stability of heterogeneous dna. *Physica A*, 419:328–334, 2014.

[12] E. Helfand. Numerical integration of stochastic differential equations. *Bell System Technical Journal*, 58:2289, 1979.

[13] H. S. Greenside and E. Helfand. Numerical integration of stochastic differential equations 2. *Bell System Technical Journal*, 60:1927, 1981.

[14] D. Svozil, P. Hobza, and J. Sponer. Comparison of intrinsic stacking energies of ten unique dinucleotide steps in a-rna and b-dna duplexes. can we determine correct order of stability by quantum-chemical calculations? *Journal of Physical Chemistry B*, 114:1191, 2010.

[15] J. Sponer, K. E. Riley, and P. Hobza. Nature and magnitude of aromatic stacking of nucleic acid bases. *Physical Chemistry Chemical Physics*, 10:2595, 2008.

[16] O. G. Berg, R. B. Winter, and P. H. Hippel. Diffusion-driven mechanisms of protein translocation on nucleic acids. *Biochemistry*, 20:6929–6948, 1981.

[17] H. Huang and P. Lindblad. Wide-dynamic-range promoters engineered for cyanobacteria. *Journal of Biological Engineering*, 2013.