Universidad Nacional de Ingeniería

Facultad de Ciencias

TESIS:

La presión mutacional termodinámica en el mecanismo de replicación del ADN y sus efectos en la evolución molecular biológica

Para obtener el titulo profesional de Licenciado en Física

Elaborado por:

Dr. Mirko Juan Zimic Peralta

ORCID: 0000-0002-7203-8847

Asesor:

Dr. Germán Yuri Comina Bellido

ORCID: 0000-0003-2114-0486

LIMA - PERÚ

2024

Citar/How to cite	Zimic Peralta [1]
Referencia/Reference Estilo/Style: IEEE (2020)	[1] M. Zimic Peralta, "*La presión mutacional termodinámica en el mecanismo de replicación del ADN y sus efectos en la evolución molecular biológica*" [Tesis]. Lima (Perú): Universidad Nacional de Ingeniería, 2024.

Citar/How to cite	(Zimic, 2024)
Referencia/Reference Estilo/Style: APA (7ma ed.)	Zimic, M. (2024). *La presión mutacional termodinámica en el mecanismo de replicación del ADN y sus efectos en la evolución molecular biológica*. [Tesis, Universidad Nacional de Ingeniería]. Repositorio institucional Cybertesis UNI.

Dedicatoria

Dedico este trabajo a todos mis colegas y estudiantes que a lo largo de estos años han brindado su apoyo y entusiasmo por este trabajo, contribuyendo con enriquecedoras discusiones, que hicieron comprender mejor la teoría de la presión mutacional termodinámica.

Lo dedico también a Paty, mi familia, la cual siempre me brindó el soporte, comprensión, y el tiempo para trabajar en investigación.

Lo dedico de manera especial a mis hijos Alen y Lean, para motivarlos en el fascinante mundo de la investigación y la ciencia. Tengo la esperanza de que Lean sea un investigador curioso y sea feliz contribuyendo con el avance de la ciencia.

Vamos Lean!

Agradecimientos

A mi familia por su soporte y su paciencia en tolerar mi ausencia por largos tiempos mientras desarrollé este trabajo.

Al Dr. German Comina por su asesoría y apoyo en la presente tesis.

A mi siempre recordado y admirado profesor Holger Valqui, con quien discutí por primera vez, los fundamentos de este trabajo y de quien recibí importantes consejos.

Al Dr. Jorge Arévalo, quien sembró la semilla del mundo biológico en mi, y con quien discutí los primeros pasos de la presión mutacional termodinámica.

A mis estudiantes, quienes con mucho entusiasmo y esfuerzo permitieron avanzar los distintos pasos que han guiado este estudio.

Resumen

El proceso de replicación y reparación del ADN es crucial en la generación de mutaciones y la evolución del ADN. Utilizando la distribución de Boltzmann para un conjunto canónico, se puede calcular la probabilidad de ingreso de nucleótidos libres al sitio de replicación, afectada por factores como la energía potencial molecular, la abundancia de nucleótidos, la secuencia de ADN cercana, y la temperatura. Estos factores incluyen la energía de los puentes de hidrógeno y las interacciones electrostáticas.

Este enfoque permite calcular las probabilidades de mutaciones ocurridas durante la replicación del ADN, ofreciendo información sobre la acumulación de nucleótidos a lo largo de la evolución. Hay dos métodos de estudio: simulaciones computacionales y aproximaciones analíticas. Nuestro estudio previo incluyó una simulación de Monte Carlo que estimó energías y predijo un aumento en el contenido de guanina-citosina (GC) en la molécula de ADN. Esta simulación se correlacionó con evidencias experimentales de secuencias genómicas, especialmente en Kinetoplastidia y Plasmodium, y permitió explicar el fenómeno del "codon bias" de una mantera natural.

En este trabajo, presentamos una evaluación analítica basada en la distribución de Boltzmann y las energías potenciales del sistema. Nuestra nomenclatura y análisis matemático-estadístico predicen un incremento de concentración de GC en el tiempo. Los resultados teóricos coinciden con las simulaciones anteriores, apoyando la idea de una "presión mutacional termodinámica" que guía la evolución y podría explicar el "codon bias", un misterio en la biología moderna.

Palabras clave — Mutación, evolución, termodinámica, distribución de Boltzmann, ADN.

Abstract

The process of DNA replication and repair is crucial in the generation of mutations and DNA evolution. Using the Boltzmann distribution for a canonical ensemble, one can calculate the probability of free nucleotide entry into the replication site, affected by factors such as molecular potential energy, nucleotide abundance, nearby DNA sequence, and temperature. These factors include hydrogen bond energy and electrostatic interactions.

This approach allows the calculation of the probabilities of mutations occurring during DNA replication, providing information on the accumulation of nucleotides throughout evolution. There are two methods of study: computational simulations and analytical approaches. Our previous study included a Monte Carlo simulation that estimated energies and predicted an increase in guanine-cytosine (GC) content in the DNA molecule. This simulation correlated with experimental evidence from genomic sequences, especially in Kinetoplastidia and Plasmodium, and allowed us to explain the codon bias phenomenon in a natural way.

In this work, we present an analytical evaluation based on the Boltzmann distribution and potential energies of the system. Our nomenclature and mathematical-statistical analysis predict an increase of GC concentration over time. The theoretical results agree with previous simulations, supporting the idea of a "thermodynamic mutational pressure" that guides evolution and could explain the "codon bias", a mystery in modern biology.

Keywords - Mutation, evolution, thermodynamics, Boltzmann distribution, DNA.

Tabla de Contenido

Dedicatoria iii
Agradecimientos iv
Resumen v
Abstract vi
Tabla de Contenido vii
Lista de Tablas ix
Lista de Figuras x
Introducción xii
CAPITULO I. PARTE INTRODUCTORIA DEL TRABAJO 1
1.1 Generalidades 1
1.2. Descripción del problema de investigación 2
1.3 Objetivos del estudio 4
1.3.1 Objetivo general. 4
1.3.2 Objetivos específicos. 4
1.4. Hipótesis 5
1.5. Antecedentes investigativos 6
1.5.1. Predicción de Mutaciones del ADN y Modelos de Evolución Utilizando Principios Físicos 6
1.5.2 Uso de la Distribución de Boltzmann en Modelos de Fenómenos Biológicos 7
1.5.3 La presión mutacional termodinámica del ADN: un posible factor en la evolución del genoma 8
CAPITULO II. MARCO TEÓRICO Y CONCEPTUAL 13
2.1. Marco teórico 13
2.1.1 Introducción al Ácido Desoxirribonucleico (ADN) 13
2.1.2 Clasificación del ADN 15
2.1.3 Mutaciones 15
2.1.4 Agentes mutagénicos 17
2.1.5 Mecanismo de Replicación del ADN 19
2.1.6 Mecanismos Subyacentes en la Fidelidad de Replicación del Ácido Desoxirribonucleico (ADN): Mecanismo de Reparación del ADN 20
2.1.7 Presión de Selección Como Interacción entre Medio Ambiente y Adaptación (Fitness) 22
2.1.8 Comprendiendo la Evolución a través de la Anagénesis y Cladogénesis 23
2.2. Marco conceptual 24

CAPITULO III. DESARROLLO DEL TRABAJO DE INVESTIGACIÓN 26
3.1 Modelo termodinámico basado en la distribución de Boltzmann para estimar las probabilidades de formación de pares no canónicos (missmatches) 26
3.2. Estudio teórico-analítico sobre los efectos a largo plazo de la estabilidad termodinámica del proceso de replicación/reparación del ADN en el contenido de Guanina-Citosina (GC) 35
3.2.1. Introducción 35
3.2.2 Conceptos básicos (estructura y funcionalidad del ADN): 36
3.2.3. Premisas consideradas 43
3.2.4. Tratamiento analítico 46
CAPÍTULO IV. ANÁLISIS Y DISCUSIÓN DE LOS RESULTADOS 57
4.1. Resultados 57
4.1.1 Predicciones del modelo analítico 59
4.1.2 Evidencia del incremento de GC en un proceso de evolución in-vitro 60
4.2 Discusión 78
4.2.1. Modificación del contenido %GC durante la diversificación 82
4.2.2. Distribución de Mutaciones en el Codón 84
4.2.3. Influencia de la Secuencia Inicial 85
4.2.4. Implicaciones del Mecanismo de Reparación 85
4.2.5. El desafío del ADN no codificante "basura" 86
4.2.6. Efecto de la temperatura 87
4.2.7. Tendencia natural hacia el incremento del contenido de GC 88
Conclusiones 93
Recomendaciones 94
Referencias Bibliográficas 95
Anexos 101

Lista de Tablas

Tabla 1: *Energias de enlace estimadas para pares de bases canonicos y no-canonicos* 51

Lista de Figuras

Figura 1. Esquema simbólico de la generación de un hueco en el ADN 28
Figura 2. Incorporación de nucleótidos en la polimerización del ADN 29
Figura 3. Distribución de probabilidad para la generación de un hueco en el ADN. Distribución para el individuo 1, individuo 2 y para toda la población. Esta última resulta ser el promedio de las distribuciones de todos los individuos, haciéndose por lo tanto uniforme. 30
Figura 4. Configuraciones potenciales durante una mutación en el conjunto canónico. 32
Figura 5. Estructura de los nucleótidos. (a)Estructura general mostrando la numeración convencional de carbonos en la pentosa. (b)Bases nitrogenadas: pirimidinas y purinas. 37
Figura 6. Dirección de la hebra y enlace fosfodíester. Se observa como el ion fosfato se une al 5to carbono de la pentosa y el ion oxhidrilo al 3ro . 38
Figura 7. Molécula de ADN. (a) Estructura tridimensional, (b) Modelo elemental de replicación, (c) hebras antiparalelas mostrando 4 bp 39
Figura 8. Diagrama del proceso de replicación del ADN. 41
Figura 9. Modelo a utilizar de la replicación del ADN. Los nucleótidos fijos $xkk = 14$ forman la vecindad de referencia VR y los nucleótidos Y compiten por enlazarse a X fijo en la hebra plantilla 46
Figura 10. Variación en el tiempo de [GC](t) para 3 condiciones iniciales. (a)$[GC]_0 = 1$, (b)$[GC]_0 = 0.5$, (c) $[GC]_0 = 0$. El tiempo se da en numero de pasos (donde en cada paso ingresa un nucle´otido del medio circundante) para estar en concordancia con los resultados del algoritmo computacional propuesto en [1] (donde se usan muchos pasos para tener un proceso continuo). 54
Figura 11. Evolución de secuencias de ADN mediante una serie de PCRs bifurcadas. Un ancestro SSU rDNA clonado en pBluescript se utilizó como plantilla para la serie 1 de 70 ciclos anidados de PCR con cebadores M13. Después de los primeros 35 ciclos, los productos de la reacción se diluyeron 1:1,000 y se utilizaron como plantillas para los 35 ciclos subsiguientes, con los cebadores rDNA RIBA y RIBB. Tras 70 ciclos, los amplicones se clonaron y se

seleccionaron al azar dos clones que se utilizaron como plantillas para la siguiente serie de ciclos de PCR anidados. Las líneas de descendencia se propagan al azar, por lo que la evolución es neutral y se comporta como un proceso estocástico. Los nodos del árbol T1 a T16 indican secuencias terminales, y 1.1 a 4.8, ancestros internos. 61

Figura 12. Comparación de la filogenia real con la filogenia inferida de máxima verosimilitud 63

Figura 13. A, B, C…P. Variación del porcentaje de GC a lo largo del tiempo (orden ancestral) en las diferentes cascadas evolutivas 72

Figura 14 Efecto de la temperatura en la diversificación 88

Figura 15. Variación de [GC] respecto al número de pasos usados en la simulación. (a) Sin restricción en mutaciones puntuales. (b) Restricciones en la conservación de la familia de aminoácidos (líena gruesa) y en la identidad de aminoácido (líneas delgadas). Además se considera [GC]|0 = 22%. 90

Figura 16. Variación en el contenido de GC de genes específicos. Genes (HSP70,HGPRT,Top2,TryR) de trypanosomatidas. El orden horizontal sigue un orden filogenético específico. La especie más antigua (*T. brucei)* tiene menor contenido de GC en sus genes comparándola con las más modernas (*Crithidia* y *Leishmania*). 91

Introducción

Entender la vida y su evolución desde una perspectiva física es un desafío enorme, especialmente en comparación con estudiar la materia inanimada. Los seres vivos son sistemas termodinámicos abiertos y dinámicos, lo que los aleja de un estado de equilibrio termodinámico. Además, la complejidad de estos sistemas viene dada por la gran cantidad de variables y factores desconocidos, lo que hace difícil un enfoque riguroso. Hasta ahora, los modelos que aplican conceptos de física y química a la biología son aproximaciones que requieren simplificaciones [Schrödinger, 1944].

Se piensa que la vida surgió a partir de reacciones químicas en un entorno prebiótico, resultando en moléculas con la habilidad de replicarse a sí mismas. Este mecanismo de replicación habría sido la base para el surgimiento de la vida tal como la conocemos, incluyendo a la molécula del ADN como la fuente de información genética que se transmite de generación en generación y evoluciona en el tiempo [Gagnon, J.-S., & Hochberg, D.,2023].

A medida que evolucionaron, los organismos vivos experimentaron cambios a nivel genético (i.e. mutaciones). Algunos cambios genéticos mantuvieron las características morfológicas externas (el fenotipo) constante, proceso que se reconoce como un envejecimiento a nivel molecular. Por otro lado, otros tipos de mutaciones terminaron en cambios notables del fenotipo, resultando en la aparición de nuevas especies, proceso reconocido como diversificación de especies [Ayala, 1984]. El ADN contiene la totalidad de la información necesaria para la vida, y sus variaciones/cambios (mutaciones) conducen a su evolución. Estos cambios pueden variar en naturaleza y origen.

En nuestra investigación, queremos entender aspectos específicos de la evolución a nivel molecular a partir de principios físicos fundamentales. Buscamos desarrollar un modelo que considere la estabilidad termodinámica del proceso de replicación del ADN. Para ello, utilizaremos la distribución de Boltzmann en un ensable canónico, bajo la suposición de que el ADN interactúa térmicamente con su entorno celular en un sistema cerrado.

Consideraremos que localmente, en la vecindad inmediata del sitio activo de la ADN polimerasa, el sistema alcanza un equilibrio rápido comparado con el tiempo total de replicación. En este caso, proponemos usar un ensamble canónico para modelar este "microestado" del sistema, en lo que se puede denominar una “vecindad espacio-temporal”. De manera racional, identificamos que la molécula de ADN busca un estado de mínima energía, lo que podría manifestarse, por ejemplo, en la acumulación de pares de bases guanina-citosina. Deseamos explorar dos fuerzas motrices de la evolución: la presión de selección y lo que denominamos "presión mutacional termodinámica", siendo esta última el objetivo principal de nuestro estudio.

El presente trabajo presenta un modelo analítico probabilístico que muestra los efectos a largo plazo de la aplicación de la presión mutacional termodinámica durante la replicación del ADN. Adicionalmente se muestra evidencia experimental que contrasta las predicciones del modelo propuesto.

CAPITULO I. Parte introductoria del trabajo

1.1 Generalidades

La evolución del ADN y su reparación son procesos biológicos fundamentales que juegan un papel crucial en la diversidad y adaptabilidad de la vida. El contexto para una exploración detallada de los procesos moleculares y termodinámicos que impulsan la evolución del ADN, se desarrollará en presente trabajo, proporcionando un marco para la investigación y el análisis en biología molecular y evolutiva.

Un panorama sobre los conceptos y los mecanismos involucrados esenciales en la replicación del ADN y la generación de mutaciones para comprender la evolución a nivel molecular son:

1. Importancia de la Replicación del ADN: La replicación del ADN es un proceso vital que permite la transferencia de información genética de una generación a otra. Durante este proceso, la precisión es clave, pero también lo son los errores ocasionales, ya que estos últimos contribuyen a la variabilidad genética.
2. Generación de Mutaciones: Las mutaciones en el ADN son cambios en la secuencia de nucleótidos. Estos cambios pueden ser el resultado de errores en la replicación del ADN o de daños causados por factores ambientales. Las mutaciones son fundamentales para la evolución, ya que proporcionan la materia prima para la selección natural.
3. Mecanismos de Reparación del ADN: Los sistemas de reparación del ADN son esenciales para corregir errores y mantener la integridad genómica. Sin embargo, no todas las mutaciones son reparadas, lo que permite que algunas de ellas se acumulen y contribuyan a la evolución de las especies.
4. La Distribución de Boltzmann en la Predicción de Mutaciones: La distribución de Boltzmann, aplicada en el contexto de la replicación del ADN, permite calcular la

probabilidad de incorporación de nucleótidos en la cadena naciente. Este enfoque termodinámico ayuda a predecir cómo la energía potencial y otros factores moleculares influyen en la aparición de mutaciones. El análisis de estas probabilidades proporciona una comprensión más profunda de la dinámica evolutiva a nivel molecular.

5. Relación con la Evolución: La acumulación de mutaciones a lo largo del tiempo es un motor de la evolución. Este proceso, influenciado por factores moleculares y ambientales, resulta en la diversificación de las especies y su adaptación a diferentes entornos.
6. Modelos Computacionales y Analíticos en la Biología Evolutiva: La utilización de modelos computacionales y análisis analíticos ha revolucionado nuestra comprensión de los procesos evolutivos a nivel molecular. Estas herramientas permiten simular y predecir cambios en la composición de nucleótidos y entender mejor cómo los factores físicos y químicos influyen en la evolución.
7. Desafíos y Oportunidades en la Investigación: A pesar de los avances, quedan preguntas abiertas, como el fenómeno del 'codon bias' y la influencia de la presión mutacional termodinámica en la evolución. La investigación continua en estos campos promete profundizar nuestro entendimiento de la biología evolutiva y de los mecanismos moleculares que la sustentan.

1.2. Descripción del problema de investigación

El estudio de la evolución del ADN, particularmente a través de los procesos de replicación y reparación, es fundamental para comprender cómo las variaciones genéticas impulsan la diversidad biológica y la adaptación. Sin embargo, la complejidad de estos procesos y su interacción con diversos factores moleculares y ambientales presentan desafíos significativos en la biología molecular y evolutiva. Este problema de investigación se centra en desentrañar estos mecanismos a nivel detallado, utilizando un enfoque interdisciplinario que combina la biología molecular, la termodinámica y las matemáticas computacionales.

El problema fundamental a abordar en el presente estudio es cómo la estabiidad termodinámica del proceso de replicación de ADN, evaluado a través de la distribución de Boltzmann, puede ser usado para predecir el patrón de mutaciones que se acumulan con el tiempo evolutivo, y qué consecuencias tiene en las características de los genomas. Específicamente, el problema de investigación incluye:

1. Generación y Acumulación de Mutaciones: Uno de los aspectos centrales de este problema es entender cómo las mutaciones se generan y acumulan durante la replicación del ADN. A pesar de los sistemas de corrección y reparación, los errores en la replicación no son completamente evitables y son una fuente crucial de variabilidad genética. Sin embargo, el proceso exacto y los factores que determinan qué errores se conservan y cuáles se corrigen siguen siendo poco claros.
2. Rol de la Distribución de Boltzmann en la Replicación del ADN: La aplicación de la distribución de Boltzmann para predecir la probabilidad de incorporación de nucleótidos en la cadena de ADN es un área emergente de estudio. La teoría sugiere que las probabilidades de incorporación de nucleótidos están influenciadas por la energía potencial del sistema y otros factores termodinámicos. Sin embargo, la aplicación práctica de esta teoría en sistemas biológicos complejos y su impacto en la tasa y naturaleza de las mutaciones requiere una investigación más profunda.
3. Implicaciones en la Evolución: Entender cómo las mutaciones se acumulan y afectan la evolución es esencial para explicar la diversidad y adaptabilidad de las especies. Este problema abarca no solo el análisis de los mecanismos moleculares subyacentes sino también su relevancia en un contexto evolutivo más amplio.
4. Modelos Computacionales y Análisis Analíticos: La creación y aplicación de modelos computacionales y análisis analíticos para predecir y simular el proceso evolutivo del ADN es un aspecto crucial. Estos modelos permiten explorar cómo diferentes variables, como la temperatura y la concentración de nucleótidos, influyen en la replicación y evolución del ADN.

5. Exploración del 'Codon Bias' y la Presión Mutacional Termodinámica: El fenómeno del 'codon bias' y el concepto de una 'presión mutacional termodinámica' son áreas intrigantes que requieren investigación adicional. Comprender estos aspectos podría ofrecer nuevas perspectivas sobre cómo las preferencias en la secuencia de nucleótidos emergen y se mantienen a lo largo de la evolución.
6. Integración de Evidencia Experimental y Teoría: Un desafío importante es integrar los hallazgos experimentales con las teorías y modelos computacionales. Esta integración es clave para validar las hipótesis y asegurar que los modelos reflejen con precisión los sistemas biológicos complejos.

El objetivo de este problema de investigación es, por lo tanto, proporcionar una comprensión más completa de cómo los procesos moleculares y termodinámicos interactúan durante la replicación y reparación del ADN, y cómo estos procesos influyen en la evolución de las especies. Al abordar estas preguntas, se espera avanzar significativamente en nuestra comprensión de la biología molecular y evolutiva.

1.3 Objetivos del estudio

1.3.1 Objetivo general.

En base a una aproximación analítica, el presente estudio busca evaluar el efecto de la estabilidad termodinámica del proceso de replicación del ADN, sobre la probabilidad de ocurrencia de mutaciones y sus efectos a largo plazo durante la evolución.

1.3.2 Objetivos específicos.

Proponer un modelo de presión mutacional termodinámico basado en la distribución de Boltzmann para estimar las probabilidades de formación de pares no-canónicos (missmatches), durante la replicación del ADN.

Desarrollar una aproximación analítica basada en la estabilidad termodinámica del proceso de replicación del ADN y evaluar los efectos a largo plazo en el tiempo, en el contenido de Guanina-Citosina en las secuencias de ADN en el límite de tiempo infinito.

Buscar evidencia experimental que contraste las predicciones del modelo de presión mutacional termodinámica planteado.

1.4. Hipótesis

El proceso de replicación/reparación del ADN es intrínsecamente un proceso dinámico y fuera del equilibrio. Este proceso se lleva a cabo en una célula que es un sistema termodinámico abierto, donde hay flujo constante de materia y energía. La ADN polimerasa es una enzima que cataliza la reacción de polimerización del ADN, mediante la incorporación de nucleótidos. Su actividad se regula estrechamente a través de múltiples mecanismos, que van desde la disponibilidad de nucleótidos hasta señales de control del ciclo celular.

Por lo tanto, de forma estricta desde el punto de vista de la termodinámica estadística, el proceso de replicación del ADN no esta enmarcado en un sistema en equilibrio térmico, y por lo tanto no es un candidato ideal para el uso directo de la distribución de Boltzmann en cualquier tipo de ensamble termodinámico en equilibrio.

Sin embargo, el modelamiento de la probabilidad de que un nucleótido específico se incorpore en el sitio activo de la ADN polimerasa, puede darse considerando una aproximación para simplificar el sistema y hacerlo viable desde el punto de vista termodinámico.

Aproximación Local de Equilibrio: Consideraremos que localmente, en la vecindad inmediata del sitio activo de la polimerasa, el sistema alcanza un equilibrio rápido

comparado con el tiempo total de replicación. En este caso, proponemos usar un ensamble canónico para modelar este "microestado" del sistema, en lo que se puede denominar una “vecindad espacio-temporal”.

1.5. Antecedentes investigativos

1.5.1. Predicción de Mutaciones del ADN y Modelos de Evolución Utilizando Principios Físicos

El estudio de las mutaciones del ADN es crucial para entender tanto la evolución biológica como la adaptabilidad de los organismos. Mientras que la genética y la bioinformática han proporcionado insights valiosos, el uso de principios físicos para modelar y predecir mutaciones está ganando terreno como un enfoque complementario.

Termodinámica y ADN

La idea de aplicar principios termodinámicos a la biología no es nueva. Schrödinger, en su libro "What is Life?", planteó la cuestión de cómo los sistemas biológicos podrían ser entendidos desde una perspectiva física. Estudios más recientes han explorado cómo la termodinámica puede usarse para entender las estructuras de las moléculas biológicas, incluido el ADN [Dill & Bromberg, 2010].

Modelos Estocásticos y Teoría de la Información

Modelos estocásticos, como las cadenas de Markov, han sido usados para modelar la secuencia de ADN y prever mutaciones [Lynch, 2007]. La teoría de la información también ha sido aplicada para entender la variabilidad y la complejidad de las secuencias de ADN [Adami, 2004].

Simulaciones Computacionales

Las simulaciones basadas en dinámica molecular ofrecen una forma de incorporar explícitamente principios físicos en la predicción de las estructuras del ADN y las posibles mutaciones [Dror et al., 2012]. Estas aproximaciones permiten entender cómo los principios físicos están siendo aplicados en el estudio de la evolución y la predicción de mutaciones en el ADN. Cada una aborda aspectos diferentes, desde la termodinámica y la teoría de la información hasta la mecánica cuántica y la dinámica molecular, lo que muestra la interdisciplinariedad del campo. Mientras que estos enfoques son prometedores, hay desafíos significativos en su aplicación, incluyendo la complejidad computacional y la necesidad de datos experimentales para validar los modelos [Bialek, 2012].

En conclusión, la utilización de principios físicos para entender y predecir mutaciones del ADN está en una etapa emergente pero prometedora. A medida que la tecnología avanza, es probable que estos métodos se vuelvan más sofisticados y útiles para abordar preguntas en evolución y genética.

1.5.2 Uso de la Distribución de Boltzmann en Modelos de Fenómenos Biológicos

La distribución de Boltzmann ha encontrado aplicaciones diversas fuera de la física pura, y uno de estos ámbitos es la biología. La descripción estadística de sistemas a través de esta distribución permite analizar fenómenos biológicos a nivel molecular y celular, desde interacciones proteicas hasta dinámicas de poblaciones de células.

Termodinámica de Interacciones Moleculares

Un área donde la distribución de Boltzmann se ha empleado ampliamente es en la modelización de la afinidad entre ligandos y receptores. La fijación de un ligando a un receptor puede modelarse en términos de un sistema que busca minimizar su energía libre, tal como se describe en la ley de acción de masas y el modelo de Langmuir [Cantor, C. R., & Schimmel, P. R. 1980].

Plegamiento de Proteínas

La distribución de Boltzmann también se ha utilizado para modelar el plegamiento de proteínas. Dada una cadena de aminoácidos, la forma final de la proteína está determinada por el estado de mínima energía libre, y se han desarrollado algoritmos estocásticos basados en la distribución de Boltzmann para predecir estas estructuras [Dill, K. A., et.al. 2008].

Redes Neuronales y Codificación de Información

En neurociencia, la distribución de Boltzmann se ha utilizado para modelar la distribución de estados en una red neuronal, especialmente en máquinas de Boltzmann restringidas [Ackley, D. H., et.al. 1985].

Aunque la aplicación de la distribución de Boltzmann en biología es poderosa, tiene sus limitaciones, incluida la suposición de equilibrio termodinámico, que puede no ser válida para todos los sistemas biológicos [Klumpp, S., & Hwa, T. 2008]. En conclusión la distribución de Boltzmann es una herramienta útil en la biología computacional y teórica, y su uso promete seguir arrojando luz sobre la complejidad inherente a los sistemas biológicos.

1.5.3 La presión mutacional termodinámica del ADN: un posible factor en la evolución del genoma

En un estudio previo, realizamos un estudio para evaluar la estabilidad termodinámica del proceso de replicación del ADN y sus consecuencias en la evolución genómica. Dicho estudio consistió de una simulación computacional, utilizando un algoritmo de Monte Carlo, empleando probabilidades estimadas por la distribución de Boltzmann para un ensamble canónico (Apéndice 1). El sesgo en el uso de codones es una característica de los organismos vivos. El origen de este sesgo podría explicarse no sólo por factores externos,

sino también por la naturaleza de la propia estructura del ácido desoxirribonucleico (ADN). Hemos desarrollado un programa de simulación de mutaciones puntuales de secuencias codificantes, en el que la sustitución de nucleótidos sigue criterios termodinámicos. Para ello hemos calculado las energías de enlace de hidrógeno y electrostática de pares de bases no canónicas en una vecindad de 5 pb. Aunque la tasa de formación de pares de bases no canónicas es extremadamente baja, dichos pares se producen con preferencia hacia una sustitución por guanina (G) o citosina (C) en lugar de adenina (A) o timina (T), debido a consideraciones termodinámicas. Esta característica, según el programa de simulación, debería traducirse en un aumento del contenido en GC del genoma a lo largo del tiempo evolutivo. Además, también se predice un sesgo de los codones hacia un mayor uso de GC. El análisis de la secuencia de ADN de los genes del linaje Trypanosomatidae corroboró la hipótesis de que la presión mutacional termodinámica del ADN es una fuerza motriz que impulsa el aumento del contenido de GC y el sesgo de codones GC.

La incuestionable desviación observada de las mutaciones de secuencia del ADN con respecto a la aleatoriedad.

La incuestionable desviación de la aleatoriedad de las mutaciones de secuencia del ADN se analizó aquí teniendo en cuenta la posible contribución de la estructura de doble cadena del ADN. Las siguientes secciones ofrecen apoyo teórico al postulado de que la presión mutacional termodinámica es un factor relevante para explicar los cambios en la secuencia del ADN a lo largo de escalas temporales evolutivas. La sección que sigue sobre el análisis de secuencias y datos compara las predicciones de las simulaciones con las observaciones del proceso microevolutivo del linaje tripanosomátido.

Predicciones de la presión mutacional termodinámica

Cuando se utilizó la distribución de Boltzmann para simular el cambio en el contenido de GC durante un largo periodo de tiempo, se observó un aumento continuo de estos nucleótidos hasta que se alcanzó una meseta. El valor de la meseta dependía de la

información de la secuencia. Así, si la secuencia no tenía ninguna función codificante y, por tanto, podía mutar libremente, la meseta alcanzaba valores superiores al 95%, aunque nunca llegaba al 100%. Sin embargo, si la simulación consideraba una secuencia codificante, la meseta observada oscilaba entre el 60% y el 75% cuando se imponía la conservación de la familia de aminoácidos. Además, cuando se impuso la restricción de mantener constante la identidad de aminoácidos, las secuencias simuladas alcanzaron antes una meseta más baja, dependiendo de cuándo se impuso la restricción. Las simulaciones aquí presentadas se realizaron considerando concentraciones equimolares de nu- cleótidos y una temperatura de 37°C; sin embargo, si se modificaban estos parámetros variaba el valor de la meseta y la velocidad necesaria para alcanzarla. A temperaturas más elevadas se obtuvo un mayor contenido de GC y se observó un aumento más rápido (datos no mostrados). Según la simulación descrita, cualquier genoma evolucionaría espontáneamente hacia un mayor contenido de GC en condiciones en las que la única fuerza que actuara sobre él fuera lo que hemos denominado presión mutacional termodinámica. Esta fuerza es independiente de la naturaleza de la estructura del ADN, una característica que la distingue de la presión de selección, que depende de muchos factores ambientales que actúan como un tamiz sobre las poblaciones. Para evaluar la validez de la presión mutacional termodinámica postulada, hemos elegido una situación en la que los organismos están expuestos a una presión de selección limitada y bastante constante. El análisis de las secuencias genéticas de tripanosomátidos depositadas en GenBank reveló un uso sesgado de codones [Alonso et al., 1992]. El primer informe afirmaba que la presión mutacional hacia GC o AT era responsable de la divergencia observada en el uso de codones. Sin embargo, cuando los miembros de los linajes mencionados anteriormente se ordenaron filogenéticamente desde los linajes más antiguos hasta los que se originaron más recientemente, basándose en la subunidad pequeña del ARN ribosómico [Maslov. et al., 1994, 1995], observamos que mostraban una clara tendencia a aumentar el contenido de GC (Fig. 2). Este hecho podría reflejar la presión mutacional termodinámica y no, como postulan Alonso et al. (1992), la

reminiscencia de genomas primigenios. El aumento global observado en el contenido de GC en las secuencias genéticas de tripanosomátidos podría ser un artefacto, resultado del equilibrio final entre genes diferentes, algunos de ellos con alto contenido de AT y otros con alto contenido de GC. Si esto fuera cierto, Leishnzania y Crithidia deberían tener una mayor proporción de genes ricos en GC, mientras que T. brucei tendría lo contrario. Alternativamente, el mayor uso de codones GC de los tripanosomátidos modernos podría reflejar la evolución de la mayoría, si no de todos sus genes, hacia un mayor contenido de GC.

Hasta ahora, hemos analizado más de 20 genes diferentes que se han descrito para 2 o más especies de tripanosomátidos. Ninguno de ellos presentaba un sesgo hacia un aumento de la TA (los datos se publicarán en otro lugar). Como ejemplo, la Fig. 3 ilustra 4 genes diferentes que demuestran un enriquecimiento del contenido de GC en *Leishmania* y Crithidia, mientras que la especie más antigua, T. brucei, tiene el menor contenido de GC en los genes correspondientes; *T. cruzi* ocupa una posición intermedia. Además, como era de esperar, las familias de codones (cuarteto o sexteto) de los tripanosomátidos modernos utilizaban codones más ricos en composición de GC que los codones utilizados por T. brucei (Fig. 4). El modelo presentado aquí implica que el ADN en evolución no está en equilibrio termodinámico. La presión mutacional termodinámica impulsó a las moléculas ancestrales de ADN ricas en AT hacia un estado máximo rico en GC. Debido a consideraciones estructurales y termodinámicas de las secuencias de nucleótidos, la molécula de ADN no mutó al azar, sino que se mostró hacia un aumento de GC tanto para las secuencias codificantes como para las no codificantes. Como consecuencia de esta presión mutacional termodinámica, existe una tendencia hacia un sesgo en el uso de codones GC y hacia un aumento del contenido de GC del genoma cuando se consideran escalas microevolutivas. Recientemente, [Galtier, et. al., 1999] han encontrado pruebas de que el ancestro común más reciente de los organismos vivos comenzó con un alto contenido en AT, independientemente de cualquier necesidad de estabilidad térmica. El escenario propuesto de polímeros ricos en AT tiene sentido porque los nucleótidos de

adenina son energéticamente los menos costosos de sintetizar no enzimáticamente. Además, los polímeros cortos ricos en adenina serían más estables en un entorno prebiótico, debido a las fuerzas de apilamiento. La presente hipótesis no contradice las teorías de la presión mutacional [Sueoka N., 1988] ni de la preselección [Bernardi G. et.al., 1988; Mouchiroud D. et.al,1989]. El modelo hipotetizado aquí predice que los genomas evolucionan hacia un mayor contenido de GC, pero podrían eventualmente moverse hacia estados ricos en AT si el ambiente intracelular presentara condiciones diferentes o si las presiones de selección fueran lo suficientemente fuertes como para contrarrestar la presión mutacional termodinámica. Por otra parte, las presiones de selección termodinámicas tendrían un efecto sinérgico con la presión mutacional termodinámica en el caso de los vertebrados de sangre caliente.

CAPITULO II. Marco teórico y conceptual

2.1. Marco teórico

2.1.1 Introducción al Ácido Desoxirribonucleico (ADN)

El ácido desoxirribonucleico, conocido comúnmente como ADN, es una molécula biológica esencial que se encuentra en casi todas las células vivas y es portadora de la información genética necesaria para la organización, funcionamiento y reproducción de los organismos. Esta molécula fundamental no solo alberga las instrucciones para el desarrollo de un organismo, sino que también es crucial para la herencia genética, transmitiendo características de una generación a la siguiente. El ADN se compone de dos largas cadenas de nucleótidos, cada una formada por un grupo fosfato, un azúcar de desoxirribosa y una base nitrogenada, que se entrelazan formando una estructura en forma de doble hélice. Las bases nitrogenadas, que son adenina (A), timina (T), citosina (C) y guanina (G), se emparejan específicamente entre sí para formar los "escalones" de la hélice, donde adenina siempre se une con timina y citosina siempre con guanina. Esta complementariedad de bases es fundamental, pues permite la replicación del ADN y la precisa transmisión de la información genética.

El descubrimiento de la estructura del ADN por James Watson y Francis Crick en 1953, basado en parte en los trabajos de Rosalind Franklin y Maurice Wilkins, fue un hito que transformó la ciencia moderna, abriendo las puertas a la era de la genética molecular. Desde entonces, el ADN ha sido central en numerosas aplicaciones científicas y tecnológicas, incluyendo la medicina, la forense y la biotecnología agrícola. En la medicina, por ejemplo, la comprensión del ADN ha llevado al desarrollo de la medicina personalizada, donde el análisis genético de un individuo puede guiar tratamientos y terapias específicas adaptadas a su perfil genético, mejorando la eficacia y reduciendo los riesgos y efectos secundarios.

La principal función del ADN es almacenar información que determina las características de un organismo y regular la producción de proteínas que realizan todas las funciones

biológicas clave. Esta información se organiza en unidades funcionales llamadas genes, cada una de las cuales contiene el código para la producción de una proteína específica o para la realización de una función particular en la célula. Sin embargo, no toda la secuencia del ADN está compuesta por genes; algunas regiones juegan roles reguladores, mientras que otras parecen no tener función conocida. Además, el ADN tiene un papel crucial en la regulación de sus propias funciones: determina cuándo y cómo se deben expresar los genes, asegurando que las proteínas se produzcan en el momento y lugar adecuados dentro de las células y tejidos.

Una característica notable del ADN es su estabilidad química, que permite que la información genética se conserve durante millones de años, una propiedad esencial para la transmisión de la información a lo largo de generaciones. A pesar de esta estabilidad, el ADN es susceptible a mutaciones, que son cambios en la secuencia de nucleótidos. Estas mutaciones pueden ser causadas por errores durante la replicación del ADN o por factores externos como radiación ultravioleta y sustancias químicas mutagénicas. Aunque muchas mutaciones son neutras o perjudiciales, algunas pueden conferir ventajas adaptativas a los organismos, siendo un motor de la evolución por selección natural.

El impacto del ADN va más allá de la biología y la medicina; también tiene profundas implicaciones en áreas como la justicia, donde el análisis del ADN es fundamental en la identificación forense y en la resolución de casos criminales. En el campo de la agricultura, la ingeniería genética ha permitido la creación de cultivos más resistentes a enfermedades y con mejores rendimientos, contribuyendo a la seguridad alimentaria global. Además, la investigación continua sobre el ADN está ayudando a desentrañar los misterios de enfermedades complejas, como el cáncer y las enfermedades genéticas, abriendo la puerta a nuevas y revolucionarias terapias.

En conclusión, el ADN no es solo la base física de la herencia, sino un pilar central en la ciencia y tecnología modernas. Comprender su estructura y función es crucial para los avances en salud, agricultura, forense y muchas otras disciplinas que dependen de la genética y la biología molecular. La doble hélice del ADN sigue siendo uno de los

descubrimientos más impactantes del siglo XX, con efectos duraderos que continúan transformando nuestra sociedad y nuestro entendimiento de la vida misma.

2.1.2 Clasificación del ADN

En una sola molécula de ADN, es posible categorizar tres distintas clases de ADN, cada una con atributos específicos: ADN funcional codante, ADN funcional no-codante y ADN no-funcional, también conocido como "ADN basura". El ADN funcional codante es el tipo de ADN encargado de la codificación de aminoácidos, y por ende, es crucial para determinar el fenotipo del organismo. Aunque el código genético permite cierta redundancia, mutaciones menores pueden ocurrir en este tipo de ADN sin cambiar la secuencia de aminoácidos resultante; a este tipo de mutaciones se les llama "mutaciones silenciosas". Por otro lado, el ADN funcional no-codante es aquel que cumple una función específica, ya sea directamente (ADN no transcrito) o mediante su producto de transcripción (ARN que se transcribe a partir del ADN), pero que no se traduce en una proteína. Este tipo de ADN incluye elementos vitales como tARNs, subunidades ribosomales, secuencias promotoras, secuencias potenciadoras, intrones, telómeros, centrómeros y orígenes de replicación [Lewin,1994]. Su funcionalidad, que en parte depende de la secuencia de nucleótidos, debe mantenerse constante a lo largo del tiempo debido a su importancia. Finalmente, el ADN no-funcional, o "basura", es aquel que ni codifica aminoácidos ni cumple una función específica ya sea como ADN o como ARN [Lewin,1994]. Este tipo de ADN se distingue por su falta de función operativa.

2.1.3 Mutaciones

La mutación se define generalmente como cualquier cambio en la secuencia del ADN. Este cambio puede ocurrir en las purinas y pirimidinas, que son componentes de los nucleótidos del ADN. Estas alteraciones pueden tener un impacto significativo en la información genética y están vinculadas a diversas patologías, como el cáncer y el envejecimiento celular [Lewin,1994].

Las mutaciones se pueden clasificar en diversas categorías, incluyendo sustituciones de nucleótidos (mutaciones puntuales), inserciones y deleciones de nucleótidos, y modificaciones cromosómicas. Las mutaciones puntuales se dividen en transiciones y transversiones, donde la primera implica la sustitución de una purina por otra purina o una pirimidina por otra pirimidina, y la segunda se refiere a la sustitución de una purina por una pirimidina y viceversa [Lewin,1994; Lehninger, Nelson, Cox,1993].

Es importante notar que las mutaciones pueden ser heredables o somáticas, dependiendo de la célula en la que ocurren y el tipo de reproducción del organismo. Las mutaciones en el ADN funcional o codificante generalmente tienen consecuencias más graves, como el cambio en la estructura y función de una proteína.

Una observación crítica es que ciertas áreas del ADN son más propensas a mutaciones que otras, conocidas como "hotspots". Aunque la teoría de la cinética de choques aleatorios sugiere una distribución uniforme de las mutaciones a lo largo del ADN, la realidad muestra que algunos sitios tienen una mayor frecuencia de mutaciones [Lewin,1994].

Además, las mutaciones están a menudo asociadas con bases nitrogenadas inusuales en el ADN, como la 5-metilcitosina. Estas bases modificadas son hotspots para mutaciones puntuales espontáneas, especialmente en organismos como *Escherichia coli* [Lewin,1994; Ayala,1984].

También es relevante el mecanismo de reparación del ADN que corrige los pares de bases no canónicos. Cuando este mecanismo falla, se pueden producir mutaciones puntuales. De hecho, la presencia de 5-metilcitosina y su deaminación espontánea en timina, un componente del ADN, genera un apareamiento erróneo que puede dar lugar a una mutación.

Finalmente, se ha observado que la deaminación de la citosina a uracilo es una de las razones por las cuales el ADN contiene timina en lugar de uracilo. El uracilo, producto de la deaminación, es usualmente reconocido como una base errónea y eliminado, mitigando el impacto de la deaminación. Este mecanismo de reparación de excisión de bases es vital para mantener la integridad del ADN.

Por lo tanto, entender las mutaciones, sus tipos, y sus consecuencias en los organismos vivos es crucial para la biología molecular y la medicina. Su estudio también tiene implicaciones en la comprensión de la evolución y el desarrollo de terapias génicas [Lewin,1994; Lehninger, Nelson, Cox,1993; Ayala,1984].

2.1.4 Agentes mutagénicos

El estudio del daño al ADN y sus mecanismos subyacentes es un área de investigación que ha recibido una atención significativa debido a su importancia en la comprensión de diversas patologías y procesos celulares [Nelson, 1996; Lehninger, Nelson, Cox, 1993]. Entre los diferentes factores que contribuyen a las alteraciones genéticas, los agentes mutagénicos ocupan un lugar central. Estos pueden clasificarse en tres categorías principales: agentes físicos, agentes químicos y agentes biológicos.

1. **Agentes Físicos**: Esta categoría abarca una amplia gama de elementos, desde la radiación ionizante y ultravioleta hasta altas temperaturas y campos eléctricos o magnéticos intensos. Estos factores pueden interferir con las enzimas que facilitan la replicación del ADN, alterando así la integridad genética [Kunkel TA, et.al., 1996].
2. **Agentes Químicos**: Estos agentes, como el bromuro de etidio, afectan directamente la estructura y función del ADN. Interfieren en el proceso de replicación, generando errores que pueden llevar a mutaciones [Suzuki, 1983].
3. **Agentes Biológicos**: Este grupo incluye entidades biológicas como virus que pueden insertar material genético en el huésped, provocando así mutaciones. Sin embargo, es crucial señalar que el mecanismo más influyente en la mutagénesis es el sistema endógeno de replicación y reparación del ADN. Este sistema, que opera de manera continua, es el agente mutagénico más relevante y a menudo se pasa por alto en la discusión de agentes externos [Kunkel, 1989].

Es fundamental tener un conocimiento completo de estos agentes mutagénicos para comprender los mecanismos que subyacen al daño del ADN y cómo este contribuye a la

variabilidad genética y al desarrollo de enfermedades. El entendimiento detallado de estos factores ofrece la oportunidad de desarrollar estrategias de intervención más efectivas.

En resumen, los factores mutagénicos comprenden una serie de procesos que pueden desencadenar cambios en el ADN. Entre ellos, se encuentran las radiaciones como los rayos ultravioleta, que poseen la energía necesaria para alterar los enlaces en la estructura del ADN. Este tipo de radiación puede crear anormalidades específicas, como los dímeros de ciclobutano-pirimidina, que afectan el proceso de copia del ADN. Todas las formas de vida se enfrentan a algún grado de exposición a este tipo de radiación, que se estima es responsable de aproximadamente el 10% de los daños al ADN causados por fuentes no biológicas.

Además, se ha observado que la exposición a luz ultravioleta puede incrementar significativamente ciertos componentes del ADN en algunos organismos, aunque aún no se comprende completamente el mecanismo detrás de estos cambios.

Por otra parte, ciertas sustancias químicas también pueden dañar el ADN de formas diversas, como la alteración de bases y daño oxidativo. Estos incluyen:

1. Desaminantes como el ácido nitroso.
2. Alquilantes como el dimetilsulfato.
3. Sustancias que imitan a las bases del ADN.
4. Agentes que provocan daños oxidativos, como el peróxido de hidrógeno y varios tipos de radicales.

Un fenómeno poco entendido pero relevante se relaciona con el desequilibrio en las concentraciones de nucleótidos trifosfato, que desempeña un papel crucial en ciertas células. Este desequilibrio puede llevar a mutaciones y otros cambios genéticos dramáticos en distintas células tanto procariotas como eucariotas.

En experimentos, se ha observado que las mutaciones ocurren de forma diferente dependiendo del tipo de desequilibrio de nucleótidos. Por ejemplo, en células de ratón expuestas a altas concentraciones de ATP, ciertos tipos de mutaciones son más prevalentes. Además, se ha notado que el desequilibrio en nucleótidos tiene un impacto

más significativo en el ADN mitocondrial, ya que carece de un sistema de reparación eficiente. En suma, el equilibrio de nucleótidos trifosfato es crítico para procesos como la replicación del ADN y la corrección de errores. Los desbalances pueden ser inducidos por factores externos, como la radiación ultravioleta, y tienen el potencial de provocar mutaciones.

2.1.5 Mecanismo de Replicación del ADN

El ácido desoxirribonucleico (ADN) es la molécula biológica que almacena la información genética en todos los organismos vivos. La replicación del ADN es un proceso biológico fundamental que asegura la transmisión precisa de la información genética de una célula madre a sus células hijas. La comprensión de los mecanismos moleculares subyacentes es esencial para entender cómo la vida se perpetúa y cómo se pueden producir errores que resultan en enfermedades como el cáncer.

Mecanismo General de Replicación

La replicación del ADN es un proceso semiconservativo, como fue confirmado por el experimento de Meselson-Stahl en 1958 [Meselson and Stahl, 1958]. En este proceso, cada cadena de la doble hélice de ADN sirve como molde para la síntesis de una nueva cadena complementaria.

Iniciación

La iniciación de la replicación comienza con el reconocimiento de secuencias específicas en el ADN denominadas "orígenes de replicación" por una variedad de proteínas iniciadoras. En eucariotas, la proteína ORC (Complejo de Reconocimiento del Origen) marca el sitio de inicio [Bell and Dutta, 2002].

Elongación

Durante la elongación, una enzima denominada ADN polimerasa añade nucleótidos a una cadena en crecimiento. En procariontes, como *E. coli*, la ADN polimerasa III es la principal enzima responsable de este proceso [Kornberg and Baker, 1992]. En eucariotas, diferentes

tipos de ADN polimerasas (como Pol ε y Pol δ) llevan a cabo la síntesis del ADN [Muzi-Falconi and Giannattasio, 2019].

Terminación

La replicación se completa cuando la maquinaria de replicación encuentra una señal de terminación o cuando se han replicado todas las secuencias de ADN.

Proteínas Asociadas y Maquinaria

- Helicasas: Desenrollan la doble hélice.
- Topoisomerasas: Resuelven la tensión en la doble hélice durante la replicación.
- Primasa: Sintetiza cebadores de ARN que sirven como puntos de partida para la síntesis de ADN.
- Ligasa de ADN: Cierra las discontinuidades en la columna vertebral de fosfato.

Fidelidad y Corrección de Errores

Las polimerasas de ADN tienen una alta fidelidad y poseen mecanismos de corrección de errores. Estos mecanismos incluyen la edición exonucleolítica que corrige los errores de apareamiento [Morrison et al., 1991].

2.1.6 Mecanismos Subyacentes en la Fidelidad de Replicación del Ácido Desoxirribonucleico (ADN): Mecanismo de Reparación del ADN

Para la preservación de la integridad genómica, es imperativo que los organismos biológicos repliquen su ADN con alta precisión. No obstante, la comprensión de los mecanismos fundamentales que garantizan la selección adecuada del deoxinucleótido trifosfato 5' (dNTP) durante la polimerización es limitada [Normile, 1996; Mellon, 1996]. Este documento aborda la contribución de múltiples componentes y etapas secuenciales en la fidelidad de la replicación del ADN, enfocándose en tres áreas distintas: teorías matemáticas, genética de procariotas y bioquímica.

Polimerasas del ADN

Numerosas polimerasas han sido purificadas y analizadas desde diversas fuentes biológicas. La fidelidad en la replicación del ADN se alcanza mediante un proceso tripartito.

Primero, se minimiza la incorporación de nucleótidos erróneos en la etapa de elongación de la cadena de ADN. Este fenómeno puede estar influenciado por la diferencia de energía libre entre los pares de bases correctos e incorrectos, la cual puede ser magnificada por la acción de la ADN polimerasa y otras proteínas asociadas [Loeb & Kunkel, 1982].

La segunda fase implica la capacidad de 'revisión' ('proofreading') que algunas polimerasas exhiben, donde un nucleótido incorrectamente incorporado puede ser eliminado inmediatamente después de su adición. Finalmente, existe un sistema de reparación post-replicación que contribuye a una mayor precisión en la replicación [Loeb & Kunkel, 1982].

Consideraciones Termodinámicas

Aunque las interacciones por puentes de hidrógeno favorecen la formación de pares de bases AT y GC, la diferencia energética en un ambiente acuoso no es suficiente para garantizar un bajo nivel de errores [Petruska & Sowers, 1986]. Diversos estudios han demostrado que las diferencias en las energías libres de disociación ($\Delta G°$) para los pares de bases complementarios e incorrectos en una solución acuosa varían entre 0.2 y 0.4 kcal/mol, pero pueden aumentar hasta diez veces en el sitio activo de la ADN polimerasa [Petruska & Goodman, 1988].

Metilación y Reparación del ADN

En organismos como Escherichia coli, la corrección de errores de emparejamiento post-replicativos incrementa la eficiencia general de la replicación en un factor de 10^2 a 10^3 [Lehninger, Nelson, Cox, 1993]. Este mecanismo de reparación de errores de emparejamiento se basa en la actividad de la enzima Dam metilasa, que metila específicamente las adeninas en las secuencias 5'-GATC en la hebra matriz, permitiendo así la discriminación entre la hebra recién sintetizada y la matriz.

Esencialmente, este proceso de reparación de malos apareamientos, a veces referido como reparación guiada por metilación, es capaz de corregir errores hasta 1000 pares de bases alejados de una secuencia GATC parcialmente metilada o hemimetilada.

2.1.7 Presión de Selección Como Interacción entre Medio Ambiente y Adaptación (Fitness)

La presión de selección representa el conjunto de fuerzas ambientales y ecológicas que influyen en la capacidad de un organismo para sobrevivir y reproducirse. Este concepto es fundamental para entender cómo las especies evolucionan y se adaptan a su entorno a lo largo del tiempo.

Lucha por la Existencia y Supervivencia del Más Apto

La existencia de todo ser vivo implica una competencia constante por recursos limitados y un enfrentamiento con factores ambientales desafiantes [Hartl & Clark, 1989]. Esta lucha se manifiesta tanto dentro de las especies como entre diferentes especies. Sin embargo, la competencia por la supervivencia es asimétrica; es decir, no todos los individuos ni todas las especies tienen las mismas probabilidades de sobrevivir. Los individuos mejor adaptados al entorno son los que tienen mayores probabilidades de reproducirse y transmitir sus genes [Darwin & Kebler, 1859].

Extinción y Adaptación Genética

Las especies que no pueden adaptarse al ambiente, especialmente aquellas que tienen un "alto riesgo vital" debido a la falta de mecanismos de supervivencia efectivos, tienen una alta probabilidad de extinción [Hartl & Clark, 1989]. En contraste, los organismos que experimentan mutaciones genéticas favorables tienen más posibilidades de transmitir estos rasgos a su descendencia. Sin embargo, mutaciones desfavorables también pueden surgir, y los individuos que las portan generalmente no sobreviven lo suficiente para reproducirse.

Selección Natural como Fuerza Adaptativa

La selección natural es el principal mecanismo a través del cual ocurre la adaptación. Es el proceso que actúa sobre la variabilidad genética de una población para favorecer los rasgos que mejoran la aptitud del organismo [Darwin & Kebler, 1859]. Esta fuerza evolutiva no sólo impulsa la adaptación, sino que también actúa como una barrera contra la desorganización genética.

Presión de Selección como Filtro

Una forma simplificada pero ilustrativa de entender la presión de selección es verla como un "filtro" que permite la supervivencia de ciertos individuos mientras elimina otros. Este filtro se modifica cuando cambian las condiciones ambientales, permitiendo la adaptación de nuevos organismos que antes no podían sobrevivir y, por otro lado, llevando a la extinción de aquellos que ya no se adaptan [Hartl & Clark, 1989].

La presión de selección por lo tanto se puede entender como la dinámica entre la adaptación y las fuerzas ambientales. Actúa como un mecanismo de filtro que determina qué organismos sobreviven y se reproducen, basado en su nivel de adaptación al entorno. Este proceso es esencial para la evolución y la diversidad biológica.

2.1.8 Comprendiendo la Evolución a través de la Anagénesis y Cladogénesis

El concepto de evolución es esencial para comprender la diversidad biológica y las dinámicas de cambio en las especies. Este cambio evolutivo se manifiesta en dos dimensiones principales: la anagénesis y la cladogénesis. Ambas representan formas distintas de adaptación y cambio en los seres vivos [Ayala, 1984].

Anagénesis: Evolución dentro de un Linaje

La anagénesis se refiere a los cambios que ocurren dentro de un linaje único a lo largo del tiempo. Estos cambios suelen ser el resultado de la selección natural, que favorece adaptaciones a modificaciones físicas o biológicas en el entorno [Ayala, 1984]. En un contexto de anagénesis, es probable que los cambios genéticos resulten en un fenotipo relativamente estable. Este proceso es similar al concepto de "envejecimiento de secuencias" en la evolución de secuencias de ADN, donde las modificaciones genéticas tienden a mantener el fenotipo [Futuyma D., 2013].

Cladogénesis: Diversificación de Linajes

En contraste, la cladogénesis implica la división de un linaje en dos o más nuevos linajes, lo que conduce a una mayor diversidad biológica y adaptación a múltiples nichos ecológicos [Ayala, 1984]. Este proceso culmina en la especiación, donde una especie se

divide en dos o más nuevas especies. En el contexto de la cladogénesis, es más probable que se observen cambios fenotípicos significativos, que es análogo al concepto de "diversificación de secuencias" en la evolución del ADN [Darwin & Kebler, 1859].

Anagénesis y Cladogénesis en la Evolución del ADN

Es importante señalar que aunque los conceptos de anagénesis y cladogénesis se aplican comúnmente a la evolución de organismos y poblaciones, también pueden adaptarse para describir la evolución de secuencias de ADN. En este contexto, se utilizan términos como "envejecimiento" para describir procesos anagenéticos de secuencias y "diversificación" para representar procesos cladogenéticos de secuencias [Kimura, 1983].

Interconexiones y Complejidad

Los procesos de anagénesis y cladogénesis no son mutuamente excluyentes y, de hecho, suelen ocurrir simultáneamente, impulsados por complejos mecanismos que aún se están explorando [Gould, 2002].

La evolución por lo tanto es un proceso complejo que puede comprenderse mejor a través de las dimensiones de anagénesis y cladogénesis. Estos conceptos nos ayudan a entender los mecanismos mediante los cuales las especies se adaptan, cambian y diversifican en respuesta a las fuerzas selectivas y los desafíos ambientales.

2.2. Marco conceptual

El estudio de la evolución del ADN, centrado en los mecanismos de replicación y reparación, constituye un campo fundamental en la biología molecular. Este marco conceptual se enfoca en entender cómo las mutaciones, esenciales para la evolución, se generan y acumulan en el genoma.

1. Replicación del ADN y Generación de Mutaciones: En la replicación del ADN, la incorporación de nucleótidos en la cadena naciente es un proceso crítico. Las mutaciones, cambios en la secuencia de nucleótidos, pueden surgir durante este proceso debido a errores de replicación o factores externos que dañan el ADN.

2. Teoría de la Distribución de Boltzmann en Replicación del ADN: La distribución de Boltzmann, aplicada a un conjunto canónico, se utiliza para predecir la probabilidad de que un nucleótido libre se incorpore al sitio de replicación. Esta probabilidad depende de varios factores, como la energía potencial molecular, la concentración de nucleótidos, la secuencia de ADN adyacente, y la temperatura.
3. Factores Energéticos en la Selección de Nucleótidos: La energía potencial del sistema, incluyendo la energía de puentes de hidrógeno y las interacciones electrostáticas, influye en la selección de nucleótidos durante la replicación. Estas interacciones determinan la afinidad de los nucleótidos hacia el sitio activo de la enzima ADN polimerasa.
4. Simulación Computacional y Análisis Analítico: La simulación computacional y el análisis analítico son dos enfoques empleados para estudiar la evolución del contenido de nucleótidos en el ADN. Utilizando estos métodos, se puede predecir cómo varía la composición de nucleótidos, especialmente la proporción de guanina-citosina (GC), a lo largo del tiempo evolutivo.
5. Presión Mutacional Termodinámica: Este concepto propone que hay una fuerza impulsora, derivada de factores termodinámicos, que influye en la evolución del ADN. La presión mutacional termodinámica, en interacción con la presión de selección, podría determinar las características de los genomas a lo largo de la evolución.
6. Fenómeno del 'Codon Bias': El 'codon bias' es una tendencia observada en genomas donde ciertos codones son más prevalentes que otros. Este fenómeno puede explicarse a través de las probabilidades derivadas de la distribución de Boltzmann, sugiriendo una relación con la presión mutacional termodinámica.

Este marco conceptual proporciona una base para comprender los mecanismos moleculares y termodinámicos que rigen la evolución del ADN, y ofrece un enfoque para investigar preguntas fundamentales en biología evolutiva.

CAPITULO III. Desarrollo del trabajo de investigación

3.1 Modelo termodinámico basado en la distribución de Boltzmann para estimar las probabilidades de formación de pares no canónicos (missmatches)

En el presente estudio, empleamos el marco teórico, previamente desarrollado y descito [Zimic M. et.al. 2002], que considera el uso de la distribución de Boltzmann para estimar las probabilidades de formación de missmatches, para evaluar el efecto de la estabilidad temodinámica durante el proceso de replicación del ADN, a lo largo del proceso evolutivo.

La evolución puede ser entendida a través de dos dinámicas principales: el proceso de envejecimiento y la diversificación. Las fuerzas que dirigen la evolución son la presión de selección y la termodinámica. Aunque es complicado medir la presión de selección debido a la necesidad de examinar estructuras tridimensionales y su interacción con el ambiente, la presión mutacional termodinámica es más directa y empuja al ADN hacia estados con menor energía potencial, reflejado principalmente en una acumulación de contenido GC.

Este paradigma de estabilidad termodinámica se ha aplicado en distintos campos biológicos, pero no en el ámbito de las mutaciones puntuales. Es esencial resaltar que mientras la presión de selección opera de forma aleatoria, la termodinámica se rige por la distribución de Boltzmann, siendo el principal motor evolutivo.

El ADN se puede clasificar en tres tipos: ADN codificante, ADN no codificante (comúnmente llamado "basura") y ADN funcional. Cada uno evoluciona de forma única. El ADN codificante está influenciado por ambas presiones, ya que sus proteínas interactúan con el entorno. Sin embargo, se piensa que el ADN no codificante "basura" evoluciona solamente por la presión mutacional termodinámica debido a su falta de función y ausencia de interacción con el ambiente.

Considerando la dificultad de cuantificar la presión de selección, nos enfocamos en casos donde la termodinámica predomina, en particular cuando el entorno es constante. La molécula de ADN se modelará como un conjunto canónico en un reservorio térmico, lo que

nos permite usar la distribución de Boltzmann para calcular probabilidades de mutaciones concretas.

Proponemos un modelo fisicoquímico que simula mutaciones puntuales, generando "huecos" en el ADN que son posteriormente ocupados por un nucleótido, basándose en la distribución de Boltzmann. Luego, se incorporan mecanismos de reparación para enmendar emparejamientos incorrectos.

Consideraciones para determinar la viabilidad de mutaciones en el ADN codante.

Proponemos criterios simplificados para este análisis. En el ADN no funcional, se asume que todas las mutaciones son viables. Para el ADN funcional-codificante, diferenciamos mutaciones relacionadas con envejecimiento y diversificación, suponiendo un entorno constante. Es importante recalcar que estos criterios son aplicables únicamente en contextos donde el ambiente es estable y la presión de selección es mínima. Respecto al ADN funcional no codificante, la evaluación de mutaciones es aún más desafiante, y es un campo en desarrollo.

El ADN, visto como sistema termodinámico, está estabilizado por interacciones electromagnéticas como enlaces covalentes, interacciones electrostáticas, puentes de hidrógeno, entre otros. Aunque la mecánica cuántica es la descripción formal de un sistema molecular, su complejidad hace que no sea la mejor opción para estudiar el ADN. Una alternativa es la dinámica molecular, pero tiene limitaciones, como la imposibilidad de simular la formación o ruptura de enlaces covalentes.

Por lo tanto, sugerimos una aproximación más sencilla, tratando al ADN como un sistema termodinámico en un conjunto canónico. A pesar de las simplificaciones, esta perspectiva permite obtener valores promedio de observables físicos relevantes. Dada la estabilidad de temperatura en la célula y la constancia en volumen y presión, aplicamos la distribución de Boltzmann para determinar la probabilidad de mutaciones puntuales. Deberá hacerse una aproximación al número de estados accesibles para evitar los términos entrópicos [Reif,1965; Mayer & Mayer,1963; Ter Haar,1961].

Génesis de "Huecos" en el ADN y Eventos Mutagénicos: Comparación de Velocidades

Definimos un "hueco" en el ADN como una ausencia de uno o varios nucleótidos en una de sus hebras, sin que esto implique una fractura completa de la molécula. Estos huecos pueden originarse por diferentes factores que, en general, llamaremos agentes mutagénicos. Entre los más relevantes destacan la radiación ionizante, radicales libres y colisiones térmicas a altas temperaturas. La Figura 1 ilustra de manera esquemática esta formación.

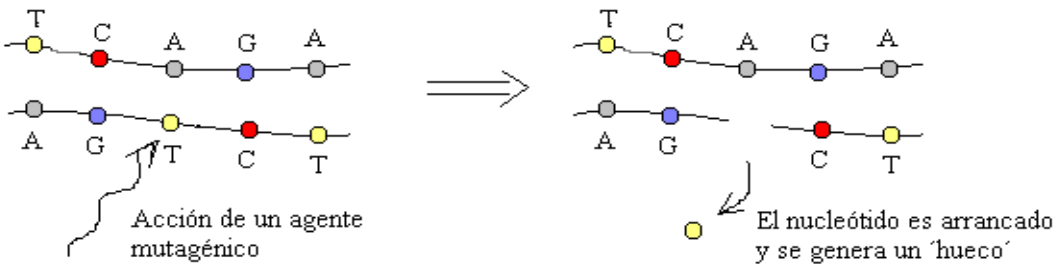


Figura 1. Esquema simbólico de la generación de un hueco en el ADN

Cuando se presenta un hueco, nucleótidos trifosfato libres, dirigidos por interacciones moleculares, lo llenan y lo sellan con la ayuda de enzimas específicas. Aunque un hueco puede no resultar en una mutación (debido a que puede ser rellenado por el nucleótido original), en ocasiones, se introduce un nucleótido no complementario según la base de Watson-Crick, lo que podría llevar a una mutación puntual, a menos que los mecanismos de reparación actúen eficientemente.

La tasa de formación de estos huecos varía según la intensidad de los agentes mutagénicos. Curiosamente, esta velocidad de formación influencia la tasa de mutaciones, aunque la última es, por lo general, más baja debido a la acción de sistemas de reparación. Durante la replicación del ADN, la enzima ADN polimerasa, espera a que el nucleótido candidato a polimerizar se acomode en la posición adecuada, sin intervenir directamente en esta acción, para finalmente verificar la complementariedad [Lewin,1994] (Figura 2). Esta importante característica de la ADN polimerasa, permite que el proceso de replicación

sea equivalente a la ocupación de huecos consecutivos, por parte de nucleótidos guiados únicamente por interacciones con la doble hebra de ADN dentro del ambiente del sitio activo de la enzima, sin la participación de factores externos.

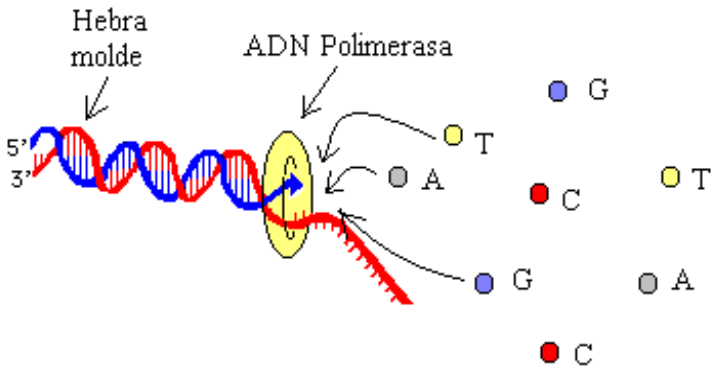


Figura 2. Incorporación de nucleótidos en la polimerización del ADN

Así, la tasa en que se generan los huecos está vinculada en gran medida a la frecuencia de replicación del ADN y, en menor medida, al ataque de agentes mutagénicos.

En la mayoría de los organismos, el ADN puede presentarse de dos formas: compacta (estado cromosómico) y relajada (conocida como "estado de ovillo") [Volkenshtein,1985]. En promedio, el ADN permanece en estado relajado el 75% del tiempo, estando activo en expresión genética [Darnell et al.,1990]. Los agentes mutagénicos atacan constantemente, sin importar el estado del ADN.

Si el ADN estuviera siempre compacto, ciertas regiones estarían más expuestas a estos agentes. En contraste, en estado relajado, toda la molécula tiene la misma probabilidad de ser atacada. Al observar a un solo individuo, esta probabilidad varía en diferentes segmentos del ADN debido a la exposición variable en su estructura relajada (Figura 3 (a) y (b)). Sin embargo, considerando a la población completa, la probabilidad promedio se distribuye uniformemente (Figura 3 (c)).

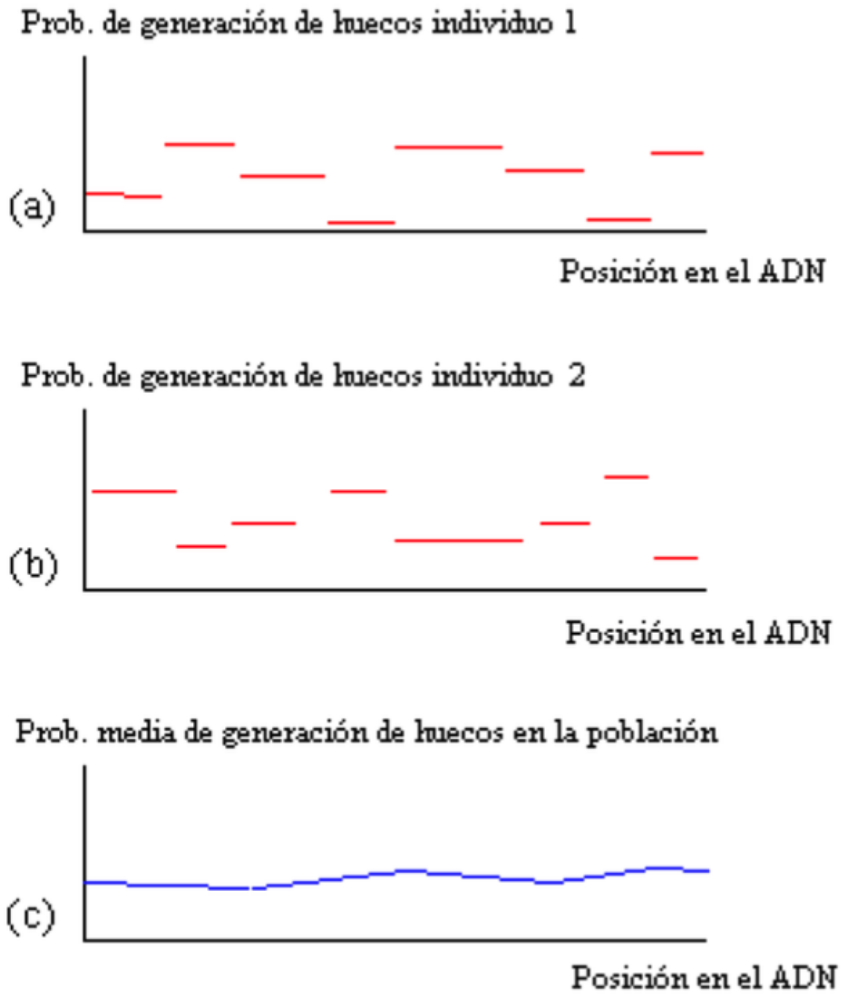


Figura 3. Distribución de probabilidad para la generación de un hueco en el ADN. Distribución para el individuo 1, individuo 2 y para toda la población. Esta última resulta ser el promedio de las distribuciones de todos los individuos, haciéndose por lo tanto uniforme.

Dado que el estado relajado prevalece más del 75% del tiempo, se puede considerar que el ADN permanece en este estado la mayoría del tiempo. Por ende, la probabilidad de que se formen huecos por agentes mutagénicos es uniformemente distribuida a lo largo del ADN. Además, la principal influencia en la tasa de mutaciones proviene del proceso de replicación, que es constante entre individuos.

Ocupación de un "Hueco" en el ADN: Presión de Selección y Presión mutacional termodinámica

Cuando se presenta un "hueco" en el ADN, los nucleótidos trifosfato libres buscan llenarlo, con el apoyo de enzimas pertinentes. Sin una influencia externa que los dirija, estos nucleótidos se guían por las interacciones con las moléculas cercanas y las colisiones generadas por la temperatura. En ausencia de influencias externas, este proceso tiende a minimizar la energía, un concepto al que nos referimos como "Presión mutacional termodinámica".

La ocupación de estos huecos, bajo la influencia de la presión mutacional termodinámica, se puede modelar considerando la probabilidad de que un determinado nucleótido ocupe ese espacio. En este contexto, nuestro sistema de interés engloba un radio de 5 pares de nucleótidos, y aplicamos la distribución de Boltzmann a un conjunto canónico.

Desde la perspectiva evolutiva, el objetivo es adaptar al organismo a su entorno. Esta dinámica adaptativa, conocida como "Presión de Selección", se vuelve esencial cuando hay cambios significativos en el ambiente. Ambas presiones, termodinámica y de selección, coexisten y juegan roles definidos en la evolución. Mientras que la primera busca optimizar la energía del ADN, la segunda actúa seleccionando las variaciones más propicias según el entorno.

Hay ocasiones en que las mutaciones cruciales para la supervivencia no son las más óptimas desde un punto de vista energético, debido a una intensa presión de selección. Nuestra investigación se enfoca en desarrollar un modelo evolutivo con mutaciones impulsadas por ambas presiones, bajo la premisa de un ambiente estable.

Veamos el proceso mediante el cual los nucleótidos libres buscan completar un hueco. En la Figura 4, ilustramos una representación sencilla, destacando tres pares de bases aledañas al hueco, mostrando su estado previo y después de ser ocupado por uno de los cuatro nucleótidos posibles. Esta representación es generalizable a cualquier número de pares.

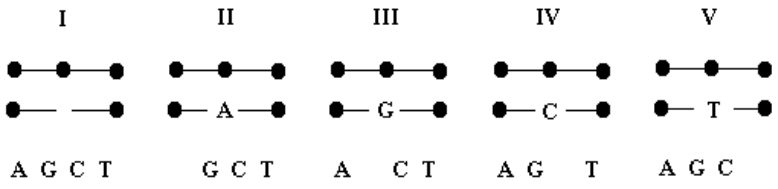


Figura 4. Configuraciones potenciales durante una mutación en el conjunto canónico.

Los diferentes escenarios, desde el I al V, corresponden a energías denotadas como E_0, E_A, E_G, E_C y E_T. Existe una relación entre el número total de configuraciones posibles y la entropía, propuesta por [Reif,1965]: $U \propto e^{S/K}$. Debido al carácter entrópico de la interacción hidrofóbica, el número de configuraciones varía según cómo cada nucleótido se acomode en el hueco. Como primera aproximación, el número de configuraciones para el escenario II es proporcional a la presencia de adeninas libres, y similarmente para los escenarios III, IV y V con relación a guaninas, citosinas y timinas respectivamente.

Con un volumen (V) constante para todos los escenarios, podemos inferir que el total de configuraciones posibles (U) se distribuye de la siguiente manera:

$$U(II) = U(E_A) = \alpha\,[dATP]\,V$$
$$U(III) = U(E_G) = \alpha\,[dGTP]\,V$$
$$U(IV) = U(E_C) = \alpha\,[dCTP]\,V$$
$$U(V) = U(E_T) = \alpha\,[dTTP]\,V$$

Donde [dNTP] representa la concentración de deoxinucleótido trifosfato, V indica el volumen total del sistema y α es una constante que encapsula los efectos hidrofóbicos. Se supone que esta constante es igual en los cuatro escenarios considerados.

Siguiendo la distribución de Boltzmann, la posibilidad de que una adenina llene el hueco corresponde a la probabilidad de que el sistema alcance una energía cercana a E_A. Análogamente, la posibilidad de que una timina ocupe el hueco es igual a la probabilidad de que el sistema adquiera una energía próxima a E_T.

Denotando el hueco como X, y la probabilidad de que este hueco sea completado por una adenina como P(X=A), entonces podemos expresar:

$$P(X=A) = P(E_A) = \alpha\,[A]\,V\,\frac{1}{Z}\,e^{-\beta E_A - a N}$$

$$P(X=T) = P(E_T) = \alpha\,[T]\,V\,\frac{1}{Z}\,e^{-\beta E_T - a N}$$

$$P(X=G) = P(E_G) = \alpha\,[G]\,V\,\frac{1}{Z}\,e^{-\beta E_G - a N}$$

$$P(X=C) = P(E_C) = \alpha\,[C]\,V\,\frac{1}{Z}\,e^{-\beta E_C - a N}$$

donde N es el número de partículas (en este caso nucleótidos trifosfato), que se intercambian en el sistema, $\beta = 1/KT$, $a = \mu / KT$, donde μ es el potencial químico, V el volumen, K la constante de Boltzmann, T la temperatura absoluta, y Z la función de partición. Se observa que la probabilidad de que un nucleótido ocupe el hueco es directamente proporcional a la concentración del mismo, por lo que la concentración de los nucleótidos trifosfato libres debe ser muy importante en un proceso evolutivo. En el presente caso, estamos asumiendo que la vecindad espacial y el intervalo de tiempo, son suficientemente pequeños de tal manera que el flujo de materia se puede despreciar.

Considerando que para una simulación de Montecarlo, lo importante son las probabilidades relativas y no las absolutas, las probabilidades arriba calculadas pueden simplificarse dividiendo en cada caso entre el menor de todos. Supongamos que para una determinada configuración, el menor valor de las probabilidades corresponda al caso en que el hueco fuera ocupado por una guanina, entonces podríamos tomar las probabilidades relativas:

P´(X=G) = P(X=G) / P(X=G) = 1

P’(X=A) = P(X=A) / P(X=G)

P’(X=T) = P(X=T) / P(X=G)

P’(X=C) = P(X=C) / P(X=G)

Es claro que ahora, tanto el volumen como el potencial químico, la constante de proporcionalidad, y la función de partición, resultan irrelevantes. Así la probabilidad relativa que el hueco sea ocupado por una adenina sería

$$P'(X=A) = \frac{[A]}{[G]}\,e^{-\beta\,(E(X=A)-E(X=G))}$$

Lo mismo ocurre para las demás posibilidades. Se observa que las probabilidades únicamente dependen de la diferencia de energía entre dos posibles configuraciones, lo cual simplifica el problema, ya que algunos tipos de energía son independientes de la configuración, y no se necesitaría calcularlos. En cualquier configuración, la energía total tiene varias contribuciones:

1- Energía de enlaces:

    1.1- Energía de enlaces covalentes

    1.2- Energía de ángulos de enlace covalente

    1.3- Energía de ángulos dihedros

2- Energía de no enlaces

    2.1- Energía electrostática

    2.2- Energía de Van der Waals

3- Energía de puentes de hidrógeno

4- Energía de interacciones hidrofóbicas

5- Energía cinética media

Para las configuraciones (II, III, IV o V), las energías identificadas por (1.1), (1.2), (1.3) permanecerán constantes, sin importar el tipo de nucleótido que llene el hueco. Esto se debe a que el enlace covalente, el enlace fosfodiester, es uniforme para los cuatro escenarios. En relación con la energía de las interacciones hidrofóbicas (ver apéndice 2), no se considerará ya que nuestra aproximación se basa en una distribución de conjunto canónico que depende solo de factores entálpicos [Davidov,1982]. Las energías mencionadas en (2.1), (2.2), (3) y (5) sí varían significativamente dependiendo del nucleótido presente en el hueco. Los cálculos para las energías del puente de hidrógeno y las interacciones multipolares se encuentran en el anexo 2. Es crucial destacar que lo relevante es la diferencia entre las energías, no su valor absoluto. Se ha demostrado que la variación de las energías totales es comparable a la diferencia de energías en la proximidad del hueco (anexo 2). Esto implica que, si designamos como E_A y E_G a las

energías totales de los sistemas II y III (es decir, las energías de la molécula completa de ADN cuando el hueco es llenado por adenina y guanina, respectivamente) y nombramos $E_{vec.\ A}$ y $E_{vec.\ G}$ a las energías de la zona circundante al hueco cuando está ocupado por adenina y guanina, podemos confirmar que:

$$(E_A - E_G) \sim (E_{vec.\ A} - E_{vec.\ G})$$

Se entiende que a medida que la vecindad aumenta en tamaño, los valores se acercan más entre sí, convergiendo cuando la vecindad engloba toda la molécula de ADN. Nuestros estudios previos han mostrado que considerar una vecindad de cinco pares de nucleótidos (dos a la derecha y dos a la izquierda del espacio vacío) es adecuado. Esto se atribuye a que las interacciones decrecen significativamente con la distancia. Así, los nucleótidos distantes del espacio vacío tienen una interacción mínima con este, aportando escasamente a la energía total. Debido a ello, para estimar la probabilidad de que un nucleótido ocupe un espacio, usaremos las probabilidades relativas y las diferencias de energía según las ecuaciones (2.1), (2.2), (3) y (5), tomando en cuenta una vecindad de cinco nucleótidos alrededor del espacio vacío.

3.2. Estudio teórico-analítico sobre los efectos a largo plazo de la estabilidad termodinámica del proceso de replicación/reparación del ADN en el contenido de Guanina-Citosina (GC).

3.2.1. Introducción

En el presente estudio buscamos desarrollar un marco analítico físico-matemático, para estudiar las probabilidades de formación de missmatches (pares no canónicos distintos de los pares Watson-Crick) durante un paso en el proceso de replicación de ADN. Se utiliza la distribución probabilística de Boltzmann para un ensamble canónico, y se plantea una cinética de primer orden para la variación del contenido de GC a lo largo de la evolución, como consecuencia de la formación de missmatches.

El punto central en el presente estudio será encontrar el porcentaje de pares canónicos GC (Guanina-Citosina) dentro del ADN en el límite de infinitas replicaciones y demostrar que existe una tendencia a alcanzar aumento de pares GC respecto de AT (Adenina-tiamina). Finalmente buscaremos, a partir de la dinámica de la variación de la concentración de pares GC en el tiempo ($[GC](t)$), buscar un lagrangiano que genere esta dinámica (principio variacional) donde la concentración será identificada con la veocidad de alguna partícula abstracta, la misma que alcanza una velocidad límite debido a la presencia de fuerzas disipativas.

Para el planteamiento del problema se utilizarán conceptos de física estadística como la distribución de Boltzmann, procesos markovianos, condición de balance detallado en el equilibrio además de los cálculos establecidos para las energías de los enlaces entre nucleótidos (enlaces de puente de hidrógeno).

3.2.2 Conceptos básicos (estructura y funcionalidad del ADN):

Los nucleótidos son moléculas que unidas adecuadamente forman la estructura de los ácidos nucleicos (ADN y ARN) y como moléculas libres cumplen además roles centrales en el metabolismo celular, sirviendo como fuentes de energía química, participando en la señalización celular y siendo incorporados dentro de importantes cofactores de reacciones enzimáticas. La habilidad de recepcionar y transmitir información genética entre generaciones es una condición fundamental para la vida y la misma es almacenada y programada dentro de una secuencia específica de nucleótidos en el ADN.

Los nucleótidos tienen 3 componentes característicos: (1)base nitrogenada, (2)pentosa y (3)fosfato. Las bases nitrogenadas se dividen según su estructura en purinas(Adenina (A) y Guanina (G)) y primidinas(Citosiana (C),Tiamina (T) y Uracilo
(U) en el ARN), entre las 2 hebras que componen la dobe hélice del ADN tenemos que los nucleótidos se encuentran enlazados por enlaces de puente de hidrógeno, el par canónico G-C mediante un triple enlace y el par A-T mediante uno doble con energías dadas por

$E(G, C) = -0.43eV$ y $E(A, T) = -0.34eV$ respectivamente.

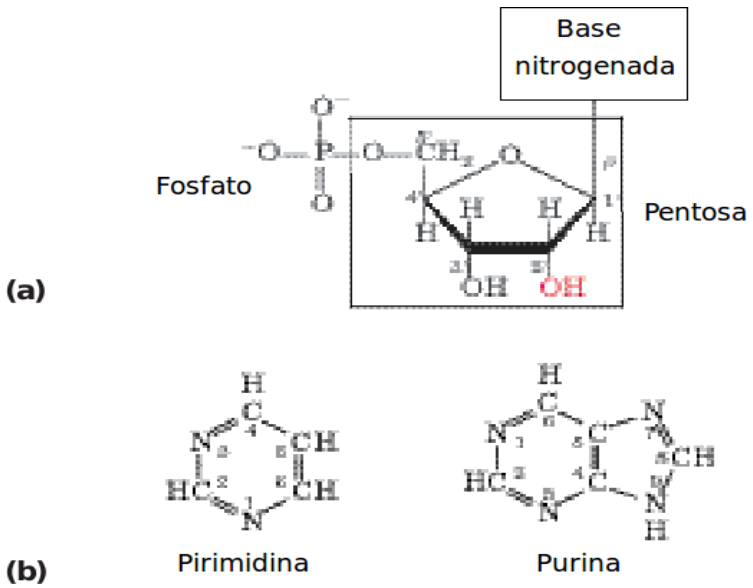


Figura 5. Estructura de los nucleótidos. (a)Estructura general mostrando la numeración convencional de carbonos en la pentosa. (b)Bases nitrogenadas: pirimidinas y purinas.

Además dentro de cada hebra del ADN (o dentro del ARN que consiste en una única hebra donde la tiamina (T) es cambiada por Uracilo (U)) tenemos que los nucleótidos se encuentran unidos por enlaces fosfodiéster, enlace covalente donde el grupo fosfato (PO_4^{-3}) del 5to carbono de la pentosa del nucleótido entrante (en el proceso de replicación) se enlaza al grupo oxhidrilo(OH) del 3er carbono de la pentosa del nucleótido que se encuentra dentro de la hebra de ADN en formación. Por lo tanto tenemos que el grupo fosfato-5' (se coloca 5' indicando que está unido al 5to carbono de la pentosa) está unido al grupo oxhidrilo-3' del siguiente nucleótido, todos estos enlaces fosfodiéster tienen la misma orientación (5' $\longrightarrow$ 3') en la hebra dándole a la misma una polaridad específica y puntos iniciales y finales distintos. Por convención, la estructura de una hebra de ácido nucleico se escribe siempre con el punto final 5' (donde tenemos un fosfato) al extremo izquierdo y el punto final 3'(donde está el oxhidrilo) al extremo derecho.

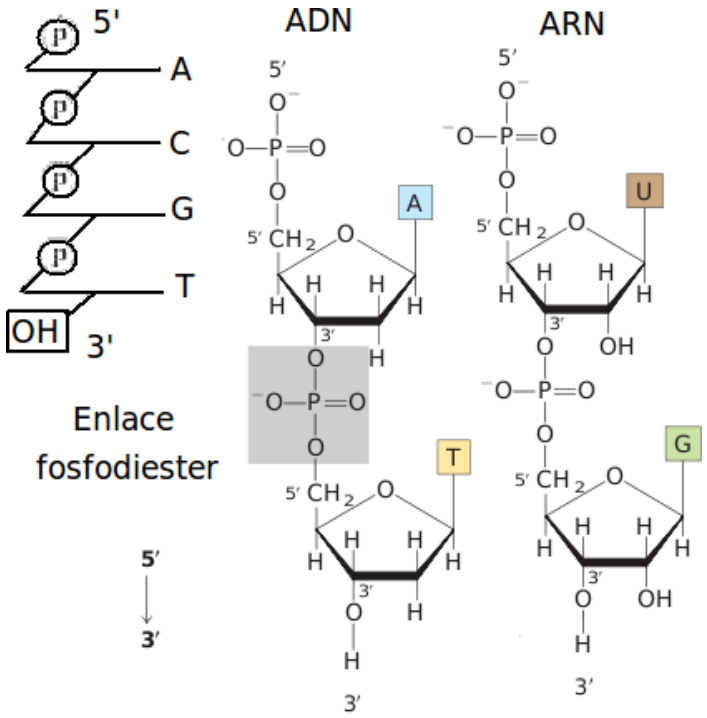


Figura 6. Dirección de la hebra y enlace fosfodíester. Se observa como el ion fosfato se une al 5to carbono de la pentosa y el ion oxhidrilo al 3ro

El descubrimiento de la estructura de doble hélice del ADN por Watson y Crick en 1953 (Watson & Crick, 1953), ha sido muy importante para la ciencia dando paso al surgimiento de nuevas disciplinas e influenciando el curso de muchas ya establecidas. Nuestro entendimiento presente de almacenamiento y uso de información genética se basa en este trabajo y el esquema de cómo la información genética es procesada por la célula para producir las proteínas requeridas para su funcionamiento es un prerequisito para la discusión de cualquier tema en el área de bioquímica.

La estructura primaria del ADN es su estructura covalente y arreglo de nucleótidos, estudios anteriores al descubrimiento de la doble hélice indicaban ya la presencia de una sustancia ácida en el núcleo celular asociada de alguna forma a la herencia (véanse los estudios de Miescher en 1868). Sin embargo, la más clara evidencia de que el ADN es el portador de información genética se obtuvo en exper- imentos en los cuales un microorganismo virulento transformaba uno no virulento en virulento pasándole al segundo su ADN mediante un proceso llamado de trans- formación (descubierto por Griffith en 1928 trabajando con cepas

de Streptococcus pneumoniae). A fines de la década de 1940 un descubrimiento clave se obtiene del trabajo de Erwin Chargaff [Chargaff & Davidson, 1955], descubriendo que los 4 nucleótidos se encuentran en diferentes proporciones y que en todo organismo celular el número de Adeninas es igual al de Tiaminas y del Citosinas igual al de Guaninas (por tanto iguales con- centraciones $[A] = [T]$, $[G] = [C]$). Experimentos posteriores usando el método de difracción de rayos X [R. Franklin y M. Wilkins] permitieron obtener un patrón de la fibra de ADN del cual se deduce su helicidad con 2 periodicidades. Con estas evi- dencias Watson y Crick construyen un modelo tridimensional asumiendo que las dos hebras del ADN son antiparalelas (una tiene dirección 5' → 3' y la otra 3' → 5') y complementarias (según la secuencia de nucleótidos), la estructura se sujeta por los puentes de hidrógeno e interacciones electrostáticas en el apilamiento de bases(como escalones de la doble hélice).

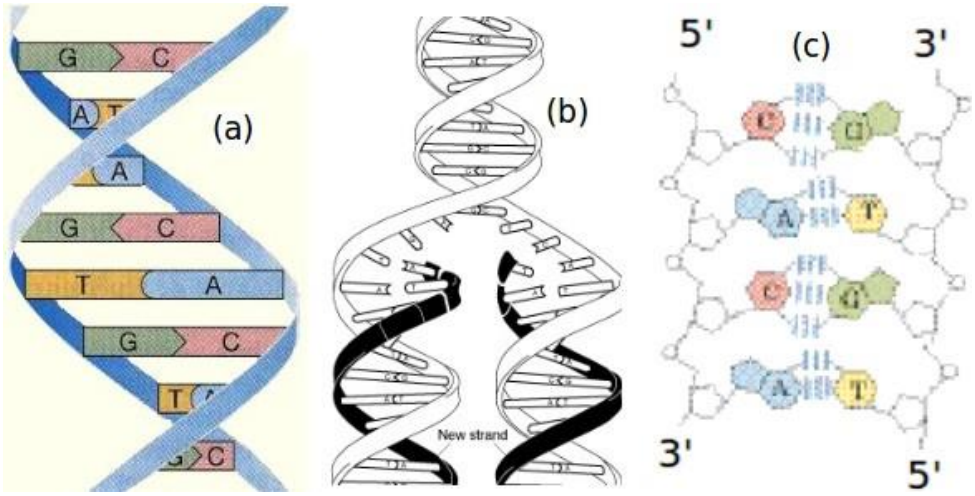


Figura 7. Molécula de ADN. (a) Estructura tridimensional, (b) Modelo elemental de replicación, (c) hebras antiparalelas mostrando 4 bp.

El modelo sugiere inmediatamente un mecanismo para la transmisión de infor- mación genética: se puede replicar la estructura separando las 2 hebras que sirven como plantillas para la síntesis de 2 moléculas de ADN, el llenado de nucleótidos es tal que se unen formando pares canónicos (G se une a C y T se une a A en ausencia de mutaciones puntuales). A continuación revisaremos brevemente el mecanismo de transcripción y replicación del ADN para plantear las hipótesis correspondientes.

La transcripción es el proceso en el que se crea una hebra de ARN complemen- taria a una secuencia dada de ADN, esta secuencia es leída por la ARN-polimerasa, la misma que produce una hebra de ARN complementaria y antiparalela la cual incluye Uracilo (U) en los sitios donde Tiamina (T) hubiera ocurrido en una hebra complementaria de ADN. El proceso consiste básicamente en el desdoblamiento de una hebra de ADN por la enzima helicasa, luego procede el llenado de nucleótidos complementarios a la hebra plantilla de ADN leyendo la misma en dirección 3' → 5' para que el llenado de la nueva hebra de ARN se de en la dirección convencional 5' → 3' y proceda de manera continua.

En organismos procariotas la transcripción ocurre en el citoplasma mientras en eucariotas debe ocurrir necesariamente en el núcleo donde se encuentra el ADN. Cuando la hebra de ARN ha sido armada entonces sigue el proceso de traducción en el cual la molécula de ARN es usada como plantilla por el ribosoma para sintetizar una proteína. El código genético es leído en unidades de 3 nucleótidos (1 codón) por unidad de tiempo que sintetizan un aminoácido, estos aminoácidos se unen en una cadena polipéptida por la formación de enlaces péptidos y forman la proteína requerida por la célula.

El proceso de replicación del ADN, por otra parte, consiste en crear 2 copias idénticas de una molécula de ADN donde cada hebra del ADN inicial sirve como plantilla para la síntesis de las nuevas hebras complementarias. La enzima helicasa desdobla el ADN inicial de modo que tenemos 2 hebras a ser llenadas, la primera es llamada línea principal (leading strand) y se orienta según 3' → 5' la cual es llenada con nucleótidos complementarios de manera continua (según la dirección convencional 5' → 3') gracias a la acción de la enzima ADN-polimerasa III. La segunda, conocida como línea atrasada (lagging strand) no puede ser llenada de manera continua debido a que se orienta según 5' → 3' y el llenado de nucleótidos debe proceder de forma discreta (para obedecer la dirección convencional), debido a ello la hebra se forma en piezas denominadas fragmentos de Okazaki de modo que el llenado pueda proceder de manera continua en cada uno de estos fragmentos. Para la formación de estos fragmentos deben intervenir 3 principales enzimas:

La enzima ARN primasa deposita secciones primarias de ARN (primer) que permiten la síntesis de los fragmentos de Okazaki en la dirección 5' → 3', la ADN polimerasa III realiza el llenado de nucleótidos en cada fragmento, luego la ADN polimerasa I reemplaza las secciones primarias de ARN por ADN y finalmente la enzima ligasa junta los fragmentos de Okazaki para completar el proceso de formación de la nueva doble hélice sobre la hebra atrasada.

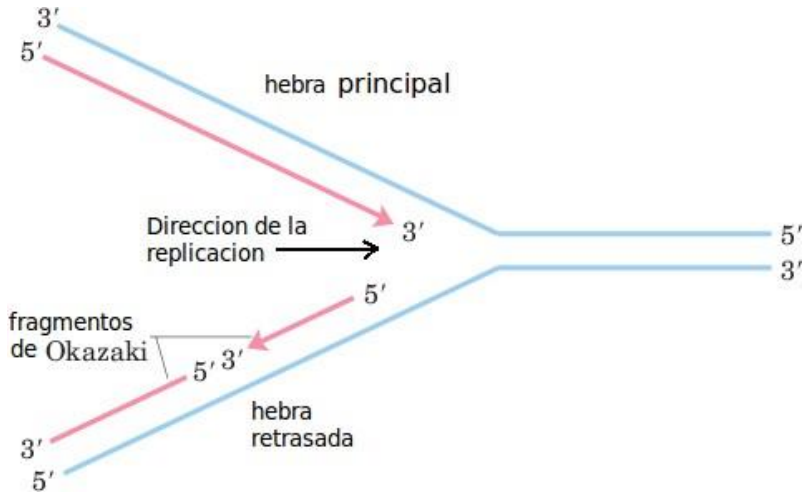


Figura 8. Diagrama del proceso de replicación del ADN.

Ahora bien, luego de terminado el proceso de replicación y considerando que los nucleótidos que se enlazan en cada doble hélice siempre lo hacen en pares canónicos (G se une a C y T a A), los cuales son más probables de ocurrir por tener las energías más negativas y por tanto ser enlaces más estables, deberíamos esperar que las 2 nuevas moléculas de ADN sean idénticas a la anterior y de ese modo habremos clonado el ADN inicial. Sin embargo, evidencias experimentales indican que la con- centración de nucleótidos en el ADN de los organismos celulares tiende a variar luego de muchas generaciones (replicaciones) lo cual permite la evolución de las especies. Debido a ello es que se considera que durante el mecanismo de replicación pueden entrar nucleótidos que no sean complementarios canónicos con probabilidad menor pero no nula, cuando esto ocurre el mecanismo de reparación intercambia el nucleótido de la hebra inicial (plantilla) por el complementario canónico del

nucleótido que ha ingresado y por tanto la nueva doble hélice resulta ser diferente que el ADN inicial (ya no sería una copia idéntica) mostrando una variación en la concentración de nucleótidos como es de esperarse.

Cuando se forma un par de nucleótidos no canónicos diremos que durante el proceso de replicación del ADN ha ocurrido una mutación puntual, esta se origina por múltiples causas como agentes mutagénicos físicos (radiación UV, temperaturas muy elevadas, rayos X),químicos (moléculas que alteran la forma de hélice del ADN) o estructurales (bases en forma tautomérica); sin embargo, el origen de estas mu- taciones se puede explicar (en ausencia de los agentes fisicoquímicos citados) por la natruraleza intrínseca del ADN y es esta la causa principal de la evolución del genoma bajo condiciones ambientales normales.

Para ello se han calculado en trabajos anteriores las energías de enlace de todos los pares de nucleótidos (incluyendo los no-canónicos) y es de esperar que se formen con mayor probabilidad los pares que tengan energías de enlace más negativas. Debido a ello, y considerando que la región de análisis es una vecindad de 5 pares de bases (donde ocurre el llenado de nucleótidos y donde las interacciones son apreciables), tendremos que esta vecindad se comporta como un sistema que intercambia energía con un foco térmico y obedece a la distribución de Bolztmann. Por lo tanto la fuerza interna que favorece la creación de más pares de GC sobre AT luego de un número considerable de replicaciones se le denomina presión termodinámica y para plantear el problema se buscan las tasas a las cuales el mecanismo de reparación intercambia el nucleótido de la hebra plantilla por el complemento canónico del que ha ingresado. Aquí pueden ocurrir 2 situaciones: transiciones, donde se reemplaza un purina por otra e igual para el caso de pirimidinas, y transversiones, donde se intercambia una purina por una pirimidina y viceversa. Las primeras son más probables y su ocurrencia es de un orden de magnitud mayor que las segundas debido a que presentan mayor estabilidad estructural.

3.2.3. Premisas consideradas

Para poder construir un modelo elemental de la evolución en el tiempo de la concentración de nucleótidos debemos asumir ciertas condiciones básicas.

*Consideraremos que una región de 5 bp (pares de bases) dentro del ADN define nuestro sistema, el mismo que puede tomar diferentes estados (configuraciones) definidos por una vecindad de referencia V_R, un nucleótido X fijo en la hebra inicial (que sirve como plantilla) al cual se le une por enlace de puente de hidrógeno otro nucleótido Y, que se encuentra libre en el medio circundante y puede tomar a priori 4 valores ($Y = A, T, C, G$) si tenemos fijados X y V_R.

*La vecindad de referencia (primer) contiene 4 nucleótidos aleatorios en una hebra (los 2 primeros fijan biunivocamente a sus respectivos complementarios canónicos en la hebra que se está formando) y por tanto tenemos $4^4 = 256$ posibles valores (número total de vecindades de referencia sin contar al nucleótido X del ADN ni al Y que ingresa desde el medio circundante).

*Tanto X como Y toman 4 valores posibles cada uno y por tanto si consideramos a la terna (V_R, X, Y) como un estado o configuración del sistema de 5bp, tenemos un total de $4^6 = 4096$ estados posibles a tomar.

*Como cada hebra de ADN contiene un número elevado de bp, podemos entonces considerar el sistema de vecindades de 5bp como un sistema canónico. Por lo tanto, se puede calcular la probabilidad de tomar un estado dado $V_R, X, Y >$ a partir de la energía de ésta configuración $E = E(V_R, X, Y)$ según la distribución de Boltzmann:

$$P\{(V_R, X, Y)\} = P\{E = E(V_R, X, Y)\} = \frac{\Omega(V_R, X, Y)}{Z} e^{\{-\beta E(V_R, X, Y)\}} \qquad (1)$$

siendo Z la función de partición (transfromada de Laplace discreta del número de estados)

$$Z = \sum_{V_R} \sum_{X} \sum_{Y} \Omega(V_R, X, Y) e^{\{-\beta E(V_R, X, Y)\}}, \qquad \beta = 1/(k_B T)$$

y Ω el número de estados o configuraciones (V_R, X, Y) que tienen energía fija $E = E(V_R, X, Y)$; es decir , Ω es el grado de degeneración de esta energía. Para poder escribir de forma explícita este factor consideramos que el mismo es proporcional al número de nucleótidos libres del tipo Y (N_Y) y un factor ($\eta(V_R, X)$) que exprese la multiplicidad de vecindades V_R y nucleótidos X presentes en el ADN (concentración de los mismos dentro del ADN y no libres fuera en la vecindad). El uso de la distribución canónica se justifica también por el hecho de que al considerar una vecindad de solo 5bp se reducen las interacciones con el nucleótido entrante a los primeros vecinos (como en el modelo de Ising) y esto es en esencia la propiedad markoviana.

Al haber cierto número de nucleótidos libres disponibles N_Y tenemos que el número de estados accesibles será proporcional a la cantidad de estos, de la misma forma al haber mayor número de nucleótidos de tipo X dentro del ADN entonces su concentración contribuye a la degeneración de la terna para cada V_R fija:

$$\Omega(V_R, X, Y) \simeq N_Y * \eta(V_R, X)$$

*Cuando ingresa un nucleótido Y que no es es complementario canónico de X ($Y \neq X^C$) entonces el mecanismo de reparación realiza el intercambio de X por el complementario de Y ($X \rightarrow Y^C$) con probabilidad $p(X \rightarrow Y^C)$ (originándose una mutación en la hebra de ADN) y por la fórmula de probabilidad total (siendo X e Y fijos):

$$p(X \Rightarrow Y^C) = \sum_{V_R} p\left(X \Rightarrow Y^C \middle| (V_R, X, Y)\right) P\{V_R, X, Y\}$$
$$= \sum_{V_R} \Re(V_R, X, Y) \frac{\Omega(V_R, X, Y)}{Z} e^{\{-\beta E(V_R, X, Y)\}} \qquad (2)$$

siendo $R(V_R, X, Y)$ la probabilidad de cambiar X por el complemento de Y dada una vecindad de referencia fija V_R.

Sin embargo, si separamos la variable *X* porque nos interesa enfocarnos en la concentración de nucleótidos dentro del ADN podemos escribir

$$P\{V_R, X, Y\} = P(V_R, Y|X = x)P(X = x) = \frac{\Omega(V_R, Y|X)}{Z_x} e^{\{-\beta E(V_R, Y|X)\}} P(X = x)$$

Donde: (3)

$Z_x = \sum_{V_R} \sum_Y \Omega(V_R, X, Y) e^{\{-\beta E(V_R, X, Y)\}}$, $\beta = \frac{1}{k_B T}$, $Z = \sum_x Z_x$

*Tenemos que $A^C = T$, $G^C = C$ y $(X^C)^C = X$, $\forall X = A, T, G, C$.

*Al calcular las tasas de transición de GC a AT y viceversa

($k_1 = p(AT \rightarrow GC)$, $k_2 = p(GC \rightarrow AT)$) solo serán consideradas las transiciones mas no las transversiones y por tanto se considerarán solo 2 sumandos en cada una.

*Se considera la regla de Chargaff para nucleótidos dentro del ADN:

$[A] = [T] = \frac{1}{2}[AT], [G] = [C] = \frac{1}{2}[GC]$

*El llenado procede en forma tal que al ingresar un nucleótido *Y* se avanza hacia la derecha (la polimerasa sintetiza en la dirección de 5´ → 3´ de modo que el ingreso del nucleótido en el siguiente paso dependerá solo del nucleótido que ingresó en el paso anterior (las interacciones tienen corto alcance). Esta propiedad markoviana permite plantear una ecuación maestra para la concentración [*X*](*t*)

*Se considera solo el llenado en la línea principal, asumiendo que el resultado será el mismo en la hebra atrasada donde se forman los fragmentos de Okazaki. Las interacciones con la ADN polimerasa serán despreciadas y la interacción de hidrógeno entre *X* e *Y* será mucho más relevante que las electrostáticas entre estos y la vecindad de referencia.

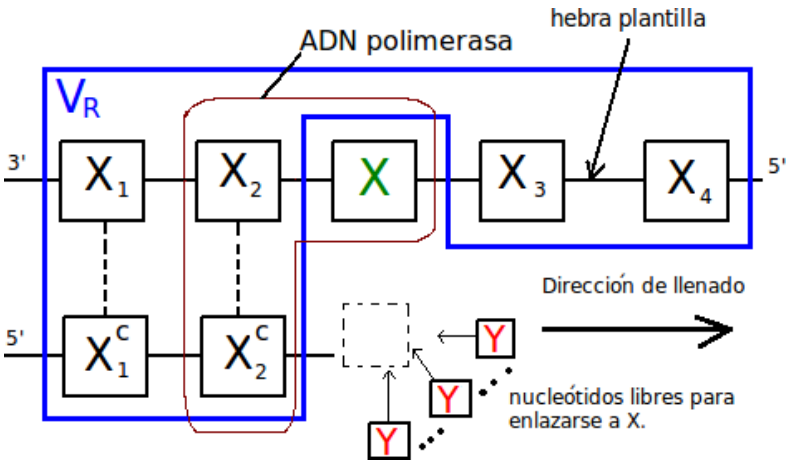


Figura 9. Modelo a utilizar de la replicación del ADN. Los nucleótidos fijos $\{x_k\}_{k=1}^4$forman la vecindad de referencia VR y los nucleótidos Y compiten por enlazarse a X fijo en la hebra plantilla.

3.2.4. Tratamiento analítico

Encontremos una expresión sencilla para $p(x \rightarrow y^C)$ (fijamos $V_R = v_R$, $X = x$, $Y = y$), para ello consideramos que dentro del sistema (v_R, x, y) la interacción de hidrógeno entre x e y es más intensa que la de los mismos con el resto de la vecindad (interacción de tipo electrostática) y el mecanismo de reparación es independiente de la vecindad de referencia. Por tanto haremos las aproximaciones:

$$\Re(v_R, x, y) \simeq \Re(x, y), \; E(v_R, x, y) \simeq E(x, y) \tag{4}$$

entonces escribimos

$$p(x \Rightarrow y^C) \simeq \frac{N_y}{Z} \Re(x, y)[\sum_{v_R} \eta(v_R, x) e^{(-\beta E(x,y))}] = \frac{N_Y}{Z} \Re(x, y)\sigma(x, y) \tag{5}$$

Siendo

$$\sigma(x, y) = e^{(-\beta E(x,y))} \sum_{v_R} \eta(v_R, x) = e^{(-\beta E(x,y))} \mu(x)$$

donde $\mu(x)$ es la distribución marginal de X (se ha sumado sobre las vecindades) y nos da la probabilidad de que $X = x$, de hecho esta es precisamente la concentración de nucleótidos X dentro del ADN. Por tanto:

$$p(x \Rightarrow y^C) \simeq \frac{N_y}{Z}\, \Re(x,y)\mu(x)e^{(-\beta E(x,y))} \qquad (6)$$

Ahora consideremos que las concentraciones de [AT] y [GC] varían de acuerdo a estas mutaciones, entonces esperamos que existan tasas de transición de un par a otro y viceversa. Es decir:

$$[AT] \xrightarrow{k_1} [GC]\ ,\quad [GC] \xrightarrow{k_2} [AT] \qquad (7)$$

siendo k_1 y k_2 las probabilidades de transición por unidad de tiempo (tasas); es decir, la probabilidad de pasar de un tipo de nucleótido a otro en un paso (una replicacion). Tenemos entonces:

$$p(AT \Rightarrow GC) = k_1\ ,\ \ p(GC \Rightarrow AT) = k_2 \qquad (8)$$

Debido a que la probabilidad de encontrar nucleótidos del tipo GC es proporcional a la concentración de los mismos en cada replicación, entonces la ecuación maestra para la probabilidad $P([GC], t) \sim [GC](t)$ también es satisfecha por $[GC](t)$:

$$\frac{d[GC](t)}{dt} = k_1[AT] - k_2[GC] \qquad (9)$$

y en el equilibrio (d[GC]/dt=0) tenemos la condición de balance detallado:

$$k_1[AT]_{eq} = k_2[GC]_{eq} \qquad (10)$$

si llamamos a $[GC] = \mu_{GC}$ y $[AT] = \mu_{AT}$, ($\mu_{GC} + \mu_{AT} = 1$):

$$\mu_{GC}|_{eq} = k/(k+1), \qquad k = \frac{k_1}{k_2}$$

Ahora, desde que las concentraciones μ_{GC} y μ_{AT} dependen del tiempo, esperamos que las tasas $\{k_i\}_{i=1,2}$

$$k_1 = p(AT \Rightarrow GC) = p(A \Rightarrow G) + p(A \Rightarrow C) + p(T \Rightarrow G) + p(T \Rightarrow C)$$

$$k_2 = p(GC \Rightarrow AT) = p(G \Rightarrow A) + p(G \Rightarrow T) + p(C \Rightarrow A) + p(C \Rightarrow T)$$

De la ecuación (5) tenemos que $p(X \rightarrow Y^C)$ es proporcional al factor de Boltzman, de modo que las energías $E(x, y)$ más negativas contribuirán más a este término. Además estamos considerando que ocurren transiciones mas no transversiones, por lo tanto nos quedamos con solo 2 sumandos:

$$k_1 = p(AT \Rightarrow GC) \simeq p(A \Rightarrow G) + p(T \Rightarrow C)$$

$$\simeq \frac{N_C}{Z}\,\Re(A,C)\mu(A)e^{(-\beta E(A,C))} + \frac{N_G}{Z}\,\Re(T,G)\mu(T)e^{(-\beta E(T,G))} \qquad (11)$$

$$k_2 = p(GC \Rightarrow AT) \simeq p(G \Rightarrow A) + p(C \Rightarrow T)$$

$$\simeq \frac{N_T}{Z}\,\Re(G,T)\mu(G)e^{(-\beta E(G,T))} + \frac{N_A}{Z}\,\Re(C,A)\mu(C)e^{(-\beta E(C,A))} \qquad (12)$$

Ahora, como el ADN consiste en dos hebras donde tenemos concentración de nucleótidos en cada una, esperamos que para $[GC] = \mu_{GC}$ tengamos en promedio igual distribuidas μ_G y μ_C en ambas hebras (regla de Chargaff) que forman la doble hélice (igual para μ_{AT}):

$$\mu_G = \mu_C = \frac{1}{2}\mu_{GC}, \qquad \mu_A = \mu_T = \frac{1}{2}\mu_{AT}$$

y por tanto podemos conocer cómo dependen de las tasas k_i de las concentraciones μ_y (y por tanto del tiempo):

$$k_1 = \left(\frac{\zeta_{AC}+\zeta_{TG}}{2}\right)(\mu_{AT}),\ k_2 = \left(\frac{\zeta_{GT}+\zeta_{CA}}{2}\right)(\mu_{GC}) \qquad (13)$$

donde

$$\zeta_{XY} = \frac{N_Y}{Z}\,\Re\,(X, Y)e^{-\beta E(X,Y)} \qquad (14)$$

Reemplazando en la ecuación maestra para μ_{GC}:

$$\frac{d\mu_{GC}}{dt} = \left[\frac{\zeta_{AC}+\zeta_{TG}}{2}\right](\mu_{AT})^2 - \left[\frac{\zeta_{GT}+\zeta_{CA}}{2}\right](\mu_{GC})^2 \qquad (15)$$

Denotando:

$$\lambda_1 = [\frac{\zeta_{AC}+\zeta_{TG}}{2}],\ \lambda_2 = [\frac{\zeta_{GT}+\zeta_{CA}}{2}] \quad (16)$$

entonces la ecuación diferencial para $\mu_{GC}(t)$ será:

$$\frac{d\mu_{GC}}{dt} = (\lambda_1 - \lambda_2){\mu_{GC}}^2 - (2\lambda_1)(\mu_{GC}) + \lambda_1 \quad (17)$$

y por tanto considerando $\mu_{GC}(0) = [GC]_0$ en $t = 0$ integramos la expresión anterior

$$\frac{1}{2\sqrt{\lambda_1\lambda_2}}\ln[\frac{2(\lambda_1-\lambda_2)\mu_{GC} - 2\lambda_1 - 2\sqrt{\lambda_1\lambda_2}}{2(\lambda_1-\lambda_2)\mu_{GC} - 2\lambda_1 + 2\sqrt{\lambda_1\lambda_2}}]\,|_{[GC]_0}^{[GC]_{(t)}} = t$$

despejando $\mu_{GC}(t)$:

$$\mu_{GC}(t) = \frac{\Theta(t)\left(-2\lambda_1+2\sqrt{\lambda_1\lambda_2}\right)+\left(2\lambda_1+2\sqrt{\lambda_1\lambda_2}\right)}{2(\lambda_1-\lambda_2)(1-\Theta(t))} \quad (18)$$

donde:

$$\Theta(t) = \exp\{\left(2\sqrt{\lambda_1\lambda_2}\right)t + \ln(C_0)\} \quad (19)$$

$$C_0 = \frac{2(\lambda_1-\lambda_2)[GC]_0-2\lambda_1-2\sqrt{\lambda_1\lambda_2}}{2(\lambda_1-\lambda_2)[GC]_0-2\lambda_1+2\sqrt{\lambda_1\lambda_2}} \quad (20)$$

Restaría hallar los valores adecuados a las energías $E(x,y)$, nucleótidos libres N_y y el valor experimental de $R_{(x,y)}$ (para todos los x e y) y poder calcular ζ_{xy} y por tanto $\{\lambda_i\}_{i=1,2}$.

Problema simplificado:

Si consideramos que $N_y = cte.\ \forall\ y$, $\Re(x, y) = cte.\ \forall\ x, y$, y que $E(x, y) = E(y, x)$ entonces tendremos que $\lambda_1 = \lambda_2 = \lambda$

$$\mu_{GC}(t) = \frac{1}{2}(1 - e^{-2\lambda t}) \quad (21)$$

lo cual en un tiempo sufientemente grande no nos da el valor esperado de μ_{GC} por encima de 0.5. Debido a esto es que debemos buscar otro método más adecuado de planteamiento. Observamos que la simplificación conduce a una indeterminación del resultado (18) que solo es válido si $\lambda_1 = \lambda_2$, requiriendo por tanto los valores exactos de los parámetros.

Para esto buscaremos partir de la condición de balance detallado, la cual predice un valor para μ_{GC} en el equilibrio:

$$((\mu_{GC})_{eq})/((\mu_{AT})_{eq}) = k_1/k_2 = \frac{(\zeta_{AC} + \zeta_{TG})(\mu_{AT})_{eq}}{(\zeta_{GT} + \zeta_{CA})(\mu_{GC})_{eq}}$$

y por tanto

$$[\frac{\mu_{GC}}{\mu_{AT}}]_{eq} = \sqrt{\frac{\zeta_{AC}+\zeta_{TG}}{\zeta_{GT}+\zeta_{CA}}} \simeq 1 \quad (22)$$

donde se ha usado la aproximación ulterior $\zeta_{XY} = \zeta_{YX}$ y vemos que en el equilibrio tanto [*GC*] como [*AT*] tienden a tener el mismo valor 1/2, lo cual no predice cambios evolutivos. Debido a esto buscamos otro planteamiento que nos pueda brindar un resultado diferente.

Variante al tratamiento:

Para ello separamos el evento conjunto (v_R, *x*, *y*) del evento particular $X = x$ para poder tener a éste último como factor multiplicativo adicional, de esta forma consideramos que la ocupación de nucleótidos de tipo *X* en el ADN es un evento importante porque determina justamente la concentración de nucleótidos ligados (internos) que nos interesa calcular en el equilibrio, la modificación viene dada al utilizar la ecuación (3) en la expresión para

$p(x \Rightarrow y^c)$

$$P(Y = y, V_R, X = x) = P(Y = y, V_R | X = x)P(X = x) \quad (23)$$

$$P(Y = y, V_R = v_R | X = x) = \frac{N_y \eta(v_R, x)}{Z_x} e^{(-\beta E(v_R, x, y))} \mu_x \quad (24)$$

$$P(X = x) = \mu_x = \sum_{v_R} \eta(v_R, x) \quad (25)$$

$$Z_x = \sum_{v_R} \sum_y \Omega(v_R, x, y) e^{-\beta E(v_R, x, y)}, \; Z = \sum_x Z_x \quad (26)$$

con las simplificaciones anteriores respecto a la independencia de $E(x, y)$ y $\Re(x, y)$ respecto de la vecindad de referencia $v_R(x, y)$

$$p(x \Rightarrow c(y)) = \frac{N_y}{Z_x} e^{-(\beta E(x,y))} \Re(x, y) (\mu_x)^2 \quad (27)$$

Expresando las tasas de transición en función del cuadrado de las concentraciones

$$k_1 \simeq [\frac{N_C}{4Z_A} \Re(A, C) e^{-\beta E(A,C)} + \frac{N_G}{4Z_T} \Re(T, G) e^{-\beta E(T,G)}] (\mu_{AT})^2 \quad (28)$$

$$k_2 \simeq [\frac{N_T}{4Z_G} \Re(G, T) e^{-\beta E(G,T)} + \frac{N_A}{4Z_C} \Re(C, A) e^{-\beta E(C,A)}] (\mu_{GC})^2 \quad (29)$$

Antes de continuar buscamos hallar una aproximación para Z_x ($\forall x = A, T, G, C$):

$$Z_x = \sum_{v_R} \eta(v_R, x) [\sum_y N_y e^{-\beta E(v_R, x, y)}] \simeq [\sum_y N_y e^{-\beta E(x,y)}] \mu_x$$

Para poder hallar los factores de Boltzmann y por tanto las correspondientes Z_x nos valemos de la siguiente tabla de energías, previamente calculadas (Zimic M. et.al., 2003), considerando la energía entre los enlaces de hidrógeno como interacción entre dipolos (en primer orden) más una constante aditiva λ:

Tabla 1

Energias de enlace estimadas para pares de bases canonicos y no-canonicos

Base pair	eV	kcal/mole
A–T	−0·34	−7·84
G–C	−0·43	−9·91
C–A	0·21	4·84
T–G	0·13 (I) −0·40 (II)	2·99 (I) −9·22 (II)
T–C	−0·14	−3·22
A–G	−0·27	−6·22*
T–T	0·02 (I) −0·16 (II)	0·46 (I) −3·68 (II)
C–C	−0·01 (I) −0·14 (II)	−0·23 (I) −3·22 (II)
A–A	0·23	5·30*
G–G	0·54	12·45*

Energías calculadas considerando arreglo de dipolos con una incertidumbre en la energía de λ = 0.12eV ; el asterisco (*) indica el acomodo de los pares de base (libres) purina-purina. (I) y (II) corresponden a 2 arreglos geométricos distintos de los pares de base (bp).

Podemos entonces encontrar Z_x, $\forall x$:

$$Z_A \simeq [e^{-\beta E(A,T)}N_T + e^{-\beta E(A,C)}N_C + e^{-\beta E(A,G)}N_G + e^{-\beta E(A,A)}N_A]\mu_A \quad (30)$$

$$Z_T \simeq [e^{-\beta E(T,A)}N_A + e^{-\beta E(T,G)}N_G + e^{-\beta E(T,C)}N_C + e^{-\beta E(T,T)}N_T]\mu_T \quad (31)$$

$$Z_G \simeq [e^{-\beta E(G,C)}N_C + e^{-\beta E(G,T)}N_T + e^{-\beta E(G,A)}N_A + e^{-\beta E(G,G)}N_G]\mu_G \quad (32)$$

$$Z_C \simeq [e^{-\beta E(C,G)}N_G + e^{-\beta E(C,A)}N_A + e^{-\beta E(C,T)}N_T + e^{-\beta E(C,C)}N_C]\mu_C \quad (33)$$

y de la tabla anterior podemos hallar todos los factores de Boltzmann considerando valores específicos para las energías:

$$e^{-\beta E(A,T)} = 5.06 * (10)^5, \qquad e^{-\beta E(G,C)} = 1.637 * (10)^7,$$
$$e^{-\beta E(A,C)} = 2.998 * (10)^{-4},$$
$$e^{-\beta E(T,G)} = 5.137 * (10)^6,$$

$$e^{-\beta E(T,C)} = 223.2, \qquad e^{-\beta E(A,G)} = 3.3864 * (10)^4, \qquad e^{-\beta E(T,T)} = 0.462,$$
$$e^{-\beta E(C,C)} = 1.472,$$
$$e^{-\beta E(A,A)} = 1.385 * (10)^{-4}, \qquad e^{-\beta E(G,G)} = 8.72 * (10)^{-10}.$$

además tenemos que $N_y = cte$ ($\forall y$), luego:

$$Z_A \simeq (5.398 * (10)^5) * N_y\mu_A \quad (34)$$

$$Z_T \simeq (5.643 * (10)^6) * N_y \mu_T \quad (35)$$

$$Z_G \simeq (2.154 * (10)^7) * N_y \mu_G \quad (36)$$

$$Z_C \simeq (1.637 * (10)^7) * N_y \mu_C \quad (37)$$

reemplazando en las ecuaciones (28) y (29) y considerando que $\Re$ (*x, y*) = *cte* , ($\forall$ x, y) encontramos para las tasas k_1 y k_2:

$$k_1 = (0.91) * \Re * \frac{\mu_{AT}}{2}, \; k_2 = (0.238) * \Re * \frac{\mu_{GC}}{2} \quad (38)$$

luego la ecuación diferencial para $\mu_{GC}(t)$ (ecuación maestra) será:

$$\frac{d\mu_{GC}}{dt} = (0.91) * \frac{\Re}{2} * (1 - \mu_{GC})^2 - (0.2385) * \frac{\Re}{2} * (\mu_{GC})^2$$

$$= \Re[(0.3359) * {\mu_{GC}}^2 - (0.91) * \mu_{GC} + 0.455] \quad (39)$$

tenemos entonces que integrar una ecuación del tipo

$$\int \frac{d\mu}{(a\mu^2 + b\mu + c)} = \frac{1}{\sqrt{\Delta}} \ln[\frac{2a\mu + b - \sqrt{\Delta}}{2a\mu + b + \sqrt{\Delta}}] = t + cte, \qquad \Delta > 0.$$

esto se cumple puesto que

$$a = 0.336 * \Re, \qquad b = -0.91 * \Re,$$

$$c = 0.455 * \Re, \Delta = \sqrt{b^2 - 4ac} = 0.217 * \Re > 0$$

despejando $\mu(t)$

$$\mu(t) = \frac{C_0 e^{(\sqrt{\Delta})t}\left(\sqrt{\Delta} + b\right) + (\sqrt{\Delta} - b)}{2a(1 - C_0 e^{(\sqrt{\Delta})t})} = \frac{C_0\left(\sqrt{\Delta} + b\right) + (\sqrt{\Delta} - b)e^{-(\sqrt{\Delta})t}}{2a(e^{-(\sqrt{\Delta})t} - C_0)}$$

siendo C_0 una cosntante dependiente de las condiciones iniciales. Tenemos entonces la solución de la ecuación (39)

$$\mu_{GC}(t) = \frac{-0.445C_0 + 1.375e^{-(0.465)t}}{0.672(e^{-(0.465)t} - C_0)} \tag{40}$$

donde vemos además que

$$\lim_{t\to\infty}[\mu_{GC}(t)] = 0.66096 \tag{41}$$

graficaremos $\mu(t)$ para 3 condiciones inciales:

$$\mu_{GC}(0) = 0(C_0 = 3.084),\ \ \mu_{GC}(0) = 0.5(C_0 = 9.449),\ \ \mu_{GC}(0) = 1(C_0 = -3.112) \tag{42}$$

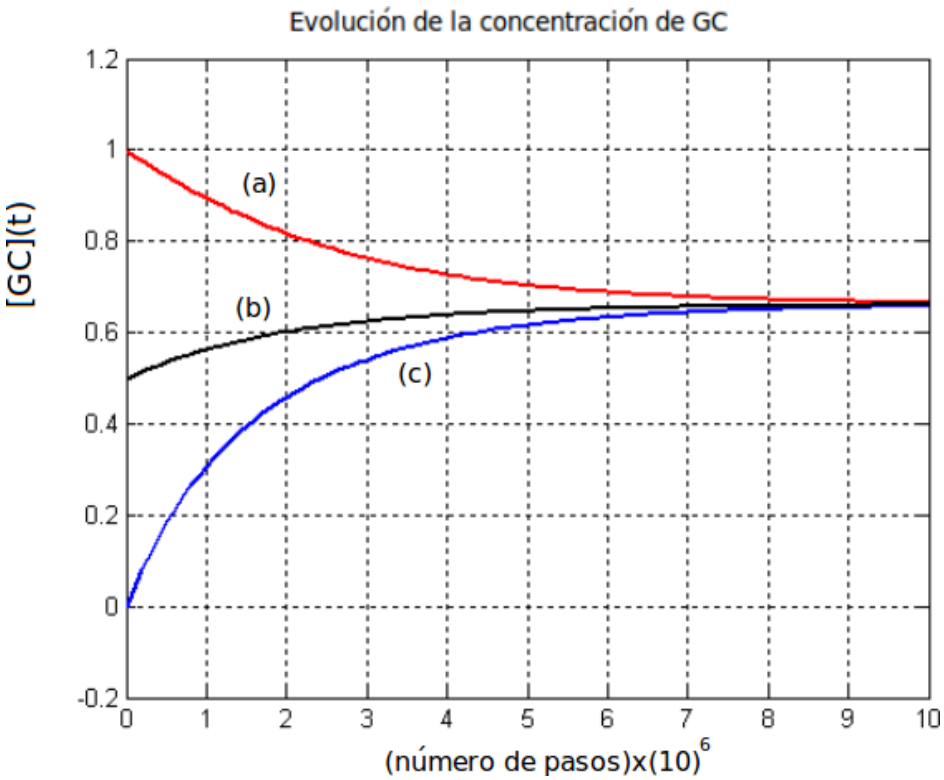


Figura 10. Variación en el tiempo de [GC](t) para 3 condiciones iniciales. (a)$[GC]_0$ = 1, (b)$[GC]_0$ = 0.5, (c) $[GC]_0$ = 0. El tiempo se da en nu´mero de pasos (donde en cada paso ingresa un nucle´otido del medio circundante) para estar en concordancia con los resultados del algoritmo computacional propuesto en [1] (donde se usan muchos pasos para tener un proceso continuo).

En la Figura 10 se observa que para cualquier concentración inicial de nucleótidos que se tome la misma tiende a estabilizarse cerca de un mismo valor $\mu_{GC\ eq}$ = 0.66096, esta tendencia a un valor límite en $[GC]$ también se ha podido comprobar com- putacionalmente y experimentalmente (estudios en genes específicos de Trypanosomatidae, como mostraremos

más adelante. Este resultado también puede deducirse a partir de la condición de balance detallado.

Consideremos un caso ligeramente más simplificado en el cual para cada función de partición Z_x (ecuaciones (30)-(33)) tomamos solo los factores de Boltzmann correspondientes a transversiones, de modo que para cada Z_x tenemos solo 2 términos.

$$Z_A \simeq [e^{-\beta E(A,T)} N_T + e^{-\beta E(A,C)} N_C]\mu_A \tag{43}$$

$$Z_T \simeq [e^{-\beta E(T,A)} N_A + e^{-\beta E(T,G)} N_G]\mu_T \tag{44}$$

$$Z_G \simeq [e^{-\beta E(G,C)} N_C + e^{-\beta E(G,T)} N_T]\mu_G \tag{45}$$

$$Z_C \simeq [e^{-\beta E(C,G)} N_G + e^{-\beta E(C,A)} N_A]\mu_C \tag{46}$$

de aquí se ve que al sustituir estas aproximaciones en las tasas k_i (considerando sólo transversiones como en el caso anterior) las mismas vuelven a depender linealmente de las concentraciones pero con una pequeña variación en la dependencia, (usamos las aproximaciones $N_y = cte$, $\Re\,(x, y) = cte.$, $Z_x = cte.$ $\forall x, y.$).

$$k_1 \simeq (\frac{\Re}{2})\mu_{AT}[\frac{e^{-\beta E(A,C)}}{e^{-\beta E(A,C)}+e^{-\beta E(A,T)}} + \frac{e^{-\beta E(T,G)}}{e^{-\beta E(T,G)}+e^{-\beta E(T,A)}}] \tag{47}$$

$$k_2 \simeq (\frac{\Re}{2})\mu_{GC}[\frac{e^{-\beta E(G,T)}}{e^{-\beta E(G,T)}+e^{-\beta E(G,C)}} + \frac{e^{-\beta E(C,A)}}{e^{-\beta E(C,A)}+e^{-\beta E(C,G)}}] \tag{48}$$

y de la condición de balance detallado:

$$[\frac{\mu_{GC}}{\mu_{AT}}]_{eq} \simeq \sqrt{\frac{[\frac{e^{-\beta E(A,C)}}{e^{-\beta E(A,C)}+e^{-\beta E(A,T)}}+\frac{e^{-\beta E(T,G)}}{e^{-\beta E(T,G)}+e^{-\beta E(T,A)}}]}{[\frac{e^{-\beta E(G,T)}}{e^{-\beta E(G,T)}+e^{-\beta E(G,C)}}+\frac{e^{-\beta E(C,A)}}{e^{-\beta E(C,A)}+e^{-\beta E(C,G)}}]}} \tag{49}$$

Considerando los parámetros $T = 300^o K$ tenemos β=38.63eV^{-1}, y como son conocidos los valores de las energías

$$E(A,C) = 0.21eV, E(A,T) = -0.34eV, E(T,G) = -0.4eV, E(G,C) = -0.43eV$$

$$[\frac{\mu_{GC}}{\mu_{AT}}]_{eq} \simeq 1.9522 \Rightarrow [GC]_{eq} \simeq 0.66127 > 0.5 \quad (50)$$

prediciendo una tendencia hacia la formación de más pares GC respecto de AT muy próxima a la anterior (0.66).

Si simplificamos más la expresión (49) considerando los factores de energía de AT y GC como los que contribuyen más, entonces

$$[\frac{\mu_{GC}}{\mu_{AT}}]_{eq} \simeq \sqrt{\frac{e^{-\beta E(A,T)}(e^{-\beta E(A,C)} + e^{-\beta E(T,G)})}{e^{-\beta E(G,C)}(e^{-\beta E(G,T)} + e^{-\beta E(C,A)})}} = \exp\{\frac{\beta}{2}(E(A,T) - E(G,C))\} \simeq 5.69 \quad (51)$$

y por tanto un valor más alto de $[GC]_{eq}$.

$$[GC]_{eq} \simeq 0.85 \quad (52)$$

De esta forma se ve que la variante al tratamiento resulta fundamental para obtener una asimetría respecto de las concentraciones, lo cual indica que se debe separa el sistema de los nucleótido fijos X que se encuentran dentro del ADN y por tanto tienden a cierto valor estable en el equilibrio (separándolos en GC y AT). Son estos los que se espera que puedan ir cambiando y de allí la importancia de tratar su distribución por separado según se ve en (23). Hay que tener en cuenta que en el planteamiento analítico no se ha considerado presión de selección y por tanto uno podría pensar que el resultado (52) es el correcto y debemos considerar la ligadura de presión de selección de alguna forma para bajar el valor límite de GC a 0.66.

CAPÍTULO IV. Análisis y discusión de los resultados

4.1. Resultados

El proceso de replicación y reparación del ADN, es el factor de tipo biológico más imporante en la generación de mutaciones y por consecuencia en la evolución del ADN.

La distribución de Boltzmann para un conjunto canónico, bajo condiciones en las cuales el entorno del sistema de replicación del ADN se evalúa dentro de una vecindad espacial y temporal (es decir dentro de un instante de tiempo breve y en un volumen del espacio limitado), puede utilizarse para calcular la probabilidad de que un nucleótido libre ingrese al sitio de replicación previo a la formación del enlace fosfodiester y como consecuencia, la polimerización de la hebra de ADN naciente.

La probabilidad de que alguno de los cuatro nucleótidos libres pueda ingresar al sitio de replicación, dependerá de diversos factores, dentro de ellos, los más relevantes corresponden a la energía potencial del sistema molecular, la abundancia de los nucléotidos libres, la secuencia de ADN del entorno al sitio de replicación, y la temperatura. Dentro de los términos más relevantes de la energía potencial, destacan la energía de puentes de hidrógeno, y las energías electrostáticas dadas por las interacciones intramoleculares a nivel de cargas y dipolos eléctricos.

Considerando estas condiciones, es posible calcular las probabilidades de que en un instante dado durante la replicación del ADN, un nucleótido libre cualquiera pueda ingresar al sitio de replicación de la enzima ADN polimerasa. El conocimiento de estas probabilidades permite estudiar ciertos aspectos del proceso evolutivo, relacionados a la acumulación de nucleótidos en el ADN como resultado de un número grande de procesos de replicación, que se dan durante un proceso evolutivo.

Existen al menos dos maneras de realizar estos estudios. (1) A partir de una simulación computacional, y (2) a partir de un estudio analítico. En ambos casos podemos predecir las

características del contenido de nucleótidos en el ADN, en los límites de muy largos tiempos de procesos evolutivos.

En un estudio previo, desarrollamos y estudiamos una simulación computacional de mutaciones puntuales, utilizando los cálculos de las probabilidades de ocupación de nucléotidos en el sitio de replicación del ADN, y la distribución de Boltzmann para un conjunto canónico en una simulación de Monte Carlo. Para calcular las probabilidades dadas por la distribución de Boltzmann, en aquel trabajo realizamos un cálculo para estimar las energías de puentes de hidrógeno, así como las energías electrostáticas, asumiendo la naturaleza dipolar de los enlaces covalentes, permitiéndonos estimar la energía potentical del sistema molecular descrito. La simulación predijo teóricamente un aumento en el contenido de guanina-citosina (GC) en la molécula de ADN, con la tendencia de alcanzar un equilibrio en el tiempo. En dicha simulación, se logró comprender el efecto en el proceso evolutivo, de algunos parámetros como la temperatura, la concentración de nucleótidos libres, el contenido inicial de GC en la secuencia, y la cinética que puede tener el mecanismo de repación del ADN. Las predicciones computacionales realizadas en aquella oportunidad, fueron contrastadas con evidencia experimental de secuencias genómicas de diversas especies cuya historia evolutiva es conocida. De manera específica se estudio el linaje de *Kinetoplastidia* y *Plasmodium* para los cuales se determinó sesgo del uso de codones (codon bias) de manera experimental.

La evidencia experimental no solo confirma la predicción teórica de buscar un aumento del contenido de GC a lo largo del tiempo de evolución, sino que explica de manera natural el fenómeno del "codon bias", el cual surge como una simple consecuencia de las preferencias naturales de que ciertas mutaciones se vean favorecidas por medio de las probabilidades calculadas a partir de la distribución de Boltzmann.

El proceso de simulación descrito nos llevó a proponer la existencia de una Presión mutacional termodinámica, que actúa como un ‘driving force’ de la evolución, la cual junto con la Presión de Selección, terminan definiendo las características de los genomas de las especies a lo largo de la evolución.

En el presente trabajo mostramos una evaluación analítica, basada en las probabilidades estimadas por la distribución de Boltzmann para un sistema canónico, las energías potenciales electrostáticas y de puentes de hidrógeno del sistema molecular. Hemos introducido una nomenclatura adecuada para describir los distintos aspectos moleculares del sistema, así como los procesos más importantes durante la replicación del ADN. Empleamos un análisis teórico matemático-estadístico, para comprender la solución estiacionaría en el límite de un tiempo infinito, para la ecuación diferencial lineal de primer orden en el tiempo, la cual se asume que explica en primera aproximación, la cinética de la variación del contenido de GC a lo largo del tiempo, en un genoma que viene experimentando múltiples procesos de replicación durante un proceso evolutivo.

Considerando una serie de premisas y estimaciones de las energías de puentes de hidrógeno y energías potenciales electrostáticas, así como abundancias relativas de nucleótidos en el entorno celular, logramos estimar las concentraciones de GC en los límites cuando el tiempo tiende al infinito.

Como resultado de este trabajo, encontramos que las estimaciones teóricas se acercan mucho a los valores predichos en las simulaciones computacionales en nuestro estudio previo, lo cual constituye una evidencia importante que apoya la postulación de la existencia de una "presión mutacional termodinámica" que estaría guiando el proceso evolutivo, y sería capaz de explicar el fenómico del "codon bias", el cual es una de las incógnitas más importantes en la biología que no se ha logrado explicar de una manera completa.

4.1.1 Predicciones del modelo analítico

El estudio analítico planteado, muestra la posibilidad de estimar las probabilidades de formación de un par no-canónico y subsecuentemente, la generación de una mutación puntual durante el proceso de replicación del ADN.

Las probabilidades de formación de pares no-canónicos, resulta depender de varios factores. Entre los más importantes destacan la energía de interacción fronta tipo puente

de hidrógeno, entre la base entrante y el nucleótido complementario, así como también la energía potencial electrostática entre el nucleótido entrante y los demás nucleótidos más cercanos. Como se demostró en nuestros estudios previos, esta energía potencial se puede estimar con una suficentemente alta precisión, al incluir una vecindad de 5 pares de bases (2 nucleótidos a la izquierda y 2 nucleótidos a la derecaha del nucleótido entrante). Finalmente, la temperatura y la concentración de los nucleótidos libres determinan también las probabilidades de que un hueco sea ocupado por uno de los cuatro nucelótidos posibles.

Considerando que la variación de la concentración de GC, varía a través de una cinética de primer orden en el tiempo, el modelamiento a través de una ecuación diferencial, permite estimar la solución del contenido de GC esperado en el límite de un tiempo infinito. Así mismo, la solución analítica muestra una variación exponencial con un coportamiento asintótico en el tiempo.

Considerando ciertos escenarios y premisas, y haciendo uso de las energías de los pares de nucleótidos (entrante y complementario), calculados en nuestro estudio previo, se llega a estimar el contenido de GC en el límite cuando el tiempo tienen a infinto. El valor del contenido de CG esperado en el límite cuando el tiempo tiende al infinito, resulta ser igual al contenido de GC en el equilibrio durante un proceso evolutivo de largo tiempo. El modelo analítico propuesto en este estudio, predice un valor del contenido de GC en el equilibrio de 66-85%.

4.1.2 Evidencia del incremento de GC en un proceso de evolución in-vitro

La filogenética experimental permite comprender los procesos fundamentales de cambio de nucleótidos en especies evolutivamente emparentadas. Se han desarrollado diversos algoritmos y criterios para inferir filogenias de genes y especies a partir de datos de secuencias, siendo el método de máxima verosimilitud, que utiliza modelos explícitos de evolución de secuencias, uno de los más populares y en crecimiento. Sin embargo, muchos modelos asumen matrices de probabilidades de cambios reversibles, donde la probabilidad de un cambio es igual a la de su evento inverso. Un estudio previo presentó un método

basado en PCR serial que genera un conjunto de datos de secuencias en evolución neutral, apoyando el uso de la máxima verosimilitud para determinar la topología correcta, la reconstrucción de secuencias ancestrales y estimaciones de fechas de divergencia.

De esta manera, buscando desarrollar un método de reconstrucción filogenética, basado en un método de máxima verosimilitud y un modelo estadístico reversible de 12 parámetros (esto significa, un modelo en que cada mutación X->Y, para todo X,Y={A,T,G,C} tiene una probabilidad única de ocurrencia), un estudio previo, generó información evolutiva a partir de datos experimentales correspondientes a un proceso evolutivo in-vitro [Sanson G. et.al., 2002].

En dicho estudio se generó una filogenia conocida mediante un método de PCR bifurcado en cuatro etapas. La secuencia ancestro (SSU rDNA) evolucionó in vitro durante 280 ciclos de PCR anidado, resultando en 15 secuencias ancestro y 16 terminales determinadas.

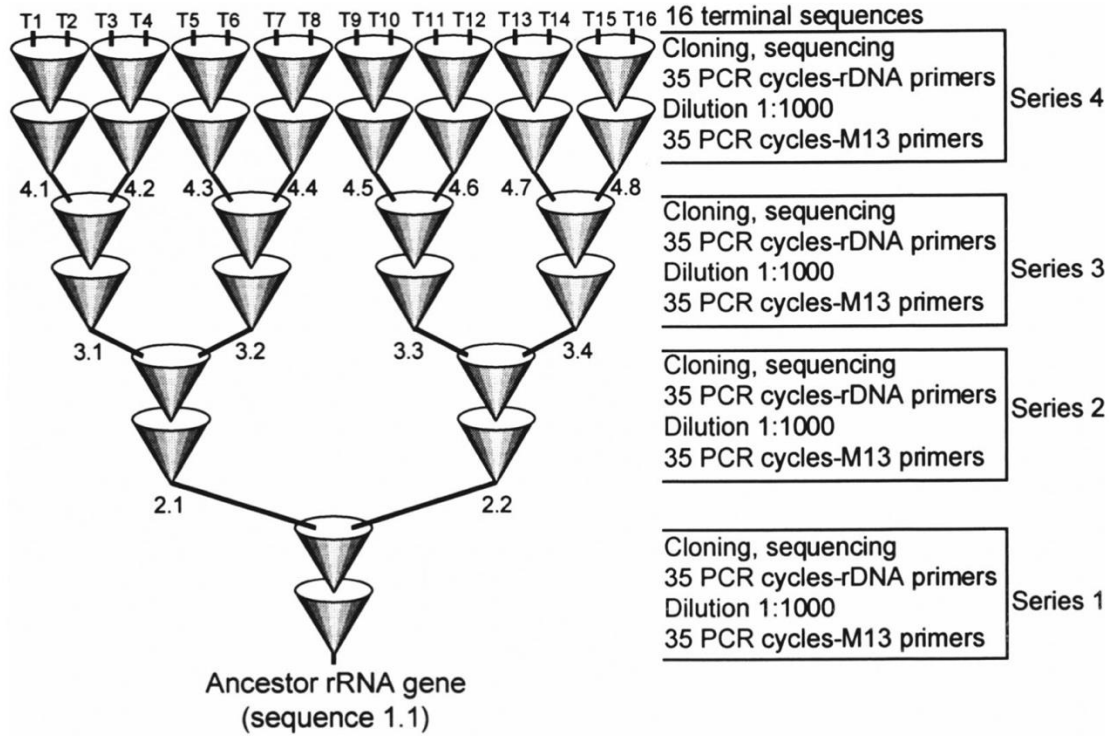


Figura 11. Evolución de secuencias de ADN mediante una serie de PCRs bifurcadas. Un ancestro SSU rDNA clonado en pBluescript se utilizó como plantilla para la serie 1 de 70 ciclos anidados de PCR con cebadores M13. Después de los primeros 35 ciclos, los productos de la reacción se diluyeron 1:1,000 y se utilizaron como plantillas para los 35 ciclos subsiguientes, con los cebadores rDNA RIBA y RIBB. Tras 70 ciclos, los amplicones se clonaron y se seleccionaron al azar dos clones que se utilizaron como plantillas para la siguiente serie de ciclos de PCR anidados. Las líneas de descendencia se propagan al azar, por lo que la evolución es neutral y se comporta como un proceso estocástico. Los nodos del árbol T1 a T16 indican secuencias terminales, y 1.1 a 4.8, ancestros internos.

El análisis de las secuencias finales permitió reconstruir con precisión la topología y las longitudes de las ramas de la filogenia real. Se estimaron fechas de divergencia y secuencias ancestrales con mínimo error, principalmente debido a inserciones y eliminaciones. Los patrones de sustitución no son descritos por modelos reversibles, por lo que se calculó una matriz de sustitución basada en las observaciones. A diferencia de estudios anteriores, aquí las mutaciones ocurrieron de forma neutral sin la intervención de un agente mutagénico. Estos hallazgos brindan apoyo experimental a filogenias y estimaciones de fechas de divergencia, basados en secuencias de ADN que evolucionan neutralmente. Las preferencias de sustitución observadas son coherentes con el alto contenido de G+C del genoma de Thermus aquaticus, sugiriendo que el método empleado simula la evolución del ADN en un organismo termofílico.

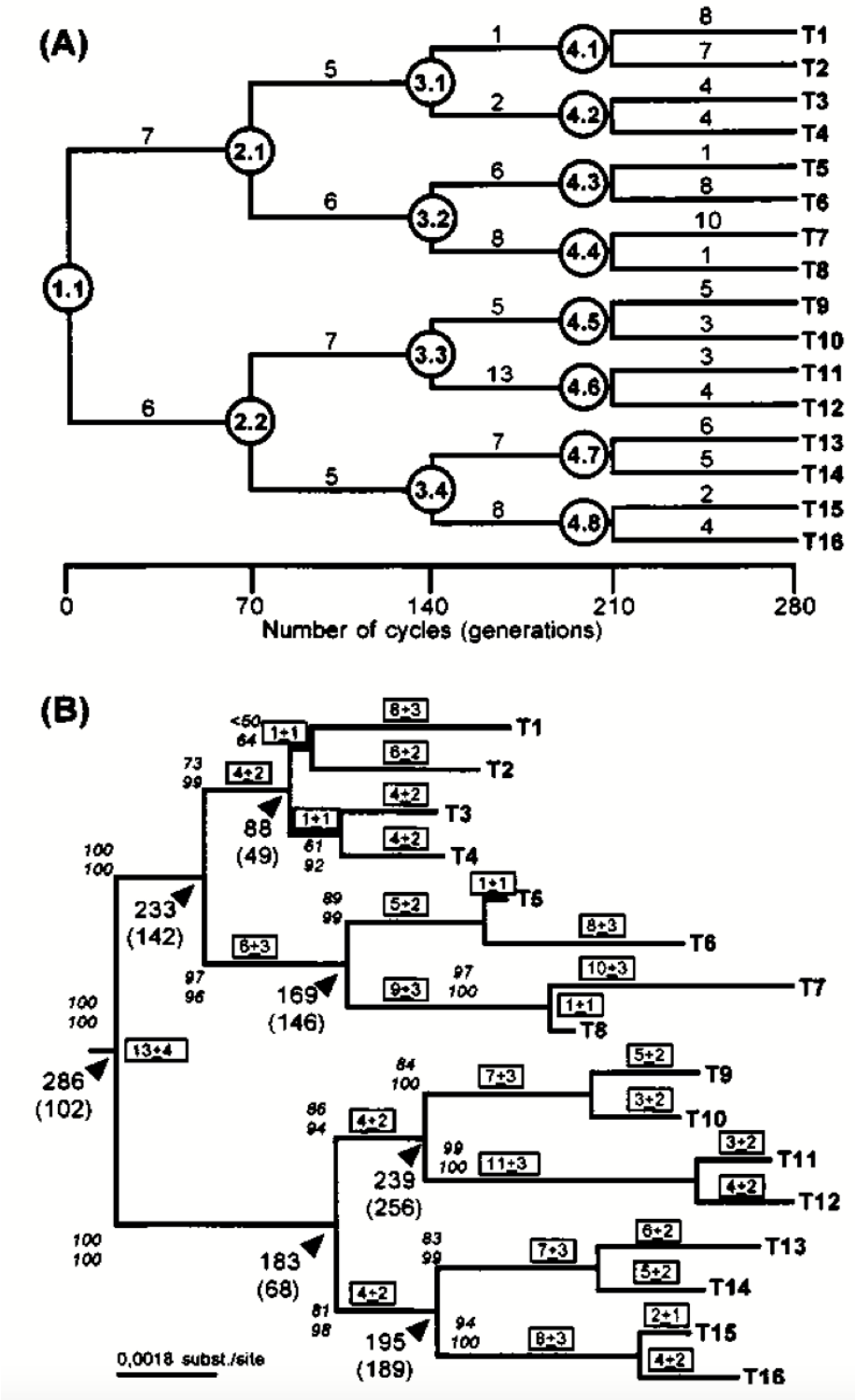


Figura 12. Comparación de la filogenia real con la filogenia inferida de máxima verosimilitud (Felsenstein 1981; Posada y Crandall 1998; Swofford 1998). La evolución in vitro de PCR en serie resultó en la topología representada (A) con longitudes de rama variables cuyos ancestros (1.1 a 4.8, encerrados en círculos) y secuencias terminales (T1 a T16) fueron secuenciados en su totalidad. La barra de escala indica el número de ciclos entre internodos y nodos del árbol. La filogenia inferida (B) tiene una topología idéntica al árbol real (A) y 9 de 30 longitudes de rama fueron estimadas correctamente. Los números en cuadros indican las longitudes de las ramas (número de sustituciones), los números en cursiva representan el porcentaje de un

determinado clúster en 100 réplicas bootstrap, con reestimación de parámetros en cada réplica bootstrap (arriba) y sin reestimación en cada réplica (abajo). Los números debajo de las flechas indican la divergencia estimada (ciclos atrás), con el rango de intervalo de confianza bajo-alto (entre paréntesis) según calculado por el análisis cuarteto de máxima verosimilitud (Rambaut y Bromham 1998). El número de sustituciones, con errores estándar correspondientes, en el árbol inferido (B) se calculó multiplicando las longitudes de rama (en sustituciones por sitio) por el número total de posiciones (2,238 bp).

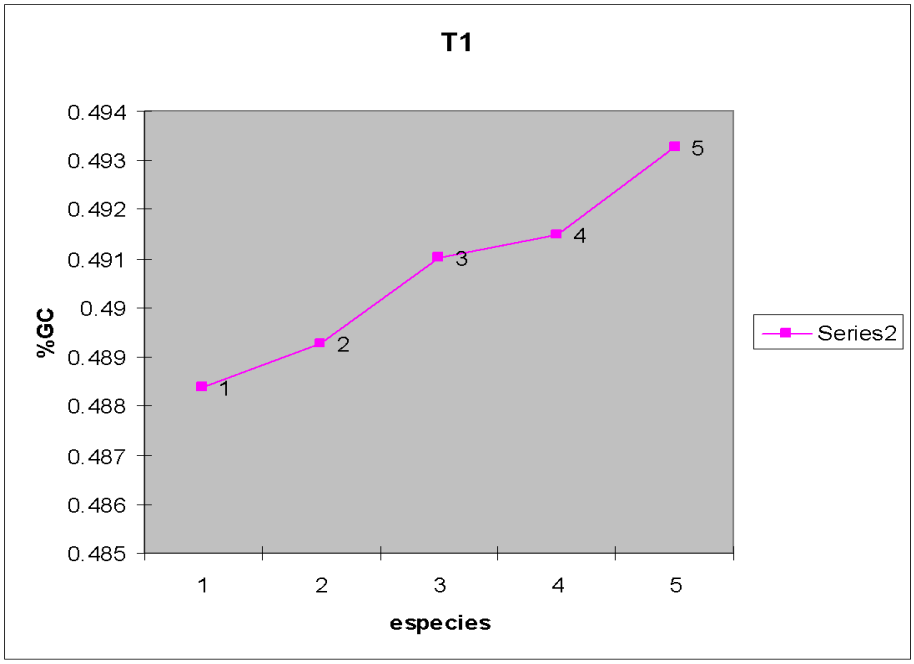
El estudo de Sanson GF. et.al., nos ofrece una oportunidad muy importante para generar evidencia experimental que permita contrastar las predicciones de una teoría de presión mutacional termodinámica, tanto a partir de los modelos de simulación in-silico, como a partir de un estudio analítico.

El estudio en cuestión, permite conocer las secuencias de cada uno de los nodos (extantes y extintos) para la filogenia completa. Eso significa que podemos identificar las diversas cascadas evolutivas (rutas con dirección ancestral: padre, hijo, nieto, bisnieto, tataranieto, etc.), con el conocmiento exacto de las secuencias de ADN de cada uno de estos individuos. De esta manera podemos evaluar cómo viene cambiando el contenido de GC a lo largo de las cascadas evolutivas, las cuales correlación con el tiempo evolutivo.

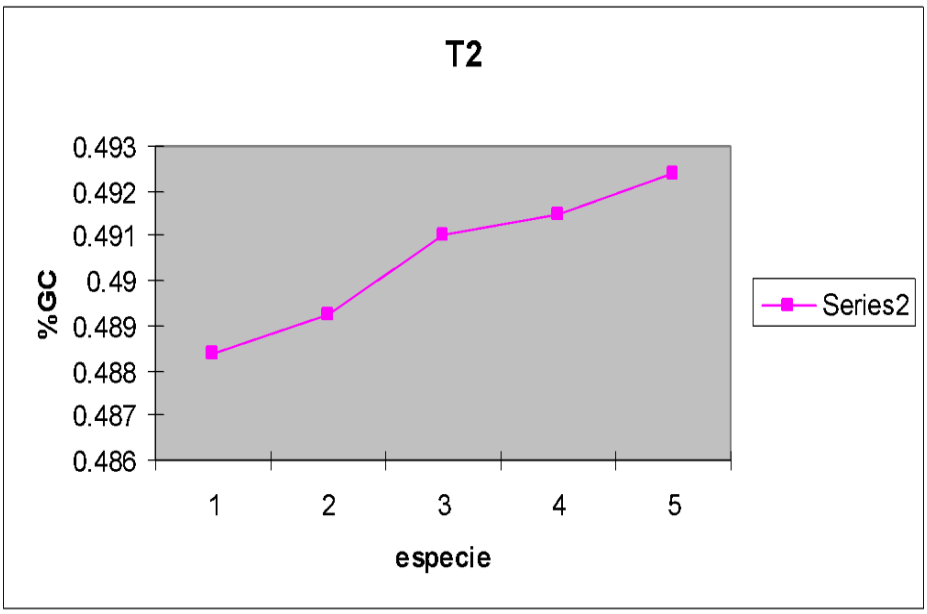
Es importante destacar, que en la naturaleza es muy dificil encontrar este tipo de información, por que generalmente, los organismos correspondientes a los nodos extintos, ya no existen, y por lo tanto sus secuencias de ADN son desconocidas. Sin embargo, en un ensayo de evolución in-vitro, los nodos internos (nodos extintos), estan disponibles para se secuenciados y poder averiguar así las caraceterísticas de las secuencias de ADN. Tomando las secuencias de los nodos internos y externos, y calculando el contenido de GC de cada una de ellas, podemos ver la variación del contenido de GC a lo largo de esta breve simulación de evolución in-vitro (i.e. a lo largo del tiempo evolutivo).

Los resultados son los siguientes:

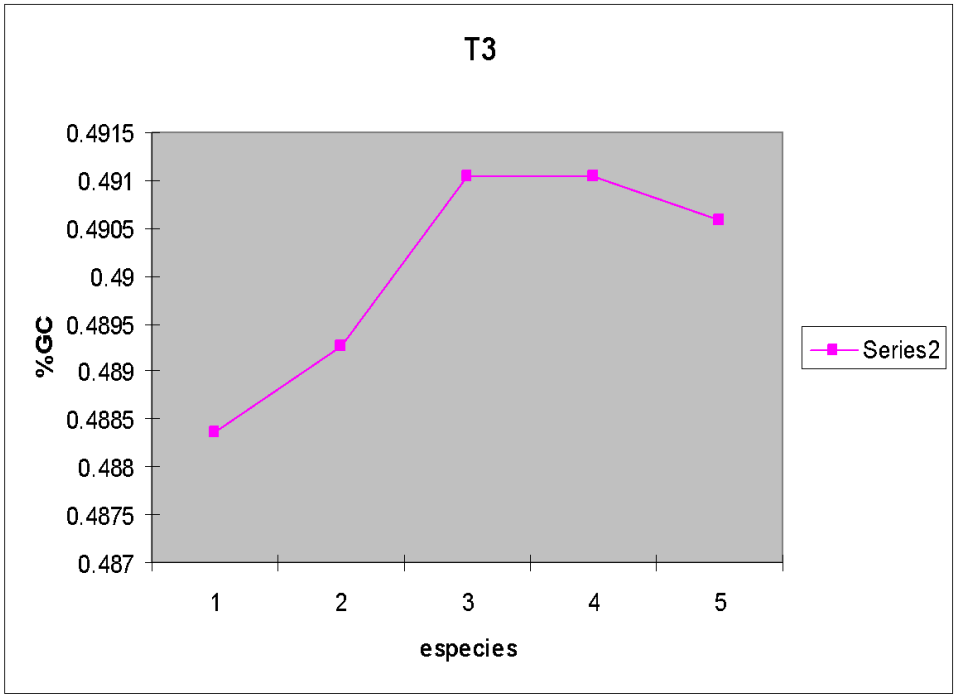
A


T1
0.494
0.493
0.492
0.491
0.49
0.489
0.488
0.487
0.486
0.485
%GC
1
2
3
4
5
1
2
3
4
5
especies
Series2

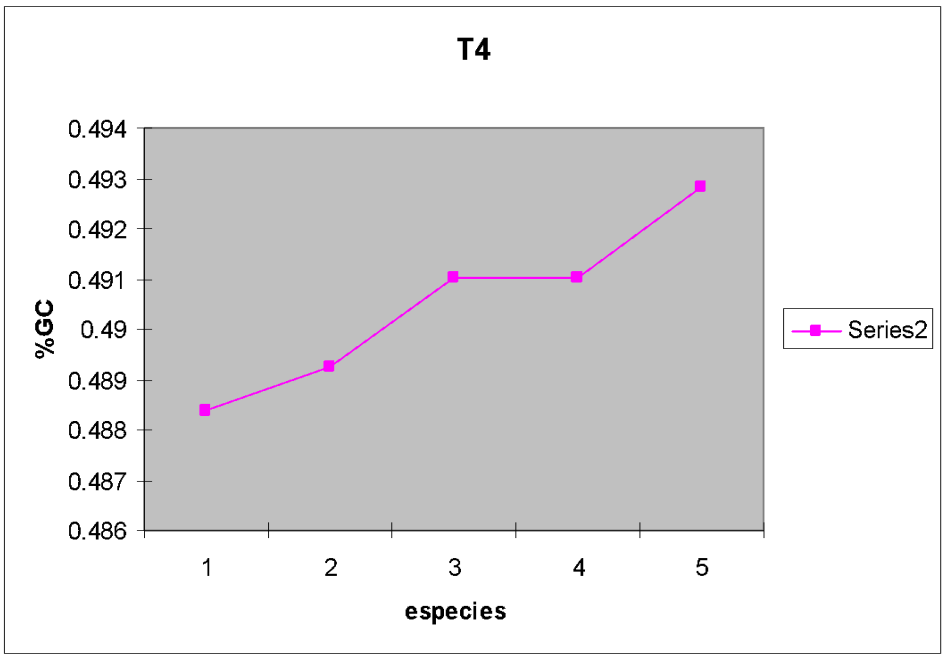
B


T2
0.493
0.492
0.491
0.49
0.489
0.488
0.487
0.486
%GC
1
2
3
4
5
especie
Series2

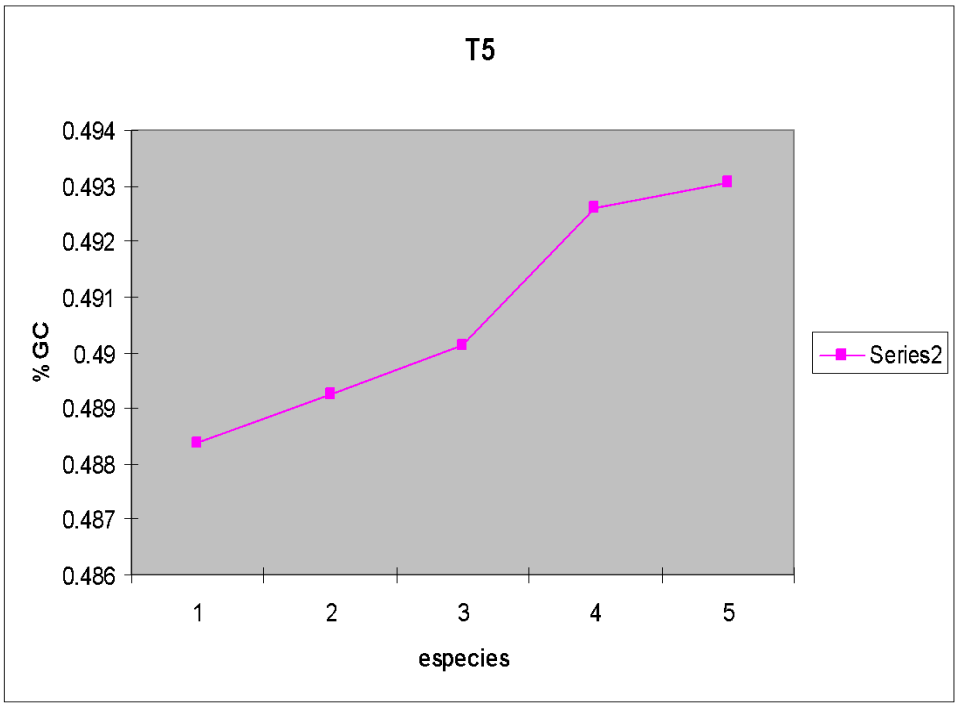
C


T3
0.4915
0.491
0.4905
0.49
0.4895
0.489
0.4885
0.488
0.4875
0.487
%GC
1
2
3
4
5
especies
Series2

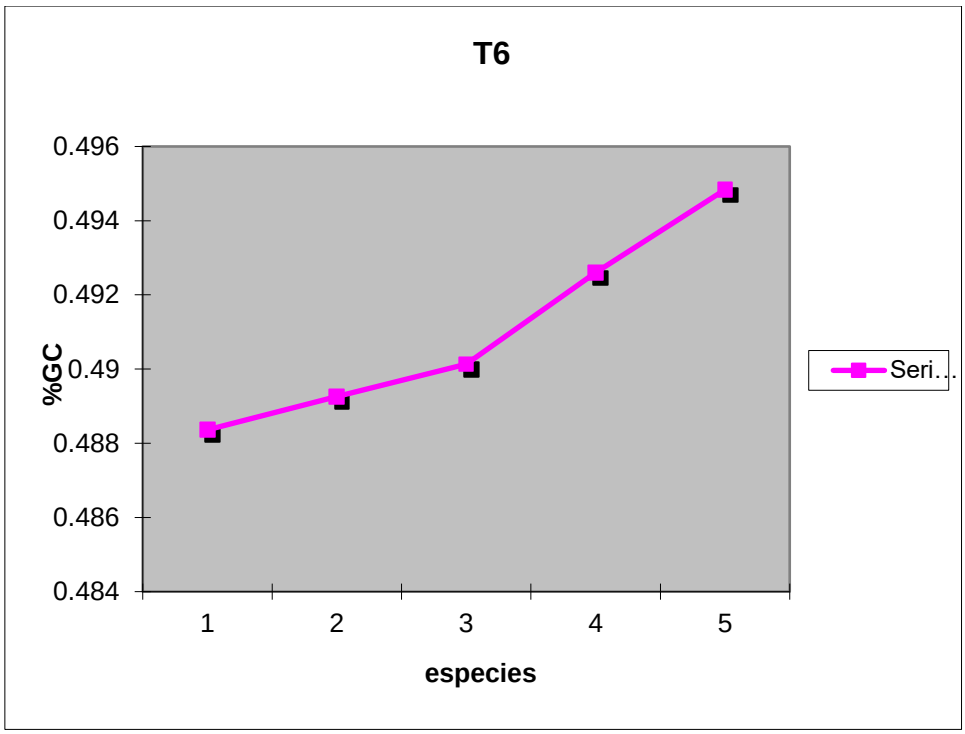
D


T4
0.494
0.493
0.492
0.491
0.49
0.489
0.488
0.487
0.486
%GC
1
2
3
4
5
especies
Series2

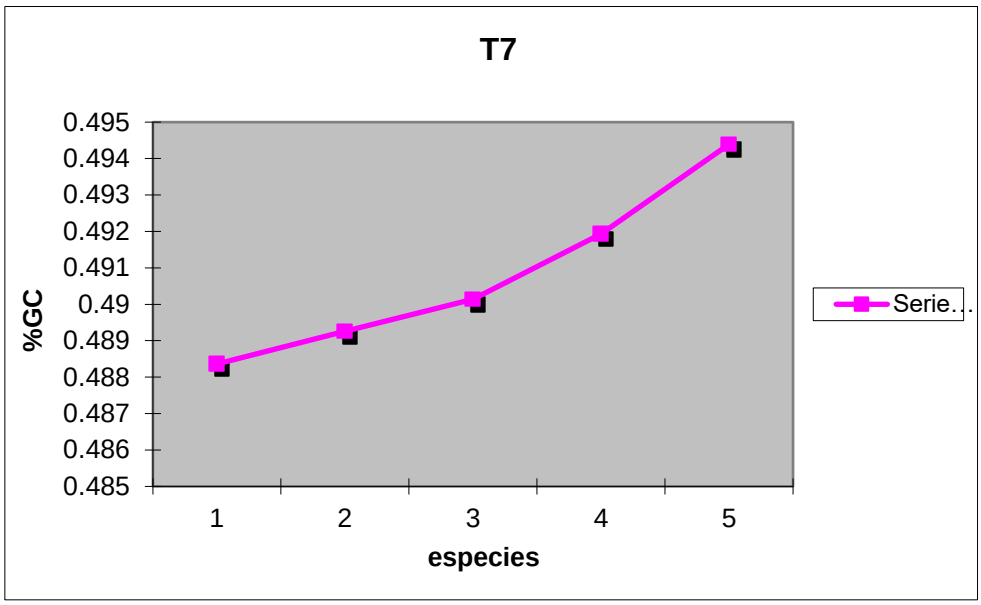
E


T5
0.494
0.493
0.492
0.491
0.49
0.489
0.488
0.487
0.486
% GC
1
2
3
4
5
especies
Series2

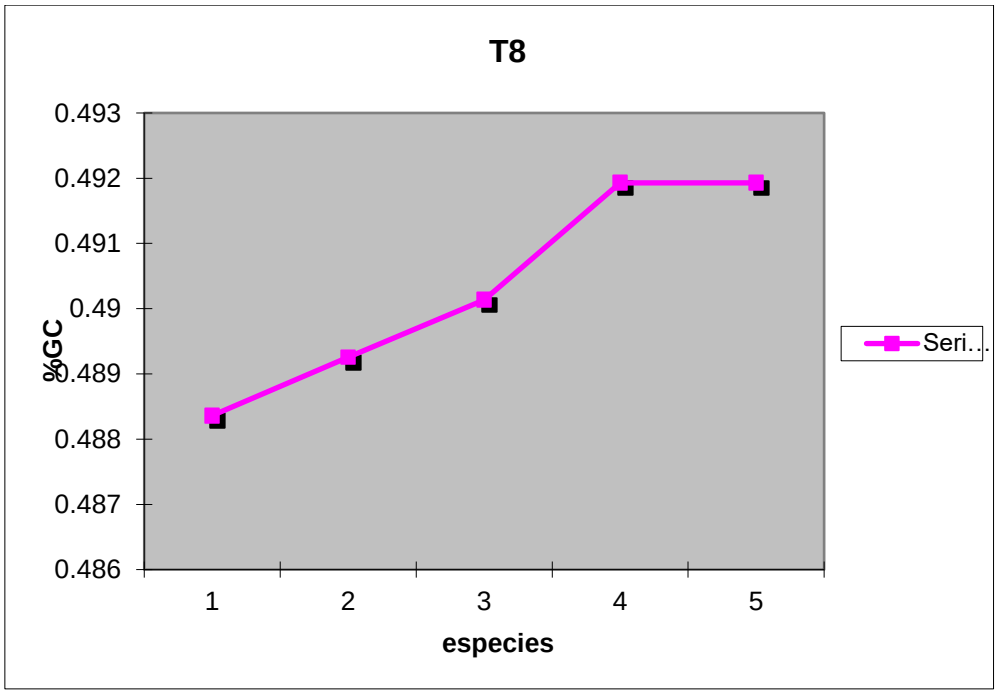
F


T6
0.496
0.494
0.492
0.49
0.488
0.486
0.484
%GC
1
2
3
4
5
especies
Seri…

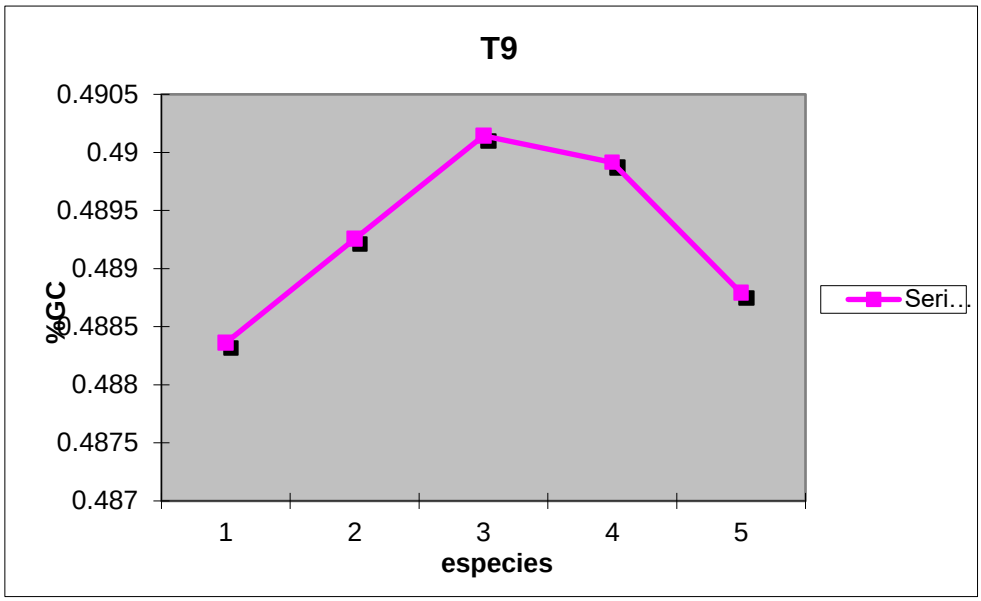
G


T7
0.495
0.494
0.493
0.492
0.491
0.49
0.489
0.488
0.487
0.486
0.485
%GC
1
2
3
4
5
especies
Serie...

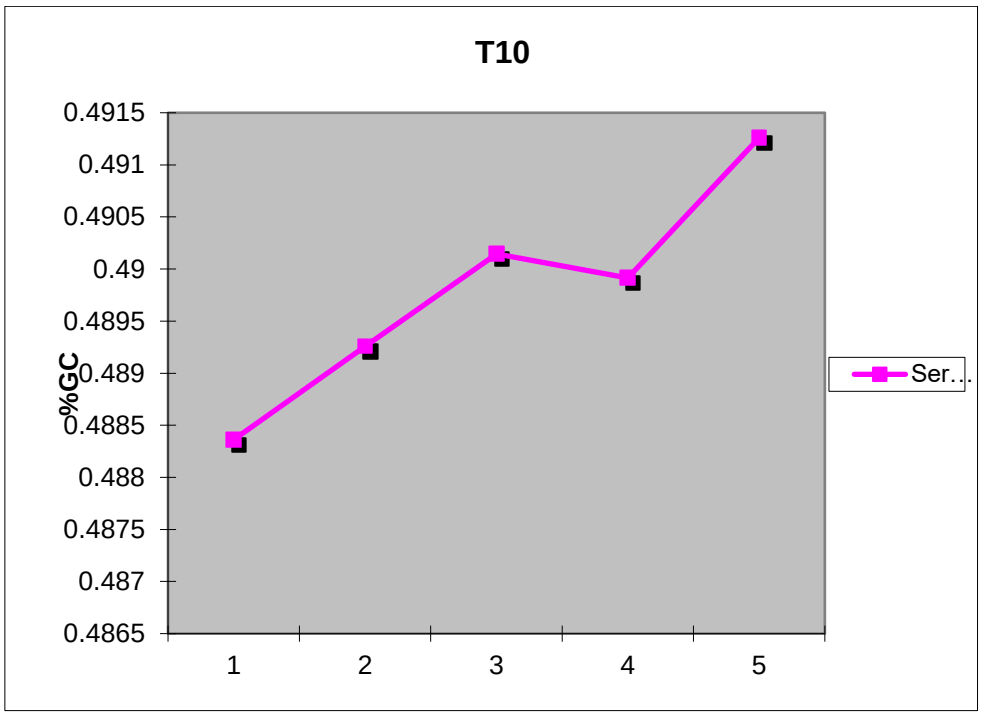
H


T8
0.493
0.492
0.491
0.49
0.489
0.488
0.487
0.486
%GC
1
2
3
4
5
especies
Seri...

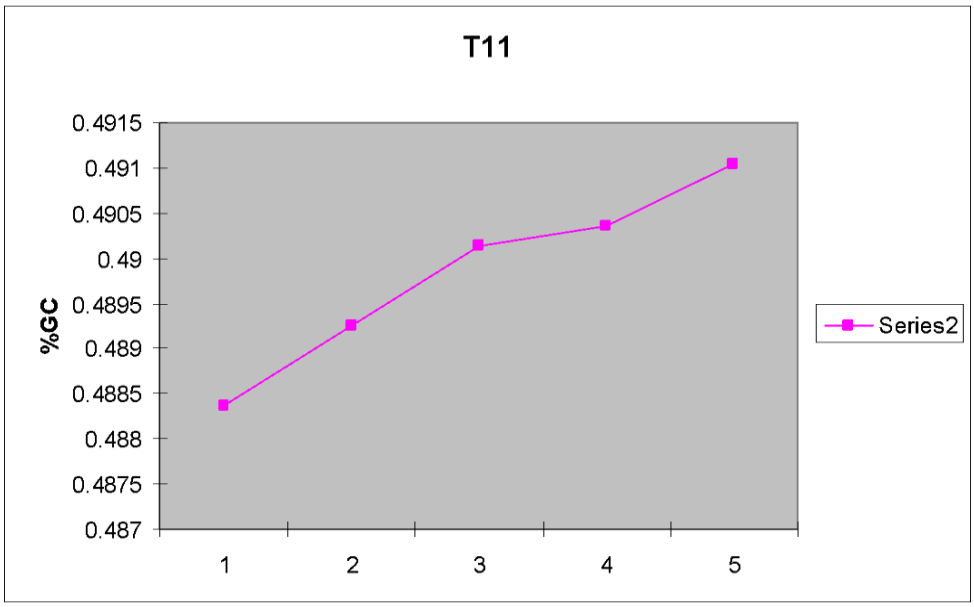
I


T9
0.4905
0.49
0.4895
0.489
0.4885
0.488
0.4875
0.487
%GC
1
2
3
4
5
especies
Seri...

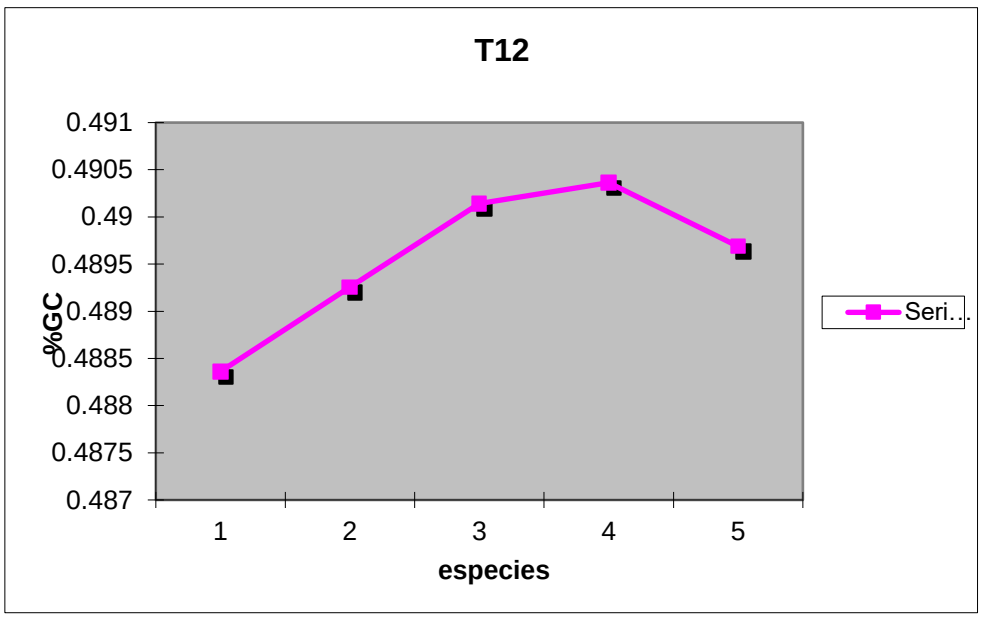
J


T10
0.4915
0.491
0.4905
0.49
0.4895
0.489
0.4885
0.488
0.4875
0.487
0.4865
%GC
1
2
3
4
5
Ser...

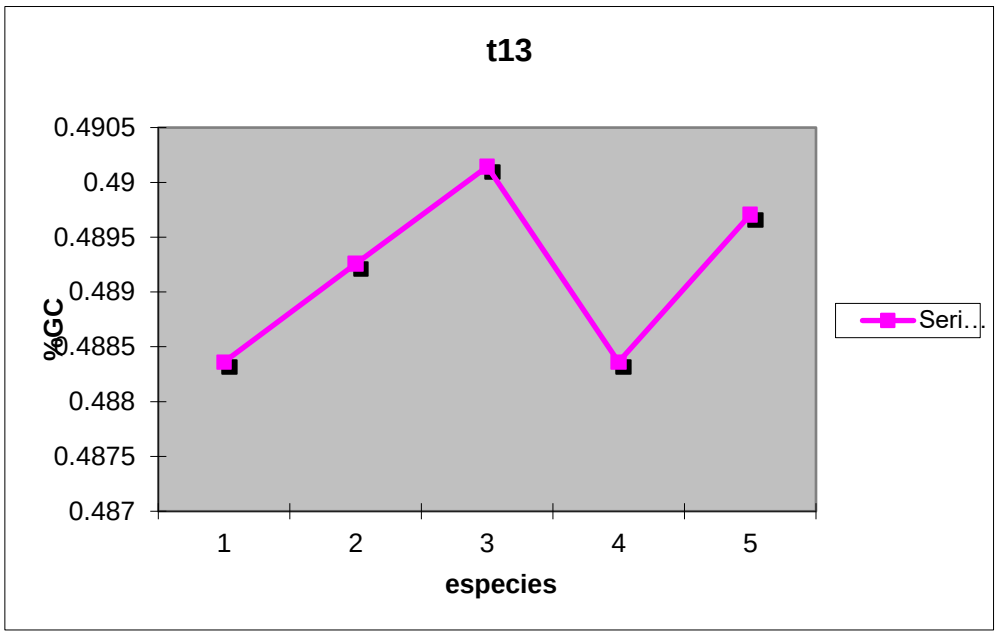
K


T11
0.4915
0.491
0.4905
0.49
0.4895
0.489
0.4885
0.488
0.4875
0.487
%GC
1
2
3
4
5
Series2

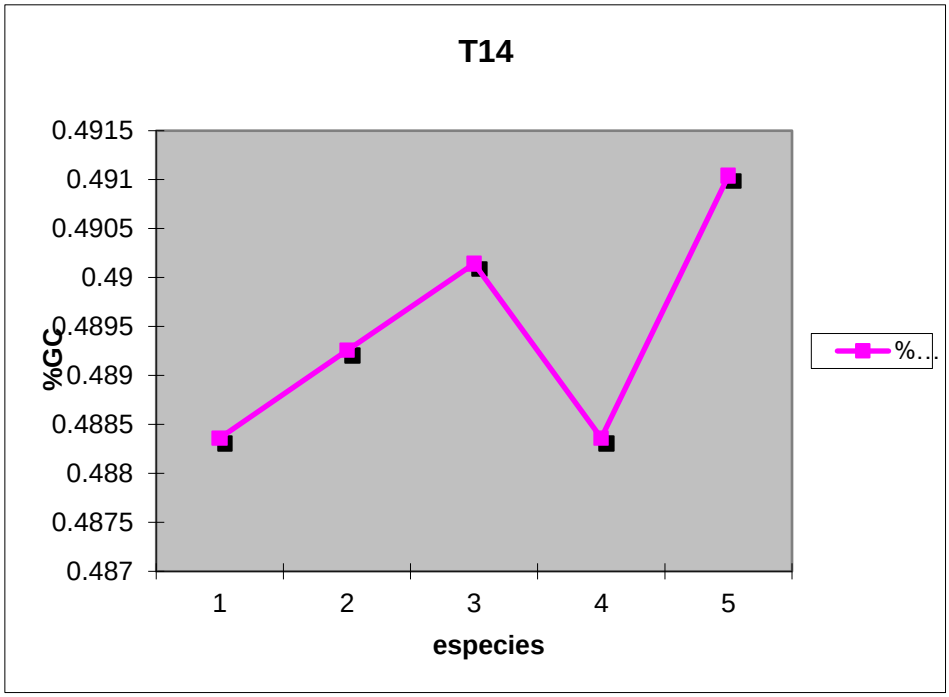
L


T12
0.491
0.4905
0.49
0.4895
0.489
0.4885
0.488
0.4875
0.487
%GC
1
2
3
4
5
especies
Seri...

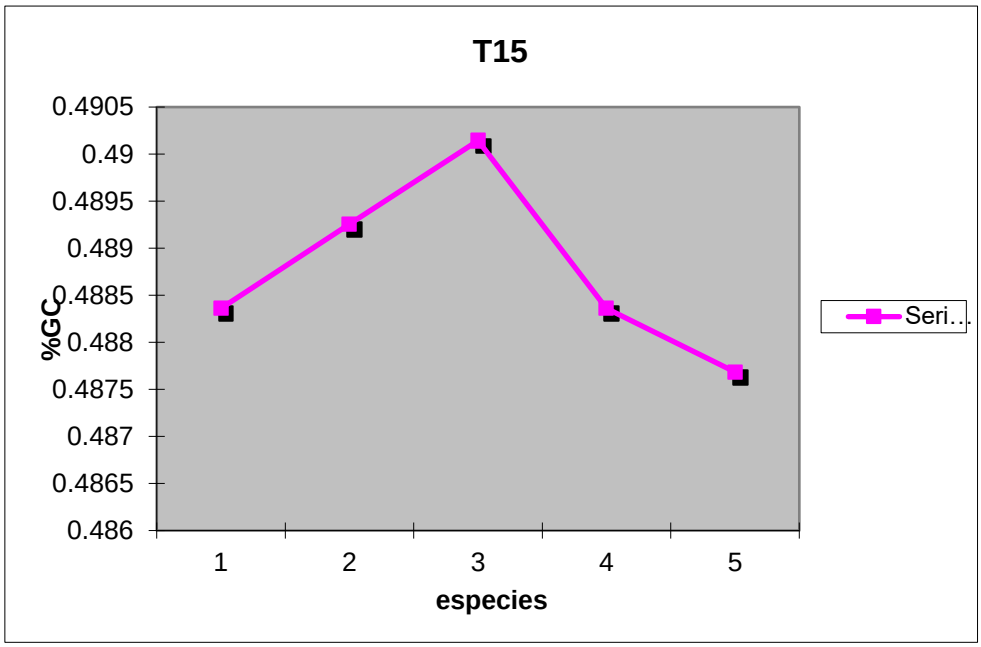
M


t13
0.4905
0.49
0.4895
0.489
0.4885
0.488
0.4875
0.487
%GC
1
2
3
4
5
especies
Seri…

N


T14
0.4915
0.491
0.4905
0.49
0.4895
0.489
0.4885
0.488
0.4875
0.487
%GC
1
2
3
4
5
especies
%…

O



P

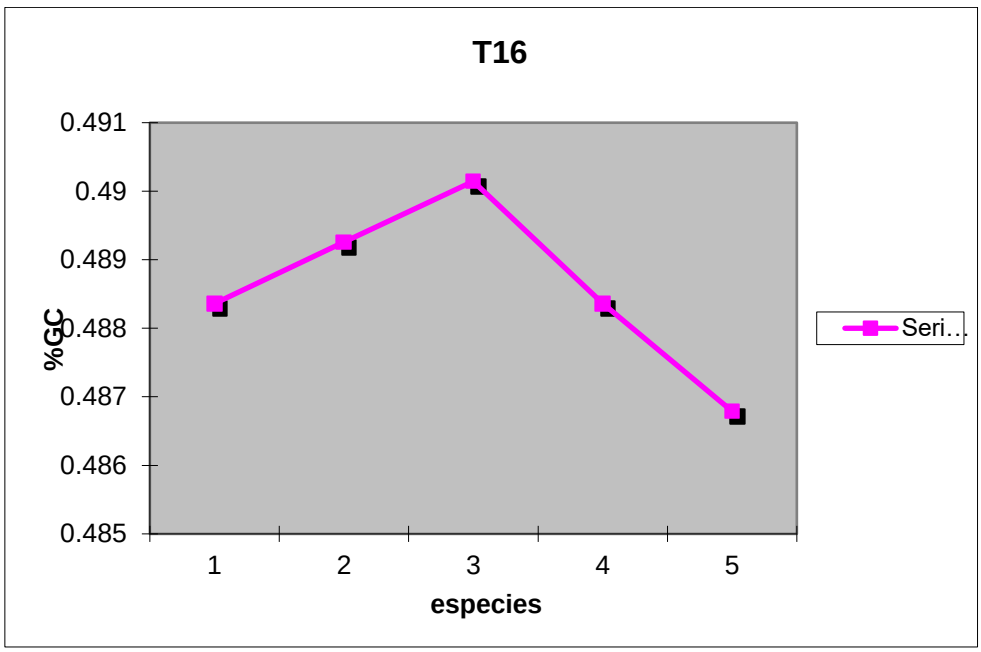


Figura 13. A, B, C…P. Variación del porcentaje de GC a lo largo del tiempo (orden ancestral) en las diferentes cascadas evolutivas

El análisis estadístico para verificar la hipótesis de que hay una relación lineal entre el porcentaje de GC y el tiempo evolutivo determinado por el orden acnestral en la cascada filogenética, realizado a través de una regresión lineal, muestra una mayoría de cascadas evolutivas con evidencia de una relación lineal con pendiente positiva significativa, mostrando evidencia de un incremento del contenido de GC a lo largo del tiempo evolutivo. A continuacion se muestran las regresiones lineales para cada una de las 16 cascadas cuyas variaciones se aprecian en la figura 13.

```
. regress gc tiempo if id==1(T1)

      Source |       SS       df       MS              Number of obs =       5
-------------+------------------------------           F(  1,     3) =  115.39
       Model |  .000014512     1  .000014512           Prob > F      =  0.0017
    Residual |  3.7728e-07     3  1.2576e-07           R-squared     =  0.9747
-------------+------------------------------           Adj R-squared =  0.9662
       Total |  .000014889     4  3.7223e-06           Root MSE      =  .00035

------------------------------------------------------------------------------
          gc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tiempo |   .0012047   .0001121    10.74   0.002     .0008478    .0015616
       _cons |   .4870667   .0003719  1309.54   0.000      .485883    .4882504
------------------------------------------------------------------------------

. regress gc tiempo if id==2(T2)

      Source |       SS       df       MS              Number of obs =       5
-------------+------------------------------           F(  1,     3) =   85.28
       Model |  .000010512     1  .000010512           Prob > F      =  0.0027
    Residual |  3.6978e-07     3  1.2326e-07           R-squared     =  0.9660
-------------+------------------------------           Adj R-squared =  0.9547
       Total |  .000010882     4  2.7205e-06           Root MSE      =  .00035

------------------------------------------------------------------------------
          gc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tiempo |   .0010253    .000111     9.23   0.003      .000672    .0013786
       _cons |   .4874254   .0003682  1323.74   0.000     .4862536    .4885973
------------------------------------------------------------------------------


. regress gc tiempo if id==3(T3)

      Source |       SS       df       MS              Number of obs =       5
-------------+------------------------------           F(  1,     3) =    6.36
       Model |  3.8651e-06     1  3.8651e-06           Prob > F      =  0.0861
    Residual |  1.8244e-06     3  6.0815e-07           R-squared     =  0.6793
-------------+------------------------------           Adj R-squared =  0.5724
       Total |  5.6896e-06     4  1.4224e-06           Root MSE      =  .00078

------------------------------------------------------------------------------
          gc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tiempo |   .0006217   .0002466     2.52   0.086    -.0001631    .0014065
       _cons |   .4881878   .0008179   596.88   0.000     .4855848    .4907907
------------------------------------------------------------------------------
```

```
. regress gc tiempo if id==4(T4)

      Source |       SS       df       MS              Number of obs =       5
-------------+------------------------------           F(  1,     3) =   53.96
       Model |  .000011452     1  .000011452           Prob > F      =  0.0052
    Residual |  6.3669e-07     3  2.1223e-07           R-squared     =  0.9473
-------------+------------------------------           Adj R-squared =  0.9298
       Total |  .000012088     4  3.0221e-06           Root MSE      =  .00046

------------------------------------------------------------------------------
          gc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tiempo |   .0010701   .0001457     7.35   0.005     .0006065    .0015338
       _cons |   .4872909   .0004832  1008.53   0.000     .4857533    .4888286
------------------------------------------------------------------------------

. regress gc tiempo if id==5(T5)

      Source |       SS       df       MS              Number of obs =       5
-------------+------------------------------           F(  1,     3) =   59.12
       Model |  .000016154     1  .000016154           Prob > F      =  0.0046
    Residual |  8.1966e-07     3  2.7322e-07           R-squared     =  0.9517
-------------+------------------------------           Adj R-squared =  0.9356
       Total |  .000016973     4  4.2433e-06           Root MSE      =  .00052

------------------------------------------------------------------------------
          gc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tiempo |    .001271   .0001653     7.69   0.005     .0007449     .001797
       _cons |   .4868665   .0005482   888.09   0.000     .4851218    .4886112
------------------------------------------------------------------------------

. regress gc tiempo if id==6(T6)

      Source |       SS       df       MS              Number of obs =       5
-------------+------------------------------           F(  1,     3) =   54.00
       Model |  .000026535     1  .000026535           Prob > F      =  0.0052
    Residual |  1.4742e-06     3  4.9141e-07           R-squared     =  0.9474
-------------+------------------------------           Adj R-squared =  0.9298
       Total |  .000028009     4  7.0022e-06           Root MSE      =   .0007

------------------------------------------------------------------------------
          gc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tiempo |   .0016289   .0002217     7.35   0.005     .0009235    .0023344
       _cons |   .4861505   .0007352   661.23   0.000     .4838107    .4884903
------------------------------------------------------------------------------
. regress gc tiempo if id==7(T7)

      Source |       SS       df       MS              Number of obs =       5
-------------+------------------------------           F(  1,     3) =   52.48
       Model |  .000021704     1  .000021704           Prob > F      =  0.0054
    Residual |  1.2407e-06     3  4.1357e-07           R-squared     =  0.9459
-------------+------------------------------           Adj R-squared =  0.9279
       Total |  .000022944     4  5.7361e-06           Root MSE      =  .00064

------------------------------------------------------------------------------
          gc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tiempo |   .0014732   .0002034     7.24   0.005      .000826    .0021204
       _cons |   .4863951   .0006745   721.14   0.000     .4842486    .4885416
------------------------------------------------------------------------------
```

```
. regress gc tiempo if id==8(T8)

      Source |       SS       df       MS              Number of obs =       5
-------------+------------------------------           F(  1,     3) =   51.52
       Model |  9.6127e-06     1  9.6127e-06           Prob > F      =  0.0056
    Residual |  5.5970e-07     3  1.8657e-07           R-squared     =  0.9450
-------------+------------------------------           Adj R-squared =  0.9266
       Total |  .000010172     4  2.5431e-06           Root MSE      =  .00043

------------------------------------------------------------------------------
          gc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tiempo |   .0009804   .0001366     7.18   0.006     .0005458    .0014151
       _cons |   .4873806    .000453  1075.86   0.000     .4859389    .4888223
------------------------------------------------------------------------------

. regress gc tiempo if id==9(T9)

      Source |       SS       df       MS              Number of obs =       5
-------------+------------------------------           F(  1,     3) =    0.34
       Model |  2.2891e-07     1  2.2891e-07           Prob > F      =  0.5993
    Residual |  2.0033e-06     3  6.6777e-07           R-squared     =  0.1025
-------------+------------------------------           Adj R-squared = -0.1966
       Total |  2.2322e-06     4  5.5805e-07           Root MSE      =  .00082

------------------------------------------------------------------------------
          gc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tiempo |   .0001513   .0002584     0.59   0.599    -.0006711    .0009737
       _cons |   .4888393   .0008571   570.37   0.000     .4861118    .4915669
------------------------------------------------------------------------------

. regress gc tiempo if id==10(T10)

      Source |       SS       df       MS              Number of obs =       5
-------------+------------------------------           F(  1,     3) =   27.20
       Model |  4.1648e-06     1  4.1648e-06           Prob > F      =  0.0137
    Residual |  4.5928e-07     3  1.5309e-07           R-squared     =  0.9007
-------------+------------------------------           Adj R-squared =  0.8676
       Total |  4.6241e-06     4  1.1560e-06           Root MSE      =  .00039

------------------------------------------------------------------------------
          gc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tiempo |   .0006454   .0001237     5.22   0.014     .0002516    .0010391
       _cons |   .4878512   .0004104  1188.81   0.000     .4865452    .4891572
------------------------------------------------------------------------------
. regress gc tiempo if id==11(T11)

      Source |       SS       df       MS              Number of obs =       5
-------------+------------------------------           F(  1,     3) =   80.13
       Model |  4.1544e-06     1  4.1544e-06           Prob > F      =  0.0029
    Residual |  1.5554e-07     3  5.1847e-08           R-squared     =  0.9639
-------------+------------------------------           Adj R-squared =  0.9519
       Total |  4.3100e-06     4  1.0775e-06           Root MSE      =  .00023

------------------------------------------------------------------------------
          gc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tiempo |   .0006445    .000072     8.95   0.003     .0004154    .0008737
       _cons |   .4878976   .0002388  2043.02   0.000     .4871376    .4886576
------------------------------------------------------------------------------
```

```
. regress gc tiempo if id==12(T12)

      Source |       SS       df       MS              Number of obs =       5
-------------+------------------------------           F(  1,     3) =    3.78
       Model |  1.4100e-06     1  1.4100e-06           Prob > F      =  0.1471
    Residual |  1.1189e-06     3  3.7298e-07           R-squared     =  0.5575
-------------+------------------------------           Adj R-squared =  0.4101
       Total |  2.5289e-06     4  6.3222e-07           Root MSE      =  .00061

------------------------------------------------------------------------------
          gc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tiempo |   .0003755   .0001931     1.94   0.147    -.0002391    .0009901
       _cons |   .4884357   .0006405   762.55   0.000     .4863973    .4904742
------------------------------------------------------------------------------

. regress gc tiempo if id==13(T13)

      Source |       SS       df       MS              Number of obs =       5
-------------+------------------------------           F(  1,     3) =    0.43
       Model |  3.2060e-07     1  3.2060e-07           Prob > F      =  0.5580
    Residual |  2.2268e-06     3  7.4227e-07           R-squared     =  0.1259
-------------+------------------------------           Adj R-squared = -0.1655
       Total |  2.5474e-06     4  6.3685e-07           Root MSE      =  .00086

------------------------------------------------------------------------------
          gc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tiempo |   .0001791   .0002724     0.66   0.558     -.000688    .0010461
       _cons |   .4886285   .0009036   540.76   0.000     .4857528    .4915042
------------------------------------------------------------------------------

. regress gc tiempo if id==14(T14)

      Source |       SS       df       MS              Number of obs =       5
-------------+------------------------------           F(  1,     3) =    1.74
       Model |  1.9894e-06     1  1.9894e-06           Prob > F      =  0.2784
    Residual |  3.4223e-06     3  1.1408e-06           R-squared     =  0.3676
-------------+------------------------------           Adj R-squared =  0.1568
       Total |  5.4116e-06     4  1.3529e-06           Root MSE      =  .00107

------------------------------------------------------------------------------
          gc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tiempo |    .000446   .0003378     1.32   0.278    -.0006289    .0015209
       _cons |   .4880946   .0011202   435.72   0.000     .4845296    .4916595
------------------------------------------------------------------------------

. regress gc tiempo if id==15(T15)

      Source |       SS       df       MS              Number of obs =       5
-------------+------------------------------           F(  1,     3) =    0.49
       Model |  5.0588e-07     1  5.0588e-07           Prob > F      =  0.5362
    Residual |  3.1283e-06     3  1.0428e-06           R-squared     =  0.1392
-------------+------------------------------           Adj R-squared = -0.1477
       Total |  3.6341e-06     4  9.0854e-07           Root MSE      =  .00102

------------------------------------------------------------------------------
          gc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tiempo |  -.0002249   .0003229    -0.70   0.536    -.0012526    .0008027
       _cons |   .4894364    .001071   456.99   0.000     .4860281    .4928448
------------------------------------------------------------------------------
```

```
. regress gc tiempo if id==16(T16)

      Source |       SS       df       MS              Number of obs =       5
-------------+------------------------------           F(  1,     3) =    1.07
       Model |  1.6325e-06     1  1.6325e-06           Prob > F      =  0.3768
    Residual |  4.5725e-06     3  1.5242e-06           R-squared     =  0.2631
-------------+------------------------------           Adj R-squared =  0.0175
       Total |  6.2050e-06     4  1.5513e-06           Root MSE      =  .00123

------------------------------------------------------------------------------
          gc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tiempo |   -.000404   .0003904    -1.03   0.377    -.0016465    .0008384
       _cons |   .4897947   .0012948   378.27   0.000      .485674    .4939154
------------------------------------------------------------------------------
```

Si analizamos la información complete comprendida en las 16 cascadas, por medio de un modelo linearl del tipo Ecuaciones de Estimación Generalizada (GEE), tanto para un modelo de correlación intercambiable o un modelo de correlación independiente, se obtiene una pendiente positiva significativamente diferente de cero, con un alto nivel de significancia.

```
. xtgee gc tiempo, family(gauss) link(id) corr(exchangeable) i(id)

Iteration 1: tolerance = 9.327e-16

GEE population-averaged model                   Number of obs      =        80
Group variable:                         id      Number of groups   =        16
Link:                             identity      Obs per group: min =         5
Family:                           Gaussian                     avg =       5.0
Correlation:                  exchangeable                     max =         5
                                                Wald chi2(1)       =     64.45
Scale parameter:                  1.49e-06      Prob > chi2        =    0.0000

------------------------------------------------------------------------------
          gc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tiempo |    .000693   .0000863     8.03   0.000     .0005238    .0008622
       _cons |   .4878589   .0003169  1539.59   0.000     .4872378    .4884799
------------------------------------------------------------------------------
```

```
. xtgee gc tiempo, family(gauss) link(id) corr(ind) i(id)

Iteration 1: tolerance = 1.119e-16

GEE population-averaged model                   Number of obs      =        80
Group variable:                        id       Number of groups   =        16
Link:                            identity       Obs per group: min =         5
Family:                          Gaussian                      avg =       5.0
Correlation:                  independent                      max =         5
                                                Wald chi2(1)       =     51.66
Scale parameter:                 1.49e-06       Prob > chi2        =    0.0000

Pearson chi2(80):                    0.00       Deviance           =      0.00
Dispersion (Pearson):            1.49e-06       Dispersion         =  1.49e-06

------------------------------------------------------------------------------
          gc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tiempo |    .000693   .0000964     7.19   0.000      .000504     .000882
       _cons |   .4878589   .0003198  1525.64   0.000     .4872321    .4884856
------------------------------------------------------------------------------
```

Este análisis confirma que la evidencia experimental obtenida a partir del ensayo de evolución in-vitro conducido por Sanson et. Al., apoya a la teoría de la presión evolutiva termodinámica, mostrando un aumento estadisticamente significativo del contenido de GC a lo largo del tiempo evolutivo.

4.2 Discusión

El proceso de replicación y reparación del ADN, es el factor de tipo biológico más imporante en la generación de mutaciones y por consecuencia en la evolución del ADN.

La distribución de Boltzmann para un conjunto canónico, bajo condiciones en las cuales el entorno del sistema de replicación del ADN se evalúa dentro de una vecindad espacial y temporal (es decir dentro de un instante de tiempo breve y en un volumen del espacio limitado), puede utilizarse para calcular la probabilidad de que un nucleótido libre ingrese al sitio de replicación previo a la formación del enlace fosfodiester y como consecuencia, la polimerización de la hebra de ADN naciente.

La probabilidad de que alguno de los cuatro nucleótidos libres pueda ingresar al sitio de replicación, dependerá de diversos factores, dentro de ellos, los más relevantes corresponden a la energía potencial del sistema molecular, la abundancia de los

nucléotidos libres, la secuencia de ADN del entorno al sitio de replicación, y la temperatura. Dentro de los términos más relevantes de la energía potencial, destacan la energía de puentes de hidrógeno, y las energías electrostáticas dadas por las interacciones intramoleculares a nivel de cargas y dipolos eléctricos.

Considerando estas condiciones, es posible calcular las probabilidades de que en un instante dado durante la replicación del ADN, un nucleótido libre cualquiera pueda ingresar al sitio de replicación de la enzima ADN polimerasa. El conocimiento de estas probabilidades permite estudiar ciertos aspectos del proceso evolutivo, relacionados a la acumulación de nucleótidos en el ADN como resultado de un número grande de procesos de replicación, que se dan durante un proceso evolutivo.

Existen al menos dos maneras de realizar estos estudios. (1) A partir de una simulación computacional, y (2) a partir de un estudio analítico. En ambos casos podemos predecir las características del contenido de nucleótidos en el ADN, en los límites de muy largos tiempos de procesos evolutivos.

En un estudio previo, desarrollamos y estudiamos una simulación computacional de mutaciones puntuales, utilizando los cálculos de las probabilidades de ocupación de nucléotidos en el sitio de replicación del ADN, y la distribución de Boltzmann para un conjunto canónico en una simulación de Monte Carlo. Para calcular las probabilidades dadas por la distribución de Boltzmann, en aquel trabajo realizamos un cálculo para estimar las energías de puentes de hidrógeno, así como las energías electrostáticas, asumiendo la naturaleza dipolar de los enlaces covalentes, permitiéndonos estimar la energía potentical del sistema molecular descrito. La simulación predijo teóricamente un aumento en el contenido de guanina-citosina (GC) en la molécula de ADN, con la tendencia de alcanzar un equilibrio en el tiempo. En dicha simulación, se logró comprender el efecto en el proceso evolutivo, de algunos parámetros como la temperatura, la concentración de nucleótidos libres, el contenido inicial de GC en la secuencia, y la cinética que puede tener el mecanismo de repación del ADN. Las predicciones computacionales realizadas en aquella oportunidad, fueron contrastadas con evidencia experimental de secuencias genómicas de

diversas especies cuya historia evolutiva es conocida. De manera específica se estudio el linaje de *Kinetoplastidia* y *Plasmodium* para los cuales se determinó sesgo del uso de codones (codon bias) de manera experimental.

La evidencia experimental no solo confirma la predicción teórica de buscar un aumento del contenido de GC a lo largo del tiempo de evolución, sino que explica de manera natural el fenómeno del "codon bias", el cual surge como una simple consecuencia de las preferencias naturales de que ciertas mutaciones se vean favorecidas por medio de las probabilidades calculadas a partir de la distribución de Boltzmann.

El proceso de simulación descrito nos llevó a proponer la existencia de una Presión mutacional termodinámica, que actúa como un ‘driving force’ de la evolución, la cual junto con la Presión de Selección, terminan definiendo las características de los genomas de las especies a lo largo de la evolución.

En el presente trabajo mostramos una evaluación analítica, basada en las probabilidades estimadas por la distribución de Boltzmann para un sistema canónico, las energías potenciales electrostáticas y de puentes de hidrógeno del sistema molecular. Hemos introducido una nomenclatura adecuada para describir los distintos aspectos moleculares del sistema, así como los procesos más importantes durante la replicación del ADN. Empleamos un análisis teórico matemático-estadístico, para comprender la solución estiacionaría en el límite de un tiempo infinito, para la ecuación diferencial lineal de primer orden en el tiempo, la cual se asume que explica en primera aproximación, la cinética de la variación del contenido de GC a lo largo del tiempo, en un genoma que viene experimentando múltiples procesos de replicación durante un proceso evolutivo.

Considerando una serie de premisas y estimaciones de las energías de puentes de hidrógeno y energías potenciales electrostáticas, así como abundancias relativas de nucleótidos en el entorno celular, logramos estimar las concentraciones de GC en los límites cuando el tiempo tiende al infinito.

Como resultado de este trabajo, encontramos que las estimaciones teóricas se acercan mucho a los valores predichos en las simulaciones computacionales en nuestro estudio

previo, lo cual constituye una evidencia importante que apoya la postulación de la existencia de una "presión mutacional termodinámica" que estaría guiando el proceso evolutivo, y sería capaz de explicar el fenómico del "codon bias", el cual es una de las incógnitas más importantes en la biología que no se ha logrado explicar de una manera completa.

El presente trabajo muestra un tratamiento analítico del proceso de formación de pares no-canónicos (missmatches), guiado por la distribución de Boltzmann para un ensamble canónico. La cinética de primer orden que representa a la variación del contenido de GC en el genoma a lo largo del tiempo evolutivo, encuentra una solución a la ecuación diferencial que modela el proceso, que se puede aproximar cuando el tiempo tiende al infinito. Dicha solución predice soluciones asintóticas que muestran la tendencia a alcanzar un equilibrio en la medida que el tiempo evolutivo transcurre. Las soluciones en el equilibrio, muestran valores asintóticos del contenido de GC, los cuales dependen de la naturaleza física fundamental de los pares no-canónicos (missmatches o pares distintos a los Watson-Crick), asociada a la energía potencial electrostática, la cual se estima por sus dos principales contribuyentes: la energía tipo puente de hidrógeno entre el nucleótido entrante y su opuesto en la hebra madre, y la energía de interacción entre el nucleótido entrante y los demás nucleótidos que se encuentran en una vecidad de cinco pares de bases (dos pares a la izquierda y dos pares a la derecha del nucleótido entrante). Los contenidos de GC estimados en la solución asintótica, se acercan a los valores en equilibrio que se obtienen a partir de una simulación computacional basada en un algoritmo de Monte Carlo, guiado por la distribución de probabilidades estimadas por la distribución de Boltzmann y las mismas energías mencionadas.

El presente estudio muestra que la hipótesis de la presión mutacional termodinámica, sugiere que durante la evolución, de no haber una presión de selección considerable, se espera que las mutaciones que aparezcan correspondan a aquellas termodinámicamente más favorables, cuya consecuencia esperada es un gradual incremento del contenido de GC a lo largo del tiempo, hasta alcanzar un valor asintótico de equilibrio. Este valor del

contenido de GC de equilibrio va a depender de la temperatura, por lo cual, los organismos termófilos, mesófilos y sicrófilos, tendrían distintos niveles de GC asintóticos alcanzados en el equilibrio. El contenido de GC inicial propio de los organismos ancestrales a partir de los cuales se inicia un proceso evolutivo, no afectan el valor del contenido de GC de equilibrio, en los casos en que se este frente a un proceso de diversificación (cladogénesis). En cambio, durante un proceso de envejecimiento (anagénesis), el valor del contenido de GC de equilibrio, si depende del contenido de GC inicial.

4.2.1. Modificación del contenido %GC durante la diversificación

A lo largo de las simulaciones de eventos de diversificación (evolución secuencial), en ausencia de una presión selectiva, se manifiesta una propensión a aumentar el %GC hasta estabilizarse. Se identificó una progresión similar en la trayectoria de los Kinetoplastidia y del Plasmodium. También se observó en la evolución de las superfamilias KV3 y KV4 de los canales iónicos de potasio, mientras que en el gen de proteasa del HIV no se observó dicho comportamiento. Esto sugiere que tras un periodo evolutivo específico, no se detectarían modificaciones en el contenido de guanina-citosina. No obstante, esto no implica un cese evolutivo, pues mantener un %GC estable no restringe la emergencia de mutaciones puntuales. Estas mutaciones se circunscribirían únicamente a reorganizar nucleótidos en la cadena de ADN, preservando un %GC invariable. Esto implica que si hubiera una fase estabilizada en la diversificación, la capacidad de cambio genético, relacionada con las mutaciones puntuales, sería considerablemente menor que en etapas evolutivas iniciales, donde el genoma tendría mayor adaptabilidad. Aunque la probabilidad de conformar un par GC supera a la de un par AT, el contenido de GC no puede maximizarse hasta el 100%. Lo que acontece es la llegada a un punto de equilibrio. El nivel de GC logrado en la evolución del ADN no codificante es superior al alcanzado en la evolución del ADN codificante, esto a causa de las limitaciones impuestas sobre mutaciones para garantizar la viabilidad del ADN codificante. Dichas limitaciones emergen de manera intrínseca debido a la presión selectiva, especialmente en situaciones donde la

presión selectiva es inexistente. La estabilización del GC y sus particularidades están influenciadas tanto por aspectos termodinámicos como por un fenómeno de codificación. Mientras que las fuerzas termodinámicas propenden al aumento del contenido de guanina-citosina tendiendo a estabilizarse, es el fenómeno de codificación, basado en si la secuencia es codificante o no codificante, el que establece el nivel de estabilización. La proyección realizada revela similitudes con las evidencias experimentales, ya que *T.brucei* presenta un %GC aproximado al 44%. Se observa una correspondencia entre ambas proyecciones, aunque es inviable establecer una analogía precisa, ya que la temporalidad de estas evidencias es incierta, debido a que los organismos analizados no han sido fechados adecuadamente. No obstante, esta semejanza confirma la coherencia del modelo propuesto. Al examinar el sesgo de codón en los Kinetoplastidia, se hallan indicios de una influencia termodinámica, pues se manifiesta una inclinación a utilizar codones con mayor contenido GC para codificar distintos aminoácidos. Se evidencia un comportamiento análogo en el sesgo de codón humano, el cual evidencia una predilección por codones con alto contenido GC. Esto señala que el sesgo de codón refleja la influencia termodinámica en la evolución biológica. Consideramos viable determinar una secuencia temporal o al menos distancias temporales relativas en la línea evolutiva de los Kinetoplastidia. Al buscar una coincidencia entre proyecciones, es posible estimar temporalidades evolutivas de organismos basándose en datos temporales existentes. A este respecto, existen métodos para confirmar esta hipótesis, como el uso del algoritmo DNAmlk del software Phylip. Sin embargo, se debe abordar con precaución, ya que dicho método presupone un ritmo molecular homogéneo en todos los organismos, lo cual es difícil de confirmar, y la secuencia del ADN utilizado puede influir en los resultados. En la simulación presentada, se observa el proceso de envejecimiento y diversificación concurrentemente. Las curvas de envejecimiento no se intersectan, pero sí convergen cerca del punto de estabilización. Esto garantiza que, independientemente de la duración del envejecimiento, siempre se nota una variabilidad en el contenido de GC de los organismos que persisten, lo que puede ser útil para entender mejor las relaciones filogenéticas. Al evaluar el %GC de una especie

que ha perdurado hasta nuestros días, realmente se está midiendo el valor presente y no el que tenía al surgir hace eones. Las especies experimentan un proceso de envejecimiento post-emergencia, lo cual determina su supervivencia ante variaciones ambientales. Esto facilita la posibilidad de múltiples eventos de diversificación desde un ancestro común en diferentes épocas. Este fenómeno puede ilustrarse mediante una proyección basada en nuestro modelo. Imaginemos una especie que se diversifica generando a la especie A, la cual se diversifica y da origen a la especie B. Ambas, A y B, experimentan envejecimiento y A origina su versión envejecida A´. Es plausible que, al igual que A originó a B, A´ pueda diversificarse y originar a una especie C.

Esta consideración es intrigante, ya que las rutas de diversificación se multiplican, implicando no solo al organismo en cuestión sino también su posición temporal. Postulamos que este es el proceso que predomina en la evolución biológica, equivalente a lo que se observa en árboles filogenéticos como una bifurcación, similar a lo que sucede con Critidia y Leptomona. No obstante, confirmar estos eventos con base en datos genómicos es un desafío considerable.

4.2.2. Distribución de Mutaciones en el Codón

Según el modelo propuesto, las mutaciones de un solo punto dentro de la región codificante tienden a localizarse principalmente en la tercera posición del codón. Tal fenómeno en el diseño evolutivo concuerda plenamente con los datos experimentales previamente obtenidos. Esto respalda y legitima las premisas adoptadas.

Las mutaciones de un solo punto y la acumulación de bases GC se concentran primordialmente en la tercera posición del codón, esto debido a que el código genético muestra su degeneración predominante en este lugar. Por ende, para mantener el aminoácido o su grupo, y garantizar la viabilidad de la mutación, es imprescindible que ésta se manifieste principalmente en la tercera posición del codón.

A través del estudio de secuencias de ADN, hemos confirmado que en la evolución de los Kinetoplastidia, desde el *T.brucei* hasta *Crithidia*, la gran mayoría de las proteínas

mantienen el tipo de aminoácido, y en menor medida, el aminoácido per se. Para este propósito, se empleó un software desarrollado en Turbo Pascal que permite ejecutar un análisis matricial del tipo “Harr Plot” entre dos secuencias de ADN relacionadas, confirmando lo previamente expuesto.

4.2.3. Influencia de la Secuencia Inicial

Diversos factores de la evolución del ADN codificante y el ADN no-codificante denominado "basura" parecen estar vinculados con las características inherentes a la secuencia de ancestros primitivos.

En relación a la evolución del ADN codificante, al comenzar la simulación con secuencias con diferentes proporciones de GC, se perciben variaciones en los porcentajes de saturación. Esto puede atribuirse a dos razones: el contenido intrínseco de GC y la prevalencia de aminoácidos ya sea degenerados o no en la secuencia. Dado que en una simulación de diversificación, el modelo busca mantener el aminoácido o su grupo, resulta más sencillo admitir mutaciones si la proteína es predominantemente de aminoácidos degenerados. Lo opuesto sucede si la proteína es mayoritariamente de aminoácidos no degenerados. En cuanto al contenido inicial de GC, postulamos que su impacto en la diversificación es limitado. Repitiendo diversas simulaciones, observamos que los porcentajes de saturación tienden a situarse entre un rango del 60% al 65%.

Con respecto al ADN no-codificante "basura", la situación es más directa. Sin restricciones mutacionales y con solo la presión mutacional termodinámica actuando, se espera un valor de saturación consistente en todos los escenarios, siempre y cuando los nucleótidos estén distribuidos de manera equitativa y equimolar.

4.2.4. Implicaciones del Mecanismo de Reparación

El proceso de reparación del ADN es vital para neutralizar los efectos de agentes mutagénicos y desviaciones durante la replicación. Sin embargo, aún se desconoce en gran medida la cinética de corrección de errores asociada a este mecanismo. En este

estudio, hemos estimado el impacto potencial que podría tener esta cinética en ciertos aspectos de la diversificación.

Mediante simulaciones, se ha evidenciado que el mecanismo de reparación del ADN juega un papel crucial en la tasa de variación del contenido de GC. A través de la simulación de procesos de diversificación de una secuencia aleatoria con dos cinéticas de reparación diferentes, se ha observado que, aunque el valor de saturación de GC se mantiene relativamente constante, lo que varía es la velocidad a la que cambia el contenido de GC. Dicho valor de saturación está determinado por la distribución de Boltzmann y por las energías de interacción entre las bases, así como la concentración de nucleótidos disponibles. En consecuencia, un mecanismo de reparación puede acelerar o retardar la tendencia al incremento del contenido de GC, modificando el trayecto pero no el resultado final.

4.2.5. El desafío del ADN no codificante "basura"

En relación con la evolución del ADN no codificante, denominado "basura", el modelo sugiere un aumento progresivo del porcentaje de GC, hasta estabilizarse en un nivel alto. Este nivel de estabilización del GC, que alcanza un 96%, se logra más rápidamente y es superior al observado en la diversificación del ADN codificante (Figura 15). Esto se atribuye a que, en la evolución del ADN no codificante, las mutaciones no enfrentan restricciones, considerándose todas ellas viables.

Dado que la evolución del ADN no codificante está regida exclusivamente por factores termodinámicos, este proceso debería ser similar en organismos que han evolucionado a la misma temperatura, dado que la distribución de Boltzmann está directamente influenciada por este factor. Esto lleva a la hipótesis de un reloj molecular universal para todos los seres vivos, dependiente únicamente de la temperatura ambiental. Teóricamente, medir el promedio de guanina-citosina en la región no codificante podría ofrecer una estimación de la antigüedad del organismo. Sin embargo, esta idea no es completamente precisa, ya que, a pesar de la falta de presión selectiva, hay que considerar la influencia de

agentes mutagénicos exógenos en el medio ambiente, como la radiación ionizante, que puede inducir mutaciones adicionales en el ADN no codificante. Así, una especie expuesta a altos niveles de radiación acumularía mutaciones a un ritmo diferente al de especies en ambientes menos hostiles.

El concepto de ADN no codificante "basura" sigue siendo un tema de debate. La ausencia de funciones conocidas en ciertas regiones del ADN no permite afirmar con certeza que estas sean efectivamente "basura".

4.2.6. Efecto de la temperatura

La temperatura sin duda debe tener un efecto importante en la evolución, ya que este parámetro está incluido en la distribución de Boltzmann. Al realizar simulaciones de diversificación para una misma secuencia inicial a 37°C y 97°C (Figura 14), se nota claramente que a 97°C la tasa de mutaciones es mayor. Sin embargo, ambas curvas llegan al mismo valor de saturación cercano a 61%.

Esta aceleración de la velocidad de mutación con la temperatura se explica por su relación con la distribución de Boltzmann. Analizando la probabilidad de que un espacio vacante en la cadena de ADN sea ocupado por una adenina, se encuentra que la relación entre las probabilidades de que cualquier par de bases ocupe un espacio tiende a igualarse a medida que la temperatura aumenta.

$$P'(X=A) = \frac{[A]}{[G]} e^{-\beta \left(E(X=A)-E(X=G)\right)}$$

se prueba que la relación entre las probabilidades que el hueco sea ocupado por cualquier par de bases, tiende a 1 conforme la temperatura aumenta. Es decir

$$\frac{P'(X=M)}{P'(X=N)} \text{ ----------> } 1 \text{ , cuando T aumenta.}$$

En altas temperaturas, la probabilidad de incorporación de una base incorrecta se equipara a la de una base correcta, resultando en un incremento notable de las mutaciones puntuales.

Al observar la diversificación a diferentes temperaturas, se ve que a temperaturas más altas (T1) la velocidad inicial de mutación es mayor, incluso superando el contenido de GC de una secuencia inicial con mayor contenido de GC, pero a una temperatura más baja (T2).

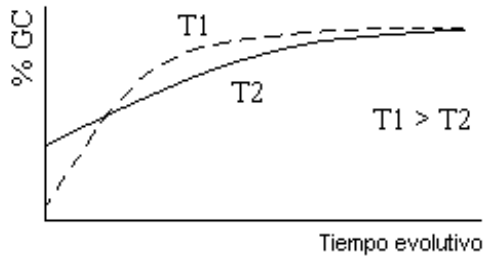


Figura 14 Efecto de la temperatura en la diversificación

Además, las altas temperaturas aumentan la frecuencia e intensidad de los choques térmicos, contribuyendo al incremento en la tasa de mutación.

Esto podría explicar por qué ciertos organismos termófilos, como las arqueobacterias, tienen un alto contenido de GC, a pesar de ser un linaje evolutivamente antiguo. Sin un análisis exhaustivo, las arqueobacterias, con un ancestro más antiguo que el de T. brucei, tienen un porcentaje de GC mayor, cercano al de *Leishmania,* una especie más reciente evolutivamente hablando.

4.2.7. Tendencia natural hacia el incremento del contenido de GC

Los resultados de nuestro estudio previo sobre la simulación computacional que busca modelar un proceso de evolución in-silico, se describe en detalle en los apéndices 1 y 2. Consideramos importante mencionar en este espacio una descripción detallada de dichos

resultados, en beneficio de desarrollar un mejor contraste y discusión. De acuerdo a los resultados de la simulación computacional, el valor de $\mu_{GC}(t)$ y su valor en el equilibrio luego de aproximadamente 10 x10^6 huecos evaluados (equivalente a 100,000 generaciones), el sistema tiende a estabilizarse hacia un valor en el equilibrio de $[GC]_{eq} = 0.66$. Esto fue corroborado en nuestro estudio previo) donde se utiliza el método de Monte Carlo basándose en la distribución canónica de Boltzmann como medida de probabilidad. En esta simulación se consideran 2 escenarios: en primer lugar secuencias no codantes/no funcionales de ADN y luego secuencias codantes donde se impone conservación de la familia e identidad de aminoácidos. Dicha simulación se basa en leer una hebra inicial de ADN para luego reconocer la vecindad inicial e ir llenando la siguiente hebra de nucleótidos (del tipo Y) donde la probabilidad de ingreso de cierto nucleótido se calcula con la distribución de Boltzmann y un algoritmo de Monte Carlo determina el resultado. Entonces lo que en cada paso se realiza es generar un agujero o sitio donde ingresan los nuevos nucleótidos y debido a esto hay una variación en la concentración de los mismos. En la simulación se considera una hebra de 1000bp y generan 20$(10)^6$ sitios (de esta forma se generan aproximadamente 20$(10)^3$ generaciones, en un proceso evolutivo simulado. El programa permite imponer restricciones a las mutaciones, en el caso de ADN no codante se admiten todas las posibles mutaciones y en el caso del ADN codante, se consideran 2 tipos de mutaciones: neutrales (conservan la identidad de aminoácido) que contribuyen al proceso de envejecimiento o anagénesis, y las que conservan la familia de aminoácidos y que contribuyen al proceso de diversificación o cladogéndesis. Mostramos a continuación las gráficas de la simulación de la variación del contenido de GC respecto al número de pasos en la simulación, donde ya se observa la convergencia hacia $[GC]|_{eq}$ para 10 $(10)^6$ pasos.

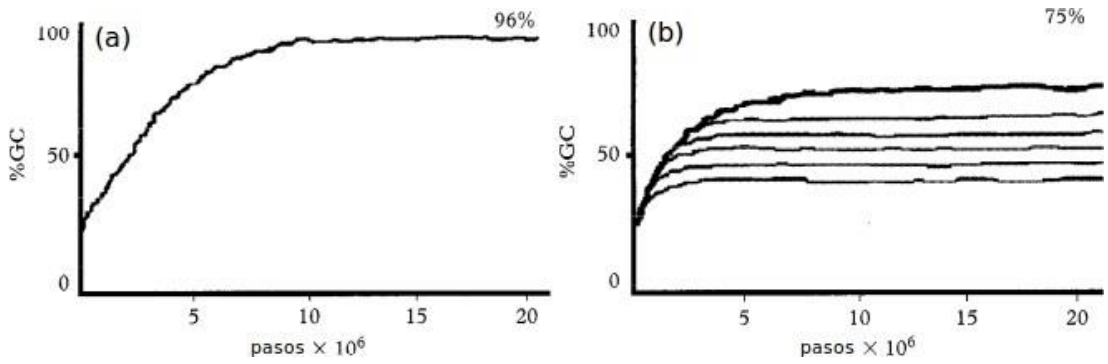


Figura 15. Variación de [GC] respecto al número de pasos usados en la simulación. (a) Sin restricción en mutaciones puntuales. (b) Restricciones en la conservación de la familia de aminoácidos (líena gruesa) y en la identidad de aminoácido (líneas delgadas). Además se considera [GC]|0 = 22%.

Esta Figura de la simulación computacional es bastante análoga a la figura correspondiente basada en el tratamiento analítico dado en líneas más arriba. Hay que observar adicionalmente que el valor en equilibrio de GC cuando no se consideran ligaduras es de aproximadamente 95% lo cual resulta alejado del valor teórico 0.66, lo que podría argumentarse por la formación de fracciones de ADN basura durante el proceso computacional. Esto resulta curioso considerando que en el tratamiento analítico no se imponen restricciones a priori y solo se manejan tasas de transición. En la parte (b) se puede observar que el porcentaje en equilibrio resulta ser menor lo cual se acerca más al valor teórico, podría suponerse también que el mecanismo analítico utilizado basado en tasas de transicion ya introduce de forma implícita las ligaduras.

Observemos además que el resultado analítico se obtiene asumiendo que no existe interacción entre los nucleótidos X e $\bar{Y}$ y la vecindad de referencia, lo que estaría por verse entonces es si al introducir alguna interacción se obtiene información adicional como por ejemplo si de esta forma se podría probar que $[A] = [T]$ y $[G] = [C]$ dentro de cada hebra de ADN como se postula en algunos textos. En el trabajo previo [Zimic M. et. Al., 2003], se evalúa la validez del postulado de la presión mutacional termodinámica escogiendo una situación en la cual los organismos son expuestos a una presión de selección limitada y

constante. Análisis de secuencias de genes de trypanosoma revelan este sesgo en la concentración de nucleótidos en los codones. El alto contenido de GC respecto de AT en las trypanosomas modernas es un reflejo de la evolución de la mayoría de sus genes hacia un contenido mayor de GC, hipótesis que se sustenta por la presencia de la presión termodinámica. En dicho estudio previo fueron analizados más de 20 genes diferentes que se han sido reportados para 2 o más especies de tripanosomatidas, ninguno de ellos presenta un sesgo hacia la formación de AT, se observa en un diagrama 4 genes diferentes que muestran que las especies más antiguas tienen menor contenido de GC en sus genomas.

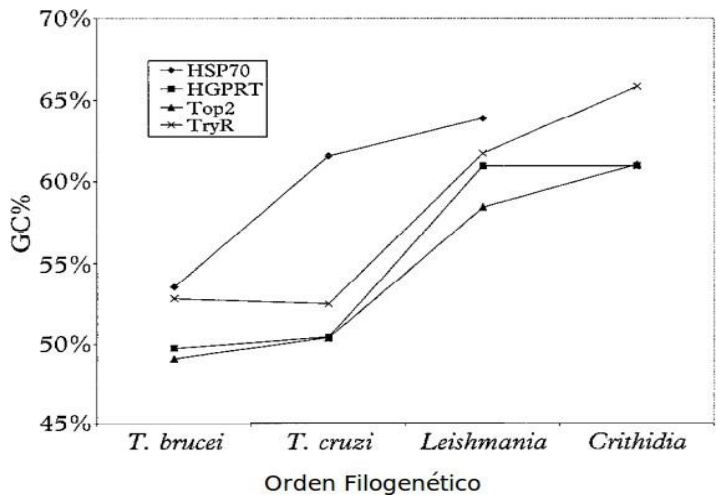


Figura 16. Variación en el contenido de GC de genes específicos. Genes (HSP70,HGPRT,Top2,TryR) de trypanosomatidas. El orden horizontal sigue un orden filogenético específico. La especie más antigua (*T. brucei)* tiene menor contenido de GC en sus genes comparándola con las más modernas (*Crithidia* y *Leishmania*).

El modelo de simulación presentado implica que el ADN en evolución busca alcanzar un equilibrio termodinámico, asociado a llegar a un contenido de GC que converge asintoticamente en el tiempo, siendo la presión mutacional termodinámica el factor que impulsa moléculas ancestrales con mayor concentración de AT hacia estados de máximo contenido de GC. En el ADN no se dan entonces mutaciones completamente aleatorias sino que existe cierta fuerza de deriva que desplaza el valor medio final por encima de 0.5. Las hipótesis presentadas no contradicen la presión mutacional direccional [Sueoka,1988] ni la presión de

selección [Bernardi, 1988], pudiendo tratarse esta última como una ligadura al sistema. Ha sido interesante el poder verificar esta tendencia que se observa exper- imentalmente considerando la distribución de Boltzmann para calcular las tasas de mutaciones y a partir de allí proponer una ecuación maestra para la concentración de nucleótidos.

Conclusiones

1. El análisis teórico físico-matemático, por el cual se estima una solución asintótica cuando el tiempo tiende al infinito, para el contenido de Guanina-Citosina (GC), muestra que se alcanza un valor asintótico en equilibrio, luego de un incremento monotónico a lo largo del tiempo evolutivo.
2. Las evidencias experimentales correspondientes a secuencias que se han sometido a un proceso de evolución in-vitro, avalan la presión mutacional termodinámica que conduce a un aumento en el contenido de guanina-citosina (GC) en el ADN a lo largo de la evolución. De este modo, la evolución se puede entender como una interacción entre la presión de selección y la presión mutacional termodinámica.
3. Con respecto al ADN no codificante, bajo nuestras hipótesis, evoluciona principalmente bajo la influencia de la presión mutacional termodinámica. Como resultado, el contenido GC aumenta, alcanzando un nivel estacionario que es superior al observado en la diversificación.
4. La temperatura influye directamente en las probabilidades de mutación. A temperaturas más elevadas, el ritmo de aumento del contenido GC es más pronunciado. Este fenómeno parece haber sido observado en las archaeobacterias.
5. En el marco evolutivo, es posible que dos fuerzas principales orientan el proceso: la presión de selección y la presión mutacional termodinámica. En general, la primera puede favorecer mutaciones que no son necesariamente las más estables desde un punto de vista termodinámico.
6. La presencia de un mecanismo no neutral de corrección de errores puede modificar la velocidad de aumento del contenido GC durante la diversificación, conduciendo eventualmente a un equilibrio de contenido GC.

Recomendaciones

1. Al utilizar la distribución de Boltzmann para un conjunto canónico, solamente se han tenido en cuenta términos entálpicos, omitiendo aspectos entrópicos como el efecto hidrofóbico. Si bien emplear la distribución de Boltzmann ofrece la ventaja de considerar únicamente términos entálpicos, surge el desafío de determinar el número exacto de estados posibles, que en determinadas situaciones es sencillo estimar. Un enfoque más exacto para definir el número de estados posibles, cuando la energía se encuentra cerca de un valor determinado, debería considerar efectos entrópicos. Hemos propuesto una primera aproximación basado en la proporcionalidad con la concentración de nucleótidos trifosfato libres, pero sugerimos una evaluación más detallada de este aspecto o el uso de la función de energía libre de Gibbs.
2. Al recurrir a la distribución de Boltzmann para un conjunto canónico, suponemos que el sistema en estudio (ADN) está en equilibrio con su entorno térmico (el resto de la célula). Aunque esta premisa puede no ser estrictamente cierta, debido a que las mutaciones pueden ocurrir a una velocidad tal que el tiempo de relajación no es adecuado, esta aproximación ha demostrado ser efectiva en otros campos, como en el plegamiento de proteínas. No obstante, una revisión más precisa del problema implicaría emplear técnicas de la termodinámica fuera del equilibrio. Consideramos que sería valioso investigar en esta dirección, para entender la aplicabilidad de las aproximaciones de equilibrio en tales escenarios.

Referencias Bibliográficas

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). "A Learning Algorithm for Boltzmann Machines". Cognitive Science, 9(1), 147-169.

Adami, C. (2004). "Information Theory in Molecular Biology". Physics of Life Reviews, 1(1), 3-22.

Alonso Guillermina (1992). Trypanosomatidae Codon Usage and GC Distribution. Mem Inst. Oswaldo Cruz, Rio de Janeiro, Vol.87 , (4):517-523.

Ayala F. J., Kiger J. (1984) Genética Moderna. Editorial Acribia – Omega D.L.

Bell, S.P., Dutta, A. (2002). "DNA Replication in Eukaryotic Cells". Annu. Rev. Biochem. 71:333–374.

Bernardi G, Mouchiroud D, Gautier C, Bernardi G. Compositional patterns in vertebrate genomes: conservation and change in evolution. J Mol Evol. 1988 Dec-1989 Feb;28(1-2):7-18. doi: 10.1007/BF02143493. PMID: 3148744.

Bialek, W. (2012). "Biophysics: Searching for Principles". Princeton University Press.

Cantor, C. R., & Schimmel, P. R. (1980). "Biophysical Chemistry". W. H. Freeman and Company.

Chargaff and Davidson (1955) The Nucleic Acids, Academic Press

Darnell, Lodish, Molecular Baltimore, (1990) Cell Biology 2nd edition Scientific American Books

Darwin, Charles, and Leonard Kebler. On the origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life . London: J. Murray, 1859.

Dill, K. A., & Bromberg, S. (2010). "Molecular Driving Forces: Statistical Thermodynamics in Biology, Chemistry, Physics, and Nanoscience". Garland Science.

Dill, K. A., Ozkan, S. B., Shell, M. S., & Weikl, T. R. (2008). "The Protein Folding Problem". Annual Review of Biophysics, 37, 289-316.

Dror, R. O., Dirks, R. M., Grossman, J. P., Xu, H., & Shaw, D. E. (2012). "Biomolecular Simulation: A Computational Microscope for Molecular Biology". Annual Review of Biophysics, 41, 429-452.

Futuyma, Douglas J., 1942-, Evolution. Sunderland, Massachusetts U.S.A, Sinauer Associates, Inc. Publishers, 2013.

Gagnon, J.-S., & Hochberg, D. (2023). Conditions for the origin of homochirality in primordial catalytic reaction networks. Scientific Reports, 19 June 2023

Galtier, N., Tourasse, N., & Gouy, M. (1999). A nonhyperthermophilic common ancestor to extant life forms. Science, 283(5399), 220-221. doi: 10.1126/science.283.5399.220.

Gould, S. J. (2002). "The Structure of Evolutionary Theory". Harvard University Press.

Hartl D, and Clark A. (1989) Principles of Population Genetics. 2nd edition. Sinauer Associates

Kimura, M. (1983). The Neutral Theory of Molecular Evolution. Cambridge: Cambridge University Press.

Klumpp, S., & Hwa, T. (2008). "Stochasticity and traffic jams in the transcription of ribosomal RNA: Intriguing role of termination and antitermination". Proceedings of the National Academy of Sciences, 105(42), 18159-18164.

Kornberg, A., Baker, T.A. (1992). "DNA Replication". W.H. Freeman and Company.

Kunkel TA, Loeb LA. On the fidelity of DNA replication. Effect of divalent metal ion activators and deoxyrionucleoside triphosphate pools on in vitro mutagenesis. J Biol Chem. 1979 Jul 10;254(13):5718-25. PMID: 376517.

Kunkel Thomas (1989) On the fidelity of ADN replication: effect of dNTPs pools on in vitro mutagenesis. Proc. Natl. Acad. Sci. Vol 254 13 5718-25.

Lehninger (1971) Biochemistry 4th printing. Worth

Lehninger, Nelson, Cox (1993) Principles of Biochemistry second edition. Worth

Lewin Benjamin (1994) Genes V. Oxford Press

Loeb, L. A. Kunkel, T. A. (1982) Annu. Rev. Biochem. 52, 429-457

Lynch, M. (2007). "The Origins of Genome Architecture". Sinauer Associates.

Maslov D. A. (1994). Evolution of RNA editing in Kinetoplastid protozoa. Nature. Vol.368.

Maslov D. A. (1995) Evolution of Parasitism in Kinetoplastid Protozoa. Parasitology Today vol. II,no. I.

Mayer J. and Mayer M. (1963) Statistical Mechanics. John Wiley and Sons

Mellon Isabel (1996) Transcription - Coupled Repair Deficiency and Mutations in Human

Mismatch Repair Genes. Science. Vol.272. April.

Meselson, M., Stahl, F.W. (1958). "The Replication of DNA in Escherichia Coli". Proc. Natl. Acad. Sci. 44(7):671–82.

Morrison, A., Sugino, A., Kornberg, A. (1991). "The 3' → 5' exonuclease of DNA polymerase I of Saccharomyces cerevisiae". Journal of Biological Chemistry. 266:5616–5620.

Mouchiroud D, Gautier C, Bernardi G. The compositional distribution of coding sequences and DNA molecules in humans and murids. J Mol Evol. 1988;27(4):311-20. doi: 10.1007/BF02101193. PMID: 3146641.

Muzi-Falconi, M., Giannattasio, M. (2019). "Molecular Mechanisms of DNA Replication". Frontiers in Molecular Biosciences. 6:29.

Normile Dennis (1996) Impact of DNA Replication Errors Put to the Test. Science. Vol.272. May.

Petruska J., Goodman, M. (1988) Proc. Natl. Acad. Sci. USA 85, 6252-6256

Petruska J. , Sowers L.C. (1986) Proc. Natl. Acad. Sci. USA 83, 1559-1562

Reif F. (1965) Fundamentos de física estadística y térmica. McGraw - Hill Book Company.

Rieper, E., Anders, J., & Vedral, V. (2011). "Quantum entanglement between the electron clouds of nucleic acids in DNA". The Journal of Physics: Condensed Matter, 23(6).

Sanson GF, Kawashita SY, Brunstein A, Briones MR. Experimental phylogeny of neutrally evolving DNA sequences generated by a bifurcate series of nested polymerase chain reactions. Mol Biol Evol. 2002 Feb;19(2):170-8. doi: 10.1093/oxfordjournals.molbev.a004069. PMID: 11801745.

Schrödinger, E. (1944). "What is Life? The Physical Aspect of the Living Cell". Cambridge University Press.

Sueoka N. Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci U S A. 1988 Apr;85(8):2653-7. doi: 10.1073/pnas.85.8.2653. PMID: 3357886; PMCID: PMC280056.

Suzuki K. (1983) UV- induced imbalance of the deoxyribonucleoside triphosphate pool in Escherichia coli. Mutat Res 122 293-8. Abstrats.

Ter Haar D. (1961) Elements of Statistical Mechanics. Holt, Rinehart and Winston. New York.

WATSON, J. D., CRICK, F. H. C. (1953), A structure for Deoxyribose Nucleic Acid. Nature, 171 (4356): 737–738.

Zimic MJ, Guerra D, Arévalo J. DNA thermodynamic pressure: a potential contributor to genome evolution. Trans R Soc Trop Med Hyg. 2002 Apr;96 Suppl 1:S15-20. doi: 10.1016/s0035-9203(02)90046-5. PMID: 12055830.

Anexos

Anexo 1: Modelo termodinámico basado en la distribución de Boltzmann para estimar las probabilidades de formación de pares no canónicos (missmatches) 1
Anexo 2: Simulación computacional mediante un algoritmo de Monte Carlo para modelar un proceso evolutivo guiado por la distribución de Boltzmann 22
Anexo 3: Evidencia experimental que permite contrastar las predicciones del modelo de simulación 36
Anexo 4: Publicación científica -DNA thermodynamic pressure: a potential contributor to genome evolution. 44

Los anexos 1,2, y 3 corresponden al estudio previo relacionado a las simulaciones computaciones de los procesos evolutivos, bajo el uso de la presión mutacional termodinámica y la distribución de Boltzmann en una simulación tipo Monte Carlo. Dichos estudios concluyeron en la tesis de Maestría en Bioquímica por la Universidad Peruana Cayetano Heredia, del autor de la presente tesis. Dichos resultados fueron publicados en un artículo científico, mostrado en el anexo 4.

Anexo 1: Modelo termodinámico basado en la distribución de Boltzmann para estimar las probabilidades de formación de pares no canónicos (missmatches)

Cálculo de la energía electrostática de una vecindad de cinco nucleótidos en la molécula de ADN

Las interacciones que estabilizan la doble cadena del ADN son fundamentalmente enlaces covalentes e interacciones débiles tales como puentes de hidrógeno, interacciones electrostáticas, interacciones de Van der Waals, interacciones magnéticas e interacciones hidrofóbicas, y adicionalmente los iones presentes en la solución, ejercen efectos.

Las interacciones entre las moléculas de agua y moléculas apolares son considerablemente más débiles que las interacciones entre las mismas moléculas de agua (puentes de hidrógeno). La introducción de moléculas apolares en agua, lleva a la destrucción de puentes de hidrógeno, por lo que el agua trata de "expulsar" a estas moléculas y poder restablecer los puentes de hidrógeno. Usualmente se dice que estas moléculas son expulsadas por el efecto de las "fuerzas hidrofóbicas", sin embargo las "fuerzas hidrofóbicas" no son en realidad un tipo de interacción molecular, son simplemente un fenómeno de cooperatividad debido a los puentes de hidrógeno entre las moléculas de agua y a una relativamente débil afinidad entre estas moléculas apolares mediante interacciones electrostáticas y de Van der Waals. Es por esta razón que se sugirió cambiar el término "interacción hidrofóbica" por el de "hidratación hidrofóbica".

El efecto hidrofóbico en sí aparece como una consecuencia entrópica por parte del agua, debido a un reacomodo de su estructura molecular y no por un efecto entálpico, es decir el efecto hidrofóbico contribuye con la energía libre a nivel de un cambio entrópico [Nemethy & Scheraga,1961; Volkenshtein,1985; Davidov,1982].

Como se indica en el capítulo 2, la probabilidad relativa de que un hueco sea ocupado por un nucleótido puede considerarse independiente del potencial químico y del volumen del sistema. Usando el mismo ejemplo del capítulo 2, se tiene que la probabilidad de que el hueco sea ocupado por una adenina resulta ser igual a la siguiente expresión:

$$P'(X=A)=\frac{[A]}{[G]}e^{-\beta\left(E(X=A)-E(X=G)\right)}$$

Es decir, la probabilidad de ocupación depende de la diferencia de energía total entre dos configuraciones. La energía total de una configuración tiene varias contribuciones, siendo las más importantes, los enlaces covalentes, las interacciones electrostáticas, de Van der Waals, magnéticas, los puentes de hidrógeno, las interacciones hidrofóbicas y la energía cinética.

Como se ha discutido, el hecho de que la probabilidad dependa de la diferencia de energía total simplifica el problema, ya que las energías independientes del tipo de nucleótido (enlaces covalentes y energía cinética), se cancelan. Es decir, es suficiente incluir las contribuciones de las interacciones de puente de hidrógeno, electrostáticas, Van der Waals, magnéticas e hidrofóbicas.

Para el cálculo de las probabilidades se empleará la distribución de conjunto canónico en función sólo de términos entálpicos, por lo cual el efecto hidrofóbico no será incluido [Reif,1965; Davidov,1982].

Debido a que se asume una configuración rígida para la molécula de ADN, las interacciones de Van der Waals y las magnéticas (ver apéndice 2), son mucho menos

importantes que las electrostáticas. Por lo tanto sólo se considerarán las interacciones de puente de hidrógeno y electrostáticas.

A continuación describiremos la metodología que hemos desarrollado para el cálculo de estas energías. La energía de puente de hidrógeno se calcula mediante una aproximación clásica, la cual muestra una buena correlación con los datos experimentales, mientras que la contribución electrostática se calcula a partir de una distribución de carga eléctrica de cada nucleótido de la molécula de ADN inmersa en un medio dieléctrico (agua).

Se demuestra más adelante, que no es necesario calcular la energía total del ADN, sino únicamente la energía de una *vecindad alrededor del hueco*, por lo cual debemos conocer únicamente la distribución de carga para la vecindad del hueco en el ADN.

La energía electrostática, de una distribución continua de carga □(r), inmersa en un medio de permisividad eléctrica ε se calcula por medio de la expresión siguiente.

$$\frac{1}{4\pi\varepsilon}\int\frac{\rho(\bar{r}_1)\rho(\bar{r}_2)dv_1dv_2}{|\bar{r}_1-\bar{r}_2|}$$

En nuestro caso, el medio dieléctrico es una mezcla de agua, trazas de etanol y sales, y tiene un valor cercano a 80, de acuerdo a lo calculado en el apéndice 5.

Para el cálculo de la energía electrostática de una vecindad de nucleótidos cualesquiera, se siguen los siguientes pasos :

1.- Se calcula las coordenadas espaciales de todos los átomos en cada uno de los cuatro nucleótidos, a partir de datos geométrico-cristalográficos [Chargaff & Davidson,1955].
2.- Se determina los momentos dipolares eléctricos de todos los enlaces covalentes [Durrant, 1972; Gilman, 1958; Weast, 1978] de cada nucleótido, y se calcula las coordenadas espaciales de cada uno de ellos.
3.- Se calcula la distribución de carga eléctrica de la vecindad, incorporando las distribuciones parciales de carga de todos los nucleótidos. Esto se consigue a partir de la información acerca de los valores de los momentos dipolares y las longitudes de enlace, asumiendo el modelo simple de un dipolo de dos cargas puntuales. De esta manera, la distribución de carga calculada, resulta ser discreta.
4.- Se calcula la energía de interacción coulombiana , entre todos los nucleótidos de la vecindad anterior, considerando un medio de constante dieléctrica igual a la del agua.

Las figuras A1.1, A1.2 y A1.3 muestran el arreglo geométrico de las bases nitrogenadas, la pentosa y el grupo fosfato respectivamente, a partir del cual, calculamos las coordenadas espaciales de cada uno de los átomos. Las flechas indican la dirección de los dipolos eléctricos generados por los enlaces covalentes [Chargaff & Davidson,1955; Jordan,1955]

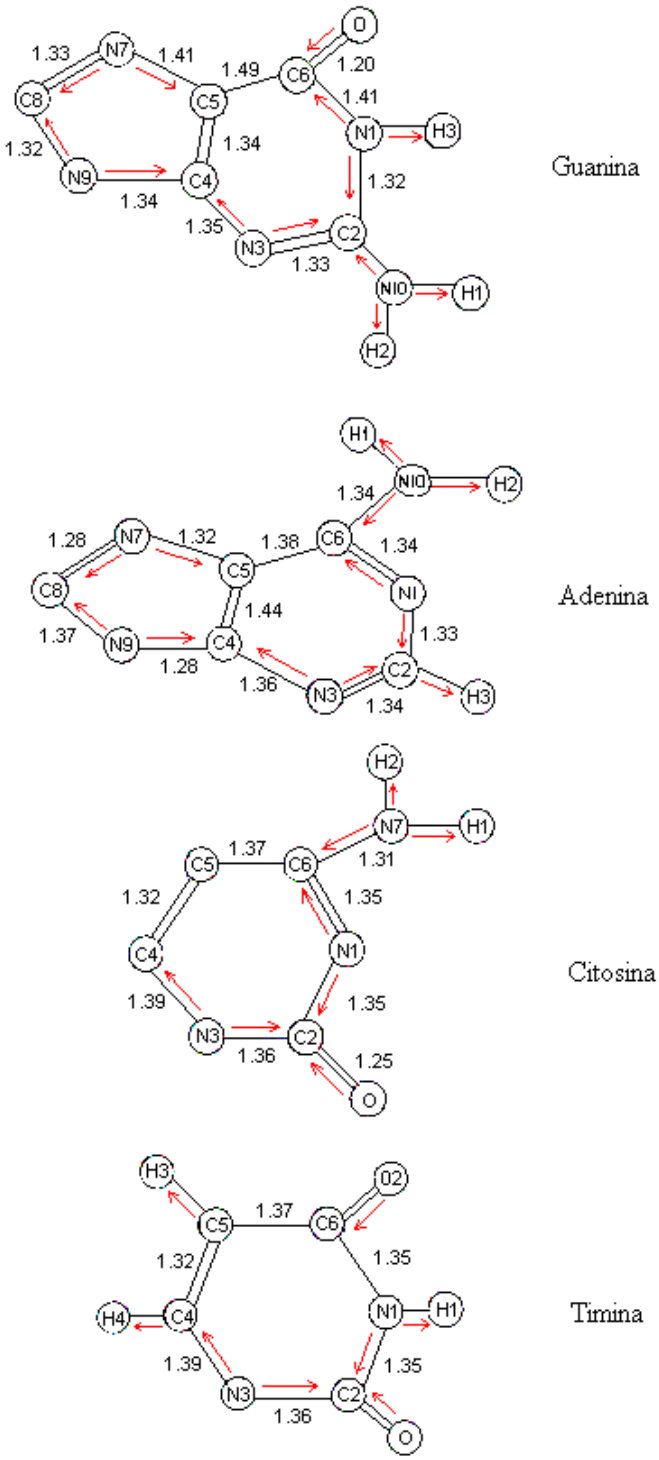


Figura A1.1 Longitudes de enlace, ángulos de enlace y momentos dipolares en las bases nitrogenadas

Algunas longitudes de enlace y momentos dipolares eléctricos utilizados, se muestran en la tabla A1.1.

Enlace	Longitud (Angstroms)	Momento dipolar (10^{-30} C.m)
C-N	1.47	0.73
N-H	1.01	4.44
C-H	1.09	1.3
C=O	1.20	7.7
C=N	-	3.0
P=O	1.63	9.0
P-O	1.63	3.0
O-H	1.20	4.98
O-C	-	2.5

Tabla A1.1 Longitud y momento dipolar de algunos enlaces covalentes

De acuerdo a estudios de difracción con rayos X, la base nitrogenada es una estructura aproximadamente plana, y está posicionada casi perpendicularmente al plano de la deoxiribosa. Igualmente en el azúcar, los átomos C4, O, C1 y C2 son coplanares, mientras que el C3 se separa una distancia de 0.5 Å de dicho plano [Chargaff & Davidson,1955] (Figura A1.2).

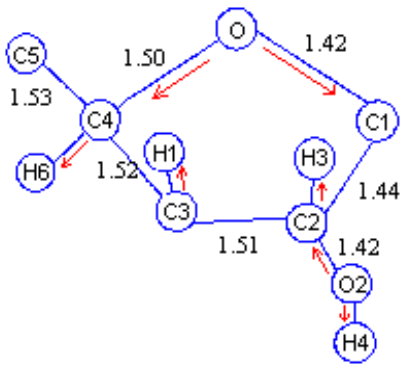


Figura A1.2 Longitudes de enlace y momentos dipolares en la deoxiribosa

La geometría del grupo fosfato resulta ser muy similar a un tetraedro, que en nuestro caso lo hemos considerado regular [Chargaff & Davidson,1955] (Figura A1.3).

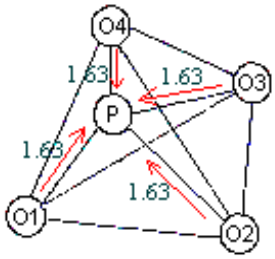


Figura A1.3 Longitudes de enlace y momentos dipolares en el grupo fosfato

Coordenadas espaciales de los átomos de cada nucleótido:

A partir de la información geométrica anterior, se calcularon las coordenadas de los átomos con respecto a un sistema de referencia cuyo origen es la proyección del átomo C3 del azúcar sobre el plano formado por C4 , O, C1 y C2 también en el azúcar. A este origen lo denominamos punto 'P' .

La línea que une P con C2 define la dirección del eje X (+), y la línea que une P con C3 la dirección del eje Z (-). El eje Y se define automáticamente exigiendo que el sistema de coordenadas sea ortogonal y dextrógiro.

La unión de la base nitrogenada con el azúcar se da a través del enlace entre C1 del azúcar y N9 de la base nitrogenada si es guanina o adenina, o con N3 de la base nitrogenada (BN), en caso de timina o citosina. La dirección de este enlace es paralela al eje Z, mientras que el enlace de N3 con C2 de las BN es paralelo al eje X. La unión del grupo fosfato con el azúcar se da a través del enlace entre C5 del azúcar y el O2 del fosfato, cuyo vector de unión es (-0.889, -0.501, 1.001)Å en el sistema de coordenadas definido anteriormente. Nótese que el plano formado por O1, O2 y O3 del fosfato, es paralelo al plano XY.

Con esta información, se calcularon las coordenadas de cada uno de los átomos en cada nucleótido (tablas A1.2, A1.3, A1.4, y A1.5): (las unidades están en Å)

Citosina

	COORDENADAS			
ATOMO	X	Y	Z	
C3	0	0	-0.5	* AZUCAR
C2	1.4248	0	0	
C1	1.9471	1.3418	0	
O	0.8988	2.2973	0	
C4	-0.4305	1.8001	0	
C5	-1.5123	1.8001	1.0818	
H8	-1.05981	0.970789	-0.82931	
H5	2.57841	0.71228	-0.829312	
H1	0	-0.770748	0.270748	
H3	1.4248	-0.770748	0.770748	
O2	1.4248	-1.01118	-1.01118	
H4	1.4248	-1.0118	-1.97118	
O4	-3.73255	1.88799	4.25534	* FOSFATO
P	-3.73255	1.88799	2.6253	
O3	-3.73255	3.40477	2.082	
O1	-5.0634	1.0998	2.082	
O2	-2.4017	1.0998	21.082	
N3	1.9471	1.3418	0.72	* BASE NIT.
C2	3.3071	1.3418	0.72	
N1	4.0424	1.3418	1.8522	
C8	3.3471	1.3418	3.0094	
C5	1.9773	1.3418	3.0333	
C4	1.3173	1.3418	1.651	
N7	4.38855	1.3418	4.04885	
H1	5.99855	1.3418	4.04885	
H2	4.04111	1.3418	4.99794	
H3	1.20855	1.3418	3.80405	
H4	0.2273	1.3418	1.8901	
O	4.15583	1.3418	-0.19853	

Tabla A1.2 Coordenadas atómicas de la citosina

Timina

ATOMO	X	Y	Z	
C3	0	0	-0.5	* AZUCAR
C2	1.4248	0	0	
C1	1.9471	1.3418	0	
O	0.8988	2.2973	0	
C4	-0.4305	1.8001	0	
C5	-1.5123	1.8001	1.0818	
H8	-1.05981	0.97078	-0.82931	
H5	2.57641	0.71226	-0.829312	
H1	0	-0.77074	0.270748	
H3	1.4248	-0.77074	0.770748	
O2	1.4248	-1.01118	-1.01118	
H4	1.4248	-1.0118	-1.97118	
O4	-3.7325	1.88799	4.25534	* FOSFATO
P	-3.7325	1.88799	2.8253	
O3	-3.7325	3.40477	2.082	
O1	-5.0834	1.0998	2.082	
O2	-2.4017	1.0998	2.082	
N3	1.9471	1.3418	1.47	* BASE NIT.
C2	3.3071	1.3418	1.47	
N1	4.0424	1.3418	2.8022	
C8	3.3471	1.3418	3.7594	
C5	1.9773	1.3418	3.7833	
C4	1.3173	1.3418	2.601	
H3	1.20855	1.3418	4.55405	
H4	0.2273	1.3418	2.8401	
O	4.15583	1.3418	0.82147	
H1	5.0524	1.3418	2.8022	
O2	4.19583	1.3418	4.80793	

Tabla A1.3 Coordenadas atómicas de la timina

Adenina

	COORDENADAS			
ATOMO	X	Y	Z	
C3	0	0	-0.5	* AZUCAR
C2	1.4248	0	0	
C1	1.9471	1.3416	0	
O	0.8968	2.2973	0	
C4	-0.4305	1.6001	0	
C5	-1.5123	1.6001	1.0818	
H6	-1.05981	0.970789	-0.62931	
H5	2.57641	0.71228	-0.32931	
H1	0	-0.77075	-0.32931	
H3	1.4248	-0.77075	0.270746	
O2	1.4248	-1.01116	0.770746	
H4	1.4248	-1.0116	-1.01116	
O4	-3.73255	1.86799	-1.97116	
P	-3.73255	1.86799	4.25534	* FOSFATO
O3	-3.73255	3.40477	2.6253	
O1	-5.0634	1.0996	2.082	
O2	-2.4017	1.0996	2.082	
C2	5.10409	1.3416	2.082	
N1	6.8088	1.3416	-0.15273	* BASE NIT.
C6	5.32863	1.3416	0.975171	
C5	3.95209	1.3416	2.22615	
C4	3.18901	1.3416	2.32245	
N3	3.69841	1.3416	1.10117	
N7	3.0861	1.3416	-0.14518	
C8	1.89929	1.3416	3.31862	
N9	1.9471	1.3416	2.83918	
N10	6.36805	1.3416	1.47.3.2656	
H1	6.02261	1.3416	4.21469	
H2	7.37805	1.3416	3.2656	
H3	5.87564	1.3416	-0.92347	

Tabla A1.4 Coordenadas atómicas de la adenina

Guanina

	COORDENADAS			
ATOMO	X	Y	Z	
C3	0	0	-0.5	*AZUCAR
C2	1.4248	0	0	
C1	1.9471	1.3418	0	
O	0.8988	2.2973	0	
C4	-0.4305	1.8001	0	
C5	-1.5123	1.8001	1.0818	
H8	-1.05981	0.970789	-0.82931	
H5	2.57841	0.71228	-0.829312	
H1	0	-0.770748	0.270748	
H3	1.4248	-0.770748	0.770748	
O2	1.4248	-1.01118	-1.01118	
H4	1.4248	-1.0118	-1.07118	
O4	-3.73255	1.88799	4.25534	* FOSFATO
P	-3.73255	1.88799	2.8253	
O3	-3.73255	3.40477	2.082	
O1	-5.0834	1.0998	2.082	
O2	-2.4017	1.0998	2.082	
N1	5.79848	1.3418	0.9233	* BASE NIT.
C8	5.3182	1.3418	2.24828	
C5	3.91707	1.3418	2.1994	
C4	3.20892	1.3418	1.08304	
N3	3.75804	1.3418	-0.17028	
C2	5.08029	1.3418	-0.17103	
N7	3.08814	1.3418	3.28504	
C8	1.85502	1.3418	2.78882	
N9	1.9471	1.3418	1.47	
N10	8.09974	1.3418	-1.21048	
H1	7.10974	1.3418	-1.21048	
H2	5.15085	1.3418	-1.55592	
H3	8.80842	1.3418	0.9233	
O	8.16473	1.3418	3.09881	

Tabla A1.5 Coordenadas atómicas de la guanina

Cada enlace covalente en un nucleótido genera un pequeño dipolo eléctrico con un valor característico (tabla A1.1). Los dipolos no nulos surgen del enlace covalente entre átomos de distinta electronegatividad. En las figuras A1.1, A1.2 y A1.3 se muestran unas flechas que representan los dipolos eléctricos, cuya dirección va desde el átomo más electronegativo hacia el menos electronegativo.

Distribución de los dipolos eléctricos en cada nucleótido:

A partir de las posiciones de los átomos en cada nucleótido, los dipolos eléctricos se disponen en el punto medio de la línea que une los átomos que forman el enlace. Con estas consideraciones se calcula el origen, dirección y módulo de los dipolos eléctricos de cada uno de los nucleótidos, los cuales se muestran en las tablas A1.6, A1.7, A1.8 y A1.9 :

CITOSINA

	POSICIÓN DEL DIPOLO			DIRECCIÓN DEL DIPOLO		
ENLACE	X	Y	Z	PX	PY	PZ
N3-C2	2.6271	1.3416	0.6200	0.7300	0.0000	0.0000
N3-C4	1.6322	1.3416	1.2050	-0.3460	0.0000	0.6428
N1-C2	3.6748	1.3416	1.1861	-0.3976	0.0000	-0.6122
N1=C6	3.6947	1.3416	2.3308	-1.5451	0.0000	2.5715
N7-C8	3.8668	1.3416	3.4291	-0.5162	0.0000	-0.5162
N7-H1	4.8915	1.3416	3.9488	4.4400	0.0000	-0.0002
N7-H2	4.2138	1.3416	4.4234	-1.5185	0.0000	4.1723
C5-H3	1.5919	1.3416	3.3187	-0.9192	0.0000	0.9192
C4-H4	0.7723	1.3416	1.7901	-1.3000	0.0000	0.0000
O=C2	3.7314	1.3416	0.1957	-5.4447	0.0000	5.4447
N9DC1	1.9471	1.3416	0.3100	0.0000	0.0000	-0.7300
AZUCAR						
O-C1	1.4219	1.8193	0.0000	1.8488	-1.6828	0.0000
O-C4	0.2332	1.9487	0.0000	-2.2132	-1.1626	0.0000
C4-H8	-0.7452	1.2854	-0.3147	-0.7506	-0.7505	-0.7506
C1-H5	2.2618	1.0269	-0.3147	0.7506	-0.7506	-0.7506
C3-H1	0.0000	-0.3854	-0.1146	0.0000	-0.9192	0.9192
C2-H3	1.4248	-0.3854	0.3854	0.0000	-0.9193	0.9192
O2-C2	1.4248	-0.5055	-0.5055	0.0000	1.7678	1.7678
O2-H4	1.4248	-1.0112	-1.4912	0.0000	0.0000	-4.9800
FO2DC5	-1.9570	1.3653	1.5819	1.5669	0.8305	-1.7621
FOSFATO						
O4=P	-3.7326	1.8680	3.4403	0.0000	0.0000	-9.0000
O1-P	-4.3980	1.4835	2.3537	2.4495	1.4145	0.9990
O2-P	-3.0671	1.4835	2.3537	-2.4495	1.4145	0.9990
O3-P	-3.7326	2.6364	2.3537	0.0000	-2.8284	0.9990

Tabla A1.6 Posición y dirección de los dipolos en la citosina

TIMINA

	POSICIÓN DEL DIPOLO			DIRECCIÓN DEL DIPOLO		
ENLACE	X	Y	Z	PX	PY	PZ
N3-C2	2.6271	1.3416	0.6200	0.7300	0.0000	0.0000
N3-C4	1.6322	1.3416	1.2050	-0.3460	0.0000	0.6428
N1-C2	3.6748	1.3416	1.1861	-0.3976	0.0000	-0.6122
N1-C6	3.6947	1.3416	2.3308	0.3760	0.0000	0.6257
O2=O6	3.7714	1.3416	3.3337	-5.4447	0.0000	-5.4447
C5-H3	1.5919	1.3416	3.3187	-0.9192	0.0000	0.9192
C4-H4	0.7723	1.3416	1.7901	-1.3000	0.0000	0.0000
O=C2	3.7314	1.3416	0.1957	-5.4447	0.0000	5.4447
N1-H1	4.5474	1.3416	1.7522	4.4400	0.0000	0.0000
N3DC1	1.9471	1.3416	0.3100	0.0000	0.0000	-0.7300
AZUCAR						
O-C1	1.4219	1.8193	0.0000	1.8488	-1.6828	0.0000
O-C4	0.2332	1.9487	0.0000	-2.2132	-1.1626	0.0000
C4-H8	-0.7452	1.2854	-0.3147	-0.7506	-0.7505	-0.7506
C1-H5	2.2618	1.0269	-0.3147	0.7506	-0.7506	-0.7506
C3-H1	0.0000	-0.3854	-0.1146	0.0000	-0.9192	0.9192
C2-H3	1.4248	-0.3854	0.3854	0.0000	-0.9193	0.9192
O2-C2	1.4248	-0.5055	-0.5055	0.0000	1.7678	1.7678
O2-H4	1.4248	-1.0112	-1.4912	0.0000	0.0000	-4.9800
FO2DC5	-1.9570	1.3653	1.5819	1.5669	0.8305	-1.7621
FOSFATO						
O4=P	-3.7326	1.8680	3.4403	0.0000	0.0000	-9.0000
O1-P	-4.3980	1.4835	2.3537	2.4495	1.4145	0.9990
O2-P	-3.0671	1.4835	2.3537	-2.4495	1.4145	0.9990
O3-P	-3.7326	2.6364	2.3537	0.0000	-2.8284	0.9990

Tabla A1.7 Posición y dirección de los dipolos en la timina

GUANINA

	POSICIÓN DEL DIPOLO			DIRECCIÓN DEL DIPOLO		
ENLACE	X	Y	Z	PX	PY	PZ
N3=C2	4.4081	1.3416	0.6794	3.0000	0.0000	-0.0018
N3-C4	3.4815	1.3416	1.2964	-0.2969	0.0000	0.6669
N1-C2	5.4293	1.3416	1.2261	-0.4082	0.0000	-0.6052
N9-C4	2.5770	1.3416	2.1165	0.6947	0.0000	-0.2244
N9-C6	1.9011	1.3416	2.9784	-0.0509	0.0000	0.7282
N7=C6	2.4716	1.3416	3.8859	-2.7815	0.0000	-1.1239
N7-C5	3.5026	1.3416	3.5922	0.4430	0.0000	-0.5802
N1-C6	5.5573	1.3416	2.4358	-0.2497	0.0000	0.6860
N10-C2	5.5800	1.3416	0.1592	-0.5162	0.0000	0.5162
N10-H1	6.6047	1.3416	-0.3605	4.4400	0.0000	0.0000
N10-H2	5.6252	1.3416	-0.5332	-4.1722	0.0000	-1.5186
N1-H3	6.3034	1.3416	1.7733	4.4400	0.0000	0.0000
O=C6	5.7405	1.3416	3.5225	-5.4447	0.0000	-5.4447
N9DC1	1.9471	1.3416	1.160	0.0000	0.0000	-0.7300
AZUCAR						
O-C1	1.4219	1.8193	0.0000	1.8488	-1.6828	0.0000
O-C4	0.2332	1.9487	0.0000	-2.2132	-1.1626	0.0000
C4-H8	-0.7452	1.2854	-0.3147	-0.7506	-0.7505	-0.7506
C1-H5	2.2618	1.0269	-0.3147	0.7506	-0.7506	-0.7506
C3-H1	0.0000	-0.3854	-0.1146	0.0000	-0.9192	0.9192
C2-H3	1.4248	-0.3854	0.3854	0.0000	-0.9193	0.9192
O2-C2	1.4248	-0.5055	-0.5055	0.0000	1.7678	1.7678
O2-H4	1.4248	-1.0112	-1.4912	0.0000	0.0000	-4.9800
FO2DC5	-1.9570	1.3653	1.5819	1.5669	0.8305	-1.7621
FOSFATO						
O4=P	-3.7326	1.8680	3.4403	0.0000	0.0000	-9.0000
O1-P	-4.3980	1.4835	2.3537	2.4495	1.4145	0.9990
O2-P	-3.0671	1.4835	2.3537	-2.4495	1.4145	0.9990
O3-P	-3.7326	2.6364	2.3537	0.0000	-2.8284	0.9990

Tabla A1.8 Posición y dirección de los dipolos en la guanina

ADENINA

	POSICIÓN DEL DIPOLO			DIRECCIÓN DEL DIPOLO		
ENLACE	X	Y	Z	PX	PY	PZ
N1-C2	5.4560	1.3416	1.2612	-0.3865	0.0000	-0.6193
N1-C6	5.5687	1.3416	2.4507	-1.0750	0.0000	2.8008
N3-C2	4.4013	1.3416	0.7010	3.0000	0.0000	-0.0161
N3-C4	3.4437	1.3416	1.3280	-0.2762	0.0000	0.6757
N7-C5	3.5191	1.3416	3.6705	0.4789	0.0000	-0.5509
N7-C6	2.4927	1.3416	3.9289	-2.7816	0.0000	-1.1237
N9-C6	1.9232	1.3416	3.0046	-0.0255	0.0000	0.7296
N9-C4	2.5681	1.3416	2.1356	0.6998	0.0000	-0.2078
N10-C6	5.8483	1.3416	3.5958	-0.5162	0.0000	-0.5162
N10-H1	6.1953	1.3416	4.5901	-1.5186	0.0000	4.1722
N10-H2	6.8731	1.3416	4.1156	4.4400	0.0000	0.0000
C2-H3	5.4903	1.3416	0.3119	0.9192	0.0000	-0.9192
N9DC1	1.9471	1.3416	1.1600	0.0000	0.0000	-0.7300
AZUCAR						
O-C1	1.4219	1.8193	0.0000	1.8488	-1.6828	0.0000
O-C4	0.2332	1.9487	0.0000	-2.2132	-1.1626	0.0000
C4-H8	-0.7452	1.2854	-0.3147	-0.7506	-0.7505	-0.7506
C1-H5	2.2618	1.0269	-0.3147	0.7506	-0.7506	-0.7506
C3-H1	0.0000	-0.3854	-0.1146	0.0000	-0.9192	0.9192
C2-H3	1.4248	-0.3854	0.3854	0.0000	-0.9193	0.9192
O2-C2	1.4248	-0.5055	-0.5055	0.0000	1.7678	1.7678
O2-H4	1.4248	-1.0112	-1.4912	0.0000	0.0000	-4.9800
FO2DC5	-1.9570	1.3653	1.5819	1.5669	0.8305	-1.7621
FOSFATO						
O4=P	-3.7326	1.8680	3.4403	0.0000	0.0000	-9.0000
O1-P	-4.3980	1.4835	2.3537	2.4495	1.4145	0.9990
O2-P	-3.0671	1.4835	2.3537	-2.4495	1.4145	0.9990
O3-P	-3.7326	2.6364	2.3537	0.0000	-2.8284	0.9990

Tabla A1.9 Posición y dirección de los dipolos en la adenina

Distribución de carga eléctrica:

A partir de los momentos dipolares **p** en cada uno de los nucleótidos, calculamos una distribución de carga eléctrica aproximada, empleando el hecho de que el momento dipolar puede entenderse como dos cargas de signo opuesto (q), separadas por una distancia (d), de manera que **p =** q **d**. De esta manera, si escogemos un valor apropiado para 'd', es posible calcular las cargas parciales 'q'. Si hacemos esto con todos los dipolos de cada nucleótido, obtenemos una distribución de carga eléctrica aproximada para la molécula [Davidov,1982].

Por cada dipolo aparecen 2 cargas de igual magnitud pero de signos opuestos, las cuales se ubican a lo largo de la dirección del dipolo separadas por una distancia 'd' por ahora arbitraria.

En primera instancia se escogió como valor para ‘d’, la mitad de la distancia entre los núcleos que participan en el enlace covalente, obteniendo así una distribución de carga determinada. Luego se escogieron distintos valores para ‘d’, mayores y menores que el antes mencionado, obteniendo en cada caso distintas distribuciones de carga eléctrica. En todos los casos, la energía potencial electrostática calculada como a continuación se describe, dio aproximadamente el mismo valor, lo cual indica que el valor escogido para “d” no es crítico para el cálculo de la energía.

De esta manera se escogió el valor de ‘d’ como la mitad de la distancia que separa los núcleos, y se calcularon las coordenadas y magnitudes de cada una de las cargas para cada uno de los cuatro nucleótidos, generando así una función de distribución discreta de carga eléctrica para cada nucleótido. Las distribuciones de carga eléctrica se encuentran en los archivos QA.dat, QG.dat, QC.dat, y QT.dat en la biblioteca de software de la Unidad de Bioquímica Teórica y Estructuras Moleculares de la UPCH.

Cálculo de la energía electrostática:

El hecho de que la probabilidad de ocupación dependa únicamente de la diferencia de energías, tiene dos implicancias importantes: Una de ellas que se menciona arriba, es que las interacciones independientes del tipo de nucleótido son irrelevantes, la segunda que se menciona en el capítulo 2, es que no es necesario calcular la energía de interacción de toda la molécula de ADN, sino únicamente la correspondiente a una vecindad de tamaño apropiado, alrededor del hueco. Esto último se prueba a continuación.

Para calcular la energía potencial del ADN a nivel de nucleótidos, deben considerarse todas las posibles interacciones entre ‘pares de nucleótidos’, pudiendo ser estos cercanos como lejanos. Así si se tuviera un ADN de 4 nucleótidos de largo, como el esquema que se muestra,

$$\begin{array}{lll} 5' & N_1 - N_2 - N_3 - N_4 & 3' \\ 3' & N_5 - N_6 - N_7 - N_8 & 5' \end{array}$$

se tiene que la energía potencial a nivel de nucleótidos de este sistema, es simplemente la suma de las energías de interacción entre todos los posibles ‘pares de nucleótidos’, es decir

$$\begin{aligned} E_{potencial} = {} & E_{12} + E_{13} + E_{14} + E_{15} + E_{16} + E_{17} + E_{18} + \\ & E_{23} + E_{24} + E_{25} + E_{26} + E_{27} + E_{28} + \\ & E_{34} + E_{35} + E_{36} + E_{37} + E_{38} + \\ & E_{45} + E_{46} + E_{47} + E_{48} + \\ & E_{56} + E_{57} + E_{58} + \\ & E_{67} + E_{68} + \\ & E_{78} \end{aligned}$$

donde E_{ij} es la energía de interacción entre los nucleótidos N_i y N_j (donde $E_{ij} = E_{ji}$). Es claro que cuando una de estas 8 posiciones se convierte en un hueco (supongamos que N_6 se convierte en un hueco), la diferencia de energía involucra únicamente los términos de interacción en donde participa directamente N_6, así la diferencia de energía entre las configuraciones en que el hueco (N_6), es ocupado por una adenina y una citosina resulta ser

$$\begin{aligned} Epot_{(6 \to A)} - Epot_{(6 \to T)} = {} & E_{1A} + E_{2A} + E_{3A} + E_{4A} + E_{5A} + E_{7A} + E_{8A} \\ & - (E_{1T} + E_{2T} + E_{3T} + E_{4T} + E_{5T} + E_{7T} + E_{8T}) \end{aligned}$$

Es decir, todos los demás términos sencillamente se cancelan, quedando únicamente las interacciones entre el nucleótido que ocupa el hueco y los restantes.

En vista que estas interacciones disminuyen con la distancia, esperamos que los nucleótidos más alejados del hueco no contribuyan sustancialmente en la energía

potencial, lo cual hace pensar en la posibilidad de simplificar el problema, reemplazando la molécula total de ADN, por una vecindad apropiada, de manera que la diferencia de energía sea razonablemente igual. Como se menciona en el capitulo 2, si denominamos E_A y E_G a las energías de toda la molécula de ADN cuando el hueco es ocupado por una adenina y una guanina respectivamente, y si denominamos $E_{vec.\,A}$ y $E_{vec.\,G}$ a las energías de una "vecindad del hueco" cuando éste es ocupado por una adenina y una guanina respectivamente, es de esperar que se cumpla

$$(E_A - E_T) \sim (E_{vec.\,A} - E_{vec.\,T})$$

Para verificar la hipótesis anterior, y encontrar el tamaño ideal de la vecindad, se calculó la energía de interacción total entre el nucleótido que ocupa el hueco y los nucleótidos restantes de la vecindad para distintos tamaños de ésta, con el procedimiento que se explicará en detalle más adelante.

Para esto se implementó el programa modificado 'cargas21', el cual calcula la energía electrostática de interacción entre el nucleótido X (que ocupa el hueco), con los nucleótidos restantes en una vecindad de 1, 3, 5, 7, 9,11 , y 21 pares nucleótidos.

Se observa que comenzando con un par de nucleótidos, la energía aumenta significativamente al pasar a una vecindad de tres pares, y algo menos al pasar a una vecindad de cinco pares. Sin embargo al incrementar la vecindad desde cinco hasta siete pares de bases el aumento de energía es poco significativo, y éste se hace despreciable al considerar vecindades de 9, 11,... y 21 pares de bases (Figura A1.5).

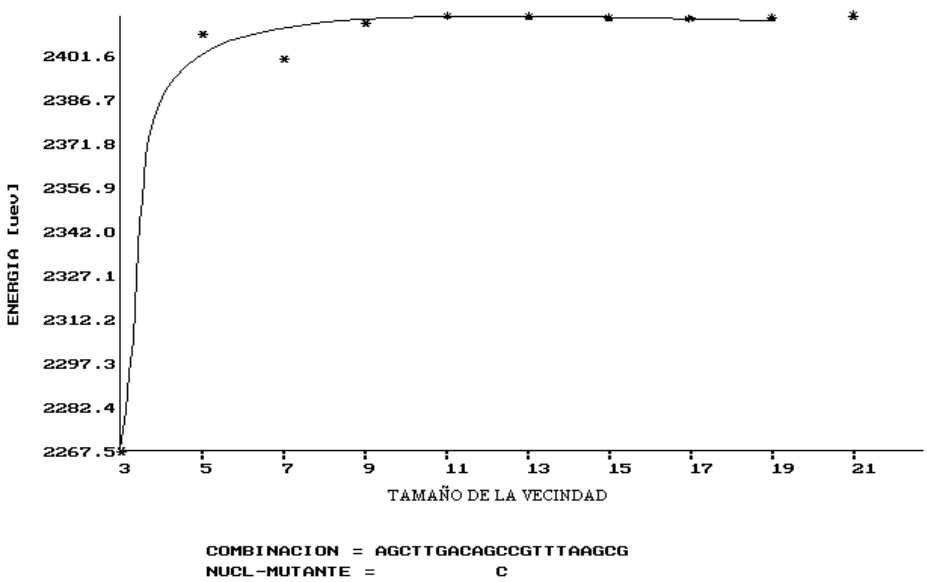


Figura A1.5 Energía de interacción entre el nucleótido X y el resto de nucleótidos para distintos tamaños de vecindad

Es evidente a partir de estos resultados, la validez de elegir una vecindad de cinco pares de bases en el calculo de la energía electrostática para determinar la probabilidad de ocupación del hueco, al cual se suma el hecho que en estas condiciones se reduce significativamente el tiempo de computación.

El procedimiento detallado para el cálculo de la energía electrostática [Espinoza & Zimic,1996], se explicará tomando como ejemplo una vecindad de 5 nucleótidos, sin que esto signifique alguna pérdida de generalidad. El sistema de interés es una vecindad de 5 pares de nucleótidos, donde el par central no es necesariamente complementario. Por ejemplo considérese la siguiente configuración:

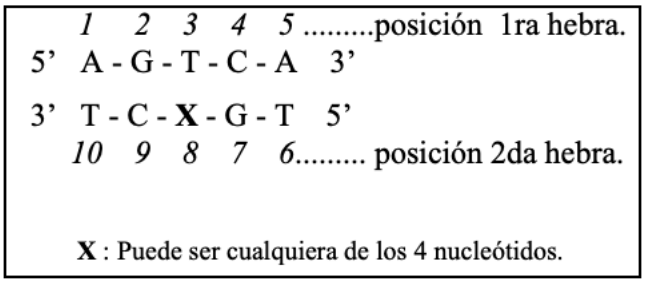


Figura A1.6 Esquema de la vecindad de cinco pares de nucleótidos

El primer paso es construir la función de distribución de carga para toda la vecindad respecto a un único sistema de referencia. Se considera que esta vecindad de cinco pares de bases se encuentra arreglada en una estructura cuasi lineal como primera aproximación. En realidad la disposición de los nucleótidos es una doble espiral, pero considerando que se requieren 10 nucleótidos para completar una vuelta, podemos hacer la aproximación de considerarlos escalonadamente alineados, de manera que cada plano del azúcar sea un peldaño y la altura del mismo la distancia de enlace entre el azúcar y su respectiva base nitrogenada.

Necesitamos las coordenadas de todas las cargas parciales de la vecindad respecto a un único sistema de referencia. Para asegurar la estructura cuasi lineal de la vecindad, se traslada a partir de la *posición 1* (Figura A1.6), cada una de las posiciones de las cargas parciales de cada nucleótido por un vector n$\mathbf{R_T}$ (n = 1,2,3,4 para el segundo, tercero, cuarto y quinto nucleótido de la Figura A1.6 respectivamente), e igualmente a partir de la *posición 6* (Figura A1.6), se traslada cada una de las posiciones anteriores por un vector n$\mathbf{R_{T1}}$ (n = 1,2,3,4 para el séptimo, octavo, noveno y décimo nucleótido respectivamente)**,** hasta formar las dos hebras. La razón de esto está en que los azúcares no se encuentran en un plano, sino mas bien en una estructura escalonada. El vector $\mathbf{R_T}$ tiene como coordenadas (- 2.955, 3.404, 3.782)Å y el vector $\mathbf{R_{T1}}$ (- 2.955, 3.404, -3.782)Å, lo cual garantiza que la distancia entre las bases nitrogenadas paralelas sea de 3.4 Å , y no exista superposición de átomos.

Una vez formadas las dos hebras independientemente, éstas se unirán de manera complementaria mediante los siguientes pasos :

a) Rotando la segunda hebra 180° alrededor del eje Z de la base (6), lo cual implica cambiar de signo a las coordenadas X e Y de las posiciones de las cargas parciales.

b) Disminuyendo las coordenadas anteriores en las bases nitrogenadas en 0.85Å para T y C y aumentando 0.85Å para G y A en la coordenada Z, se asegura que las bases nitrogenadas complementarias se encuentren a la altura y separación adecuada para formar los puentes de hidrógeno.

c) Uniendo las dos hebras por un vector $\mathbf{R_L}$ = (13.351, 2.6832, 0.00) + 4 $\mathbf{R_T}$, que va desde la posición 1 hasta la posición 6, se satisface la formación de los puentes de hidrógeno con la distancia apropiada.

Finalmente después de todos estos pasos se tiene el sistema de átomos y la distribución de carga eléctrica de la vecindad respecto a un sistema de referencia en la posición (1). La energía electrostática, para una distribución discreta de carga, como es el caso que se va a tratar, esta dada por la sumatoria.

$$E=\frac{1}{4\pi\varepsilon}\sum\frac{q_i q_j}{\left|\overline{r_i}-\overline{r_j}\right|}$$

Para obtener la energía de interacción entre nucleótidos, se debe excluir de la sumatoria aquellas interacciones entre cargas parciales correspondientes a un mismo nucleótido, ya que ellos contribuyen con la energía interna de los mismos. Con estas consideraciones se implementó el programa CARGAS, el cual calcula la energía de interacción electrostática para una secuencia de nucleótidos determinada tal como se ha descrito.

Para una vecindad de cinco nucleótidos, en donde sólo el par central no es complementario tipo Watson-Crick, se pueden dar 4^6 (4096) configuraciones diferentes. Las energías de estas 4096 diferentes vecindades fueron calculadas mediante el programa CARGAS en un procesador Intel 486-33Mhz, requiriendo un tiempo de 24 horas de computación continua aproximadamente. Los resultados se encuentran en los archivos Ene11, Ene12, Ene13, Ene14, Ene21, Ene22, Ene23, Ene24, Ene31, Ene32, Ene33, Ene34, Ene41, Ene42, Ene43 y Ene44, en la biblioteca de software de la Unidad de Bioquímica Teórica y Estructuras Moleculares. En cada uno de estos archivos se observará una tabla con un número de seis dígitos que representan a la vecindad, y a su derecha el valor de la energía en eV. Debe resaltarse que el ordene de magnitud de estas energías son de 0.01 a 0.1 veces la energía de un puente de hidrógeno, dependiendo de la secuencia de la vecindad.

En los archivos EneXY, los cinco primeros dígitos representan la secuencia de bases de la hebra intacta de cinco bases de longitud, y el sexto dígito indica el nucleótido que ocupa el hueco. Debemos remarcar que para el cálculo de la energía electrostática de las vecindades, se ha considerado a la carga negativa del grupo fosfato apantallada por la presencia de un ion monovalente situado muy cerca a ésta. La razón de esto es que si no se apantalla la carga eléctrica del grupo fosfato de cada uno de los nucleótidos de la vecindad, las energías de interacción electrostática toman un valor positivo, lo cual significa que la configuración de doble hebra sería inestable por las repulsiones mutuas entre estas cargas negativas. Sin embargo, al considerar que el grupo fosfato es apantallado por un ión monovalente muy próximo, las energías electrostáticas toman un valor negativo dando cuenta de una situación estable. Este hecho es conocido, ya que la estructura de doble hélice del ADN se desestabiliza cuando disminuye la concentración de iones monovalentes positivos.

Adicionalmente se calculó la energía electrostática por un método alternativo, considerando únicamente la interacción entre los dipolos de los enlaces covalentes, es decir los demás términos multipolares que aparecen naturalmente en la energía coulombiana fueron obviados. De esta manera con la información acerca de la posición, orientación y módulo de los dipolos, se calculó la energía de interacción dipolar electrostática. La ecuación que da la energía de interacción entre dos dipolos es [Reitz *et al*.,1980]:

$$E_{12} = \frac{1}{4\pi\varepsilon} \left[\frac{\mathbf{p1} \cdot \mathbf{p2}}{r^3} - \frac{3(\mathbf{r}.\mathbf{p1})(\mathbf{r}.\mathbf{p2})}{r^5} \right]$$

donde **p1** y **p2** son los momentos dipolares, y **r** el vector de separación. Con todas las consideraciones anteriores se implementó el programa ENERGÍA, el cuál calcula la energía dipolar para cualquier vecindad de cinco nucleótidos. Igual que antes, se excluyen las interacciones de dipolos que pertenecen a un mismo nucleótido, ya que estos contribuyen a la energía interna del mismo, y lo que se busca es sólo la energía de interacción entre nucleótidos.

Las energías calculadas por este método para las 4096 vecindades, coincidieron bastante con las calculadas por el método anterior. Esto indica que el tipo de interacción electrostática entre los nucleótidos en la doble hebra del ADN, es fundamentalmente dipolar, probablemente por que la carga eléctrica del grupo fosfato está apantallada por un catión.

Energía de puentes de hidrógeno :

Experimentalmente, la energía de puentes de hidrógeno entre las bases nitrogenadas complementarias de Watson-Crick, ha sido medida en : E(A-T) = -0.34eV (-7.837 Kcal/mol), y E(G-C) = -0.43eV (-9.912 Kcal/mol), aproximadamente [Davidov,1982]. Sin embargo no están reportados los valores de energía de puente de hidrógeno para pares de bases distintos de Watson-Crick. Pretendemos calcular la energía de puente de hidrógeno para los pares de Watson-Crick, y los distintos a estos, para ello se calculará la energía de interacción entre los dipolos que contribuyen al puente de hidrógeno. En este caso, al estar

los átomos tan cerca unos de otros, se considera que entre ellos no existen moléculas de agua, de manera que la constante dieléctrica empleada es 1.

Los valores de energía obtenidos por este método para los pares AT y GC son :

A-T = -0.22 eV (-5.071 Kcal/mol) G-C = -0.32 eV (-7.376 Kcal/mol)

los cuales son muy próximos a los valores experimentales. Lo mismo se hizo para los demás pares, obteniéndose los siguientes valores de energía de puente de hidrógeno para pares de bases aislados (tabla A1.10).

C-A =	0.329 eV	
T-G =	0.245 eV	
T-C =	-0.016 eV	*
A-G =	-0.149 eV	
T-T =	0.136 eV	*
C-C =	0.107 eV	*
A-A =	0.348 eV	
G-G =	0.659 eV	

Tabla A1.10 Energías de puente de hidrógeno

(*) En estos tres casos (pirimidina-pirimidina), se ha considerado una separación entre ambas bases igual a la que tendrían si estuvieran en posiciones complementarias en dos hebras paralelas de ADN, es decir más separados que los pares pirimidina-pirimidina libes (véase apéndice 3).

Este modelo no es el rigurosamente indicado para calcular las energías de puente de hidrógeno, sin embargo podemos notar claramente que las energías de los pares AT y GC se encuentran exactamente 0.12 eV por encima de los valores experimentales. Si consideramos que el modelo aproximado admite una corrección semiempírica como un término aditivo para ajustar los valores teóricos con los experimentales, entonces tendríamos las siguientes energías de puente de hidrógeno corregidas (tabla A1.11).

	(eV)		(Kcal/mol)		
A-T =	-0.34		-7.84		
G-C =	-0.44		-10.14		
C-A =	0.21		4.84		
T-G =	0.13(I)	-0.40(II)	2.99(I)	-9.22(II)	**
T-C =	-0.14		-3.22		
A-G =	-0.27		-6.22	*	
T-T =	0.02(I)	-0.16(II)	0.46(I)	-3.68(II)	**
C-C =	-0.01(I)	-0.14(II)	-0.23(I)	-3.22(II)	**
A-A =	0.23		5.30	*	
G-G =	0.54		12.45	*	

Tabla A1.11 Energías de puente de hidrógeno corregidas

- (*) Estas energías corresponden a pares purina-purina libres, sin embargo cuando se encuentran en una doble hebra de ADN, estos pares purina-purina perturban demasiado

la configuración de doble hebra y se vuelven energéticamente prohibitivas (ver apéndice 3),

- (**) Los casos (I) y (II) corresponden a las dos configuraciones que pueden tomar estas bases. Los casos (II) son prohibitivos en una configuración de doble hebra de ADN, mientras que los I no (Véase el apéndice 3)

Éstas serían las energías de puente de hidrógeno para pares de bases libres, salvo los casos T-T, C-C y T-C, que por tratarse de pirimidinas, se ha considerado una distancia entre bases, que asegura la linealidad de las dos hebras complementarias. Es decir, las distancias de separación entre bases en este caso, son mayores de las que serían si estos pares pirimidina-pirimidina estuvieran libres. Como se ve las energías correspondientes a los pares AT y GC coinciden bastante bien con los valores experimentales.

Los casos (II) , tal como se demuestra en el apéndice 3, no serán considerados por ser energéticamente prohibitivos.

La energía potencial de interés para la vecindad resulta ser igual a la suma de la energía de puentes de hidrógeno y la energía electrostática (usualmente denominada, energía de apilamiento). Esta energía potencial resultante determina la probabilidad de ocupación de un hueco por un nucleótido determinado.

Pares de bases distintos de Watson-Crick

Entendemos por pares de bases de Watson-Crick, a aquellos que mayoritariamente conforman la doble hélice del ADN, es decir los pares A-T y G-C. En todas las figuras de este apéndice, las líneas rojas indican fuerzas atractivas, mientras que las líneas azules indican fuerzas repulsivas. Las interacciones más importantes entre los pares AT y GC son los puentes de hidrógeno (tomadas como interacciones dipolares en nuestro modelo), O...HN y NH...N para el A-T, y NH...O, N...HN, y O...HN para el G-C (Figura A3.1).

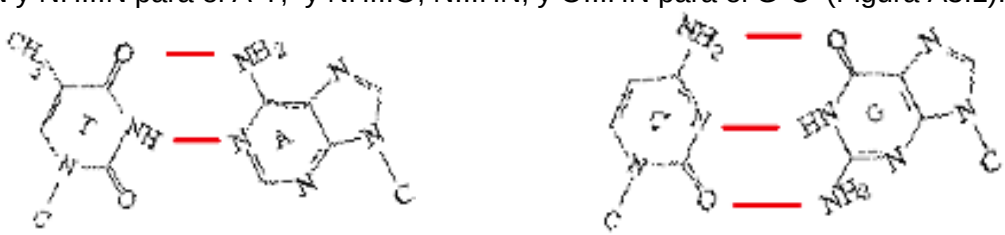


Figura A1.7 Pares de Watson-Crick A-T, G-C

Las interacciones más importantes en el par TC, son las interacciones dipolares O...HN, NH...N, O...O (Figura A1.8).

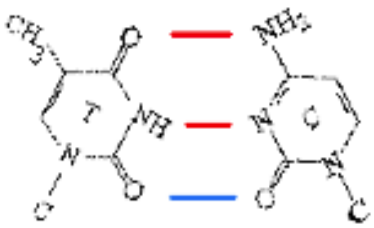


Figura A1.8 Par T-C

Las interacciones más importantes en el par AC, son las interacciones dipolares NH...HN, N...N, O...H (Figura A1.9).

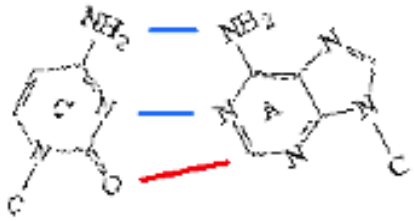
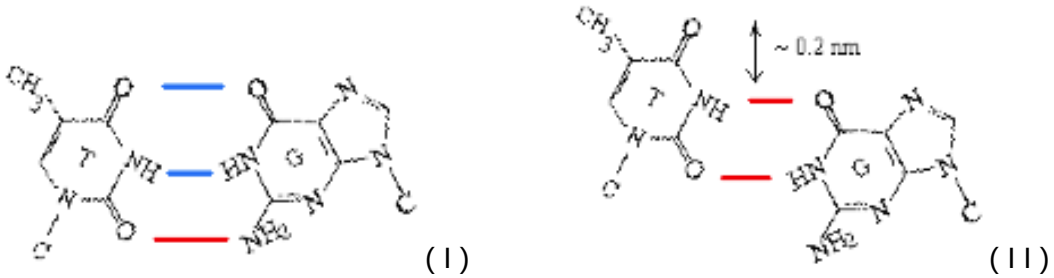


Figura A1.9 Par A-C

El par TG podría estabilizarse de dos maneras (Figura A3.4). En el primer caso (I), las interacciones dipolares serían O...O, NH...HN y O...HN, mientras que en el segundo caso (II), serían NH...O, y O...HN.



(I) (I I)

Figura A1.10 Variantes del par T-G

Sin embargo para los propósitos del presente trabajo necesitamos la configuración que adoptan estando ligados a una doble hebra de ADN. Como se aprecia en la Figura, la configuración (II) requiere un desplazamiento relativo de 0.2nm aproximadamente [Chargaff & Davidson,1955] (obviamente esto es debido a que hemos partido de la premisa de construir nuestra vecindad de cinco pares de bases como una estructura tipo escalera lineal rígida), el cual se puede conseguir estirando ciertos enlaces covalentes como los que unen la pentosa con la base nitrogenada, y girando algunos otros como el del grupo fosfato de manera que su azúcar se eleva un poco sobre el plano inicial. Definitivamente existen muchas formas de conseguir esto, sin embargo el costo energético en el mejor de los casos es demasiado grande al punto que lo vuelve casi prohibitivo. Así, teniendo en cuenta las constantes elásticas de los enlaces covalentes (aprox. 400 Kcal /mol $Å^2$), y de los ángulos de enlace (aprox. 50 Kcal / mol $grado^2$), [Van Gunsteren & Berendsen, 1996], se puede hacer un estimado de la energía necesaria para ordenar ambas bases con la configuración (II), obteniendo un valor de aproximadamente 530 Kcal / mol, lo cual lo vuelve prácticamente prohibitivo.

En el par AG, las interacciones más importantes son las dipolares O...HN, NH...N, y NH...H (Figura A1.11).

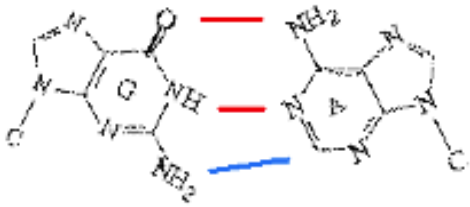


Figura A1.11 Par A-G

Los pares AA y GG como se ve en la Figura A3.6 son muy repulsivos. Las interacciones dipolares entre ellos son NH...HN, N...N, H...H, y O...O .

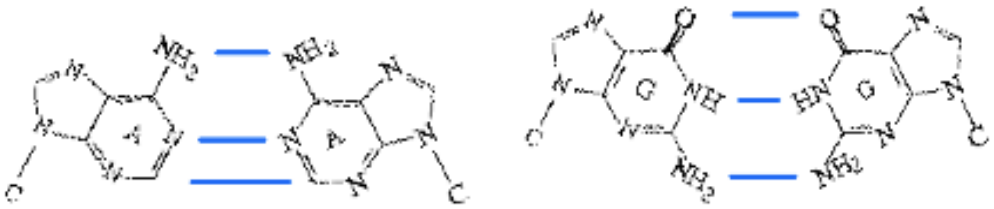
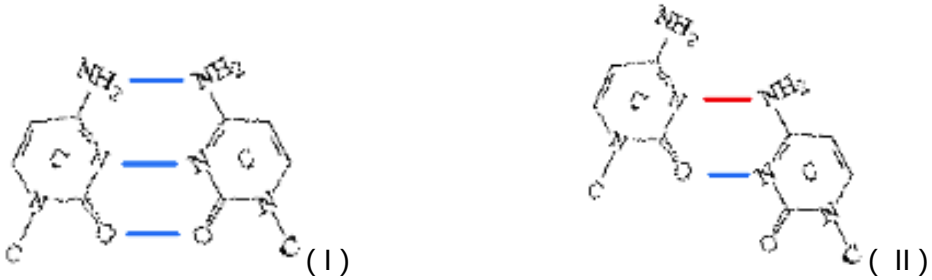


Figura A1.12 Pares A-A y G-G

En el caso de tener un par de purinas (A-G, A-A o G-G), en posiciones complementarias en una doble hebra de ADN, se requiere que la doble hebra tenga que deformarse para darles cabida. La deformación básicamente consiste en separar ambas cadenas una distancia aproximada de 0.2nm. Esto se puede conseguir de muchas maneras, estirando algunos enlaces covalentes, y variando algunos ángulos de enlace. Luego de realizar distintas combinaciones, calculamos que en promedio se requeriría aproximadamente 160 Kcal /mol para acomodar un par de purinas complementarias en una estructura de ADN lineal y rígida. Para propósitos prácticos nosotros lo vamos a considerar prohibitivo.

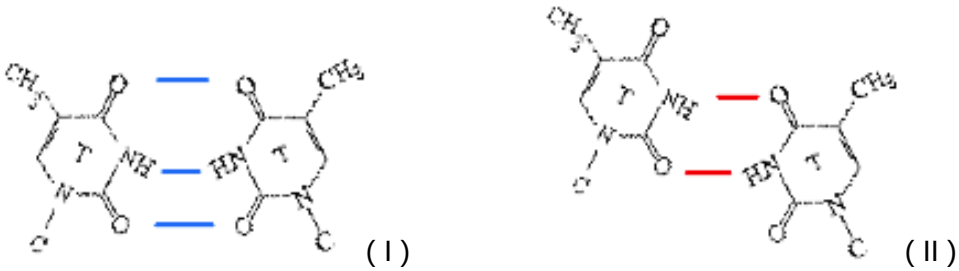
El par CC también tiene dos configuraciones (I) y (II) (Figura A1.13). Para los cálculos en el apéndice 1, se han considerado ambas configuraciones.



(I) (II)

Figura A1.13 Variantes del par C-C

El par TT también tiene dos configuraciones (I) y (II) (Figura A1.14). Para los cálculos en el apéndice 1, se han considerado ambas.



(I) (II)

Figura A1.14 Variantes del par T-T

Igual que para el caso T-G (Figura A1.10), acá también buscamos la configuración de bases complementarias en una doble hebra de ADN, por lo tanto para los pares C-C (II) y T-T (II), también se requiere efectuar un desplazamiento de aproximadamente 0.2nm, con lo cual existe un requerimiento energético muy grande que también los vuelve prohibitivos, así es más probable que se den las configuraciones tipo (I).

Anexo 2: Simulación computacional mediante un algoritmo de Monte Carlo para modelar un proceso evolutivo guiado por la distribución de Boltzmann

El modelo de simulación de mutaciones puntuales, basado en la estabilidad termodinámica de la molécula de ADN, ha sido implementado en varios códigos de programación para el compilador Turbo Pascal v.7.0 de Borland. Estos códigos efectúan diversas simulaciones, tales como procesos de mutaciones del ADN no-codante-basura, y codante, distinguiendo las contribuciones al envejecimiento y diversificación. Los códigos desarrollados, se encuentran en la biblioteca de software de la *"*Unidad de Bioquímica Teórica y Estructuras Moleculares*"* del Laboratorio de Bioquímica de la Universidad Peruana Cayetano Heredia.

En consenso, los distintos códigos de programación mencionados presentan los siguientes procedimientos:

1- Lectura y almacenamiento de secuencias de ADN en un formato particular.
2- Lectura de los parámetros con que se realiza la simulación : Temperatura, constante dieléctrica, concentración de nucleótidos trifosfato libres, cinética del mecanismo de reparación.
3- A partir de una distribución uniforme de probabilidad a lo largo de toda la secuencia, se genera un hueco en una de las dos hebras.
4- Reconocimiento de la vecindad (de cinco pares de bases), y cálculo de la energía potencial electromagnética para los cuatro casos: Cuando el hueco es ocupado por una guanina, citosina, timina y adenina.
5- Cálculo de la probabilidad que el hueco sea ocupado por cada uno de los nucleótidos, empleando la distribución de Boltzmann, y simulación de la mutación empleando un algoritmo de Monte Carlo.
6- Simulación de la acción del mecanismo de reparación, de acuerdo a la cinética asumida.
7- Determinación de la viabilidad de la mutación, de acuerdo a los criterios establecidos. En el caso de tratarse de una mutación letal, el programa simula la muerte del individuo.
8- Cálculo del porcentaje de guanina-citosina. Almacenamiento de la secuencia mutada en un archivo.
9- El proceso anterior se repite por un número determinado de veces. En promedio una simulación genera 10 millones de huecos en una secuencia de 1000 pares de bases aproximadamente. Notemos que debe existir una equivalencia entre el número de huecos generados y el tiempo real, ya que la frecuencia de generación de huecos en la naturaleza debe tener un valor promedio que en principio dependería del tipo de organismo.

Para economizar tiempo de computación, se realiza una modificación razonablemente válida, con la cual no se pierde generalidad. En lugar que el mecanismo de reparación corrija el error en la mitad de las veces, y en la otra mitad se equivoque sin tener ninguna preferencia por algún tipo de nucleótido, se modifica el algoritmo de manera que el mecanismo de reparación siempre se equivoque, y por igual con todas las bases, de esta manera se cumple la neutralidad, pero se ahorra la mitad del tiempo de simulación. Esto se ha efectuado únicamente por razones prácticas sin que afecte la naturaleza de las predicciones.

Seguidamente se muestra el diagrama de flujo consenso de los algoritmos de simulación de mutaciones puntuales basado en la estabilidad termodinámica de la molécula de ADN.

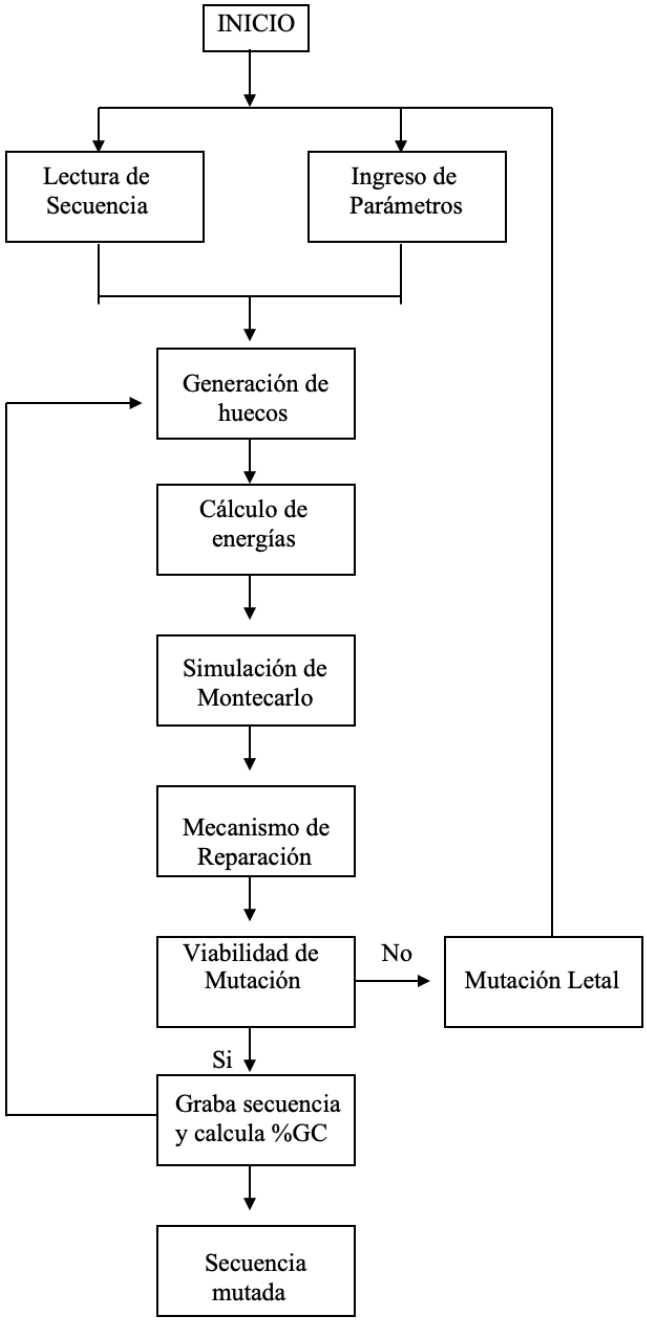


Figura A2.1 Diagrama de flujo del algoritmo de simulación de mutaciones puntuales

Criterios para determinar la viabilidad de una mutación

Partiendo de la premisa que es posible estudiar la evolución de cada tipo de ADN independientemente, sin que la evolución de uno influya sustancialmente sobre otro, se puede diferenciar la evolución del ADN codante, no-codante-basura y funcional.

Empleando el modelo simple de mutaciones puntuales, es necesario establecer criterios para determinar la viabilidad de una mutación. En principio esto depende de la región en

que se ha producido la mutación, pudiendo ser ésta, la región codante, no-codante-basura, o funcional.

En el caso de la región no-codante-basura, consideramos que toda mutación es viable, puesto que como se ha planteado, asumimos que sobre ella no actúa la presión de selección. En el caso de la región funcional, el problema es muy complicado y no hay forma por el momento de determinar la viabilidad de la mutación. El problema fundamental en este caso consistiría en determinar la estructura tridimensional de la molécula de ribonucleótidos tras la mutación y verificar su funcionalidad.

En el caso de la región codante, el problema es complicado pero admite una simplificación. Formalmente se debe determinar la estructura tridimensional de la proteína a partir de la secuencia de aminoácidos, para luego determinar si ésta es aún funcional. Sin embargo el problema se simplifica al considerar la premisa de estar ante una presión de selección nula, ya que en estas circunstancias, el organismo se encuentra adaptado.

Existen algunas evidencias que dan ciertas pautas para comprender mejor el problema de la viabilidad de mutaciones en el ADN codante. Es conocido que proteínas homólogas de distintos organismos, presentan alta homología en cuanto al tipo o familia de aminoácidos, y una baja homología en cuanto a aminoácidos solamente. Esto se debe a que la conservación de la familia de aminoácidos en la proteína, asegura la conservación de dominios y motivos, y con ello el buen plegamiento y por lo tanto la funcionalidad de la misma. Al extrapolar estas evidencias bajo la consideración de un medio ambiente estable (presión de selección muy pobre), queda claro que toda mutación en la región codante que conserva el tipo de aminoácido tiene una gran posibilidad de mantener la funcionalidad de la proteína asegurando la sobrevivencia del organismo.

En principio existen muchas mutaciones que conservan el fenotipo y contribuyen al envejecimiento (anagénesis de secuencias). De todas ellas, las mutaciones que conservan el aminoácido aseguran la conservación del fenotipo. Las mutaciones que contribuyen a la diversificación (cladogénesis de secuencias), son las que producen cambios fenotípicos y a la vez aseguran la sobrevivencia del organismo. Las más simples son aquellas que conservan la familia de aminoácidos, ya que de esta manera se asegura en la mayoría de los casos un plegamiento similar de la proteína y por lo tanto una funcionalidad adecuada.

Con estos argumentos se sugieren criterios muy simples de viabilidad de mutación en un proceso de envejecimiento o diversificación. Así, para la simulación de un proceso de envejecimiento, se exige la conservación del aminoácido, mientras que para la simulación de un proceso de diversificación, se exige la conservación de la familia de aminoácido. En ambos casos, las mutaciones viables están restringidas a aquellas que satisfacen las condiciones impuestas, lo cual es un alto costo de la simplicidad del criterio, ya que se están dejando de lado otras mutaciones distintas a las anteriores, que si pueden ser viables.

Elección de la cinética de corrección de errores más simple para el mecanismo de reparación

El sistema de reparación del ADN está formado por lo menos por tres mecanismos independientes: La reacción de polimerización, el mecanismo de "proof-reading" de exonucleasas y el sistema de reparación post-replicación. Es poco lo que se sabe acerca de la cinética de corrección de errores de estos mecanismos. Los conocimientos al respecto están limitados a valores de la eficiencia global de dichos mecanismos, entendido como el número de errores que se cometen por número de reparaciones. Sin embargo, aún no existe información respecto a la eficiencia del mecanismo de reparación a nivel de nucleótidos individuales. Esto se puede entender mejor con un ejemplo. Supongamos que un hueco en el ADN es ocupado por una base no complementaria a la cadena opuesta, formándose un par no canónico, como G-T. El mecanismo de reparación intentará corregir este mal apareamiento. En los casos más críticos, no existe manera de reconocer el nucleótido incorporado, por lo que el mecanismo de reparación debe elegir por la conservación de una de las dos bases. Evidentemente hay dos posibilidades, en un caso puede conservar G y reemplazar T por C, o conservar T y reemplazar G por A. En estas

situaciones, se desconoce lo que haría el mecanismo de reparación. Una posibilidad que no podemos descartar es que el mecanismo de reparación, por alguna razón físico-química o estérica, tenga preferencia por conservar algún tipo de bases. Un mecanismo así, definitivamente tendría un efecto muy importante sobre la abundancia de nucleótidos en el ADN durante la evolución.

Este aspecto será considerado durante las simulaciones que se efectuarán con el modelo de mutaciones puntuales, en donde se asume una cinética de reparación neutra, es decir que no tiene preferencia por algún tipo de nucleótido. Más adelante se verá el efecto de distintas cinéticas distintas de la neutral para el mecanismo de reparación, sobre un proceso de evolución simulado por un modelo de mutaciones puntuales.

Predicciones del modelo de simulación

A continuación se presentan una serie de simulaciones efectuadas con el modelo de mutaciones puntuales descrito anteriormente. En todos los casos, se ha considerado la presión de selección nula con los criterios de viabilidad descritos arriba tanto para los procesos de diversificación y envejecimiento, así como un mecanismo de reparación neutro. Todas estas simulaciones constituyen las predicciones del modelo, las cuales deberán ser confrontadas con datos de la realidad.

La leyenda que aparece en la parte superior de cada una de las figuras que a continuación se presentan, tiene el siguiente significado :

- *Nuc cambiados*: Indica el número de nucleótidos que fueron sustituidos en la secuencia original.
- *Num. Nucleótidos*: Indica el número de nucleótidos de la molécula de ADN que va a someterse a la simulación.
- *%GC total*: Indica el porcentaje de guanina-citosina en la molécula de ADN durante la simulación.
- *Mutac. Increm.*: Indica el número de huecos representadas por cada pixel en la Figura
- *Temp*: Indica la temperatura en grados Kelvin, a la cual se lleva a cabo la simulación.
- *[C], [T], [G], [A]* : Indican las concentraciones relativas de los deoxirribonucleótidos trifosfato libres.
- **.zzz*: Indica el nombre del archivo que contiene la secuencia de ADN que va a someterse a la simulación.

El eje vertical representa el porcentaje de guanina y citosina (%GC), de la molécula de ADN durante la simulación, y el eje horizontal representa el número de huecos generados. A continuación se presentan algunas predicciones del modelo en diferentes situaciones.

Diversificación del ADN codante

En la Figura 4.1, se aprecia la variación del porcentaje de guanina-citosina, (%GC), en función del número de huecos generados, para un proceso de diversificación. La molécula de ADN inicial es una secuencia de 900 nucleótidos generada aleatoriamente con un 22% de GC. La temperatura de simulación es de 37°C (310°K), y las concentraciones relativas de los cuatro nucleótidos trifosfato libres son idénticas (concentraciones equimolares). Esta simulación se realizó con el programa divers.pas.

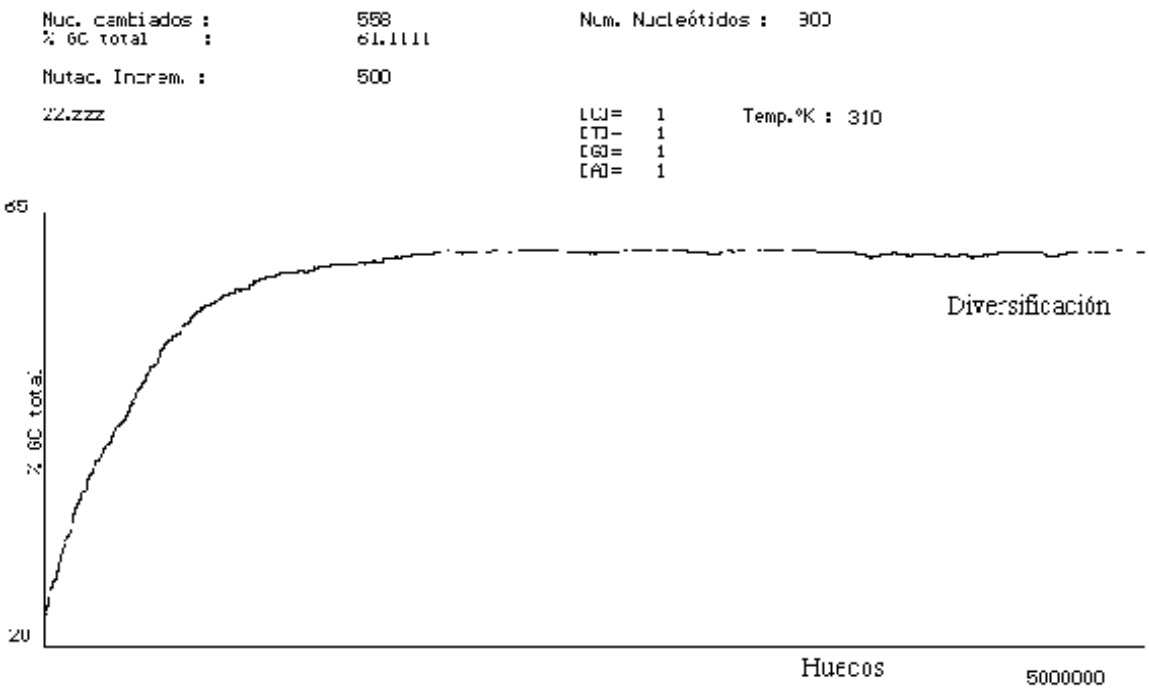


Figura A2.2 Diversificación de una secuencia codante con 22 % GC

Se observa claramente un crecimiento monótono del contenido de guanina-citosina en función del número de huecos generados, que en este caso llegan a 5 millones. Puede notarse que el contenido de guanina-citosina tiende a alcanzar un valor estacionario que se observa como una asíntota o plateau en la Figura. Nótese que este plateau de GC es de aproximadamente 61.1%. Un detalle adicional es que el plateau se alcanza recién en aproximadamente 1 300 000 huecos generados aproximadamente. En 5 millones de huecos generados, fueron sustituidos 558 nucleótidos de la secuencia original de ADN.

Se efectuó una simulación similar a la anterior, con la diferencia de que en este caso la secuencia inicial tiene un contenido de guanina-citosina próximo al 44% (Figura 4.2). Todos los parámetros se mantuvieron iguales a la simulación 5.1, para observar el efecto de la secuencia inicial. La simulación también se realizó con el programa divers.pas, obteniéndose la siguiente Figura :

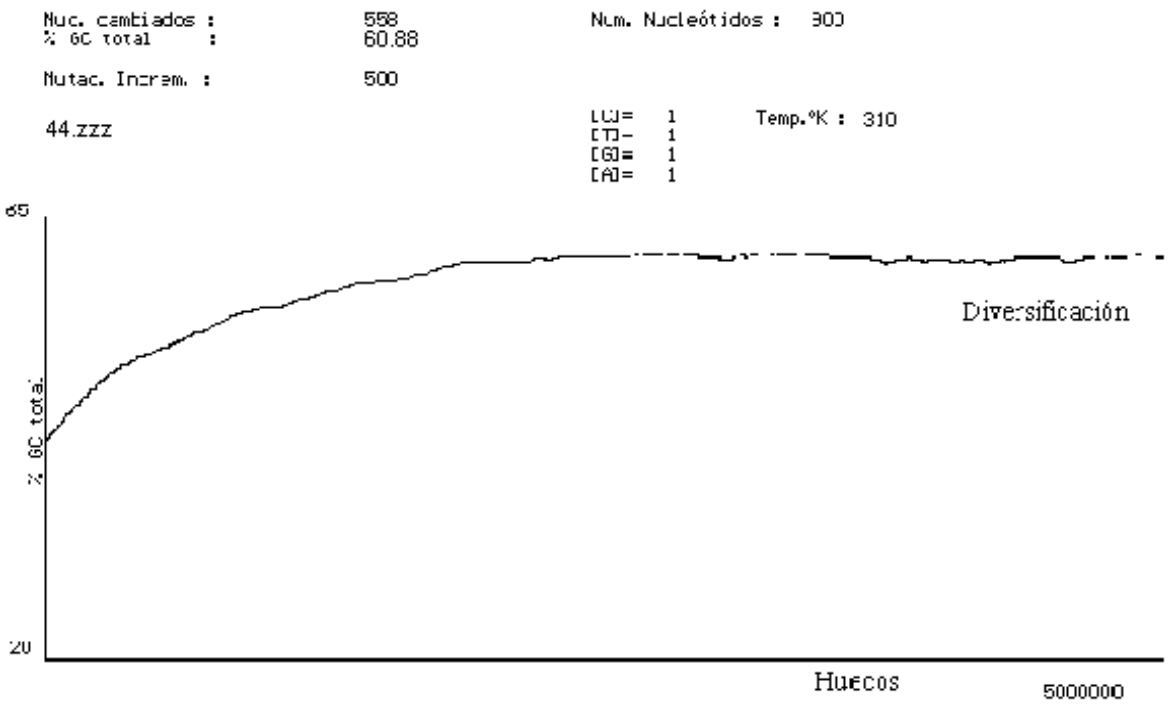


Figura A2.3 Diversificación de una secuencia codante con 44 % GC

Se observa un crecimiento monótono y asintótico, con un valor de saturación para el porcentaje de guanina-citosina cercano al 61%.

Evolución del ADN no-codante-basura

Para simular la evolución de una secuencia de ADN no-codante-basura, se considera solamente la presión mutacional termodinámica sin ningún tipo de restricción sobre las mutaciones. Para ello se implementó el programa Nocod.pas. La temperatura de simulación es de 37 °C, y las concentraciones de nucleótidos trifosfato libres son iguales. En la simulación se parte de una secuencia inicial con un 22% GC y 900 nucleótidos de longitud, obteniéndose como resultado la siguiente Figura 4.3.

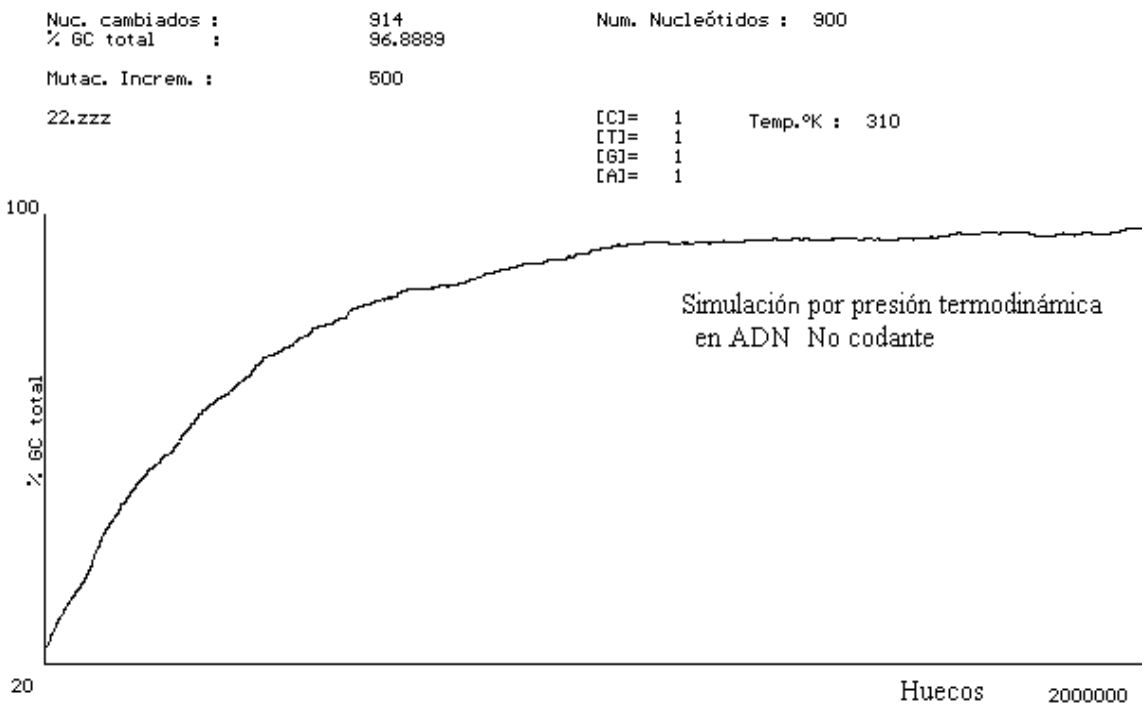


Figura A2.4 Evolución de una secuencia no-codante-basura con 22 %GC inicial

Se observa que luego de 2 millones de huecos generados, fueron sustituidos 914 nucleótidos, alcanzándose un plateau de 96.8% GC.

Se realizó una simulación similar a la anterior (Figura 4.4), con la diferencia que la secuencia inicial tiene un 50% GC. Todos los demás parámetros son iguales a la simulación anterior. Se empleó el programa Nocod.pas.

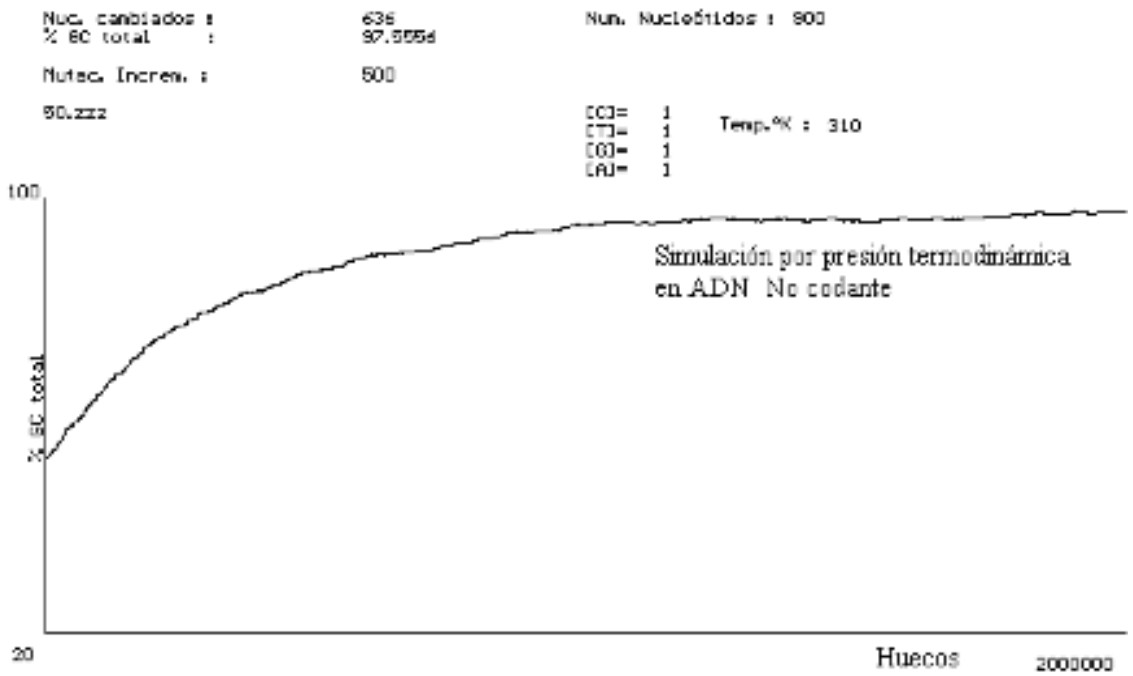


Figura A2.5 Evolución de una secuencia no-codante-basura con 50 %GC inicial

Se aprecia claramente que también hay una tendencia al aumento del contenido de guanina-citosina. En este caso el valor de saturación para el porcentaje %GC es de 97.5% aproximadamente. En esta simulación, se sustituyeron 636 nucleótidos en 2 millones de huecos generados.

Efecto de la temperatura en el proceso de diversificación

Para comprender mejor el efecto de la temperatura sobre el proceso de diversificación, se repitió la simulación 4.2 con la diferencia de que la temperatura en este caso es de 97 °C, (370°K) (Figura 4.5). Por lo demás todos los parámetros se mantuvieron iguales. El programa utilizado fue divers.pas.

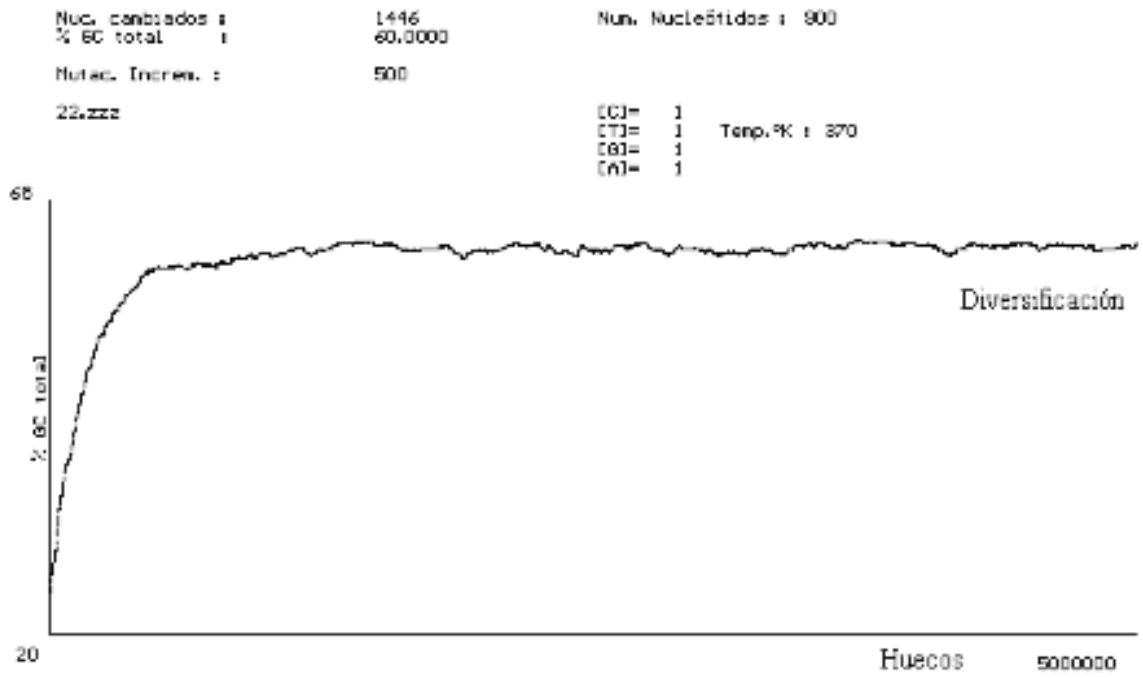


Figura A2.6 Diversificación de una secuencia codante con 22 % GC a 97° C

Claramente se observa el crecimiento asintótico. Lo sorprendente en este caso es la velocidad de crecimiento del %GC, ésta es bastante mayor que en los casos anteriores, alcanzándose el plateau rápidamente cerca a los 500 mil huecos generados. En este caso para 5 millones de huecos generado, se han sustituido 1446 nucleótidos. El porcentaje de saturación en este caso es de 60% aproximadamente.

Efecto de la concentración de nucleótidos trifosfato libres en el proceso de diversificación.

En todas las simulaciones anteriores, se ha considerado que los nucleótidos trifosfato libres se encuentran en concentraciones equimolares. En la simulación de la Figura 4.6 se tiene una curva de diversificación a 37°C para una secuencia inicial codante con 22%GC. En esta oportunidad se ha tomado la concentración de dATP libre como 2 veces mayor que el resto de nucleótidos trifosfato.

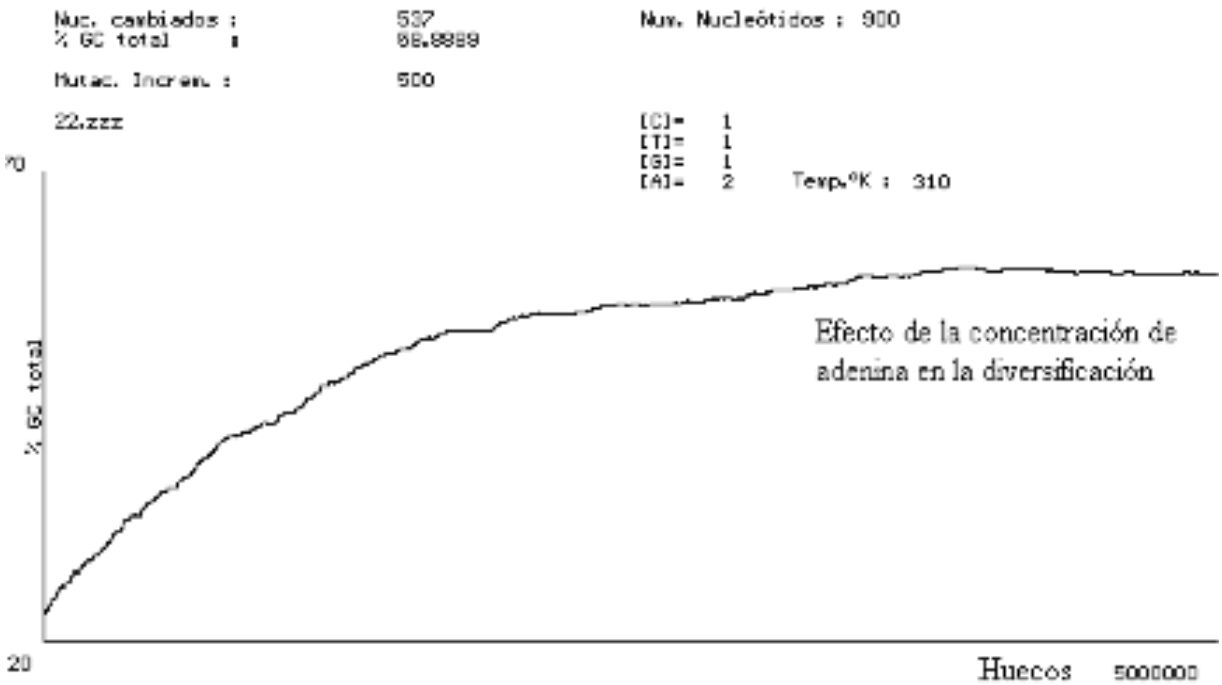


Figura A2.7 Efecto de la concentración de nucleótidos trifosfato libres

Claramente se aprecia un aumento del contenido de GC, con la tendencia a alcanzar un plateau cercano al 60%. Comparando esta curva con la Figura A2.7 se observa que existe una diferencia en la velocidad con que se alcanza el plateau. En este caso al haber más dATP libre, la velocidad de incremento del %GC es más lenta y toma más tiempo en alcanzarse el valor estacionario, aunque este último es prácticamente el mismo.

Efecto del contenido de aminoácidos degenerados en el proceso de diversificación

Para comprender mejor el efecto del contenido de aminoácidos degenerados sobre la diversificación y envejecimiento de una secuencia de ADN codante, se efectuó la simulación de la Figura A2.8, en donde la secuencia de partida es muy rica en aminoácidos degenerados y tiene un %GC cercano al 42% . Por lo demás todos los parámetros se mantienen igual que en la simulación anterior.

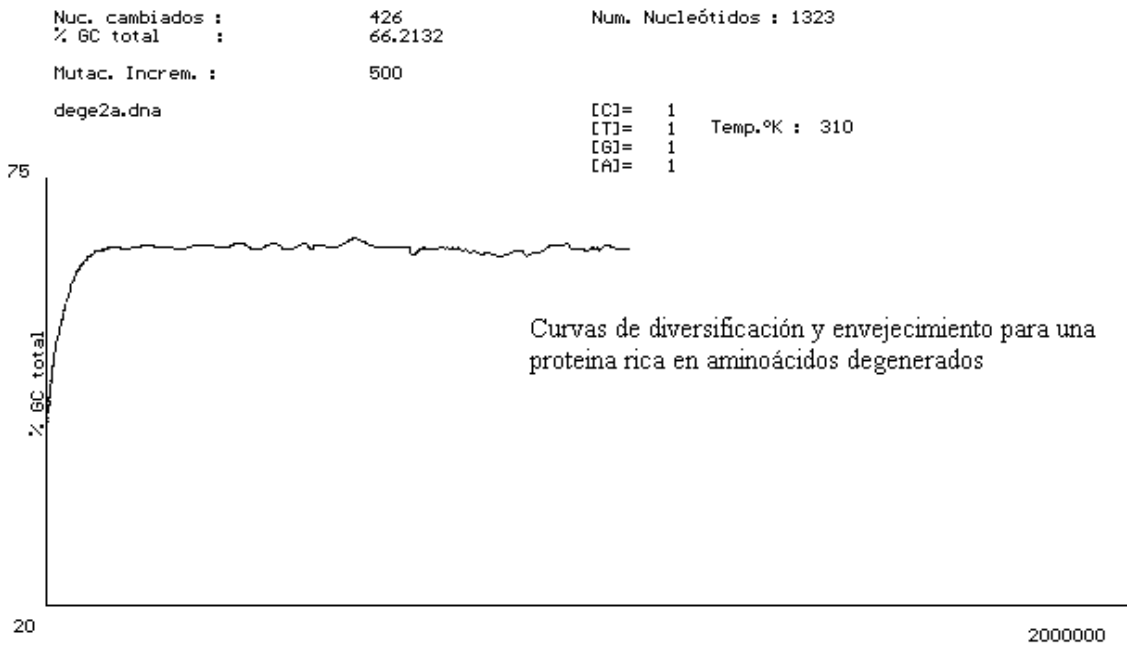


Figura A2.8 Diversificación y Envejecimiento de una secuencia codante rica en aminoácidos degenerados

Puede verse claramente que la velocidad de crecimiento de esta curva de diversificación es mucho mayor comparada con los casos anteriores en que las secuencias de partida no son tan ricas en aminoácidos degenerados como en este caso.

Efecto del mecanismo de reparación sobre la diversificación

En todas las simulaciones anteriores se ha considerado un mecanismo de reparación neutral, es decir carente de preferencia por alguna base. La Figura A2.9 es un ejemplo de ello, la cual muestra un plateau de cerca de 61% GC.

En la Figura 4.9 se tiene una diversificación que se diferencia de la Figura 4.8 únicamente en la cinética del mecanismo de reparación, que en este caso tiene una marcada preferencia de 9:1 por conservar timinas frente a los demás nucleótidos, siendo todos los demás parámetros idénticos.

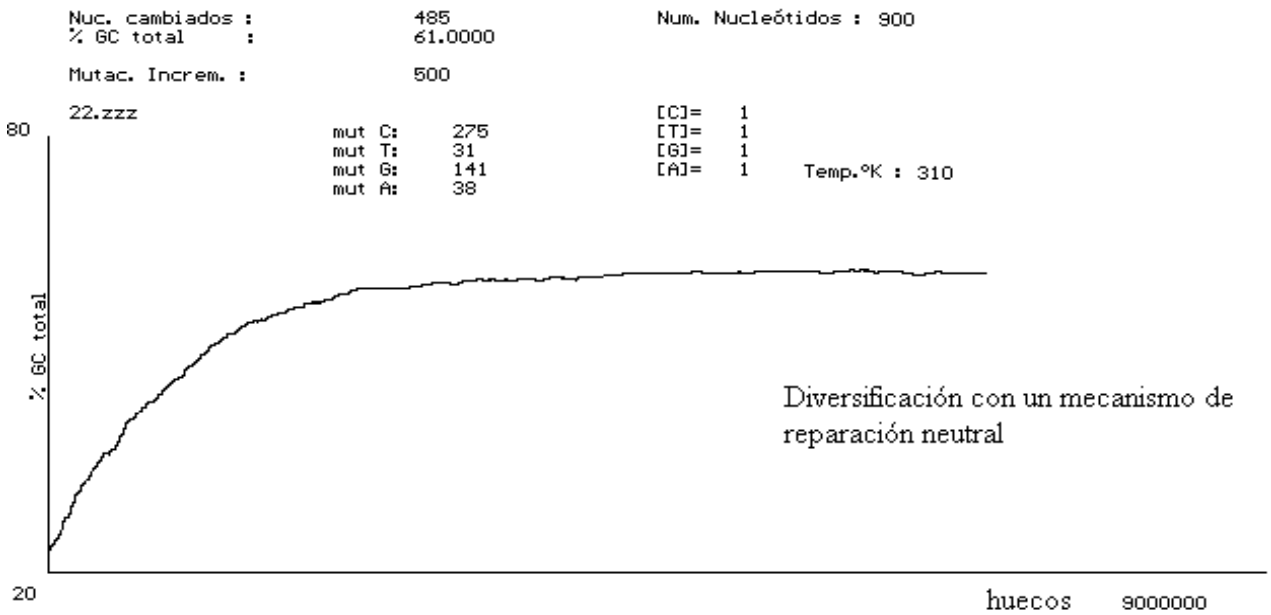


Figura A2.9. Diversificación con un mecanismo de reparación neutral

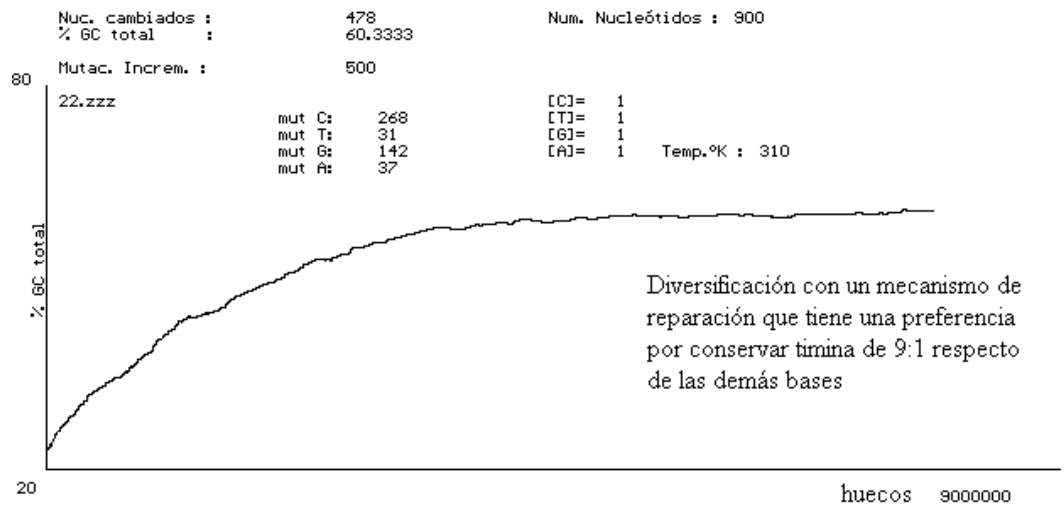


Figura A2.10. Diversificación con un mecanismo de reparación con preferencia por timina

En la diversificación de la Figura A2.10 en que se considera un mecanismo de reparación con una preferencia por conservar timina de 9 a 1, respecto de las demás bases, se observa que el plateau es muy próximo al anterior (60.3 % GC), y que la diferencia está únicamente en una marcada reducción de la velocidad de incremento del %GC.

Simulación de procesos de envejecimiento

Es posible visualizar simultáneamente un proceso de diversificación y envejecimiento. Para ello se implementó el programa Div-env.pas. En este caso se parte de una secuencia inicial la cual se somete a una simulación de diversificación, que después de un cierto tiempo es interrumpida para iniciar una simulación de envejecimiento, para luego retornar al punto de diversificación en que se quedó, y repetir el proceso varias veces. En este caso se partió de una secuencia codante de 1323 nucleótidos, con un %GC cercano al 39%. La simulación se realizó a 37°C, y las concentraciones relativas de los nucleótidos trifosfato libres fueron equimolares (Figura A2.11).

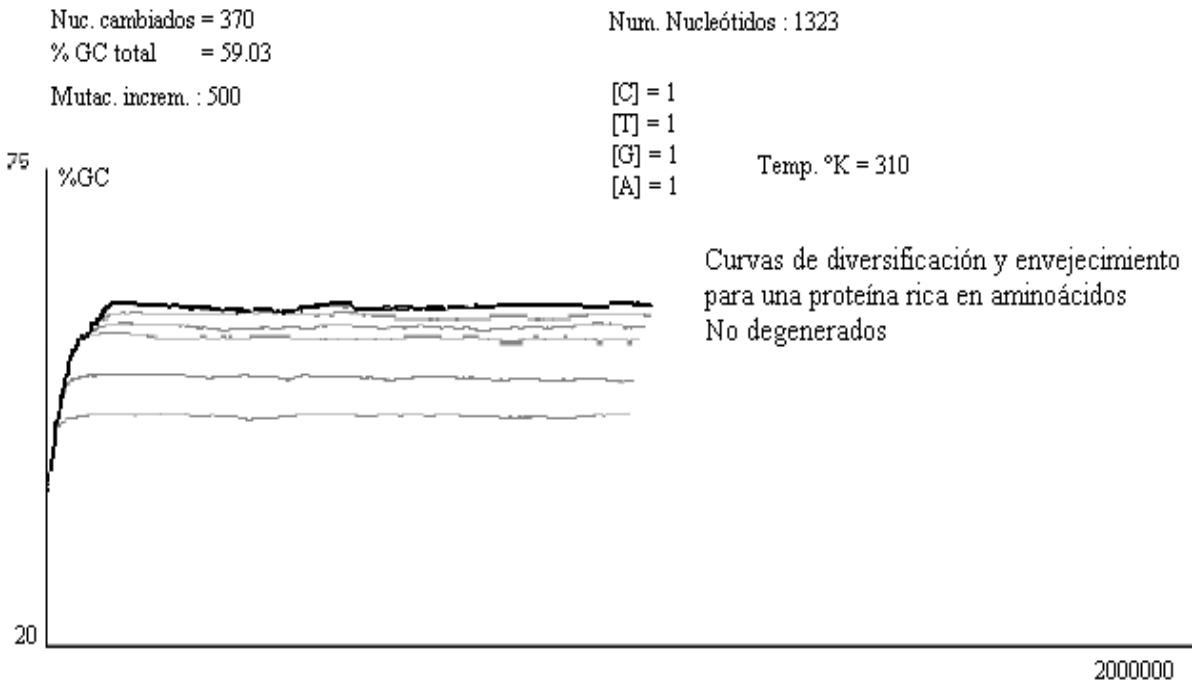


Figura A2.11 Diversificación y Envejecimiento de una secuencia codante rica en aminoácidos no degenerados

En esta simulación, la curva más externa y resaltada vista desde la izquierda representa la diversificación. Las demás curvas internas que nacen a partir de varios puntos de la curva de diversificación, representan el envejecimiento de las secuencias representadas por los puntos mencionados. La curva de diversificación representaría a un continuo de organismos que han diversificado unos de otros, y las curvas de envejecimiento, representan la variación por envejecimiento de algunos de estos.

Variación del contenido de guanina-citosina en las tres posiciones del codón durante un proceso de diversificación

Modificando parcialmente el programa divers.pas, se implementó un algoritmo para medir el contenido de guanina-citosina en las tres posiciones de codón de toda la secuencia durante la simulación de un proceso de diversificación (Figura A2.12). La secuencia inicial tiene un %GC cercano al 22%. La temperatura de simulación es de 37° C. El programa empleado en este caso es gc3pos.pas.

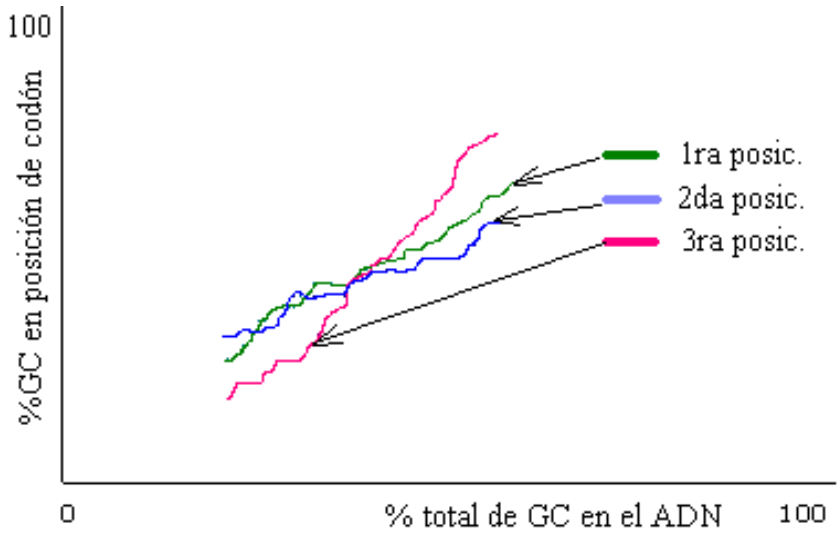


Figura A2.12. Variación del porcentaje de guanina-citosina en las tres posiciones del codón

Nótese que el eje horizontal de la curva indica el %GC total de la secuencia, que como hemos visto antes, de alguna manera es proporcional al tiempo evolutivo. En el eje vertical se tiene los porcentajes de guanina-citosina tanto para la primera, segunda y tercera posición de codón. Se nota claramente que existe una tendencia a incorporar más G y C en la tercera posición del codón registrándose así más mutaciones en esta posición, lo cual se deduce por la mayor pendiente de la curva.

Anexo 3: Evidencia experimental que permite contrastar las predicciones del modelo de simulación

La prueba de consistencia del presente modelo consiste en comparar las predicciones experimentales con información experimental de la evolución de algún linaje filogenético que reúna las condiciones impuestas: Medio ambiente casi constante, es decir presión de selección casi nula, y mecanismo de reparación neutral.

Por esta razón, se requiere analizar un linaje relativamente pequeño que se haya desarrollado en un intervalo de tiempo igualmente pequeño de manera que no se hayan producido cambios significativos en el medio ambiente, asegurando que los organismos ya se encuentran totalmente adaptados. Un posible candidato sería algún linaje de organismos superiores que haya habitado la tierra en algún período corto y estable, lo cual es muy difícil de determinar.

Una alternativa consiste en aprovechar el hecho de que los organismos superiores presentan mecanismos homeostáticos que le aseguran mantener ciertos parámetros muy conservados, por ejemplo los mamíferos poseen una temperatura corporal y pH que varía en un rango relativamente pequeño. De esta manera algunos organismos superiores hacen la vez de un medio ambiente suficientemente conservado para algunos parásitos. Por esta razón proponemos que algún linaje de parásitos sería el ideal para la verificación de este modelo.

Un segundo punto a considerar es el tipo de reproducción. Debido a que por ahora pretendemos simular únicamente procesos de mutaciones puntuales, se deben evitar eventos naturales como modificaciones cromosómicas, crossing overs, que se dan notablemente y con mayor frecuencia en organismos de reproducción sexual. Por esta razón conviene analizar algún linaje de parásitos de reproducción clonal, de manera que las modificaciones cromosómicas sean irrelevantes.

Un tercer punto ha considerar, está relacionado con la presencia de intrones. Nuestro modelo está implementado en un algoritmo que por el momento no puede discriminar intrones en una secuencia de ADN codante. Debido a que las bases de datos genómicas más abundantes no corresponden a cADN, y considerando que es algo complicado identificar la región del intrón y que la viabilidad de las mutaciones en estas estructuras son bastante complejas de estudiar, resulta práctico analizar organismos carentes de intrones.

Un cuarto e importante punto consiste en que el linaje sea lo suficientemente grande en cuanto a tiempo evolutivo de manera que haya acumulado suficientes mutaciones para observar cambios importantes.

Existe un linaje de parásitos que satisface bastante bien estas tres últimas condiciones: los *Kinetoplastidia*. Los Tripanosomátidos junto con la *Leishmania*, *Crihtidia* y otros, constituyen un linaje filogenético ideal para la verificación del modelo, ya que son organismos eucariotes inferiores que carecen de intrones [Vickerman,1994]. No hay evidencias de que los *Kinetoplastidia* presenten algún tipo de intercambio sexual notable, por lo que su reproducción es marcadamente clonal. Además este linaje es tan antiguo, que debe haber acumulado las mutaciones e información que estamos buscando [Maslov,1995; Maslov,1994].

Por estas razones los *Kinetoplastidia* constituyen el linaje que mejor satisface las condiciones impuestas por nuestro modelo.

Adicionalmente a éste, tomaremos otros linajes que no satisfacen tan bien las condiciones necesarias, y las analizaremos para tener una comparación.

Algunos aspectos de la evolución del linaje de los *Kinetoplastidia*

Para tratar de verificar las múltiples predicciones del modelo, se mostrarán datos experimentales adecuadamente organizados, de manera de contar con curvas reales equivalentes a las figuras obtenidas en las distintas simulaciones del capítulo 4.

De acuerdo a recientes trabajos [Maslov *et al*.,1994; Maslov *et al*.,1995], el linaje de los *Kinetoplastidia* está organizado filogenéticamente como lo muestra el cladograma de la Figura 5.1. Éste ha sido construido en base a subunidades ribosomales pequeñas nucleares (sRNA), empleando el algoritmo de Parsimonia [Stewart,1993; Fernandez,1993; Olsen,1993; Macintyre,1994], del programa de filogenia Phylip. El outgroup elegido es la *Euglena*.

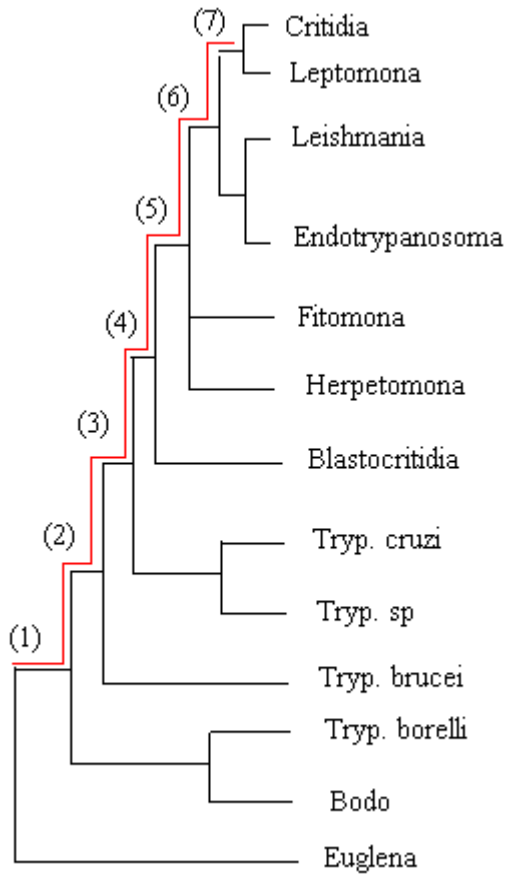


Figura A3.1 Árbol filogenético de Maslov & Simpson para el linaje de los *Kinetoplastidia* empleando Euglena como outgroup [Maslov *et al*.,1994; Maslov *et al*.,1995].

Es importante notar que en las predicciones del capítulo 4, las secuencias de ADN generadas a lo largo de la simulación descienden directamente una de otra, como resultado de mutaciones puntuales en el ancestro. La evidencia experimental necesaria para apoyar el modelo debe tener las mismas características, es decir los organismos elegidos deben pertenecer a una misma línea evolutiva o cascada filogenética. En el árbol filogenético de los *Kinetoplastidia* (Figura 5.1), se identifica la cascada filogenética marcada en doble línea. Esta cascada está compuesta por los ancestros comunes ya desaparecidos (1), (2), (3)...(7). Estos 7 organismos descienden uno de otro, mientras que otros como *Critidia* y *Leptomona* descienden de algún ancestro común teniendo un parentesco muy alto.

Estos siete organismos se encuentran desaparecidos, sin embargo podemos identificar entre los organismos que han sobrevivido, aquellos más cercanos evolutivamente a los ancestros desaparecidos. Así el organismo más cercano a (1) es *Bodo* o *Tripanosoma borelli*, el más cercano a (2) es *T. brucei*, el más cercano a (3) es *T. cruzi* o *Tripanosma*

sp, y así hasta el más cercano a (7) que es la *Critidia* o *Leptomona*. Por lo tanto, de los organismos que han sobrevivido hasta la actualidad, el mejor orden evolutivo es:

0	*Euglena*
1	*T.borelli* o *Bodo*
2	*T.brucei*
3	*T.cruzi* o *T. sp*
4	*Blastocritidia*
5	*Fitomona* o *Herpetomona*
6	*Leishmania* o *Endotripanosoma*
7	*Critidia* o *Leptomona*

Se analizaron las secuencias de ADN nuclear codante reportadas en la base de datos genómica del Entrez NCBI y se calculó el contenido total de bases G,C para cada uno de los organismos anteriores a partir de dichas secuencias. Para determinar el %GC, se empleó el software DNASIS y un algoritmo que desarrollamos en Turbo Pascal. Básicamente el cálculo se realiza agrupando a todos los nucleótidos en un mismo conjunto, a partir del cual se calcula el %GC total. Los resultados se muestran en la tabla 5.1:

Especie	%GC	(Región codante Nuclear)
T. brucei	43.50	ADN nuclear codante
T. cruzi	52.96	ADN nuclear codante
Fitomona	66.67	ADN nuclear codante
Critidia	61.80	ADN nuclear codante
Endotrypanosoma	60.92	ADN nuclear codante
Leishmania	59.76	ADN nuclear codante

Tabla 5.1 Contenido de GC de algunos *Kinetoplastidia*

Se observa que el *Endotripanosoma* tiene un %GC muy cercano a la *Leishmania*, así como la *Leptomona* con *Critidia*, lo cual es consistente con el hecho de que estos organismos están evolutivamente muy emparentados. No se encontró reportado ninguna secuencia codante nuclear para *Blastocritidia* ni el resto de organismos mencionados. La *Fitomona* tiene reportado ADN codante nuclear con un %GC bastante alto (66 %), lo cual será discutido más adelante. Con esta información se elaboró la Figura 5.2 en donde se muestra la variación del %GC en función del orden filogenético para los seis organismos, y una línea de tendencia de crecimiento logarítmico que mejor se ajusta.

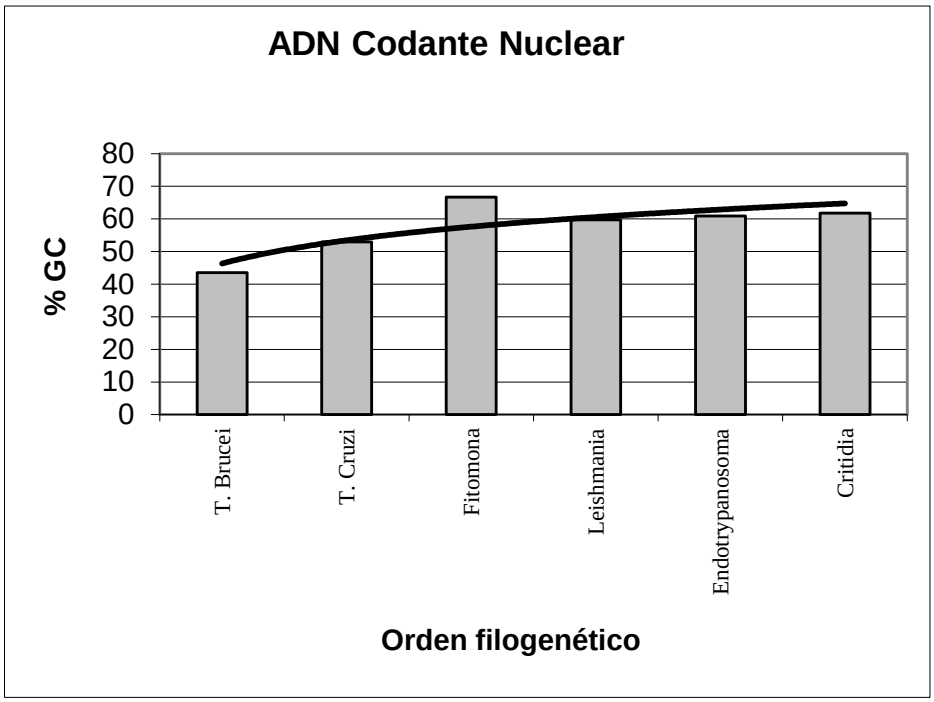


Figura A3.2. Contenido de GC vs. orden filogenético para el ADN codante nuclear

Las predicciones del modelo con que podemos comparar esta última información corresponde a la Figura 4.2, en donde se simula un proceso de diversificación a partir de una secuencia codante con 44% GC, que es equivalente al del *T. brucei* quien tiene un %GC codante nuclear cercano a 44%. Se puede notar que cualitativa y cuantitativamente ambas curvas (Figura 4.2 y 5.2), son muy similares. Nótese que el eje horizontal de la Figura 5.2 presenta a los organismos en un orden arbitrario, debido a que éstos no han sido fechados. Esto será discutido en el próximo capítulo.

Existe un trabajo similar realizado por el grupo de Alonso [Alonso *et al*.,1992], en el que se han analizado regiones codantes para ciertos genes de algunos *Kinetoplastidia*, *Leishmania* y *Critidia*. La tabla 5.2 muestra el resultado de dicho trabajo: El %GC en la primera, segunda y tercera posición de codón, así como el %GC total. El ADN considerado es codante nuclear e incluye secuencias flanqueadoras.

	% GC total	1ra	2da	3ra
T.brucei	44	57	40.8	56.9
T.cruzi	51	59.6	43.7	67.3
Leishmania	57	59.5	44.6	84.6
Critidia	58	62.7	43.1	88.9

Tabla 5.2 Contenido de GC en las tres posiciones de codón [Alonso *et al*.,1992]

A partir de estos datos se elabora la Figura 5.3: %GC en la primera segunda y tercera posición del codón, versus %GC total.

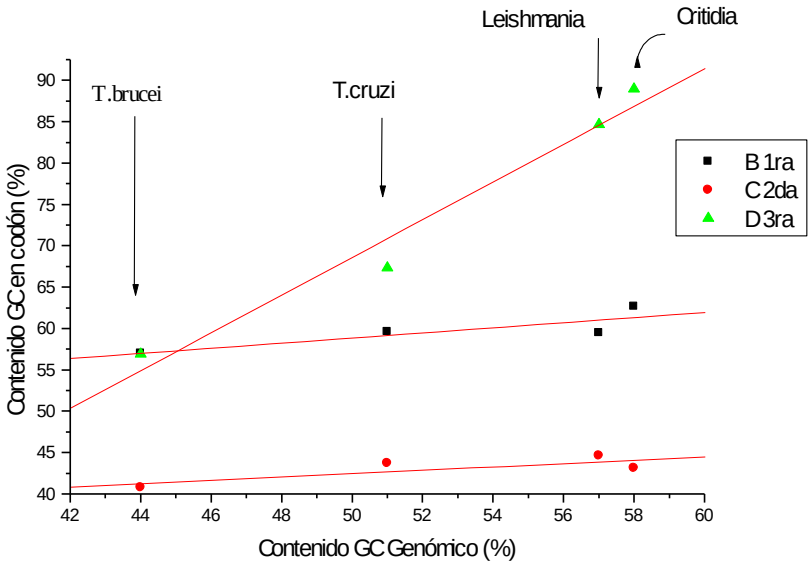


Figura A3.3. Variación del contenido GC en las posiciones de codón, en función del contenido de GC total [Alonso *et al*.,1992]

Se observa claramente que existe una relación lineal entre las variables, así como una mayor variación a nivel de la tercera posición del codón, lo cual indica que la mayoría de las mutaciones han ocurrido a nivel de esta posición. Estos datos reales correspondientes a la Figura 5.3, son equivalentes a la predicción del modelo mostrada en la Figura 4.11. El mismo grupo [Alonso *et al*.,1992], realizó un trabajo más detallado, midiendo el contenido de GC de la región codante además de las regiones flanqueadoras 5´ y 3´ individualmente (tabla 5.3).

	%GC total	%GC codante	región 5′	región 3′
T.brucei	44	51.6	40.9	43.6
T.cruzi	51	56.9	41.6	41.0
Leishmania	57	62.9	55.5	58.3
Critidia	58	64.9	44.3	50.5

Tabla 5.3 Contenido de GC en las regiones flanqueadoras [Alonso *et al*.,1992]

En esta tabla puede observarse que las regiones flanqueadoras 5´ y 3´ presentan un incremento del %GC mucho más lento que el correspondiente a la región codante. Esto hace pensar que las regiones flanqueadoras son más conservadas. Creemos que esto se debe a que las regiones flanqueadoras 5' y 3' (separadores intergénicos) desempeñan funciones importantes en la regulación post transcripcional. En *Leishmania* y Tripanosomátidos, se observa que un arreglo de varios genes se encuentran regulados por un sólo promotor, lo cual da lugar a un mARN policistrónico. Poliadenilaciones, cortes e incorporaciones de un miniexón por trans splicing, ocurren justamente a nivel de las regiones flanqueadoras, las cuales son reconocidas por enzimas especializadas. Sería de esperar que las regiones flanqueadoras 5' sean aún más conservadas que las correspondientes 3', debido a las funciones de traducción adicionales que deben cumplir. En concordancia con esta hipótesis se observa que la región flanqueadora 3' presenta una mayor tasa de variación del %GC que la región 5'.

El incremento del contenido de bases GC, tanto en el genoma total como en las tres posiciones del codón, sugiere la existencia de un bias o sesgo en el uso de codones (codon usage), por parte de las distintas especies. Esto se debe a que la única manera de incrementar el contenido de GC genómico y conservar el aminoácido o tipo de aminoácido, es sencillamente reemplazando los codones por sus equivalentes ricos en GC. A partir de la distribución de codones para los distintos aminoácidos presentados en el apéndice 4 (tabla A4.1) [Alonso *et al*.,1992], se elaboran las figuras 5.4 y 5.5, en las que se aprecia la distribución de codones de leucina y serina para *T.brucei, T.cruzi*, *Leishmania* y *Crithidia* (no están procesados aún los datos para los demás organismos). Se observa claramente la tendencia de acumular codones ricos en GC a lo largo del tiempo, ya que *T.brucei* es el organismo más antiguo, y *Critidia* el más moderno. Este comportamiento se ha observado en todos los aminoácidos para estos cuatro organismos.

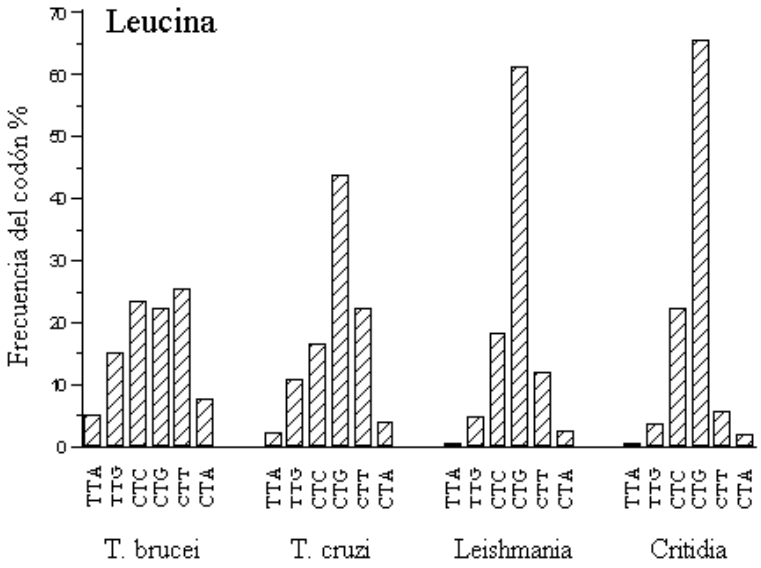


Figura A3.4. Distribución de codones para leucina

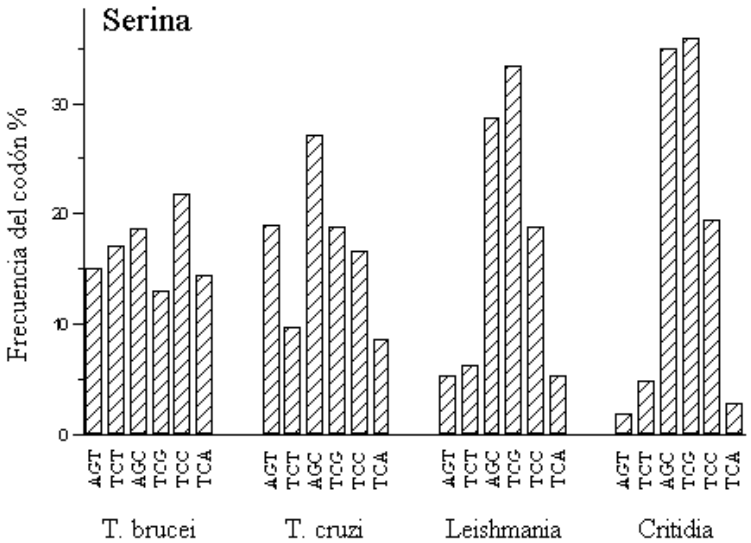


Figura A3.5. Distribución de codones para serina

En lo que respecta al ADN no-codante-basura, éste no se encuentra reportado en una cantidad razonable para los organismos de nuestro interés, ni tampoco se puede garantizar que sea realmente un ADN basura. Ante esto no se tiene evidencias para contrastar las predicciones del modelo en cuanto a la evolución del ADN no-codante-basura.

Una prueba adicional consiste en el análisis individual de algunos genes para estos organismos. Así se analizaron los genes hsp70, triosa fosfato isomerasa (TPI), y ubiquitina para tres de los organismos tratados arriba (tabla 5.4).

Organismo		HSP70	TPI	Ubiquitina
T. brucei	%GC	46.48	45.6	38.2
T. cruzi	%GC	55	No se encontró	44.7
Leishmania	%GC	63.8	56.8	65.0

Tabla 5.4 Contenido de GC en genes individuales

La información de la tabla 5.4 muestran que cada uno de los tres genes presenta un incremento sistemático del %GC, lo cual sugiere que genes individuales podrían evolucionar impulsados por la presión mutacional termodinámica, lo cual permitiría hablar de la evolución de un sólo gen.

Algunos aspectos de la evolución del linaje del *Plasmodium*

Un linaje bastante estudiado y muy interesante corresponde al genus *Plasmodium*. Varias especies de *Plasmodium* se caracterizan por tener un contenido de GC bastante bajo (27-28 %), lo cual le confiere características peculiares.

Trabajos de filogenia para el linaje del *Plasmodium* fueron realizados a partir de secuencias pequeñas de ARN ribosomal [Escalante A. & Ayala F., 1994], empleando el algoritmo DNAML del software Phylip.

Determinamos el contenido de GC para las secuencias de genes codantes nucleares de las 11 especies de *Plasmodium* analizadas por Escalante & Ayala. La información fue obtenida de la base de datos genómica del Entrez NCBI.

Código	Especie	%GC
Pfa	*P. Falciparum*	28.66
Pma	*P. Malarie*	37.70
Pvi	*P. Vivax*	46.47
Pre	*P. Reichenowi*	27.14
Pfr	*P. Fragile*	39.72
Pkn	*P. Knowlesi*	38.14
Pcy	*P. Cynomolgi*	42.99
Pbe	*P. Berghei*	30.71
Pga	*P. Gallinaceum*	29.61
Plo	*P. Lophurae*	No reportado
Pme	*P. Mexicanum*	No reportado

Tabla 5.5 %GC para el genus *Plasmodium*

No se encontraron secuencias codantes nucleares tanto para *P. Lophurae* como para *P. Mexicanum* en ninguna de las bases de datos genómicas actualizadas hasta la fecha.

Estas 11 especies de *Plasmodium* (phylum Apicomplexa), son protozoarios muy variables que parasitan reptiles, aves o mamíferos. El más virulento es el *Plasmodium falciparum*, el agente causante de la letal malaria. Por estas razones deben haber sufrido una presión de selección muy parecida a la del linaje de los *Kinetoplastidia*.

El árbol filogenético para estas especies se aprecia en la Figura 5.5, [Escalante A. & Ayala F., 1994], en donde se ha colocado el %GC de las secuencias codantes nucleares, entre paréntesis al costado del nombre de la especie.

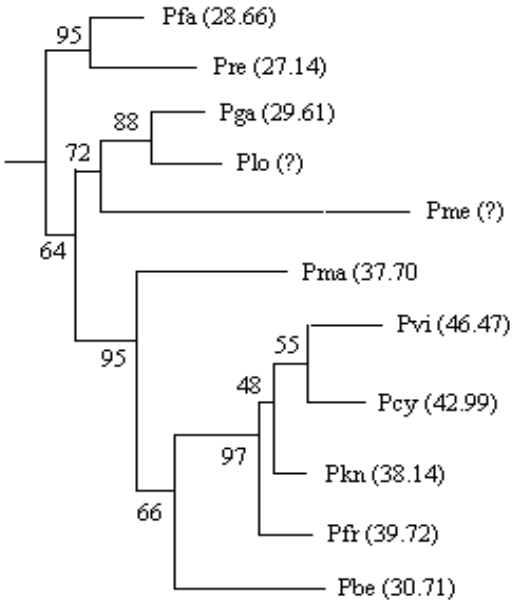


Figura A3.6. Árbol filogenético del genus *Plasmodium,* inferido a partir de secuencias de genes 18S SSU rRNA, derivado por el método N.J. del programa Phylip [Escalante A. & Ayala F., 1994]

A partir del árbol filogenético de la Figura 5.5, es posible distinguir la cascada filogenética, y ordenar a todas las especies de acuerdo a un orden evolutivo, ya que se trata de un árbol cuantitativo. Así se tiene la Figura 5.6 mostrando el %GC del ADN codante nuclear, en función del orden filogenético, y una línea de tendencia de crecimiento logarítmico que mejor se ajusta.

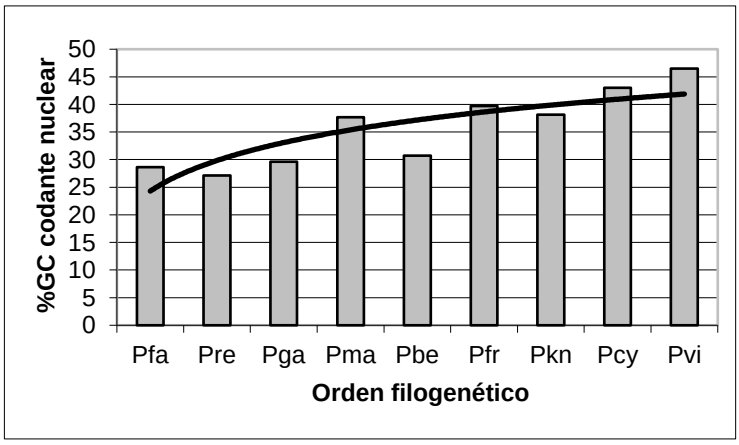


Figura A3.7. %GC codante nuclear para el linaje del *Plasmodium*

Anexo 4: Publicación científica -DNA thermodynamic pressure: a potential contributor to genome evolution.

Zimic MJ, Guerra D, Arévalo J. DNA thermodynamic pressure: a potential contributor to genome evolution. Trans R Soc Trop Med Hyg. 2002 Apr;96 Suppl 1:S15-20. doi: 10.1016/s0035-9203(02)90046-5. PMID: 12055830.

Esta publicación muestra el modelo de simulación computacional mencionado y los resultados como consecuencia de la presión mutacional termodinámica

TRANSACTIONS OF THE ROYAL SOCIETY OF TROPICAL MEDICINE AND HYGIENE (2002) 96, SUPPLEMENT 1 S1/15–S1/20

Population genetics

DNA thermodynamic pressure: a potential contributor to genome evolution

Mirko J. Zimic[1], Daniel Guerra[1] and Jorge Arévalo[1,2] *[1]Division de Bioquímica y Biología Molecular, Departamento de Ciencias Fisiológicas; [2]Instituto de Medicina Tropical 'Alexander von Humboldt,' Universidad Peruana Cayetano Heredia, Lima, Perú*

Abstract

Codon usage bias is a feature of living organisms. The origin of this bias might be explained not only by external factors but also by the nature of the structure of deoxyribonucleic acid (DNA) itself. We have developed a point mutation simulation program of coding sequences, in which nucleotide replacement follows thermodynamic criteria. For this purpose we calculated the hydrogen bond-like and electrostatic energies of non-canonical base pairs in a 5 bp neighbourhood. Although the rate of non-canonical base pair formation is extremely low, such pairs occur with a preference towards a guanine (G) or cytosine (C) rather than an adenine (A) or thymine (T) replacement due to thermodynamic considerations. This feature, according to the simulation program, should result in an increase in the GC content of the genome over evolutionary time. In addition, codon bias towards a higher GC usage is also predicted. DNA sequence analysis of genes of the Trypanosomatidae lineage supported the hypothesis that DNA thermodynamic pressure is a driving force that impels increases in GC content and GC codon bias.

Keywords: leishmaniasis, trypanosomiasis, *Leishmania donovani*, *Leishamania chagasi*, *Trypanosoma brucei*, *Trypanosoma cruzi*, *Crithidia fasciculata*, codon usage bias, DNA thermodynamic pressure, genome evolution

Introduction

After the first prokaryote and eukaryote gene sequences were compiled and analysed, it was concluded that synonymous codons were not used at random but showed considerable bias (GRANTHAM *et al.*, 1980a, 1980b). In addition, at least for unicellular organisms, codon usage was quite similar within the same genome and among taxonomically related species. This fact led to the proposal of the genome hypothesis, that each genome would have a coding strategy. This hypothesis was, however, not apparently valid when higher eukaryotes were considered (AOTA & IKEMURA, 1986), although there is some evidence for a relation between codon bias and phylogeny (MARIN *et al.*, 1989). Although these organisms showed a strong bias in codon usage, there are marked differences among genes within the same genome. This is particularly remarkable for warm-blooded vertebrates, where discrete, long stretches of deoxyribonucleic acid (DNA) with high gene density exhibited more guanosine + cytosine (GC) content than the rest of the genome (BERNARDI *et al.*, 1985; BERNARDI, 1989). The genes located in these GC-rich compartments, named isochores, showed stronger usage bias towards GC-rich codons than those present in GC-poor genome segments.

Different causes, not mutually exclusive, have been claimed to be responsible for non-random codon usage: availability of transfer ribonucleic acid (tRNA) and the nature of the codon–anticodon interaction (IKEMURA, 1981, 1982), levels of gene expression (SHARP *et al.*, 1986), context effect (BULMER, 1990; NUSSINOV, 1990), and gene location within the transcription unit (DELORME & HÉNAUT, 1991). On the other hand, GC content differences among living organisms were explained either by differences in directional mutation pressure (SUEOKA, 1988) or because thermal stability favourably selects GC-rich genome organisms to adapt them to conditions found in warm-blooded animals (BERNARDI *et al.*, 1985).

Because the double helix DNA structure is responsible for most of the relevant features of the genetic material, the present report explores the possibility that this physical structure contributes to the observered general increase in GC content during evolution and the codon usage bias towards GC. The potential contribution of nucleotide structure to favour non-random combinations was first addressed by ROWE & TRAINOR (1983). They used thermodynamic considerations in an Ising-like model of DNA, and showed how codon bias may have been imposed at the time primitive DNA first formed in the 'primordial soup'.

In the present paper, the nucleotide composition bias of living organisms was approached using Boltzmann's probability distribution and Monte-Carlo simulations. These permitted calculation of probabilities of occurrence of point mutations, and predicted a spontaneous increase in the GC content of DNA during long periods of time.

Trends observed during simulations were compared with reported data for the eukaryote trypanosomatid lineage, which constitutes a suitable microevolutionary lineage, to assess the thermodynamic pressure hypothesis that we propose. The trypanosomatids *Trypanosoma brucei*, *T. cruzi* and *Leishmania* are protozoa that infect humans. The first of these microorganisms is an extracellular parasite and the other 2 are intracellular for at least part of their life cycle in mammals. Therefore, they are within an environment subject to strict vertebrate homeostasis mechanisms. In addition, we have included *Crithidia*, a trypanosomatid that lives within insects and is therefore devoid of thermoregulatory constraints. These protozoa have clonal reproduction (REVOLLO *et al.*, 1998), their genes lack introns, and there is a good phylogenetic tree that permits a reasonable microevolutionary analysis (MASLOV *et al.*, 1994).

Materials and Methods

A computational simulation of point mutations with a Monte-Carlo algorithm based on Boltzmann probability distribution has been developed. The Boltzmann distribution was used to calculate the occurrence probability of a point mutation, considering the DNA molecule as a canonical ensemble. This assumption can be made because, within the cell, temperature, volume and pressure do not change appreciably. A point mutation is preceded by what we denote as the formation of a 'hole', which we define as a discontinuity of one or more nucleotides in one of the strands of the DNA molecule, without total disruption of the molecule. Every hole should be occupied and sealed with a new free nucleotide. Assuming that there is no external force field to

Address for correspondence: Jorge Arevalo, Laboratorio de Bioquimica, Universidad Peruana Cayetano Heredia, Avenida Honorio Delgado 430, Lima 31, Perú; fax +51 126 40535, e-mail jazz@upch.edu.pe

influence the chemical identity of the nucleotide to occupy the hole, the incoming nucleotide would be guided only by molecular forces between it and the vicinity of the hole. It would be expected that the system should, probabilistically, attain the smaller potential energy level. Although the most probable occupying nucleotide should be the one complementary to that present in the opposite strand, i.e., adenine + thymine (AT) and GC, there is a chance that non-complementary base pairs form. The force that impels the occupation of holes is here denoted the thermodynamic pressure.

Probability calculation

Once a hole (X) is formed, it could be occupied by any of the 4 free nucleotide types. The probability that a hole X is occupied by one base N is calculated with the aid of the Boltzmann distribution:

$$P(\mathrm{X}=\mathrm{N}) = U\,\mathrm{e}^{-\beta.E(\mathrm{X}=\mathrm{N})} \qquad (1)$$

where β is equal to $1/KT$ (K is Boltzmann's constant and T the absolute temperature), U is the number of accesible states, and $E(\mathrm{X}=\mathrm{N})$ is the energy of the DNA molecule when the hole is occupied by the base N. The caclulation of the number of accessible states requires analysis of the system entropy, since both magnitudes are related through

$$U \propto \mathrm{e}^{S/K} \qquad (2)$$

(REIF, 1965), where S is the entropy. Because the hydrophobic effect is entropic in nature, the number of accessible states depends in principle on the hydrophobic interaction with stacked bases in the vicinity of the hole. As a first approximation, the number of accessible states of each configuration could be considered proportional to the number of free triphosphate nucleotides in the proximity of the hole. Considering the volume (V) for the DNA–hole configuration and its nearby surroundings, the number of accessible states for the case in which a hole is occupied by a deoxynucleotide triphosphate (dNTP) can be estimated according to

$$U = \gamma[\mathrm{dNTP}]\,V \qquad (3)$$

where [dNTP] is the free triphosphate nucleotide concentration and γ is a proportionality constant that absorbs the hydrophobic effects. This was assumed to be the same for each type of nucleotide. A simplification was performed, after normalizing the probabilities with respect to one of them (i.e., $P(\mathrm{X}=\mathrm{G})$). Thus, the absolute probabilities become relative as shown in the equation 4:

$$P'(\mathrm{X}=\mathrm{G}) = P(\mathrm{X}=\mathrm{G})/P(\mathrm{X}=\mathrm{G}) = 1$$

$$P'(\mathrm{X}=\mathrm{A}) = P(\mathrm{X}=\mathrm{A})/P(\mathrm{X}=\mathrm{G}) \qquad (4)$$

The same equation was applied to the other two nucleotides. Thus, for example, the relative probability for a hole being occupied by an adenine would be

$$P'(\mathrm{X}=\mathrm{A}) = \frac{[\mathrm{A}]}{[\mathrm{G}]}\,\mathrm{e}^{-\beta(E(\mathrm{X}=\mathrm{A})-E(\mathrm{X}=\mathrm{G}))} \qquad (5)$$

The same procedure was used to calculate the probabilities for other nucleotides. In the present work, normalization was performed with respect to the highest probability. After normalization, occupation probabilities depend only on the energy difference of 2 configurations. This simplifies the problem, because some energy terms that are independent of the sequence configuration do not need to be calculated. In any sequence configuration, the total energy has several terms, which are categorized as the bond energy terms (covalent bonds, bond angles, dihedral angles) and the non-bonding terms (electrostatic energy, Van der Waals interactions, hydrogen bonds, magnetostatic energy). After normalization, it suffices to include only interaction energies between the nucleotide occupying the hole and the neighbouring bases, for which the hydrogen bond, electrostatic, Van der Waals, and magnetostatic terms should be taken into account. Because of the relative order of magnitude of these energies, only hydrogen bonds and electrostatic interactions were considered. The other forces are so weak that they can be neglected.

Electrostatic energy calculation

The electric charge distribution of each nucleotide was calculated starting from the dipolar moments (p) and distances (d) of the covalent bonds. A discrete charge distribution was estimated, where partial charges (q) of opposite sign are placed within each pair of atoms linked by a covalent bond. Since $p = qd$, the value of each partial charge and its position within the nucleotide was determined. The electrostatic energy could be calculated from the following equation:

$$E = \frac{1}{4\pi\kappa\varepsilon_0}\sum\frac{q_i q_j}{|\overline{r_i} - \overline{r_j}|} \qquad (6)$$

where r_i is the position of the partial charge q_i, and ε_0 is the vacuum permitivity. The discrete constant (κ) has been taken as 80 because of the neighbouring water. The summation domain should consider all the possible interactions among partial charges of different nucleotides, excluding those belonging to the same nucleotide. The geometry assumed for the DNA double strand was a linear arrangement of 2 parallel strands. This assumption is justified below (Results and Discussion, Vicinity size).

Hydrogen bond-like energy calculation

Hydrogen bond energies among canonical base pairs were calculated theoretically from quantum mechanics as follows: $E(\mathrm{A}-\mathrm{T}) = -0{\cdot}34$ eV ($-7{\cdot}837$ kcal/mol), and $E(\mathrm{G}-\mathrm{C}) = -0{\cdot}43$ eV ($-9{\cdot}912$ kcal/mol) (SEPRODI *et al.*, 1969). However, hydrogen bond energies for non-canonical base pairs have not yet been calculated or experimentally measured. In the present work, hydrogen bond energies for the non-canonical base pairs were estimated with a very simple, classical approach. The hydrogen bond-like energy could be taken, in a first approximation, as the electrostatic interaction energy among the dipoles (E_{12}) that participate in the hydrogen bond, corrected by an additive constant λ (equation 7). The value of λ used allowed the best fit with the canonical base pairs' theoretical energies as shown above.

$$\mathrm{E}_{12} = \frac{1}{4\pi\kappa\varepsilon_0}\left[\frac{p1.p2}{r^3} - \frac{3(r.p1)(r.p2)}{r^5}\right] + \lambda \qquad (7)$$

where $p1$ and $p2$ are the dipolar moments of the residues that contribute to the hydrogen bond, and r is the connection vector between them. The dielectric constant was considered as 1.

Computational algorithm

A point mutation simulation program has been implemented in several codes for the Turbo Pascal v. 7.0 Borland compiler. These codes enabled us to make simulations using 2 different scenarios: firstly, non-coding/non-functional DNA sequences, and, secondly, coding DNA sequences where conservation of the amino acid or of the amino acid family was imposed. In addition, the simulation considered a hypothetical repair mechinism.

The algorithm consisted of the following steps. (i) The DNA sequence and the simulation paramaters (temperature, concentrations of the free nucleotides, kinetics of the repair mechanism, working scenario) were read. (ii) Starting from a uniform probability distribution along the whole sequence, a hole was generated in one of the 2 strands. (iii) The vicinity (5 bp) was recognized and the electrostatic and hydrogen bond-like energies were calculated for the 4 possible configurations, in which the hole was occupied by either G or C, or by T or A. (iv) The probability of occupation of the hole by each of the nucleotides was calculated using the Boltz-

mann distribution. (v) A Monte-Carlo algorithm determined the nucleotide that occupied the hole. (vi) The action of a repair mechanism was simulated. An arbitrarily neutral repair mechanism was assumed (in the sense that it did not have any preference to conserve any type of base), with an efficiency of 50% (half of the time it corrected the mismatch, and the other half it made a mistake, leading in this last case to a point mutation). At this time, a mutation could have occurred. (vii) Restrictions were imposed on mutations, depending on the working scenario. In the case of the non-coding/non-functional DNA, all mutations are accepted. By contrast, 2 types of mutations were considered for coding DNA: neutral mutations (the amino acid is conserved) and mutations that conserve the amino acid family; the rest of the possible mutations were excluded by considering them lethal. (viii) GC content of the sequence was computed and plotted against the number of the generated holes. (ix) Processes (ii) to (viii) were repeated, beginning with the last sequence obtained after step (vii) for a certain number of times (*c.* 20 million holes for a sequence of approximately 1000 bp).

Analysed data

Accession numbers for the gene sequences are as follows: *HSP70*: *T. brucei*, L14477; *T. cruzi* M26595; *L. donovani*, X52314H. *HGPRT*: *T. brucei*, L10721; *T. cruzi*, L07486; *L. donovani*, L25412; *C. fasciculata*, U19968. *TOP2*: *T. brucei*, M26803; *T. cruzi*, M91165; *L. chagasi*, AF051307; *C. fasciculata* X59623. *TryR*: *T. brucei*, X63188; *T. cruzi*, M38051; *L. donovani*, Z23135; *C. fasciculata*, Z12618. Homologous blocks were detected with the Macaw (Multiple Alignment Construction & Analysis Workbench, version 2.0.5) program using the Gibbs sampling method.

Results and Discussion

The unquestionably observed deviation of DNA sequence mutations from randomness was here explored considering the potential contribution of the DNA double-stranded structure. The following sections give theoretical support to the postulate that thermodynamic pressure is a relevant factor in explaining DNA sequence changes over evolutionary time scales. The section below on sequence and data analysis compares predictions from simulations with observations from the trypanosomatid lineage microevolutionary process.

Vicinity size

The electrostatic interaction energy between the nucleotide that occupies the hole and its neighbours decreases markedly with distance. After plotting the interaction energy versus the vicinity size (i.e., a vicinity size of $2N + 1$ represents a hole plus N bp upstream and N bp downstream), it was demonstrated that only the interaction energy within a 5 bp vicinity was relevant for the total interaction energy (data not shown). A 5 bp vicinity contributed significantly to the electrostatic interaction energy between the candidate for occupying the hole and its neighbouring bases. Further away, these interactions decrease until they reach a negligible level. A 5 bp vicinity of a double helix DNA arrangement determines almost a half-turn which, for simplicity, has been considered as 2 linear, parallel strands.

Hydrogen bond energy of non-canonical base pairs

The estimated hydrogen bond-like energy values were obtained by the method described above, corrected with a λ value of 0·12 eV (Table). The energy values marked with an asterisk in the table correspond to the interaction of free purine–purine base pairs. Nevertheless, these base pairs would be forbidden because, once immersed within a linear-double stranded vicinity, they would perturb the configuration to such an extent that the configuration would become energetically unfavourable.

Cases I and II in the Table correspond to different arrangements of the nucleotides in which they are accommodated so as to participate in significant hydrogen bond-like interactions. Under the assumption of a linear double-stranded arrangement of the 5 bp vicinity, type II configurations were also energetically prohibited (data not shown). Therefore, in our model the configurations marked with an asterisk and (II) have not been considered. It is necessary to point out, however, that the DNA double helix may be flexible enough to accommodate some of the forbidden base pairs. Nevertheless, because a real and complete scenario would be extremely difficult to address, we have decided to impose a simplification at this level, which should not affect the qualitative aspects of the predictions of this model.

Table. Estimated hydrogen bond-like energies for canonical and non-canonical free base pairs[a]

Base pair	eV	kcal/mole
A–T	−0·34	−7·84
G–C	−0·43	−9·91
C–A	0·21	4·84
T–G	0·13 (I) −0·40 (II)	2·99 (I) −9·22 (II)
T–C	−0·14	−3·22
A–G	−0·27	−6·22*
T–T	0·02 (I) −0·16 (II)	0·46 (I) −3·68 (II)
C–C	−0·01 (I) −0·14 (II)	−0·23 (I) −3·22 (II)
A–A	0·23	5·30*
G–G	0·54	12·45*

[a]Calculated according to the procedure described in the text. The value of λ was taken as 0·12 eV; an asterisk (*) indicates the accommodation of free purine–purine base pairs. (I) and (II) correspond to 2 different arrangements of the base pairs; in (II) a 0·12 nm displacement perpendicular to the plane of the deoxyribose has been taken into account.

The canonical base pairs exhibit the higher values (−0·34 and −0·43 eV for AT and GC, respectively). Among the non-canonical base pairs, TG and AG, when existing as free base pairs or within a flexible structure, are the more stable (−0·40 and −0·27 eV, respectively), followed by TC, TT and CC. The relative order of these base pairs has been supported experimentally (L. Marky, personal communication).

Predictions of the thermodynamic pressure

When the Boltzmann distribution was used to simulate the change in GC content over a long period of time, continuous increase of these nucleotides was observed until a plateau was reached. The value at the plateau was dependent on the sequence information. Thus, if the sequence did not have any coding role, and therefore could mutate freely, the plateau attained values over 95%, although it never reached 100% (Fig. 1, A). However, if the simulation considered a coding sequence the observed plateau ranged between 60% and 75% when amino acid family conservation was imposed (Fig. 1, B). Moreover, when the restriction of keeping the amino acid identity constant was imposed the simulated sequences reached a lower plateau earlier, depending on when the restriction was imposed (Fig. 1, B). The simulations presented here were made considering equimolar nucleotide concentrations and a temperature of 37°C; however, if these parameters were modified the plateau value and the speed needed to reach it varied. At higher temperatures a higher GC content was obtained and a more rapid increase was observed (data not shown).

According to the simulation described above, any genome would evolve spontaneously towards a higher GC content under conditions where the only force acting on it was what we have referred to as the thermodynamic pressure. This force is self-contained in the nature of the DNA structure, a feature that distinguishes it from the selection pressure, which depends on many environmental factors that act like a sieve on populations.

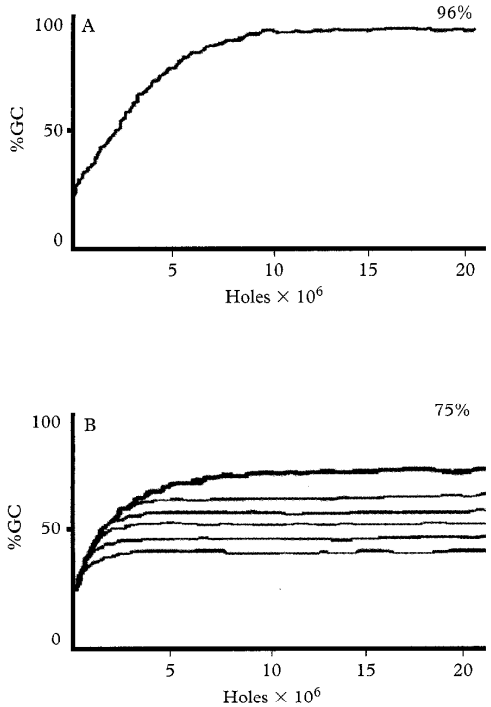


Fig. 1. Results of simulation of point mutation under themodynamic pressure. The conditions selected were an initial 22% GC content of a 1000 bp random sequence at 37°C. Equimolar concentrations of dATP, dTTP, dGTP, and dCTP and a neutral repair system were used (see text). A. No restriction was imposed on point mutations. B. Restrictions of conservation of amino acid family and amino acid identity were imposed. The bold line represents the mutational process under amino acid family conservation. Thin lines represent various mutational processes under amino acid identity conservation, with the starting sequences for each case having been arbitrarily chosen from different stages of the simulation with amino acid family restriction.

Sequence and data analysis

To assess the validity of the postulated thermodynamic pressure, we have chosen a situation where organisms are exposed to a limited and rather constant selection pressure. Analysis of the trypanosomatid gene sequences deposited in GenBank revealed biased codon usage (ALONSO *et al.*, 1992; ALVAREZ *et al.*, 1994). The first report claimed that mutational pressure towards either GC or AT was responsible for the observed codon usage divergence. Nevertheless, when the lineage members mentioned above were phylogenetically ordered from older lineages to those that originated more recently, based on the small subunit ribosomal RNA (MASLOV *et al.*, 1994), we noticed that they showed a clear trend to increase in GC content (Fig. 2). This fact could reflect the thermodynamic pressure and not, as postulated by ALONSO *et al.* (1992), the reminiscence of primeval genomes. The observed global increase in GC content in trypanosomatid gene sequences may be an artefact, resulting from the final balance between different genes, some of them with high AT content and others with high GC content. If this were true, *Leishmania* and *Crithidia* should have a higher proportion of GC-rich genes, whereas *T. brucei* would have the opposite. Alternatively, the higher GC codon usage of modern trypanosomatids could reflect the evolution of most if not all their genes towards a higher GC content. The thermodynamic pressure hypothesis supports the second of these 2 options.

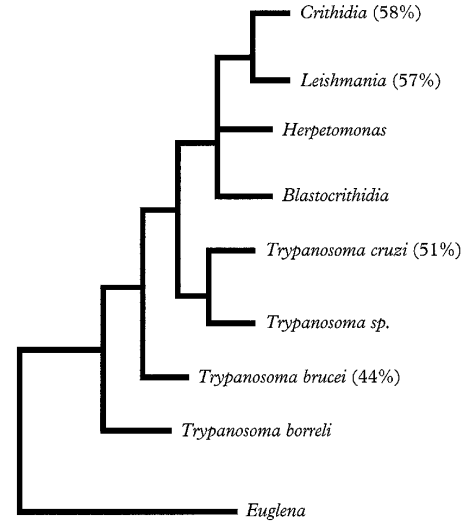


Fig. 2. Phylogenetic tree of trypanosomatids based on 18S ribosomal RNA (MASLOV *et al.*, 1994). The GC content of coding regions is shown in parentheses for some species (ALONSO *et al.*, 1992).

So far, we have analysed more than 20 different genes that have been reported for 2 or more trypanosomatid species. None of them presented a bias toward an AT increase (data to be published elsewhere). As an example, Fig. 3 illustrates 4 different genes that demonstrate enrichment of GC content in *Leishmania* and *Crithidia*, whereas the oldest species, *T. brucei*, has the lowest GC content in the corresponding genes; *T. cruzi* occupies an intermediate position. Moreover, as might be expected, the codon families (either quartet or sextet) of modern trypanosomatids used codons richer in GC composition than the codons used by *T. brucei* (Fig. 4). The model presented here implies that evolving DNA is not in thermodynamic equilibrium. Thermodynamic pressure impelled ancestral AT-rich DNA molecules towards a maximal GC-rich state. Because of structural and thermodynamic considerations of nucleotide sequences, the DNA molecule did not mutate at random but displayed

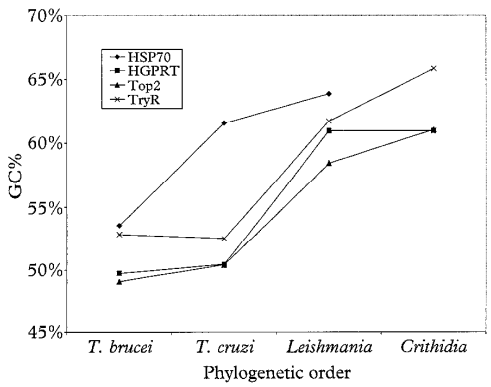


Fig. 3. Variation in GC content of specific genes (*HSP70*, *HGPRT*, *Top2*, *TryR*) of trypanosomatids. The horizontal order of the organisms follows the phylogenetic order shown in Fig. 2.

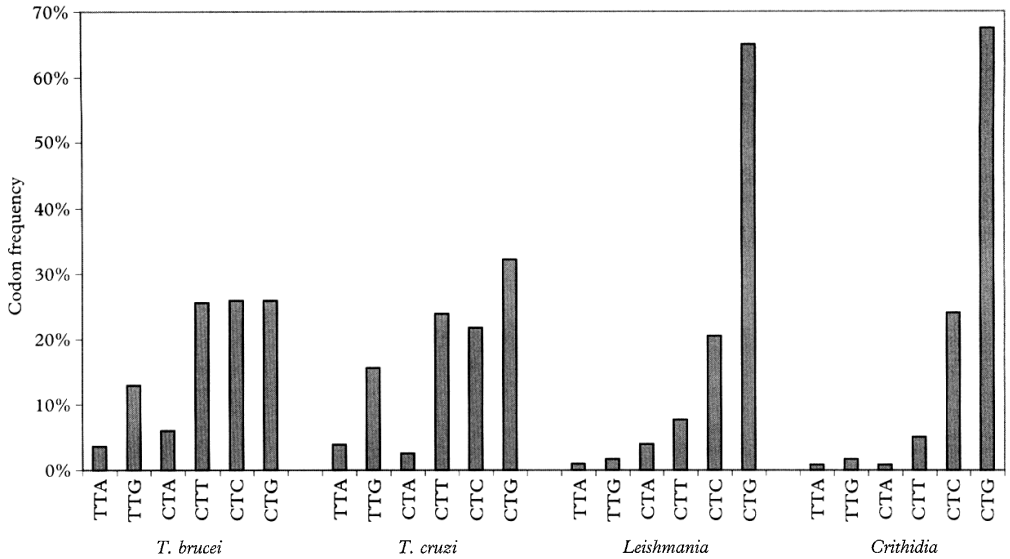


Fig. 4. Codon usage for leucine in 4 trypanosomatid species. Newer organisms show a clear tendency to use GC-rich codons. Similar results were obtained with the other amino acids (data not shown). The horizontal order of the organisms follows the phylogenetic order shown in Fig. 2.

bias towards a GC increase for both coding and non-coding sequences. As a consequence of this thermodynamic pressure, there is a trend towards a GC codon usage bias and to an increase of the GC content of the genome when microevolutionary scales are considered.

Recently, GALTIER *et al.* (1999) have found evidence that the most recent common ancestor of living organisms started with a high AT content, irrespective of any claimed need for thermal stability. The proposed scenario of AT-rich polymers makes sense because adenine nucleotides are energetically the least expensive to synthesize non-enzymatically (SCHWARTZ & BAKKER, 1989). Furthermore, short adenine-rich polymers would be more stable in a prebiotic environment, due to stacking forces.

The present hypothesis contradicts neither the mutational pressure (SUEOKA, 1988) nor the selection pressure (BERNARDI *et al.*, 1985; BERNARDI & BERNARDI, 1986) theories. The model hypothesized here predicts that genomes evolve towards a higher GC content, but they could eventually move towards AT-rich states if the intracellular environment presented different conditions or if selection pressures were strong enough to counteract the thermodynamic pressure. On the other hand, thermal selection pressures would have a synergic effect with thermodynamic pressure in the case of warm-blooded vertebrates. It is necessary to analyse a considerable number of genes and genomes of different lineages with good microevolutionary scales to test further the predictions made here.

Acknowledgements

This work was supported by a grant from CONCYTEC. We are very grateful to Holger Valqui and Oscar Moran for their critical comments, and to Claudia Machicado and José Chou for their comments and help in compiling and analysing part of the DNA sequences that support this work. Thanks also to Armando Bernui, Javier Espinoza, Jairzhinho Ramos, Luis Marky and Cristian Orrego, for their contribution to discussions, and especially to Mrs Ellen M. Pragen for her editorial work.

References

Alonso, G., Guevara, P. & Ramirez, J. (1992). Trypanosomatidae codon usage and GC distribution. *Memórias do Instituto Oswaldo Cruz*, **87**, 517–523.

Alvarez, F., Robello, C. & Vignali, M. (1994). Evolution of codon usage and base contents in kinetoplastid protozoans. *Molecular Biology and Evolution*, **11**, 790–802.

Aota, S. & Ikemura, T. (1986). Diversity in G + C content at the third position of codons in vertebrate genes and its cause. *Nucleic Acids Research*, **14**, 6345–6355.

Bernardi, G. (1989). The isochore organization of the human genome. *Annual Review of Genetics*, **23**, 637–661.

Bernardi, G. & Bernardi, G. (1986). Compositional constraints and genome evolution. *Journal of Molecular Biology*, **24**, 1–11.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. & Rodier, F. (1985). The mosaic genome of warm-blooded vertebrates. *Science*, **28**, 953–957.

Bulmer, M. (1990). The effect of context on synonymous codon usage with low codon usage bias. *Nucleic Acids Research*, **18**, 2869–2873.

Delorme, M. & Hénaut, A. (1991). Codon usage is imposed by the gene location in the transcript unit. *Current Genetics*, **20**, 353–358.

Galtier, N., Tourasse, N. & Gouy, M. (1999). A nonhyperthermophilic common ancestor to extant life forms. *Science*, **283**, 1–2.

Grantham, R., Gautier, C. & Gouy, M. (1980a). Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. *Nucleic Acids Research*, **8**, 1893–1913.

Grantham, R., Gautier, C., Gouy, M., Mercier, R. & Pavé, A. (1980b). Codon catalog usage and the genome hypothesis. *Nucleic Acids Research*, **8**, r49–r61.

Ikemura, T. (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *Journal of Molecular Biology*, **151**, 389–409.

Ikemura, T. (1982). Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. *Journal of Molecular Biology*, **158**, 573–597.

Marin, A., Betranpetit, J., Oliver, J. L. & Medina, J. R. (1989). Variation in G + C-content and the codon choice: differences among synonymous codon groups in vertebrate genes. *Nucleic Acids Research*, **17**, 6181–6189.

Maslov, D., Avila, H., Lake, J. & Simpson, L. (1994). Evolution of RNA editing in kinetoplastid protozoa. *Nature*, **368**, 345–348.

Nussinov, R. (1990). General nearest neighbour prefences in G/C oligomers interrupted by A/T: correlation with DNA

structure. *Journal of Biomolecular Structure and Dynamics*, **8**, 399–411.
Reif, F. (1965). *Fundamentals of Statistical and Thermal Physics*. New York: McGraw Hill.
Revollo, S., Oury, B., Laurent, J.P., Barnabé, C., Quesney, V., Carrière, V., Noel, S. & Tibayrenc, M. (1998). *Trypanosoma cruzi*: impact of clonal evolution of the parasite on its biological and medical properties. *Experimental Parasitology*, **89**, 30–39.
Rowe, G. & Trainor, L. E. H. (1983). A thermodynamic theory of codon bias in viral genes. *Journal of Theoretical Biology*, **101**, 171–203.
Schwartz, A. W. & Bakker, C. G. (1989). Was adenine the first purine? *Science*, **245**, 1102–1104.
Seprodi, L., Biczó, G. & Ladik, J. (1969). The effect of electric field on the electronic structure of DNA. Calculation of the polarizability and of the permanent dipole moment for the nucleotide bases and the base pairs. *International Journal of Quantum Chemistry*, **3**, 621–634.
Sharp, P. M., Tuohy, T. M. F. & Mosurski, K. (1986). Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Research*, **14**, 5125–5143.
Sueoka, N. (1988). Directional mutation pressure and neutral molecular evolution. *Proceedings of the National Academy of Sciences of the USA*, **85**, 2653–2657.