UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

Propiedades estructurales y termodinámicas del DNA y sus posibles implicaciones evolutivas

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

BIÓLOGA

P R E S E N T A :

GABRIELA SANTOS RODÍGUEZ

DIRECTOR DE TESIS:
DR. PEDRO EDUARDO MIRAMONTES VIDAL
2015

Ciudad Universitaria, D. F.

1. Datos del alumno

Santos
Rodríguez
Gabriela
56 87 43 26
Universidad Nacional Autónoma de México
Facultad de Ciencias
Biología
308598788

2. Datos del asesor

Dr.
Pedro Eduardo
Miramontes
Vidal

3. Datos del sinodal 1

Dr.
Claudia Andrea
Segal
Kischinevzky

4. Datos del sinodal 2

Dr.
Rafael
Camacho
Carranza

5. Datos del sinodal 3

Dr.
Pablo
Padilla
Longoría

6. Datos del sinodal 4

Quím.
Viviana
Escobar
Sánchez

7. Datos del trabajo escrito

Propiedades estructurales y termodinámicas del DNA y sus posibles implicaciones evolutivas.
83 páginas
2015

Índice general

Introducción **1**

Descripción de la molécula de DNA 1

Regulación de la expresión génica 15

Regulación de la expresión génica en procariontes 17

Regulación de la expresión génica en eucariontes 18

Evolución y DNA 20

Selección natural: origen e implicaciones 20

Síntesis moderna 21

Evolución molecular 22

Síntesis extendida 24

Organización del genoma 25

Estructura del genoma 25

Secuenciación de genomas 27

Secuenciación del genoma de *Escherichia coli* K-12 cepa MG1655 27

Proyecto del Genoma Humano 28

Genómica comparativa y evolución del genoma 29

Antecedentes **31**

El índice de homogeneidad, IDH 33

Justificación **36**

Hipótesis y objetivos **37**

Hipótesis 37

Objetivo general 37

Material y métodos **39**

Obtención de secuencias 39

Determinación del IDH a las secuencias de *H. sapiens* y *E. coli* K-12 MG1655 40

Resultados y discusión **43**

Conclusión y perspectivas **50**

Apéndice 1 **I**

Programa utilizado para descargar las secuencias codificantes CDS de *H.sapiens* I

Programa escrito en Python para determinar los valores de dWS, dYR y dMK IV

Programa para hacer las gráficas de dispersión de valores de IDH VIII

Programa escrito en R para la prueba de Welch X

Apéndice 2 **XII**

Parámetros utilizados en la aplicación Table Browser del UCSC Genome Browser XII

Índice de figuras

1. Bases nitrogenadas 2
2. Nucleótidos de DNA 3
3. Enlace fosfodiéster entre los nucleótidos de DNA 3
4. Esquema de dos nucleótidos y sus distancias relativas 4
5. Apareamientos tipo Watson y Crick 5
6. Esquema de la longitud total de dos nucleótidos apareados. 5
7. Surcos mayor y menor de la molécula del DNA 6
8. Esquema de la distancia entre cada grupo fosfato de la cadena de nucleótidos 7
9. Esqueleto de fosfatos de la molécula de DNA 8
10. Conformaciones A, B, Z del DNA 9
11. Esquema que muestra la estructura de un nucleosoma 11
12. Esquema que muestra las colas de histona en el nucleosoma. 12
13. Esquema que muestra fibra de 30 nm y la organización de los nucleosomas en ésta. 13
14. Esquema que muestra los diferentes niveles de empaquetamiento del DNA 14

15. Esquema que muestra los niveles de regulación de la expresión génica . . 16

16. Esquema que muestra el conjunto de elementos que forma un operon . . 18

17. Esquema que muestra los grupos funcionales disponibles en la molécula del DNA 19

18. Esquema de un elemento *Alu* 27

19. Grupos funcionales expuestos en los pares de bases del DNA en el surco mayor y menor 32

20. Resultados de Miramontes y colaboradores (1995) 35

21. Esquema que muestra la diferencia entre un ORF y una secuencia CDS . 40

22. Diagrama de flujo del programa en Python para obtener valores de IDH . 41

23. Gráfica de dispersión en tres dimensiones de los valores de IDH para ORF y CDS de *H. sapiens* 44

24. Gráfica de dispersión en tres dimensiones de los valores de IDH de los ORF de *H. sapiens* y *E.coli* 45

25. Gráfica de dispersión de en tres dimensiones de los valores de IDH para ORF de *H. sapiens* y secuencias alu 46

26. Histograma polar de los valores de IDH dWS, dYR y dMK de las secuencias Alu de *H. sapiens* 47

27. Diagrama descriptivo del programa en R para descargar secuencias CDS de *Homo sapiens* I

Propiedades estructurales y termodinámicas del DNA y sus posibles implicaciones evolutivas.

por

Gabriela Santos Rodríguez

Resumen

La molécula del DNA presenta variaciones estructurales debido a la diferente composición de bases nitrogenadas y a su interacción con el resto de elementos que la conforma. Estas variaciones estructurales son necesarias para la interacción con proteínas para la regulación de la expresión génica y para su compactación. La variación de la estructura de la molécula del DNA además de observarse en secuencias codificantes y no codificantes, se observa en los genomas de organismos de diferentes dominios.

En 1995 Miramontes y colaboradores desarrollan el índice de homegeneidad, IDH, para determinar la variación estructural de secuencias codificantes (CDS) de quince organismos reportadas hasta el momento. Las propiedades estructurales que consideraron en su análisis fueron la cantidad de enlaces de hidrógeno, si la base nitrogenada es púrica o pirimidínica y el tipo de grupo funcional que presentaban las bases nitrogenadas. De acuerdo a estas propiedades estructurales concluyeron que la molécula del DNA presenta un estilo estructural de acuerdo al dominio taxonómico del organismo del que proviene y tiene un fenotipo interno con sus propias presiones de selección y evolución, siendo la organización estructural un factor tentativo de selección.

Debido al aumento de la secuenciación de genomas completos de diferentes organismos, en este trabajo se extendió el análisis realizado por Miramontes y colaboradores en 1995. Se determinó el IDH de los $19,239$ marcos abiertos de lectura (ORF) de *Homo sapiens* del ensamblado del genoma GRCH38, los $4,294$ ORF del genoma de *Escherichia coli* K-12 MG1655 con registro en NCBI NC_000913.3. Además se determinó el IDH de $320,089$ secuencias Alu de *H. sapiens*. Los valores de IDH de las tres variaciones estructurales de las diferentes secuencias analizadas se graficaron en tres dimensiones para comprobar si se agrupaban de acuerdo a su estilo estructural. La gráfica de dispersión de las nubes de puntos de los valores de IDH de los ORF de *H. sapiens* y de *E. coli* K-12 MG1655 muestra que las secuencias se agregan de manera independiente dependiendo del organismo del que provienen. La independencia de ambos conjuntos de datos se confirmó con la prueba estadística de Welch. La gráfica de dispersión de las nubes de puntos de los valores de IDH de las secuencias Alu y los ORF de *H. sapiens* muestra que se agregan en la misma región. Sin embargo la prueba de Welch no confirmó que ambas muestras

tuvieran la misma distribución. Debido a esto se analizó la distribución de los valores de IDH de las tres variaciones estructurales de la molécula del DNA de las secuencias Alu. A partir de este análisis se encontró que la distribución de los valores de IDH para el tipo de base en las secuencias Alu presenta una distribución bimodal, por lo cual se invalida el resultado de la prueba de Welch. Por lo tanto, se concluye que las secuencias codificantes de *H. sapiens* y *E. coli* presentan un estilo estructural. Sin embargo para las secuencias codificantes y no codificantes de *H. sapiens* no podemos confirmar lo mismo. Para comprobar si las secuencias Alu y las secuencias codificantes de *H. sapiens* presentan el mismo estilo estructural se propone se realicen pruebas de Monte Carlo, así como determinar el IDH en otro tipo de secuencias no codificantes como las LINE, lo cual podría realizarse con los programas desarrollados en este trabajo. También se propone se realice este análisis con más organismos de diferentes clados y diferentes modos de vida, por ejemplo angiospermas o gimnoespermas que presentan una mayor proporción de DNA no codificantes, y organismos procariontes simbiontes de organismos eucariontes para analizar si este modo de vida influye en el estilo estructural de la molécula del DNA.

Introducción

Descripción de la molécula de DNA

El ácido desoxirribonucleico, DNA, es el repositorio de la información genética. Las secuencias de aminoácidos de todas las proteínas y las secuencias nucleotídicas del ácido ribonucleico, RNA, se encuentran codificadas por él.

El DNA está compuesto por grupos fosfato, desoxirribosa y bases nitrogenadas; éstas últimas son: la adenina, la timina, la citosina y la guanina . La adenina y la guanina son bases púricas, y la citosina y la timina son bases pirimídinicas. La adenina y la citosina presentan un grupo amino (NH_2) en el carbono 6 y en el carbono 4 respectivamente. En cambio la guanina y la timina presentan un grupo ceto en el carbono 6 y en el carbono 4 respectivamente (figura 1) (Nelson *et al.*, 2008).

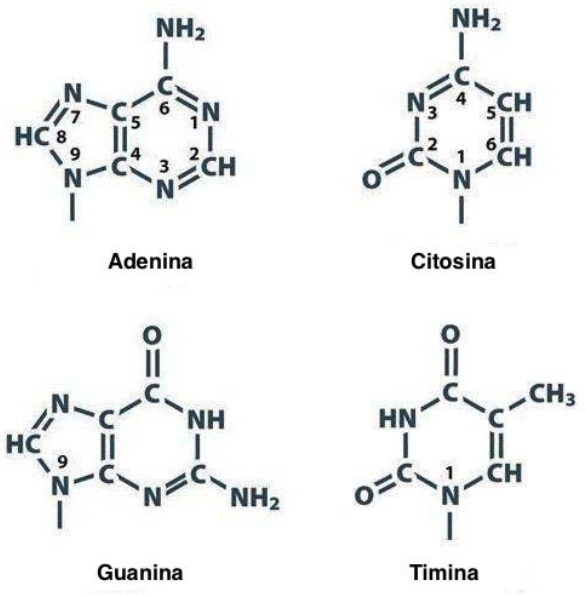

Figura 1: Bases nitrogenadas (figura modificada de Brown, 2011).

El nucleótido de DNA está formado por la unión de un grupo fosfato a una desoxirribosa y a la base nitrogenada (figura 2). Los nucleótidos están unidos en la molécula del DNA por medio de enlaces fosfodiéster. Estos enlaces son a partir de la unión del grupo fosfato 5' de un nucleótido con el grupo hidroxilo 3' de la desoxiribosa del siguiente nucleótido. Todos los enlaces fosfodiéster presentan la misma orientación en la secuencia de nucleótidos. Por lo tanto, cada hebra de DNA presenta una polaridad específica y diferentes terminaciones 5' y 3' (figura 3) (Nelson *et al.*, 2008).

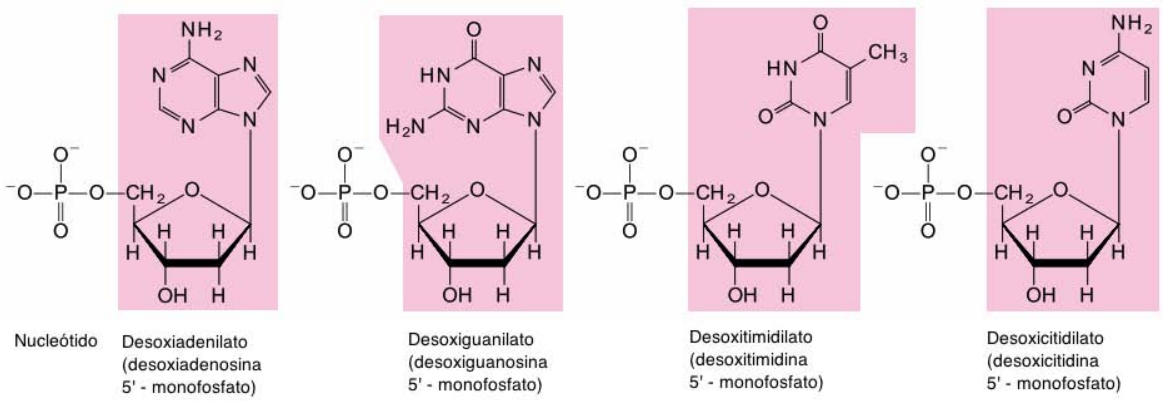

Figura 2: Nucleótidos de DNA (figura modificada de Nelson *et al.*, 2008)

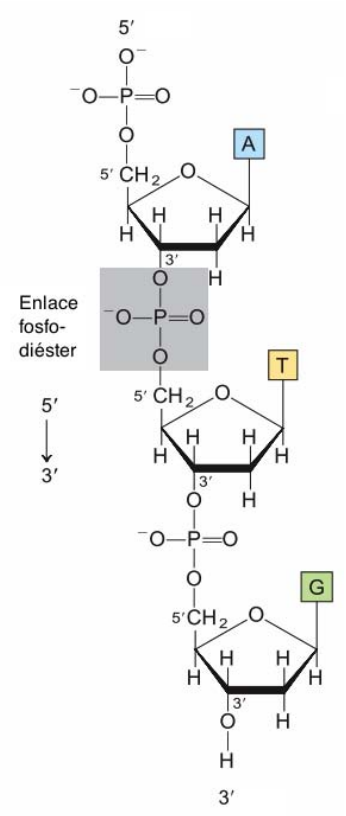

Figura 3: Enlace fosfodiéster entre los nucleótidos de DNA. Se observa la unión del grupo fosfato 5' del desoxitimidilato con el grupo hidroxilo 3' del desoxiadenilato (figura modificada de Nelson *et al.*, 2008).

Debido a la unión de los nucleótidos por medio de los enlaces fosfodiéster, las desoxi-

rribosas de dos nucleótidos unidos presentan una distancia de 6 Å ó 0.6 nm. Las bases nitrogenadas de los dos nucleótidos dejan un espacio entre sí de 2.7 Å y cada una de ellas tiene un grosor de 3.3 Å (figura 4) (Calladine *et al.*, 2004).

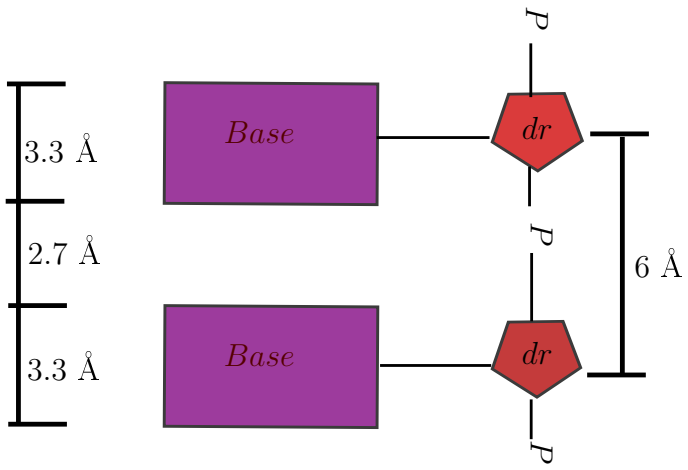

Figura 4: Esquema que muestra dos nucleótidos y sus distancias relativas. Se observa el tamaño de cada base, el espacio entre estas y la distancia entre cada desoxirribosa (dr: desoxirribosa, P : grupo fosfato) (figura modificada de Calladine *et al.,* 2004).

La unión de las bases nitrogenadas de dos hebras de DNA entre sí se da por medio de enlaces de hidrógeno. La adenina forma dos enlaces de hidrógeno con la timina y la citosina forma tres enlaces de hidrógeno con la guanina (figura 5) (Nelson *et al.*, 2008). Esta interacción específica entre las bases nitrogenadas se denomina apareamiento tipo Watson y Crick. Cuando dos cadenas de nucleótidos se aparean entre sí presentan una distancia horizontal total de 18 Å (figura 6) (Calladine *et al.*, 2004) .

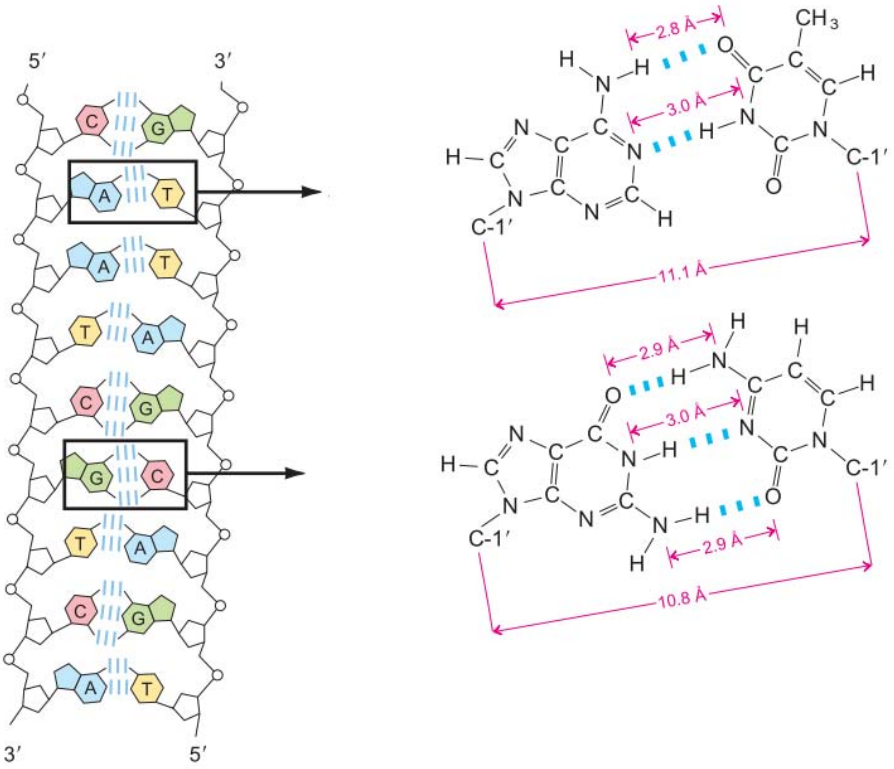

Figura 5: Apareamientos tipo Watson y Crick. Del lado izquierdo se muestra el apareamiento de dos hebras de DNA. Del lado derecho se muestran los apareamientos de las bases A-T y G-C. La A y la T presentan dos enlaces de hidrógeno, y la G y la C presentan tres enlaces de hidrógeno (figura modificada de Nelson *et al.*, 2008).

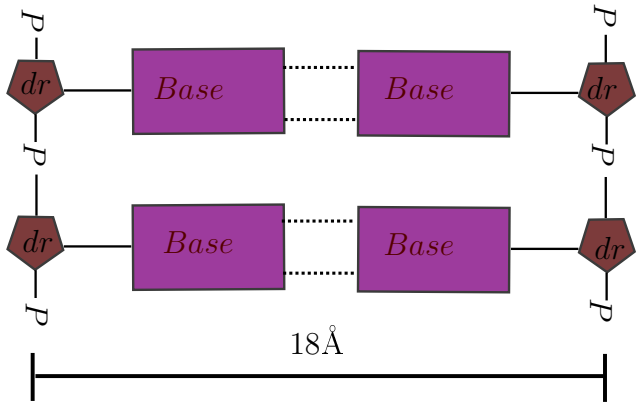

Figura 6: Esquema que muestra a las bases nitrogenadas apareadas y la longitud total de nucleótido a nucleótido de manera horizontal (figura modificada de Calladine *et al.*, 2004).

La forma helicoidal de la molécula de DNA se debe por el apilamiento de los nucleótidos con un ángulo de giro de 36º (Calladine *et al.*, 2004). Debido al apareamiento de las dos hebras de DNA, la molécula helicoidal forma dos surcos, el surco menor, de 12 Å de ancho, y el surco mayor, de 22 Å de ancho (figura 7) (Nelson *et al.*, 2008). Entre mayor sea el tamaño del surco, la región expuesta de las bases nitrogenadas es más accesible. Debido a ésto existen un mayor número de proteínas que se pueden unir a secuencias específicas que se encuentran en las bases nitrogenadas expuestas en el surco mayor.

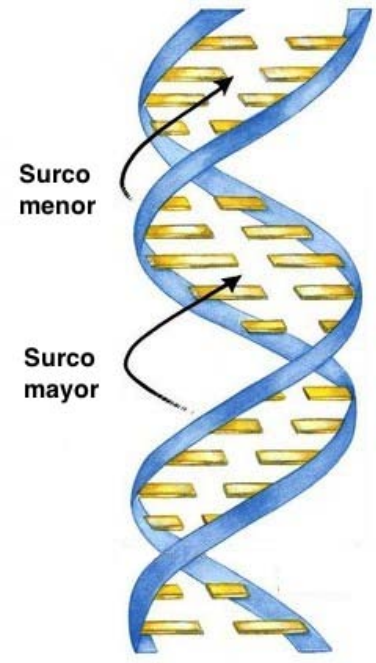

Figura 7: Surcos mayor y menor de la molécula del DNA (figura modificada de Nelson *et al.*, 2008).

El giro de los nucleótidos provoca que los grupos fosfato aledaños se encuentren a una distancia vertical de 6 Å entre sí a lo largo de la hélice y las bases nitrogenadas asciendan cada 3.3 Å (Watson y Crick, 1953) (figura 8).

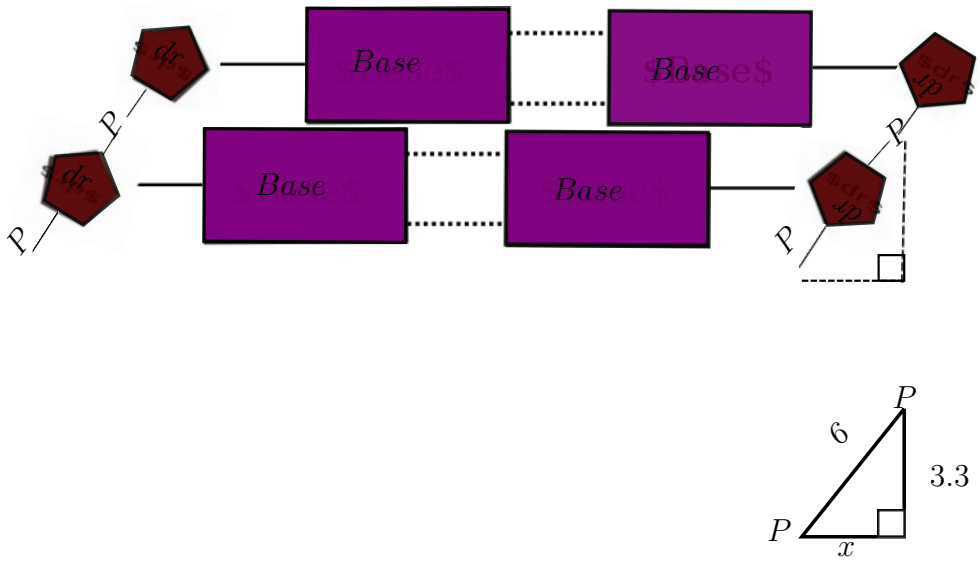

Figura 8: Esquema que muestra la distancia entre cada grupo fosfato superior e inferior de la cadena de nucleótidos, 6 Å, y la distancia entre un par de bases y el siguente, 3.3 Å (figura modificada de Calladine *et al.,* 2004).

Cada grupo fosfato de la perifería se encuentra a una distancia de 9 Å del centro del par de bases, con un ángulo de 32.3º. Por lo tanto cada grupo fosfato conforma 1/11 de giro de la hélice de la molécula del DNA (figura 9).

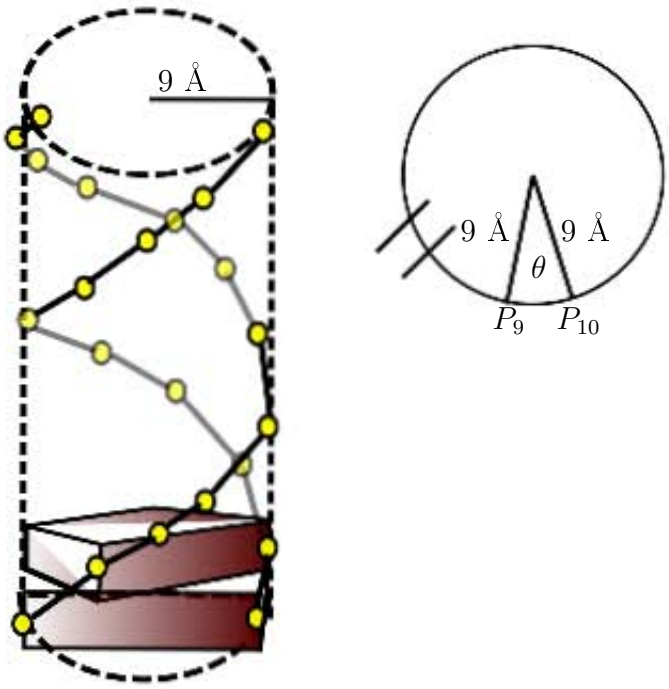

Figura 9: El esquema del lado izquierdo muestra el giro del esqueleto de los grupos fosfato alrededor de un radio de 9 Å. Del lado derecho se esquematiza la vista superior del esquema del lado izquierdo, mostrando el ángulo θ de separación entre cada grupo fosfato (figura modificada de Calladine *et al.,* 2004).

La variación de posiciones en cada uno de los elementos de la molécula del DNA provoca diferentes conformaciones. Las conformaciones mejor caracterizadas son la B, que es la descrita por Watson y Crick, la A y la Z. La conformación B es la más estable en condiciones fisiológicas, presenta un diámetro aproximado de 20 Å y una dirección de giro hacia la derecha con 10.5 grupos fosfato por giro (Watson y Crick, 1953; Franklin y Gosling, 1953). La conformación A presenta un diámetro aproximado de 26 Å y una dirección de giro hacia la derecha con 11 fosfatos por giro. Los pares de bases en la

conformación A presentan una inclinación de 20° respecto al eje vertical de la hélice. La inclinación de los pares de bases provoca que el surco mayor sea más profundo y que el surco menor sea más superficial. La conformación Z presenta un diámetro de 18 Å y una dirección de giro hacia la izquierda con 12 grupos fosfato por giro. En esta conformación el esqueleto del DNA toma una apariencia en zigzag y el surco mayor es superficial, a diferencia del surco menor que es más profundo y angosto (figura 10) (Nelson *et al.*, 2008).

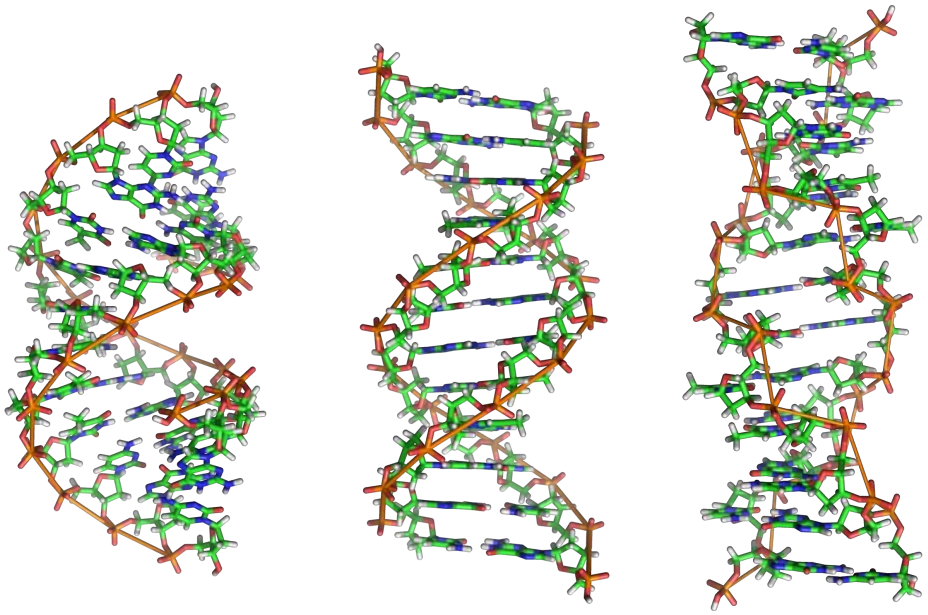

Figura 10: De izquierda a derecha se muestran las conformaciones A, B y Z de la molécula de DNA (WikimediaCommons, 2015)

La composición de la molécula del DNA puede describirse de acuerdo al contenido de GC, es decir, a la suma de la proporción de bases de G y C (Krebs *et al.*, 2009). Esta proporción varía dentro del genoma[1] de un mismo organismo y entre los genomas de diferentes organismos. Dentro de un mismo genoma, los genes presentan una mayor proporción de GC. Entre diferentes organismos, las diferentes proporciones de GC pueden deberse a la variación de la presión de selección sobre las diferentes secuencias del genoma, a un sesgo

[1]El conjunto de información que se encuentra en el DNA de un organismo. En él se encuentra codificada toda la información de las proteínas y moléculas de RNA que el organismo puede sintetizar.

mutacional y a un sesgo de recombinación asociado a la reparación del DNA (Birdsell, 2002). De acuerdo al análisis de polimorfismos y a patrones de sustitución en los genomas de mamíferos se ha demostrado que el contenido de GC ha evolucionado debido a un sesgo de la recombinación que favorece la fijación de mutaciones que cambian A y T por G y C (Spencer *et al.,* 2006; Duret y Ardnt, 2008). Por otra parte, en bacterias y arqueas el contenido de GC en los genomas se encuentra afectado por factores ambientales como la temperatura o la disponibilidad de oxígeno y nitrógeno (Foerstner *et al.*, 2005).

Compactación de la molécula del DNA

Los eucariontes compactan su genoma en cromosomas, los cuales consisten en moléculas de DNA asociadas a proteínas que empaquetan las hebras de DNA de manera más compacta. El complejo de DNA y proteínas de empaquetamiento se denomina cromatina. A diferencia de los eucariontes, las bacterias compactan su genoma en una sola molécula circular de DNA, la cual también se encuentra asociada con proteínas de empaquetamiento pero que son diferentes a las de los eucariontes (Alberts *et al.*, 2002). Las proteínas encargadas de empaquetar el DNA en las bacterias son: HU (proteína similar a histona) para enrollar, FIS (*factor for inversion stimulation*) y IHF (*integration host factor*) para doblar, y HNS (*histone-like nucleoid structuring*) para compactar. El conjunto de estas proteínas y el DNA se denomina nucleoide. La compactación del genoma en las bacterias se debe al superenrrollamiento de la molécula de DNA, es decir, que segmentos de la doble hebra de DNA se encuentran enrollados entre sí. Las topoisomerasas son las enzimas responsables del superenrrollamiento de la molécula del DNA (Hartl y Ruvolo, 2011).

En los eucariontes, las proteínas que se unen al DNA para formar cromosomas son las histonas y las proteínas no histonas. El primer nivel de compactación de la molécula de DNA se produce por la formación el complejo de histonas-DNA con un tamaño aproximado de 10 nm, el nucleosoma. Éste consiste en dos moléculas de las histonas H2A, H2B, H3 y H4, y un segmento de DNA de aproximadamente 147 pares de bases. Cada

nucleosoma se separa por una región de DNA, DNA de unión, de una longitud variable cercana a los 80 pares de bases (figura 11).

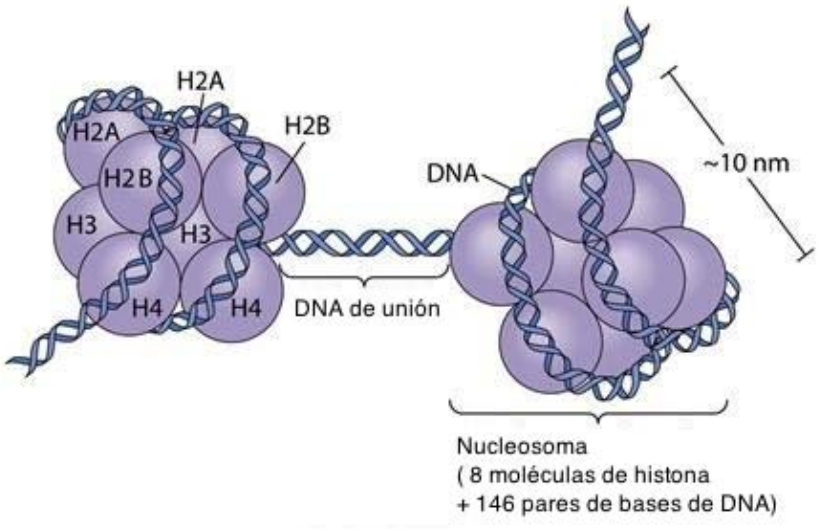

Figura 11: Esquema que muestra la estructura de un nucleosoma. Se observa el complejo de ocho histonas con el DNA y el DNA de unión (figura modificada de Klug *et al.,* 2012).

Además cada núcleo de histonas presenta una "cola" N-terminal la cual se extiende fuera del núcleo de DNA-histonas (figura 12). En la región de contacto entre el DNA y las histonas se forman aproximadamente 142 enlaces de hidrógeno. La mayoría de estos enlaces es entre los aminoácidos de las histonas y el esqueleto de azúcar-fosfato del DNA (Alberts *et al.*, 2002).

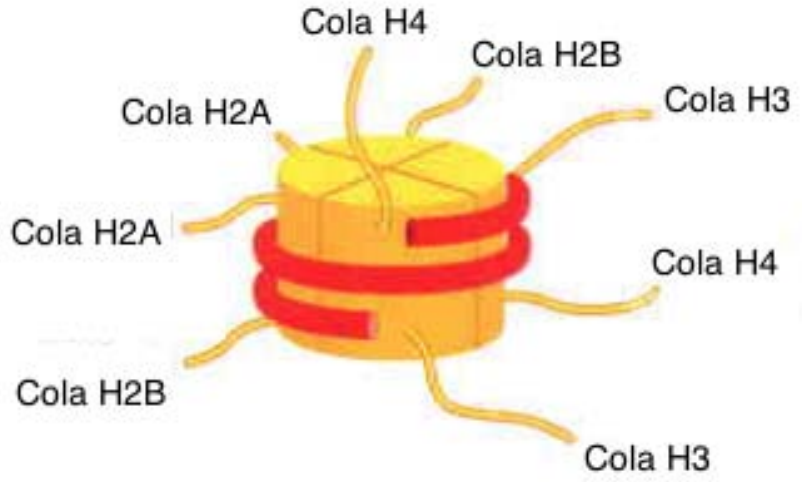

Figura 12: Esquema que muestra las colas de histona en el nucleosoma en rojo se esquematiza al DNA (figura modificada de Alberts *et al.*, 2002).

El siguiente nivel de compactación es la fibra de 30 nm. En ella los nucleosomas se encuentran empaquetados uno encima de otro. De acuerdo a imágenes de microscopía electrónica se propone que los nucleosomas se empaquetan de manera intercalada formando una estructura solenoidal. El empaquetamiento de los nucleosomas se debe a la unión entre sus "colas" de histonas, además de la presencia de la histona de unión, histona H1, la cual se une a cada nucleosoma (figura 13) (Alberts *et al.*, 2002).

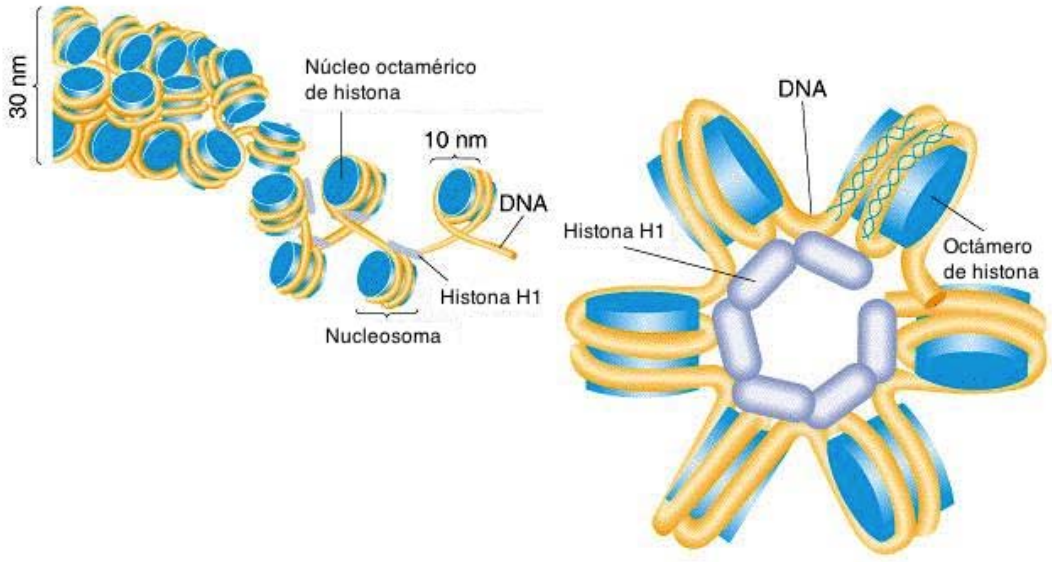

Figura 13: Esquema que muestra fibra de 30 nm y la organización de los nucleosomas en ésta. Del lado izquierdo se muestra la fibra de 30 nm conformada por el empaquetamiento de los nucleosomas. Del lado derecho se observa la organización de los nucleosomas entre sí por la histona H1 y el DNA de unión (figura modificada de Abu El-Soud, 2007).

Dentro del núcleo, la fibra de 30 nm se dobla en bucles de cromatina, cada uno con aproximadamente 100 kb. Posteriormente éstos son organizados en dominos de cromátina de 1 Mb. Durante la mitosis el empaquetamiento del DNA en la fibra de 30 nm se condensa en una cromatida de un cromosoma en metafase (figura 14) (Hartl y Ruvolo, 2011).

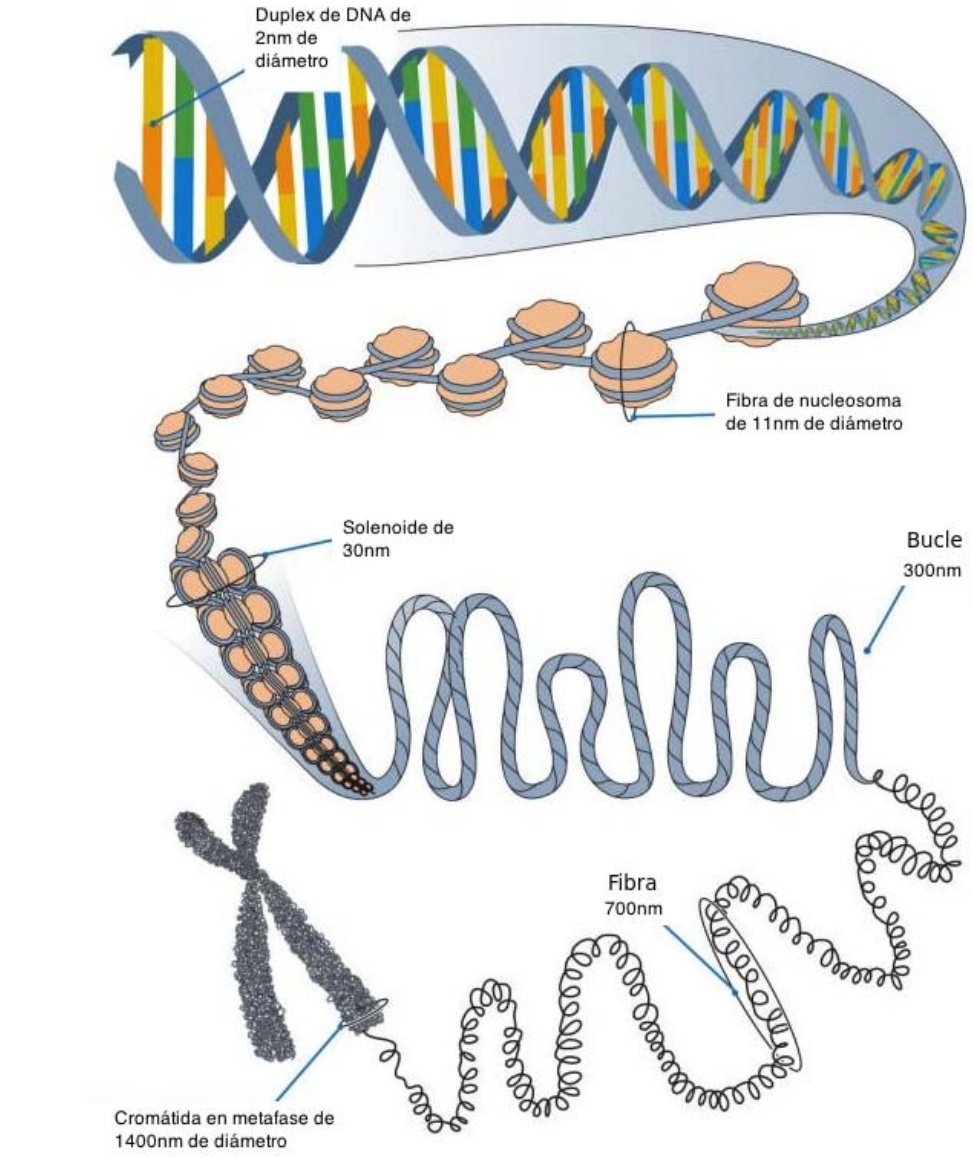

Figura 14: Esquema que muestra las diferentes niveles de empaquetamiento del DNA. La molécula del DNA se empaqueta en los nucleosomas, posteriormente éstos forman la fibra de 30 nm (solenoide). El solenoide forma bucles de cromatina de 300 nm, éstos se compactan en la fibra de 700 nm y se condensa en una cromátida (figura modificada de Hartl y Ruvolo, 2011).

En los eucariontes existen dos tipos de cromatina, la heterocromatina que se encuentra altamente condensada, y la eucromatina que se encuentra menos condensada. El DNA en la heterocromatina presenta un menor número de genes. Los genes en las regiones

de eucromatina que se compacten en heterocromatina dejarán de expresarse debido al mayor empaquetamiento. Las diferentes estructuras de cromantina se deben a las modificaciones covalentes, reversibles, de las cuatro histonas en el núcleo del nucleosoma. Estas modificaciones son mono, di y tri metilaciones, así como acetilaciones. La combinación de estas modificaciones en cada nucleosoma es un código de histonas (Alberts *et al.*, 2002).

Regulación de la expresión génica

El gen se define como un segmento de DNA que codifica la información para sintetizar un producto biológicamente funcional. Este producto es RNA y posteriormente puede sintetizarse a proteína.Los procesos celulares para que la información en la secuencia de DNA pase a RNA y a proteínas son la transcripción y la traducción (Nelson *et al.*, 2008). Estos procesos presentan una regulación y control para que los diferentes procesos celulares sean coordinados y que la cantidad necesaria de productos se sintetice en el lugar y momento indicado. La célula presenta dos redes que coordinan el funcionamiento celular, la red enzimática que manipula a los metabolitos y la red regulatoria que manipula a los genes, transcritos y proteínas. Por lo tanto, los procesos de transcripción y traducción son dinámicos (Lesk, 2013).

La regulación de la expresión génica sucede a nivel transcripcional, traduccional y post-traduccional (figura 15) (Lesk, 2013). El control transcripcional inicia al permitir la regulación sincronizada de múltiples genes que codifican a productos necesarios para actividades independientes. La transcripción es regulada por interacciones DNA-proteína, en especial aquellos dominios proteicos que componen a la RNA polimerasa (Nelson *et al.*, 2008).

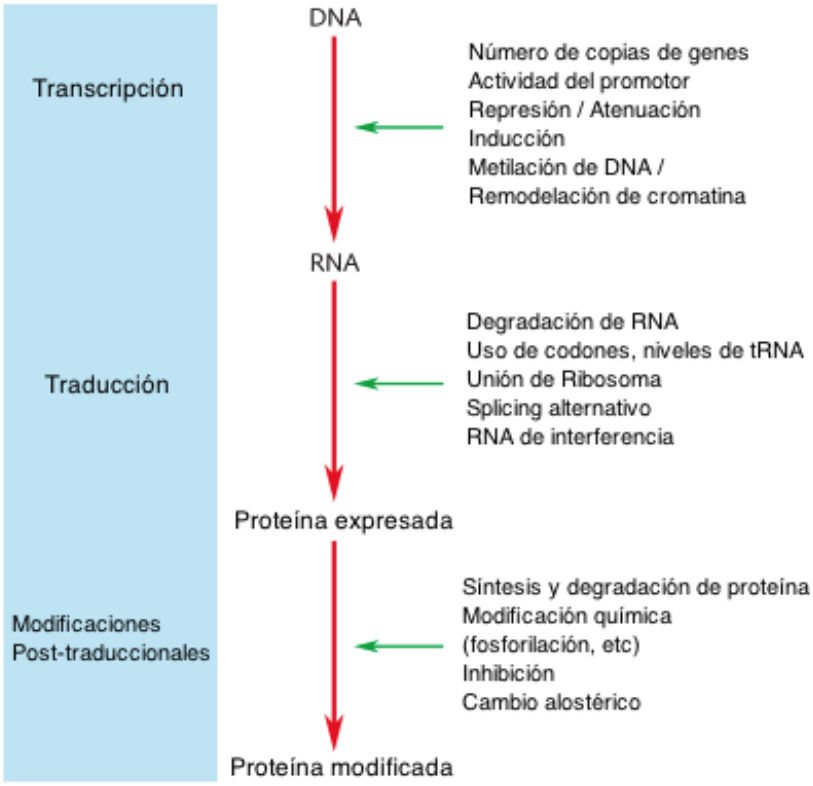

Figura 15: Esquema que muestra los diferentes niveles de regulación de la expresión génica. Se observa que durante la expresión génica existen diferentes tipo de regulación a nivel transcripcional, traduccional y post-traduccional (figura modificada de (Lesk, 2013)).

La RNA polimerasa se une al DNA e inicia la transcripción en las regionas promotoras, cuyas secuencias varían considerablemente afectando la afinidad de unión de la RNA polimerasa, y por lo tanto la frecuencia de inicio de la transcripción. El inicio de la transcripción también es regulado por proteínas que se unen a regiones promotoras o a regiones cercanas a éstas. El nivel basal de transcripción se determina por el efecto de las secuencias promotoras en la función de la RNA polimerasa y por sus factores de transcripción asociados. Existen al menos tres tipos de proteínas que regulan la transcripción por medio de la RNA polimerasa: los factores de especificidad, que alteran la especifici-

dad de la RNA polimerasa en regiones promotoras; los represores, que impiden la unión de la RNA polimerasa al promotor; y los activadores que aumentan la interacción entre la RNA polimerasa y la región promotora (Nelson *et al.*, 2008).

Regulación de la expresión génica en procariontes

En los procariontes la regulación de la expresión génica sucede, principalmente, a nivel transcripcional río arriba del inicio del gen del sitio de unión de la RNA polimerasa al DNA (región 5'). La regulación de la expresión génica es regulada por represión y activación. Los represores detienen la transcripción al bloquear el sitio de unión de la polimerasa al unirse a regiones específicas cercanas al promotor denominadas operadores (Lesk, 2013; Nelson *et al.,* 2008). Los activadores aumentan la interacción de la RNA polimerasa con el promotor. Los sitios de unión de los activadores normalmente se encuentran en regiones cercanas al promotor. En los procariontes un ejemplo de proteína que regula el inicio de la trancripción actuando como un factor de especificidad es la subunidad σ de la holoenzima de la RNA polimerasa (Nelson *et al.*, 2008).

En las bacterias la coordinación de la expresión de genes necesarios para un mismo proceso es más sencilla que en eucariontes debido a que los genes se transcriben juntos. Los RNA mensajeros (mRNA) con múltiples genes en un solo transcrito se denominan policistrónicos. La regulación de la transcripción ocurre en un solo promotor. El operón es el conjunto de genes, el promotor y las secuencias necesarias para la regulación de los genes (figura 16) (Nelson *et al.*, 2008).

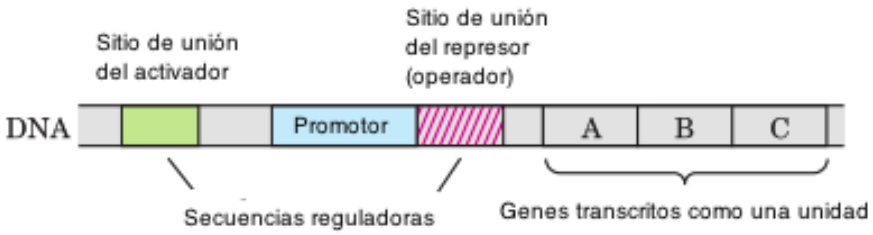

Figura 16: Esquema que muestra el conjunto de elementos que forma un operón. Se observa el sitio de unión al activador, la secuencia promotora, el sitio de unión al represor y las letras A,B y C representan los genes que se transcriben juntos (figura modificada de Nelson *et al.,* 2008)

Regulación de la expresión génica en eucariontes

Los represores y activadores, al igual que en procariontes, se unen a regiones de la secuencia de DNA afectando la transcripción, sin embargo, los sitios de unión de estas proteínas pueden ser en regiones distantes al promotor. Para el caso de los activadores, los sitios de unión en la secuencia de DNA se denominan aumentadores (*enhancers*) (Nelson *et al.*, 2008).

Los promotores en los eucariontes se encuentran generalmente inactivos, ya que el acceso a éstos está restringido por la estructura de la cromatina, a que predominan los mecanismos de regulación positiva (presencia de proteínas de activación) y a que hay una separación en tiempo y espacio de la transcripción, que se realiza en el núcleo, y la traducción, que se realiza en el citoplasma. Por lo tanto el primer paso de la expresión génica se ve regulado a nivel de remodelamiento de cromatina.[2] Los principales procesos para activar la cromatina para la transcripción son la acetilación y la desacetilación de las histonas (Nelson *et al.*, 2008).

Para que la holoenzima de RNA polimerasa II (Pol II) se una a los promotores necesita de la participación de los factores de transcripción basales, de los transactivadores de

[2]Proceso enzimático que provoca cambios estructurales en la cromatina. Las enzimas modifican covalentemente el núcleo de histonas del nucleosoma.

unión a DNA,[3] y de coactivadores.[4] (Nelson *et al.*, 2008)

El control de la expresión génica en los eucariontes es más complejo que en los procariontes, debido al mayor número de proteínas reguladoras y a la combinación de éstas. (Nature Education, Consultado 2015-28-08)

Dominios de unión al DNA de proteínas reguladoras

Las proteínas reguladoras de la expresión génica se unen a regiones específicas de la secuencia del DNA. Para que las proteínas puedan unirse de manera específica necesitan reconocer características de la molécula del DNA. Las proteínas diferencian entre los pares de bases de acuerdo a los grupos funcionales de éstos expuestos en el surco mayor (figura 17) (Nelson *et al.*, 2008).

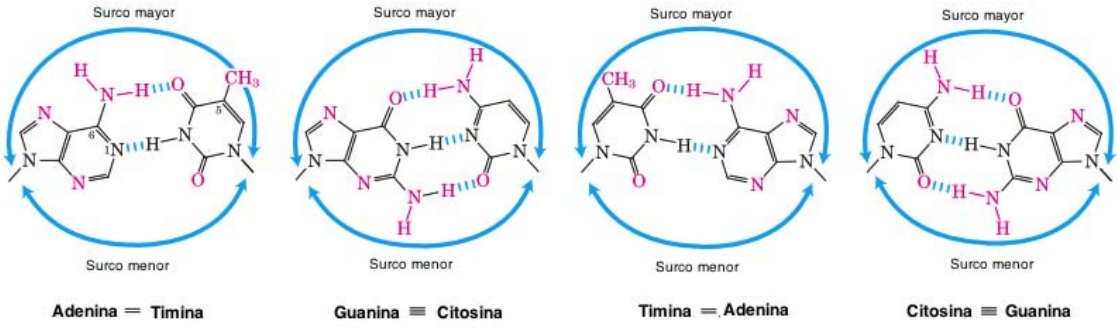

Figura 17: Esquema que muestra los grupos funcionales disponibles en la molécula del DNA. Se observa los grupos funcionales de los cuatro pares de bases en los surcos mayores y menores del DNA. Los grupos funcionales que pueden servir para el reconocimiento de pares de bases por las proteínas de unión al DNA se muestran en rojo (figura modificada de Nelson *et al.*, 2008).

Algunos ejemplos de motivos de unión a DNA por proteínas reguladoras son: hélice-giro-hélice, dedos de zinc, homeodominio y horquilla-hélice-hélice. También hay motivos con interacciones proteína-proteína para formar dímeros que interactúan con el DNA, ejemplos de éstos son: el zipper de leucina y el motivo hélice-bucle-hélice. Además de los

[3]Proteínas que se unen a enhancers para facilitar la transcripción.

[4]Proteínas necesarias para la interacción entre los transactivadores de unión al DNA y el complejo formado por la Pol II y los factores de transcripción.

dominios antes mencionados existe un motivo importante para el inicio de la transcripción que se encuentra en la proteína de unión a TATA[5], (TBP, *TATA binding protein*) (Nelson *et al.*, 2008). Igualmente las proteínas involucradas en la compactación del DNA también presentan motivos de interacción con el DNA. Las histonas presentan el motivo de plegado de histona (*histone fold*) (Xu y Morrical, 2010).

Evolución y DNA

Selección natural: origen e implicaciones

El mecanismo de selección natural propuesto por Charles Darwin postula que cambios sutiles que provean una ventaja a los organismos provocan su evolución (Darwin, 1859). La teoría de evolución de Darwin fue influenciada por dos conceptos definidos en el siglo XIX para explicar la causa de la morfología que presentan los organismos: la ley de condiciones de existencia y la ley de unidad de tipo. El concepto de ley de condiciones de existencia define que los organismos presentan la morfología que tienen debido al ambiente en el que viven, es decir, todos sus órganos y estructuras son necesarias para las circunstancias en las que se encuentran. En cambio, el concepto de unidad de tipo define que los organismos están conformados por un mismo plano estructural independiente de las funciones particulares de órganos y estructuras necesarias para el ambiente en el que se encuentren. Los naturalistas de la época debatían respecto a cuál de estos dos conceptos era que explicaba la morfología de los organismos. (Appel, 1987; Gould 2002).

Bajo este contexto, los naturalistas del siglo XIX empezaron a preguntarse la razón por la cual las formas de los organismos van cambiando. Los naturalistas que argumentaban que el concepto de condiciones de existencia es el que determina la morfología de los organismos afirmaban que las formas de los organismos cambian de acuerdo a la utilidad que le proveen. Por otra parte, los naturalistas que argumentaban que el concepto de

[5]Secuencia río arriba del sitio de inicio de la transcripción que es rica en T y A denominada caja TATA.

unidad de tipo es el que determina la morfología de los organismos, afirmaban que la utilidad de las formas es complementaria y por lo tanto los cambios en éstas no se deben a la función que realizan.

El principal postulador del concepto de la ley de condiciones de existencia fue George Cuvier. Cuvier sostenía que los organismos son intransformables y por lo tanto se encuentran diseñados de acuerdo al contexto en el que viven. La argumentación de George Cuvier se define como una visión funcionalista de los organismos, debido a que le atribuyó una mayor importancia a la especialización de los órganos. Por otra parte, Geoffroy Saint Hilaire fue el principal postulador del concepto de la ley de unidad de tipo. Saint Hilaire argumentaba que los organismos presentan un plano estructural en común y que posteriores transformaciones están restringidas a leyes estructurales impuestas por este molde. Los argumentos de Saint Hilaire se definen como una visión estructuralista de los organismos debido a que le atribuyó una mayor importancia a la organización morfológica general (Appel, 1987; Gould, 2002).

Los conceptos de la ley de condiciones de existencia y la ley de unidad de tipo se consideraban opuestos. Darwin logró conjuntarlos con su teoría evolutiva. Sin embargo, le atribuyó una mayor importancia al concepto de ley de condiciones de existencia debido a que lo consideró dentro del concepto de adaptación argumentando que las estructuras definidas dentro del concepto de ley de unidad de tipo son adaptaciones pasadas. Por lo tanto, la teoría de evolución de Darwin es una explicación funcionalista, por medio del concepto de adaptación, del proceso evolutivo (Gould, 2002). Darwin logró integrar el conocimiento biológico del siglo XIX por medio de su teoría de la selección natural.

Síntesis moderna

El estudio de la genética de poblaciones logró unir bajo la teoría darwiniana a las variaciones debido a las mutaciones con las variaciones en los organismos . A partir del estudio de la genética de poblaciones , de la genética experimental y la sistemática es

que se proponuso la síntesis moderna de la evolución.

La síntesis moderna toma su nombre del libro de Julian Huxley en 1942, *Evolution, The Modern Synthesis*. A partir del libro de Huxley se conjuntaron las leyes de Mendel con la teoría de evolución de Darwin por medio de la demostración matemática de que presiones de selección sobre pequeñas diferencias genéticas pueden provocar cambios evolutivos. El principal objetivo de la síntesis es explicar el proceso evolutivo por medio de mecanismos genéticos, por lo tanto, al inicio de su desarrollo se aceptaban otros mecanismos evolutivos además de la selección natural. Un ejemplo de estos mecanismos es la deriva génica propuesta por Sewall Wrigth, el cual es el proceso de cambio en la frecuencia de los alelos en la población debido al muestreo aleatorio de los individuos. Sin embargo, posteriormente la síntesis moderna restringió a la selección natural como el principal mecanismo evolutivo. Un ejemplo de esto es la reformulada argumentación de Sewall Wrigth del proceso de deriva génica donde expone que es un mecanismo adaptativo. Wrigth justificó lo anterior al definir que la deriva génica actúa como el productor de los elementos necesarios para el proceso de selección interdémica (Gould, 2002).

Evolución molecular

Después de la consolidación de la síntesis moderna aunado al desarrollo de la biología molecular es que se iniciaron los estudios evolutivos a nivel molecular (Koonin, 2009). La similitud en las secuencias genéticas entre los organismos muestra un origen evolutivo común y las diferencias que presentan se deben a mutaciones a lo largo de la historia evolutiva de los organismos. Las mutaciones que se heredan pueden ser letales para los organismos pero en algunos casos pueden otorgarle algún beneficio para que pueda sobrevivir mejor al ambiente en el que se encuentra, es decir, le otorgan una ventaja adaptativa. Por lo tanto, la molécula del DNA puede ser vista como un registro histórico de los cambios que han presentado los organismos desde el origen de la vida hasta la actualidad (Nelson *et al.*, 2008).

Las primeras implementaciones del análisis molecular evolutivo se realizaron por Zuckerkandl y Pauling demostrando que las secuencias de aminoácidos del citocromo c y de las globinas se encuentran altamente conservadas en animales lejanamente emparentados (Zuckerkandl y Pauling, 1965). Posteriormente definen al reloj molecular como la tasa constante de la evolución de la secuencia característica de cada proteína en ausencia de un cambio evolutivo (Zuckerkandl, 1987). Después, Carl Woese demostró la conservación de la molécula del RNA ribosomal en todas las formas celulares proponiendo su uso para el análisis filogenético (Woese y Fox, 1977).

A finales de la decada de los sesenta y principios de los setentas Mooto Kimura propusó la teoría neutral de la evolución molecular, la cual establece que la mayoría de las mutaciones fijadas a lo largo de la evolución son neutralmente selectivas (Kimura y Ota, 1974). Además apoya la hipótesis del reloj molecular de Zuckerkandl y Pauling dado que la teoría neutral de la evolución argumenta que las secuencias génicas evolucionan a una tasa relativamente constante a lo largo de intervalos de tiempo largos y toman valores semejantes en especies distintas. Por lo tanto, se concluyó que las mutaciones benéficas para el organismo que estén bajo selección natural son lo suficientemente raras como para inferir que son la principal causa del proceso evolutivo. Estudios posteriores definieron mejor la teoría neutral de la evolución al mostrar que no se necesita que una mutación sea neutral sino que sólo ejerza un ligero efecto deletéreo con la capacidad de no ser eliminado por la selección purificadora.

A la modificación de la teoría neutral se le denomina la teoría casi neutral de la evolución molecular, la cual es la que más se aleja del paradigma seleccionista enfatizado en la síntesis moderna, dado a que postula que la mayoría de las mutaciones fijadas durante la evolución no son afectadas por selección positiva (Ohta, 1992).

Síntesis extendida

Actualmente se argumenta que la teoría evolutiva necesita extenderse para tomar en cuenta los avances metodológicos y disciplinarios de la biología, incluyendo la genética molecular, la biología del desarrollo y las "ómicas" (genómica, transcriptómica, metabolómica, etc.). Una de las propuestas de la síntesis extendida es integrar diferentes enfoques de la biología evolutiva, como análisis filogenéticos, genética de poblaciones, evolución molecular y ecología evolutiva, por mencionar algunos. Normalmente la mayoría de los estudios evolutivos sólo utilizan uno de estos métodos. Esto se debe a que en el desarrollo de la síntesis moderna se considera al gen como la principal unidad de selección. Al mismo tiempo se desarrolla la tecnología para secuenciar proteínas y DNA lo que reforzó esta perspectiva. El gen se convirtió en la principal unidad para los análisis en genética de poblaciones y en el estudio de la evolución molecular. El gen se consideraba conceptualmente aislado de las influencias ambientales, los procesos regulatorios e interacciones con el resto del genoma. El estudio aislado de los genes respecto al ambiente y al resto del genoma se debía a que no se conocía la influencia de factores externos como para agregarlos al análisis y es útil para cierto tipo de preguntas porque deja de lado los factores que le imponen ruido. Sin embargo, este enfoque del gen aislado es limitante ya que los organismos se encuentran en ambientes que varían en el tiempo y el espacio, porque los alelos funcionan en diversos contextos genéticos y los genes afectan rasgos que tienen consecuencias ecológicas (Pigliucci y Müller, 2010).

Además de conjuntar diferentes enfoques en el estudio de la biología evolutiva, la síntesis extendida también propone, a partir de análisis genómicos, estudiar procesos evolutivos a través de toda la escala genética de organización. La información obtenida a partir del genoma es más precisa y menos sesgada que la información obtenida a partir de un solo gen o de un número de ellos. Cada gen presenta diferentes niveles de variación y fijación, lo que refleja las diferentes combinaciones de las fuerzas evolutivas para mantener su funcionalidad. A partir de datos genómicos se puede estudiar la interacción de diferentes

genes entre sí, lo que ha mostrado que el contexto genómico tiene un gran efecto en la expresión fenotípica. Por lo tanto se ha concluido que los genes no evolucionan de manera aislada sino en contexto del resto del genoma. Los datos genómicos nos son útiles para conjuntar las diferentes disciplinas bajo un análisis común. Por ejemplo, las secuencias genómicas facilitan el desarrollo de marcadores genéticos útiles para la genética cuantitativa, proveen información para desarrollar modelos más precisos para el estudio evolutivo de las secuencias y también proveen de información para probar el efecto de la selección en el genoma, permiten inferir cambios en las secuencias regulatorias como *enhancers* y microRNAs (miRNA) e identificar posibles diferencias genéticas que puedan explicar las diferencias fenotípicas. Los estudios evolutivos, como se mencionó anteriormente, se han enfocado en el estudio del gen. Actualmente con datos genómicos se puede acceder a los límites inferiores y superiores del gen, logrando estudiar por ejemplo, haplotipos y particiones génicas (exones, intrones, regiones transcritas pero no traducidas del gen en 5' y 3' y regiones regulatorias). A partir del aumento del número de genomas secuenciados es posible realizar estudios de genómica comparativa bajo un marco evolutivo. Esto ha permitido examinar la diversificación de la estructura del genoma y la función de los genes que lo compone. Por medio de la información obtenida de los análisis genómicos es posible comprender mejor los mecanismos evolutivos. (Pigliucci y Müller, 2010)

Organización del genoma

Estructura del genoma

A partir de la comparación de diferentes genomas se ha encontrado que hay una proporción diferente de secuencias codificantes y secuencias no codificantes en los diferentes organismos. Por ejemplo, los genomas de organismos procariontes presentan aproximadamente un 20 % (Elgar y Vavouri, 2008) de secuencias no codificantes, en cambio, organismos eucariontes como el del ser humano presenta un 98 % de secuencias no codificantes (Morris, 2012).

Existen diferentes tipos de secuencias no codificantes, como los RNAs no codificantes, los elementos regulatorios, los intrones, los pseudogenes, las secuencias repetidas, los transposones y las secuencias teloméricas. Los RNA no codificantes son aquellas secuencias que codifican a RNA ribosomal, a RNA de transferencia y a RNA pequeños como los micro RNAs que sirven como reguladores post transcripcionales. Los elementos regulatorios son secuencias que se encargan del control de la transcripción de los genes. Éstos se pueden encontrar en las regiones 5' y 3' no traducidas de los genes (UTRs, *untranslated regions*) o en los intrones. Los intrones son las secuencias no codificantes de un gen que se transcriben en la secuencia precursora del RNA mensajero y posteriormente son removidas a través del *splicing* de RNA. Los pseudogenes son secuencias que perdieron su capacidad para codificar a proteínas debido a mutaciones que previenen la transcripción del gen. Las secuencias teloméricas son las regiones repetidas del DNA al final del cromosoma que protegen el deterioro cromosomal durante la replicación del DNA (Pigliucci y Müller, 2010).

Los transposones son los principales elementos no codificantes del genoma y son elementos genéticos móviles, es decir, que pueden cambiar su posición dentro del genoma, lo que puede provocar mutaciones dependiendo del sitio genómico al cual se inserten. El tipo de mutaciones que pueden provocar son las inserciones y deleciones. Ejemplos de transposones son los elementos dispersos cortos y largos (SINE, *short interspersed elements* y LINE, *long interspersed elements*). Junto con los SINE y LINE los elementos LTR (*long terminal repeats*) y los transpoones de DNA conforman aproximadamente el 40 % del genoma humano (Gu *et al.*, 2000).

Las secuencias Alu son parte de los elementos dispersos cortos, SINE y existen cerca de 600,000 secuencias en el genoma humano (Kapitonov y Jurkal, 1996). Las secuencias Alu constan de aproximadamente 300 nucleótidos y se componen por dos promotores de RNA polimerasa III, una región intermedia rica de adeninas que divide a la secuencia de nucleótidos en dos y la termina con una cola corta de poliadeninas (figura 18) (Deininger,

2011).

Figura 18: Esquema de un elemento *Alu*. Los bloques *A* y *B* representan los componentes internos de los promotores de la RNA polimerasa III, la región rica de adeninas esquematiza por *AA* y la secuencia termina con la cola de poli adenina (imagen modificada de Deninger, 2011

.

Secuenciación de genomas

La secuenciación de genomas se considera una consecuencia de la revolución tecnológica en la biología molecular durante los años setentas. Esta revolución fue resultado del aumento del conocimiento de la molécula del DNA. Por lo tanto, históricamente, el desarrollo de la secuenciación se inició con la determinación de la estructura del DNA por James Watson y Francis Crick y el estudio del código genético. Posteriormente, a finales de los años setenta e inicio de los ochenta, se empezaron a secuenciar genomas de organismos cada vez más complejos y en 1985 y 1986 se consideró la factibilidad de secuenciar el genoma humano (HGP) (García-Sancho, 2012).

Genoma de *Escherichia coli* K-12 cepa MG1655

Escherichia coli K-12 fue uno de los primeros organismos cuyo genoma fue secuenciado, en parte debido a que es uno de los organismos modelo preferidos en el estudio de la genética, en la biología molecular y en la biotecnología (Blattner *et al.*, 1997).

El genoma de *E. coli* K-12 consiste de 4,639,221 bp en un complejo circular dúplex de DNA, 87.8 % de él consiste de genes que codifican a proteínas, 0.8 % codifica para RNAs estables, 0.7 % consiste de secuencias repetidas no codificantes y aproximadamente 11 %

codifica a secuencias regulatorias (Blattner *et al.*, 1997).

Proyecto del Genoma Humano

El principal objetivo del proyecto del genoma humano (HGP, por sus siglas en inglés) fue determinar la identidad de los 3 mil millones de nucleótidos que comprenden al genoma humano, así como caracterizar a todos los genes codificados, almacenar la información en bases públicas de datos y desarrollar y/o mejorar herramientas para analizar los datos de secuenciación. En 1988 el panel del National Research Council decidió apoyar el proyecto. En ese momento las recomendaciones del panel fueron construir mapas físicos de cada cromosoma y que se estudiaran genomas de organismos más simples antes de realizar la secuenciación de todo el genoma humano. El HGP inició oficialmente el primero de octubre de 1990 junto con la secuenciación de *Mycoplasma capricolum, Escherichia coli, Caenorhabditis elegans* y *Saccharomyces cerevisiae* (Kumar *et al.*, 2015).

El HGP ha ayudado a comprender mejor la genética humana, comprender cómo el DNA modifica el desarrollo de las especies, su evolución, biología y susceptibilidad a enfermedades. Además ésto ha provocado el desarrollo de disciplinas de investigación como la anotación y comparación de genomas. Antes del HGP se estimaba que el genoma humano presentaría de 35,000 a 150,000 genes.

En septiembre de 2003 el National Human Genome Research Institute (NHGRI) lanzó el consorcio de investigación ENCODE, the ENCyclopedia Of DNA Elements, para integrar las anotaciones del genoma humano e identificar todos los elementos funcionales. El proyecto se inició con una prueba piloto donde se probaron y compararon los diferentes enfoques de anotación a 1 % del genoma humano. Los resultados del proyecto piloto mostraron que el proceso de transcripción no sólo sucede en genes que codifican a proteínas sino que también suecede con mayor frecuencia de la esperada en RNAs no codificantes y pseudogenes. Además se encontró que los genes que codifican a proteínas presentan una complejidad poco esperada en regiones no traducidas (UTRs), en exones que se encuen-

tran hasta 200kb de distancia y en *loci* sobrepuestos o intercalados (ENCODE Project Consortium, 2004).

En 2007 el proyecto ENCODE se expandió para estudiar todo el genoma humano y actualmente se enfoca en la anotación de genes, codificantes y no codificantes para proteínas, así como de sus transcritos de RNA y regiones de reguladores transcripcionales (Birney *et al.*, 2007).

El ensamble más reciente del genoma humano es el GRCh38 de diciembre de 2013. La información de éste proviene de modelos génicos construidos a partir de los alineamientos del proteoma y cDNA humano. La tabla 1 resume la información del ensamblado del genoma GRCh38 (Flicek *et al.*, 2013)(http://www.ensembl.org/Homo_sapiens/Info/Annotation)

Tabla 1 : Ensamblado del genoma GRCh38	
Pares de bases	3,545,834,991
Genes codificantes	20,296
Genes no codificantes	25,173
Genes no codificantes pequeños	7,703
Genes no codificantes largos	14,889
Pseudogenes	14,424

La categoría de genes no codificantes pequeños se refiere a genes de menos de 200 pares de bases y la categoría de genes no codificantes largos se refiere a genes mayores a 200 pares de bases los cuales son RNAs con función biológica (Flicek *et al.*, 2013).

Genómica comparativa y evolución del genoma

La genómica comparativa ha ayudado a comprender las relaciones evolutivas entre las especies. Las similitudes y diferencias entre las secuencias genómicas sirven para inferir relaciones filogenéticas entre especies de la misma manera que los rasgos morfológicos y fisiológicos se utilizaban en el pasado para distinguir especies. Antes de poder utilizar

ensambles de genomas completos, los análisis de genómica comparativa se realizaban con un pequeño número de genes homólogos, lo cual era una importante limitación. Actualmente con la disponibilidad de genomas de diferentes especies se pueden realizar mejores comparaciones que han conllevado a resultados más informativos (Kumar *et al.*, 2015).

El análisis de las conformaciones locales de la molécula del DNA en los genomas ha mostrado que existen motivos de DNA repetidos en diferentes especies. La distribución de éstos motivos en el genoma no es al azar. Los análisis de secuencias de genomas completo[6] sugieren la inestabilidad genómica[7] debido a conformaciones diferentes a la conformación B del DNA las cuales pueden provocar cambios evolutivos rápidos. Las conformaciones del DNA diferentes a la conformación B afectan a la replicación y a la transcripción (Zhao *et al.*, 2010).

[6]Análisis de una variación genética a lo largo de todo el genoma con el objetivo de identificar su asociación a un rasgo observable.

[7]Mayor frecuencia de mutaciones en el genoma de un linaje celular.

Antecedentes

En la revisión de Travers y Muskhelishvili en 2015 acerca de la estructura y la función del DNA se concluye que la información que codifica la molécula del DNA a nivel estructural es igual de importante que la información de secuencias que codifican para la síntesis de proteínas. Esto se debe a que la información estructural permite el empaquetamiento de la molécula y la regulación de la expresión génica. La información estructural depende de las propiedades fisicoquímicas de la secuencia y de las propiedades físicas de los pares de bases consecutivos.

La interacción entre las proteínas y el DNA se debe al reconocimiento entre éstas. Los reconocimientos pueden clasificarse como directos e indirectos. En los directos las bases nitrogenadas del DNA se unen de manera específica a la proteína, por ejemplo, debido a la secuencia de nucleótidos se forman curvaturas locales en la molécula lo que provoca que en el surco mayor se encuentren más expuestos los nucleótidos que en el surco menor. Los pares de bases A-T y T-A se pueden diferenciar en el surco mayor debido al cambio del patrón de cargas parciales por al grupo metilo en la timina. De manera similar, los pares de bases de G-C y C-G son distinguibles por el cambio de posición del grupo amino en el carbono 4 de la cisteína (figura 19). En cambio, en los reconocimientos indirectos la afinidad de la unión depende del reconocimiento de la estructura del DNA y no del tipo de pares de bases que se encuentre expuesto (Travers y Muskhelishvili, 2015).

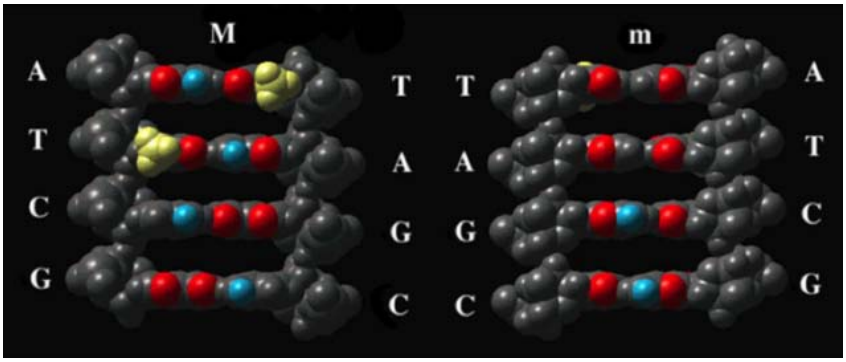

Figura 19: Grupos funcionales expuestos en los pares de bases del DNA en el surco mayor (M) y menor (m). En amarillo se observa el grupo funcional metilo de la timina, en azul los grupos funcionales básicos: en el surco mayor se observan los grupos amino 6 de la adenina y el 4 de la citosina. En el surco menor se observa el grupo amino-2 de la guanina. En rojo se encuentran los átomos de oxígeno y los nitrógenos cíclicos (Travers y Muskhelishvili, 2015).

Además de las interacciones entre proteína y DNA para la regulación de la expresión génica Travers y Muskhelishvili también mencionan la importancia de la capacidad de la molécula del DNA para cambiar su conformación para compactarse y descompactarse.

La molécula de DNA puede modelarse como una cadena lo suficientemente elástica para curvarse de acuerdo a las configuraciones necesarias para su empaquetamiento. Estas configuraciones son dependientes de las bases nitrogenadas que conformen a la secuencia de DNA (Zhurkin, 1983; Koo *et ak.,* 1986; Satchwell *et al.,* 1986; Luijsterburg *et al.,* 2008; Travers y Muskhelushvili, 2015). El empaquetamiento del DNA se debe a la interacción del nucleosoma y el DNA enrrollado en él. La compactación del DNA en el nucleosoma se facilita porque la molécula tiene direcciones de curvatura preferenciales. Esta preferencia de curvatura se debe a secuencias cortas de G-C y de A-T que van alternando. En general, las secuencias codificantes presentan una mayor proporción de G-C. Esto se debe a que los codones de los amino ácidos más abundantes también tienen una mayor proporción de G-C. Por lo tanto, las secuencias no codificantes presentan una mayor proporción de A-T. Esta distribución de los pares de bases en el genoma implica que las secuencias codificantes son menos flexibles, sin embargo en los cromosomas eucariontes las secuencias

codificantes se encuentran en los nucleosomas en mayor proporción que las secuencias codificantes. Esto posiblemente se deba a otro tipo de interacción física (Struhl y Segal, 2013), y a una consecuencia evolutiva de la interacción DNA-nucleosoma estudiada en 2015 por Xing y He, donde concluyen que existe un sesgo mutacional, favoreciendo la presencia de G-C en los genomas eucariontes.

El índice de homogeneidad, IDH

En 1995 Miramontes y colaboradores analizaron la homogeneidad estructural de la molécula del DNA en diferentes organismos a partir de un índice que valora variaciones estructurales del DNA: el índice de homogeneidad, IDH. Las propiedades estructurales que tomaron en cuenta fueron: el tipo de enlace entre los pares de bases, si las bases son púricas o pirimidínicas y si el grupo funcional que presentan es amino o ceto. Para el caso de el tipo de enlace lo definieron como fuerte o débil de acuerdo al número de enlaces de hidrógeno que unen a los pares de bases. Un enlace fuerte es aquel que presenta tres enlaces de hidrógeno y un enlace débil es el que presenta dos enlaces de hidrógeno.

A cada nucleótido de la secuencia de DNA a la que se determina el IDH se le asigna un valor de cero o de uno de acuerdo al tipo de propiedad estructural que presenta. Por ejemplo, al analizar el tipo de enlace se asigna el uno a los nucleótidos que forman enlaces fuertes y cero a los nucleótidos que forman enlaces débiles.

A partir de la secuencia binaria que proviene de la secuencia de nucleótidos de DNA se cuantifica el tipo de variación estructural en dos nucleótidos contiguos de acuerdo a sus valores de cero y uno y se cuantifica el total de ceros y unos en la secuencia. Por ejemplo, la secuencia AATGCAATGA traducida a código binario de acuerdo al tipo de enlace es 0001100010 y presenta dos pares de nucleótidos 00, un par de nucleótidos 01 y dos pares de nucleótidos 10 con un total de siete ceros y tres unos. El índice de IDH, d, se calcula a partir de estos datos. El valor de IDH para el tipo de enlace de la secuencia AATGCAATGA se calcula de la siguiente manera:

$$d = \frac{N_{00}N_{11} - N_{01}N_{10}}{N_0 N_1} = \frac{(2)(0) - (1)(2)}{(7)(3)} = -0.0952381$$

Donde: N_{ij} es el número de ij dinucleótidos ($ij = 00, 10, 01, 11$), N_0 y N_1 es el número total de ceros y de unos en la secuencia y el producto de éstos normaliza el índice al intervalo $[-1, 1]$

Cuando el IDH se calcula sobre la secuencia binaria obtenida a partir de la dicotomía fuerte-débil, se le denomina dWS y análogamente se le llama dYR y dMK para los casos purina-pirimidina y amino-ceto.

En el citado estudio, se propuso que el IDH para el caso WS es una parametrización de la energía libre (Breslauer et al, 1989) y que YR es una medida de la varianza de los ángulos estructurales, sin tener una interpretación semejante para el caso MK.

Miramontes y colaboradores en 1995 encontraron que existen características físicas de la molécula del DNA que correlacionan con el dominio taxonómico al que pertenecen las secuencias de los quince organismos que analizaron. Para su estudio analizaron los CDS (*coding secuences*) de los quince organismos disponibles hasta ese momento. A las características físicas de la molécula del DNA que correlacionan con el dominio taxonómico de las secuencias de los organismos analizados se les definieron como *estilos estructurales* de la molécula del DNA. Al graficar los valores de IDH para tipo de enlace,dWS, y tipo de base,dYR, observaron que éstos se agrupan en diferentes cuadrantes (figura 20).

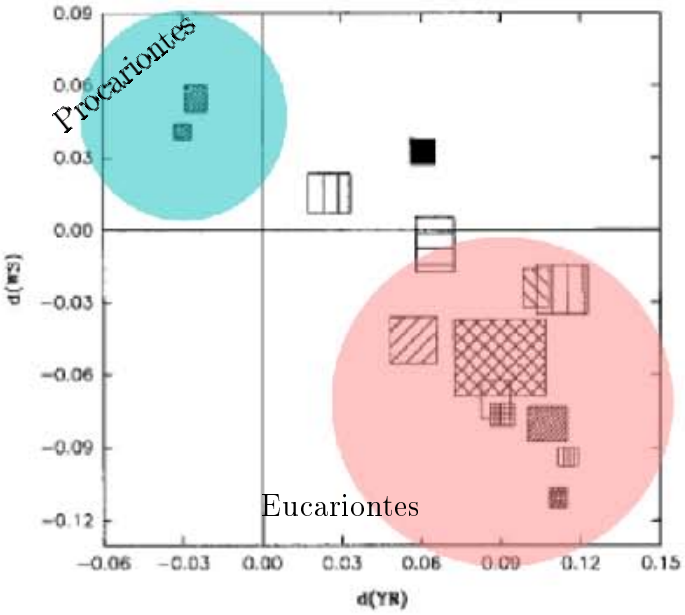

Figura 20: Gráfica de valores de IDH para tipo de base (dYR) y tipo de enlace (dWS). En el cuadrante superior izquierdo se encuentran agrupados los organismos procariontes y en el cuadrante inferior derecho se encuentran agrupados los organismos eucariontes. (imagen modificada de Miramontes *et al., 1995*

A partir de sus resultados concluyen que el genoma de los organismos analizados presenta un fenotipo interno con sus propias presiones de selección y evolución siendo la organización estructural un factor de selección tentativo.

Justificación

En 1995, Miramontes y colaboradores (Miramontes *et al.*, 1995) desarrollaron un método, el IDH, índice de homogenidad de DNA, para cuantificar la variación estructural en secuencias codificantes y encontraron que los organismos de diferentes dominios presentan estilos estructurales particulares. Aunado a ésto, la revisión de Travers y Muskhelushvili en 2015 acerca de la estructura y la función del DNA (Travers y Muskhelishvili, 2015) concluye que la información estructural en la molécula del DNA permite el empaquetamiento de la molécula.. Estudios recientes también muestran que el genoma de los procariontes y eucariontes presentan diferencias topológicas (Xing y He, 2015; Teves y Henikoff, 2014). Debido al aumento exponencial del número de genomas secuenciados de diferentes organismos se propone extender el análisis de Miramontes a un mayor número de secuencias codificantes así como a comparar los estilos estructurales de secuencias codificantes y no codificantes del mismo genoma.

Hipótesis y objetivos

Hipótesis

Las secuencias codificantes de *H.sapiens* presentan un estilo estructural diferente a las secuencias codificantes de *E. coli* K-12 MG1655, y las secuencias, codificantes y no codificantes, de *H. sapiens* presentan el mismo estilo estructural.

Objetivo general

Determinar y comparar el IDH de las secuencias codificantes y no codificantes de los genomas de *H. sapiens* y de *E. coli* K-12 MG1655.

Objetivos particulares

- Obtener secuencias codificantes y no codificantes del ensamblado del genoma de *H.sapiens* GRCh38 y del de *E. coli* K-12 MG1655.
- Determinar el IDH de las secuencias de *H.sapiens* GRCh38 y *E. coli* K-12 MG1655..
- Comparar el IDH de las secuencias codificantes de *H.sapiens* GRCh38 con el de las secuencias codificantes de *E. coli* K-12 MG1655.
- Comparar el IDH de las secuencias codificantes con el de las no codificantes de *H.sapiens* GRCh38

- Realizar una prueba estadística que permita determinar si se encuentran diferencias significativas entre los estilos estructurales de las secuencias comparadas.

Material y métodos

Obtención de secuencias codificantes y no codificantes de *H. sapiens* y *E. coli*K-12 MG1655

Los genomas utilizados en este estudio fueron el ensamblado del genoma de *H. sapiens* GRCh38 y el genoma de *E.coli* K-12 MG1655 con registro en NCBI NC_000913.3. Las secuencias de *E. coli* se obtuvieron de la base de datos Regulatory Sequence Analysis Tool (RSAT). Las secuencias de *H. sapiens* se obtuvieron de las bases de datos: CCSB Human Orfeome Collection, la base de datos Ensembl, utilizando un programa escrito en R (apéndice 1), y del UCSC Genome Browser utilizando la aplicación Table Browser (apéndice 2).

Las secuencias seleccionadas para *E. coli* K-12 MG1655 fueron $4,294$ ORF (*open reading frames*). Las secuencias codificantes seleccionadas para *H. sapiens* fueron $19,239$ ORF, y $6,915$ secuencias codificantes (CDS, *coding secuence*). Se tomaron en cuenta los ORF y las CDS de manera separada porque las secuencias ORF presentan exones e intrones, y las secuencias CDS son todos los exones concatenados que son traducidos a proteínas (figura 21). Como secuencias no codificantes se seleccionaron $320,089$ secuencias Alu, al ser el tipo de secuencia más abundante de los elementos SINE en el genoma humano (Gu *et al.*, 2000).

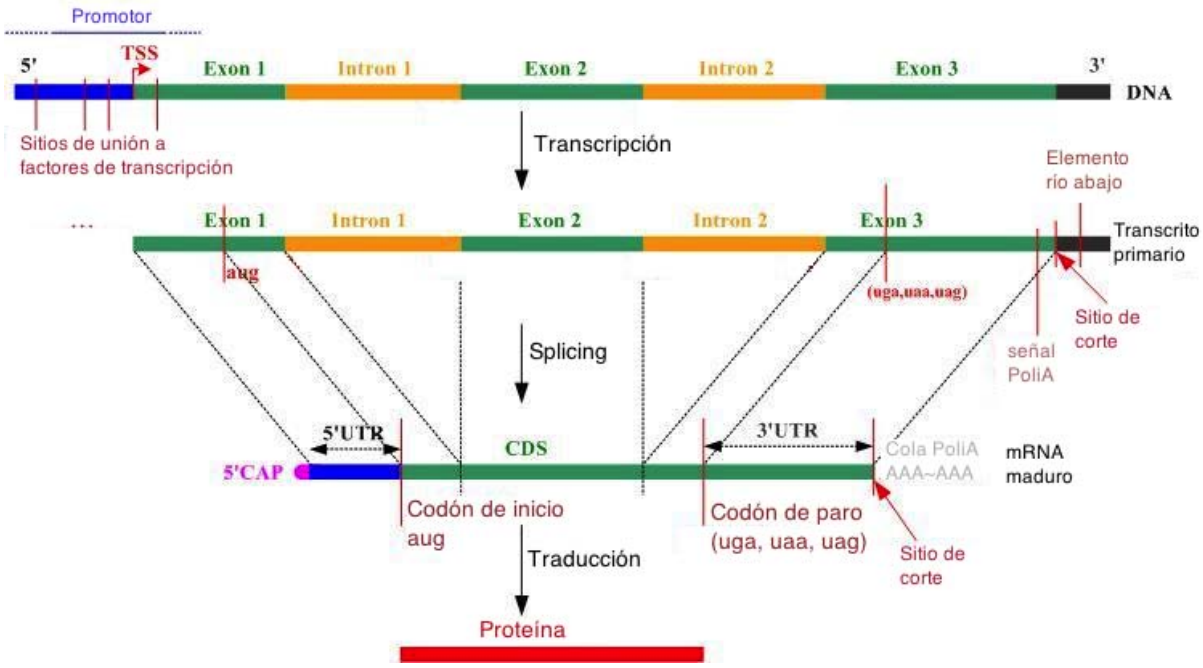

Figura 21: Esquema que muestra la diferencia entre un ORF y una secuencia CDS. La secuencia de DNA presenta la región promotora, los sitios de unión de factores de transcripción, intrones y exones. El transcrito primario presenta el codón de inicio (aug) y los codones de paro (uga, uaa y uag), exones e intrones. Después del splicing se obtiene el mRNA maduro que presenta las secuencias 5' y 3' UTR y los exones concatenados que serán traducidos a proteína (CDS) (figura modificada de www.biostars.org/p/47022/).

Determinación del IDH a las secuencias de *H. sapiens* y *E. coli* K-12 MG1655

Se escribió un programa en el lenguaje de programación Python con la biblioteca Biopython para leer las diferentes secuencias, traducirlas a código binario de acuerdo al tipo de propiedad estructural y posteriormente obtener sus valores de IDH (figura 22)(apéndice 1).

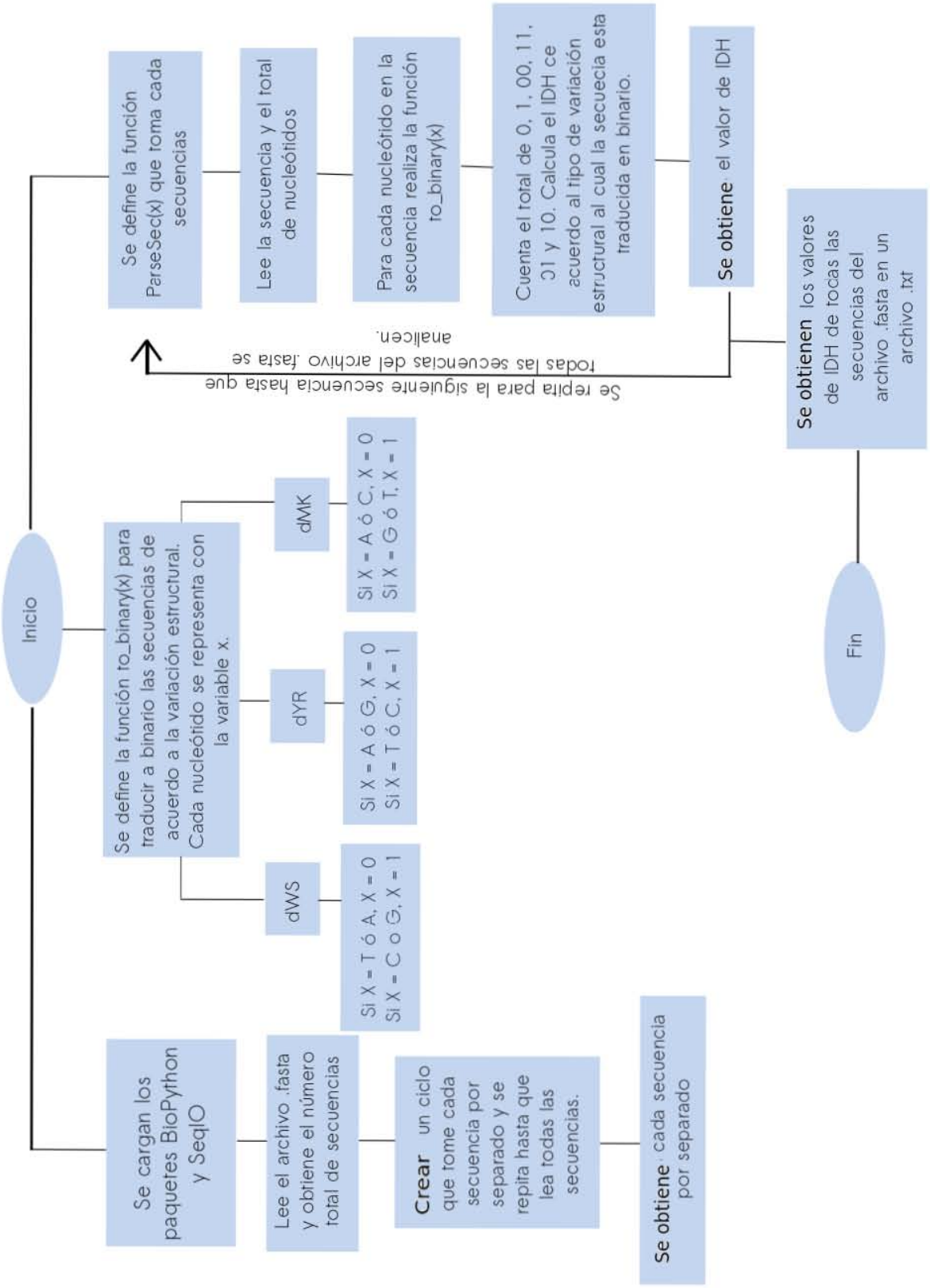

Figura 22: Diagrama de flujo que explica el programa escrito en Python para obtener los valores de IDH de dWS, dYR y dMK de las secuencias de *E. coli* y de *H. sapiens*.

El primer paso del programa fue leer las secuencias, de manera independiente, de cada organismo. Posteriormente se definió la función para traducir las secuencias a código binario de acuerdo a la propiedad estructural. Por último, se contabilizó el número de $00, 01, 11, 10$ y el total de ceros y unos de las secuencias para obtener los valores de IDH y se guardaron en un archivo de texto.

A partir de los valores de IDH con los tres tipos de propiedades estructurales, se hicieron tres diferentes gráficas de dispersión en tres dimensiones con un programa escrito en el lenguaje de programación R (apéndice 1). Para el primer análisis se compararon los ORF con los CDS de *H. sapiens*, para el segundo análisis se compararon los ORF de *H. sapiens* con los ORF de *E. coli* K-12 MG1655 y para el tercer análisis se compararon los ORF de *H. sapiens* con las secuencias Alu de *H. sapiens*.

Por último se hicieron las pruebas estadísticas de Welch con un programa escrito en el lenguaje de programación R (apéndice 1) de los valores de IDH de los tres tipos de propiedades. Las pruebas estadísticas se hicieron para evaluar si los ORF y CDS de *H. sapiens*, los ORF de *H.sapiens* y los ORF de *E. coli* K-12 MG1655 , y las secuencias Alu y los ORF de *H. sapiens* presentaban diferencias significativas.

Resultados y discusión

A las secuencias ORF de *E. coli*, los ORF, las CDS y las secuencias alu de *H.sapiens* se les determinó el IDH para dWS, dYR y dMK. A partir de los tres valores de IDH se hicieron gráficas de dispersión para comparar los estilos estructurales de la molécula del DNA de *H. sapiens* y de *E. coli*. Además se realizó la prueba estadística de Welch (Welch, 1947) para determinar si hay diferencias significativas entre los estilos estructurales de las secuencias de *H. sapiens* y de *E. coli*

Debido a que las secuencias ORF de *H.sapiens* presentan intrones se compararon sus valores de IDH con los valores de IDH de las secuencias CDS de *H. sapiens* para mostrar que presentan el mismo estilo estructural y por lo tanto, poder compararlas con los ORF de *E. coli* y las secuencias Alu.

La comparación de los valores de IDH de las secuencias CDS y ORF de *H. sapiens* muestran que pese a que los ORF contienen intrones, se comportan de la misma manera que las secuencias codificantes. La gráfica de dispersión en tres dimensiones muestra que ambas nubes de puntos se encuentran en la misma región (figura 23). El valor p de la prueba t de Welch para estas secuencias es de 0.7221 por lo que no se rechaza la hipótesis nula y se concluye que las distribuciones tienen la misma media. A partir de la gráfica de dispersión y la prueba estadística podemos afirmar que las secuencias CDS y los ORF de *H. sapiens* presentan el mismo estilo estructural.

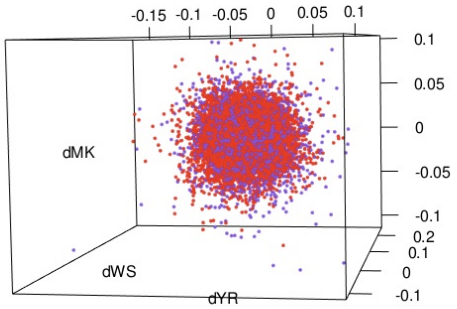

Figura 23: Gráfica de dispersión en tres dimensiones de valores de IDH para ORF (rojo) y CDS (morado) de *H. sapiens*. En el eje x se encuentran los valores de dYR, en el eje y se encuentran los valores de dWS y en el eje z se encuentran los valores de dMK.

A partir de los valores de IDH de los ORF de *E. coli* y de los valores de IDH los ORF de *H. sapiens* se construyó la gráfica de dispersión en tres dimensiones para comparar el estilo estructural de la molécula del DNA de ambos organismos. En esta gráfica (figura 24) se observa que la nube de puntos de los valores de IDH de las secuencias ORF de *H. sapiens* se agrupa de manera independiente que la nube de puntos de los valores de IDH de los ORF de *E. coli*. Ambas nubes de puntos presentan un desplazamiento en dirección opuesta. El valor p de la prueba de Welch fue de 6.57 x 10^{-16} por lo que se rechaza la hipótesis nula y se concluye que las muestras tienen diferente estilo estructural. De acuerdo a este resultado se confirma lo propuesto por Miramontes y colaboradores en 1995 de que organismos de dominios diferentes presentan un estilo estructural distinto.

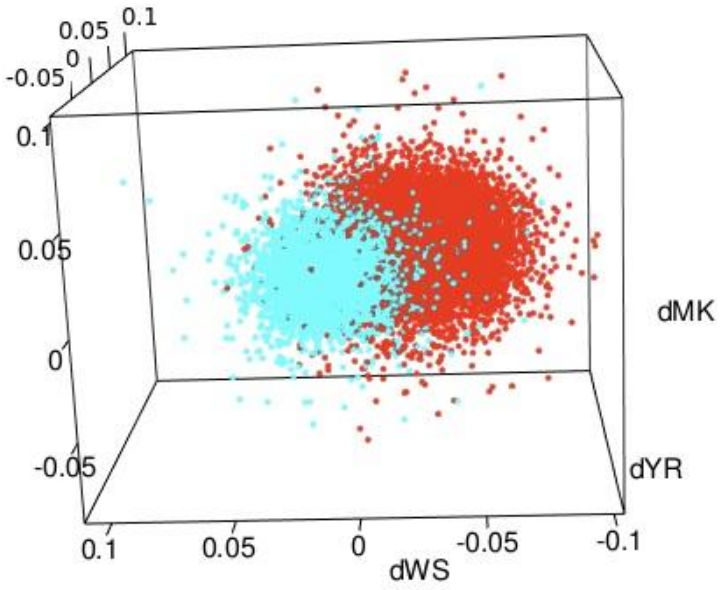

Figura 24: Gráfica de dispersión en tres dimensiones de los valores de IDH de los ORF de *H. sapiens* (rojo) y *E.coli* (cían). En el eje x se encuentran los valores de dYR, en el eje y se encuentran los valores de dWS y en el eje z se encuentran los valores de dMK.

Al comparar los valores de IDH de las secuencias Alu y los valores de IDH de los ORF de *H. sapiens* se observa que la nube de puntos de las secuencias Alu y los ORF de *H. sapiens* se agrupan en la misma región de la gráfica en tres dimensiones (figura 25) con lo cual se puede argumentar que las secuencias codificantes y no codificantes presentan el mismo estilo estructural. Sin embargo el valor p de la prueba Welch fue de 2.2 x 10^{-16} por lo que se rechaza la hipótesis nula con lo cual se puede argumentar que las secuencias no presentan el mismo estilo estructural.

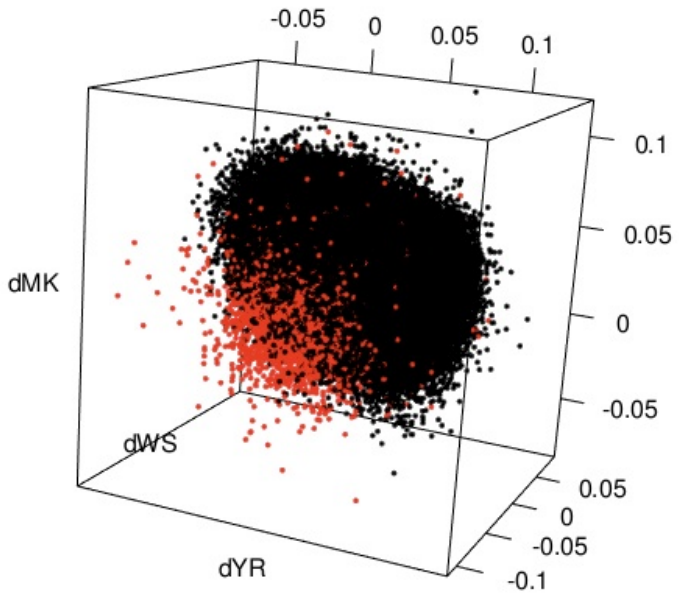

Figura 25: Gráfica de dispersión en tres dimensiones de los valores de IDH para los ORF de *H. sapiens* (rojo) y las secuencias Alu (negro). En el eje x se encuentran los valores de dYR, en el eje y se encuentran los valores de dWS y en el eje z se encuentran los valores de dMK.

Debido a los resultados anteriores, se analizó las distribuciones de los valores de IDH de las secuencias ORF y Alu de *H. sapiens* son distintas. Por lo que construimos los histogramas de los valores de dWS, dYR y dMK de las secuencias Alu (figura 26). En el histograma de los valores de dYR se observa una distribución bimodal, por lo que podemos concluir que existen dos subpoblaciones de tipos de secuencias Alu.

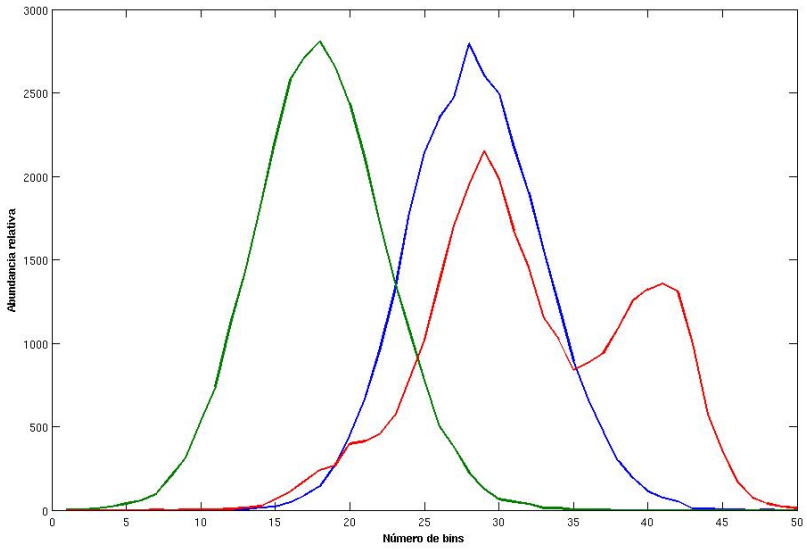

Figura 26: Histograma polar de los valores de IDH dWS (azul) , dYR (rojo) y dMK (verde) de las secuencias Alu de *H. sapiens*.

Debido a la bimodalidad de la distribución de los valores de IDH de dYR de las secuencias Alu se invalida a la prueba Welch porque ella supone que las distribuciones son normales. Consecuentemente, no podemos comparar las medias de los valores de IDH de los ORF y de las secuencias Alu de *H. sapiens*. Hasta donde sabemos, no hay pruebas paramétricas que comparen la semejanza de las medias de dos muestras cuando no se cumple el supuesto de normalidad.

La presencia de dos subpoblaciones de tipos de secuencias Alu podría explicarse debido a

la diferencia de la longitud de la cola de poli adenina. En 2002 Roy-Engel y colaboradores dividieron a los elementos Alu en las subfamilias Alu J y S; y en las subfamilias AluYb8, Alu Ya5 y Alu Ya5a2, de acuerdo a la edad relativa de inserción de las secuencias. La subfamila Alu J es la más antigua, la subfamila S es edad intermedia y el resto de las subfamilias son las más jóvenes. La longitud media de las colas de poliA's para las subfamilias Alu J y S es de 21 $\pm$ 8 y para las subfamilias AluYb8, Alu Ya5 y Alu Ya5a2 va de 26 a 30 $\pm$ 9 a 11. La diferencia de subfamilias respecto al tamaño de las colas de poli adenina se relaciona con la tendencia de mutaciones que favorece la proporción de G y C en los genomas eucariontes que menciona Xing y He en 2015 (ver antecedentes).

La diferencia de tamaños de las colas de poli adeninas podrían afectar la distribución de los valores de IDH de dYR. Recordando que índice IDH se calcula mediante la siguiente fórmula:

$$d = \frac{N_{00}N_{11} - N_{01}N_{10}}{N_0 N_1}$$

donde las adeninas toman el valor de cero cuando se traducen a binario para el tipo de base, dYR, por lo tanto N_{00} en nuestra fórmula podría ser lo suficientemente grande como para que d sea ≥ 0.

Miramontes y colaboradores en 1995 propusieron que la organización estructural del genoma de un organismo es un factor tentativo de selección y que presenta un fenotipo interno. De acuerdo a la síntesis extendida a partir de análisis genómicos, se pueden estudiar los procesos evolutivos a través de toda la escala genética de organización, enfatizando que al estudiar la interacción de los diferentes genes entre sí se puede observar su efecto en la expresión fenotípica (Pigliucci y Müller, 2010). Por lo tanto, el estudio de la organización estructural de la molécula del DNA debería de tomar en cuenta el contexto genético de las secuencias que se están analizando, es decir, la presencia de secuencias que se encuentren bajo selección positiva, las diferentes tasas de mutación y recombi-

nación, y la posición de las secuencias dentro del genoma, por mencionar algunas. De manera particular en nuestro estudio los diferentes tamaños de las colas de poli adenina en las secuencias Alu posiblemente están siendo afectados por el contexto en el que se encuentran, provocando un aumento o una disminución de la longitud de las colas de poli adenina.

Conclusión y perspectivas

La estructura de la molécula del DNA en el genoma no es uniforme. Las secuencias codificantes y no codificantes varían estructuralmente debido a la diferente composición de bases nitrogenadas y a la interacción entre los diferentes elementos que conforman a la molécula. Las diferencias estructurales que presenta la molécula son necesarias para su interacción con las proteínas. Estas diferencias estructurales también se observan en los genomas de organismos de diferentes dominios, lo cual no es de extrañarnos debido al distinto orden y proporción en que se encuentran las secuencias codificantes y no codificantes; y los diferentes mecanismos de regulación de la expresión génica y de compactación del DNA.

En el presente trabajo de acuerdo a los valores de IDH de los ORF y CDS de *H. sapiens* se concluye que pese a que los ORF presentan intrones las secuencias presentan el mismo estilo estructural. Además los valores de IDH de las secuencias codificantes de *H. sapiens*y *E. coli* se confirma que presentan un estilo estructual distinto. En cambio para las secuencias codificantes (ORF) y no codificantes (Alus) de *H. sapiens* no podemos confirmar que presenten el mismo estilo estructural debido a que los valores de dYR de las secuencias Alu presentan una distribución bimodal. Para comprobar si presentan el mismo estilo estructural se propone realizar pruebas de Monte Carlo. Estas pruebas calculan la distancia (D_0) entre las medias de las poblaciones que se quieren comparar, posteriormente se toman submuestras aleatorias de cada una de las poblaciones y se calcula nuevamente la distancia de las medias. Este procedimiento se realiza múltiples

veces con nuevas submuestras aleatorias. Por último se comparan las distancias de las medias calculadas de las submuestras con la distancia original (D_0), si la mayoría de las distancias de las submuestras es menor a D_0 se puede concluir que ambas poblaciones son iguales, es decir, presentan la misma media. Además para comprobar si las secuencias codificantes y no codificantes de *H. sapiens* presentan el mismo estilo estructural, se podría determinar el IDH de otro tipo de secuencias no codificantes como los elementos dispersos largos, LINE. Otro tipo de secuencias clasificadas como no codificantes, como los pseudogenes, podrían darnos resultados sesgados debido a que anteriormente codificaban a proteína y no podríamos considerarlos completamente como secuencias no codificantes. Los programas desarrollados en este trabajo para la descarga de secuencias y determinación del IDH podrían ser nuevamente utilizados para futuros análisis.

Debido a los resultados obtenidos de las secuencias no codificantes de *H. sapiens* proponemos que el análisis se realice en un mayor número de organismos que presenten una mayor proporción de secuencias no codificantes, por ejemplo las gimnoespermas y las angioespermas debido a que son representantes de organismos con genomas grandes donde la mayor parte de éste es no codificante. Además sería interesante analizar secuencias de organismos simbiontes para saber si el simbionte ha modificado el *estilo estructural* de su genoma debido a su interacción con el hospedero.

Apéndice 1

Programa utilizado para descargar las secuencias codificantes CDS de *H.sapiens*

A continuación se muestra el esquema que explica el programa escrito en el lenguaje de programación R para descargar las secuencias CDS de *H. sapiens* (figura 36) así como el código del mismo.

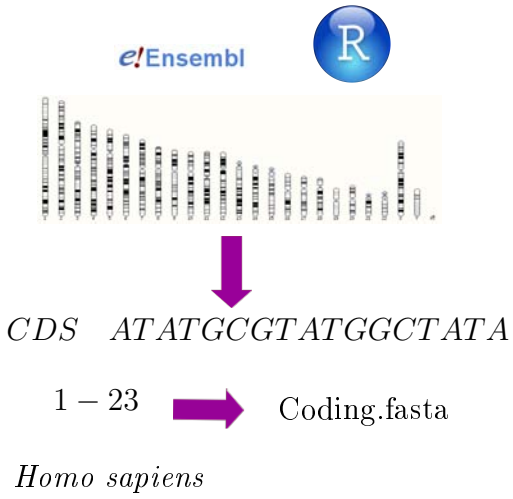

Figura 27: Diagrama que explica el programa escrito para obtener las secuencias codificantes de *H. sapiens*. A partir de la base de datos Ensembl se obtuvieron las secuencias codificantes (CDS) de *H. sapiens* por medio del lenguaje de programación R. Las secuencias se obtuvieron de cada uno de los 23 cromosomas y se guardaron en un solo archivo tipo fasta.

Código en R para descargar secuencias CDS de *H. sapiens*

```
library("biomaRt")
#browseVignettes("biomaRt")
ensembl=useMart("ensembl")
#listDatasets(ensembl)
ensembl = useMart("ensembl",dataset="hsapiens_gene_ensembl")
filters = listFilters(ensembl)
#filters
library(Biostrings)
#Secuencias codificantes

#Toma todas las secuencias codificantes de los cromosomas
# Se hace una solicitud a la base de datos de Ensembl pidiendo el tipo
#de secuencia (en este caso codificante) y la región en donde la debe de
#buscar, en nuestro caso es cada uno de los cromosomas y especificamos
#el tamaño de cada uno de los cromosomas

C1 = getSequence(chromosome=1, start=1, end=248956422,
                              type="entrezgene",seqType="coding", mart=ensembl)
C2 = getSequence(chromosome=2, start=1, end=242193529,
                                 type="entrezgene",seqType="coding", mart=ensembl)
C3 = getSequence(chromosome=3, start=1, end=198295559,
                                 type="entrezgene",seqType="coding", mart=ensembl)
C4 = getSequence(chromosome=4, start=1, end=190214555,
                                 type="entrezgene",seqType="coding", mart=ensembl)
C5 = getSequence(chromosome=5, start=1, end=181538259,
                                 type="entrezgene",seqType="coding", mart=ensembl)
```

```
C6 = getSequence(chromosome=6, start=1, end=170805979,
                                  type="entrezgene",seqType="coding", mart=ensembl)
C7 = getSequence(chromosome=7, start=1, end=159345973,
                                  type="entrezgene",seqType="coding", mart=ensembl)
C8 = getSequence(chromosome=8, start=1, end=145138636,
                                  type="entrezgene",seqType="coding", mart=ensembl)
C9 = getSequence(chromosome=9, start=1, end=138394717,
                                  type="entrezgene",seqType="coding", mart=ensembl)
C10 = getSequence(chromosome=10, start=1, end=133797422,
                                  type="entrezgene",seqType="coding", mart=ensembl)
C11 = getSequence(chromosome=11, start=1, end=135086622,
                                  type="entrezgene",seqType="coding", mart=ensembl)
C12 = getSequence(chromosome=12, start=1, end=133275309,
                                  type="entrezgene",seqType="coding", mart=ensembl)
C13 = getSequence(chromosome=13, start=1, end=114364328,
                                  type="entrezgene",seqType="coding", mart=ensembl)
C14 = getSequence(chromosome=14, start=1, end=107043718,
                                  type="entrezgene",seqType="coding", mart=ensembl)
C15 = getSequence(chromosome=15, start=1, end=101991189,
                                  type="entrezgene",seqType="coding", mart=ensembl)
C16 = getSequence(chromosome=16, start=1, end=90338345,
                                  type="entrezgene",seqType="coding", mart=ensembl)
C17 = getSequence(chromosome=17, start=1, end=83257441,
                                  type="entrezgene",seqType="coding", mart=ensembl)
C18 = getSequence(chromosome=18, start=1, end=80373285,
                                  type="entrezgene",seqType="coding", mart=ensembl)
C19 = getSequence(chromosome=19, start=1, end=58617616,
                                  type="entrezgene",seqType="coding", mart=ensembl)
```

```
C20 = getSequence(chromosome=20, start=1, end=64444167,
                                  type="entrezgene",seqType="coding", mart=ensembl)
C21 = getSequence(chromosome=21, start=1, end=46709983,
                                  type="entrezgene",seqType="coding", mart=ensembl)
C22 = getSequence(chromosome=22, start=1, end=50818468,
                                  type="entrezgene",seqType="coding", mart=ensembl)
C23 = getSequence(chromosome=23, start=1, end=156040895,
                                  type="entrezgene",seqType="coding", mart=ensembl)

#Juntamos las secuencias de cada cromosoma en un solo objeto
Coding= data.frame()
Coding=rbind(Coding, as.list(C1,C2,C3,C4,C5,C6,C7,C8,C9,
C10,C11,C12,C13,C14,C15,C16,C17,C18,C19,C20,C21,C22,C23))
#Eliminamos todas las entradas que no presenten secuencias
pos=which(Coding[,1] == "Sequence unavailable")
Coding=Coding[-pos,]
#Las secuencias obtenidas las pasamos a un archivo .fasta
library(seqinr) #Para crear archivos fasta utilizamos este paquete
write.fasta(as.list(Coding[,1]), names= as.list(Coding[,2]),
file.out="coding.fa", open="w", nbchar=60, as.string=FALSE)
```

Programa escrito en Python para determinar los valores de dWS, dYR y dMK

Función para traducir a código binario de acuerdo a dWS

```
#Para obtener dWS
```

```
from string import *
letras = ['A', 'C', 'G', 'T']
#Se define una funcion para traducir a binario
def to_binarydWS(x):
  if x=='A' or x=='T'or x=='a' or x=='t':
    return '0'
  elif x=='C' or x=='G' or x=='c' or x =='g':
    return '1'
  else:
    pass
    #print(x.id)
```

Función para traducir a código binario de acuerdo a dYR

```
#Para obtener dYR
from string import *
letras = ['A', 'C', 'G', 'T']
#Se define una funcion para traducir a binario
def to_binarydYR(x):
  if x=='A' or x=='G'or x=='a' or x=='g':
    return '0'
  elif x=='C' or x=='T' or x=='c' or x =='t':
    return '1'
  else:
    pass
    #print(x.id)
```

Función para traducir a código binario de acuerdo a dMK

```
#Para obtener dMK
from string import *
letras = ['A', 'C', 'G', 'T']
#Se define una funcion para traducir a binario
def to_binarydMK(x):
  if x=='A' or x=='C'or x=='a' or x=='c':
    return '0'
  elif x=='T' or x=='G' or x=='t' or x =='g':
    return '1'
  else:
    pass
    #print(x.id)
```

Función para hacer un ciclo en un archivo .fasta y leer cada secuencia por separado, después con la función TO_ BINARY(X) anterior se traducen las secuencias a código binario. A continuación se muestra el código para dWS. Los códigos para dYR y dMK, son equivalentes sólo se debe cambiar la función TO_BINARY (X) a la correspondiente para cada uno de ellos.

```
#Se define una funcion para hacer un ciclo sobre nuestras secuencias las cuales
#pasaran a binario con la funcion anterior


def ParseSec(x):
   seqWS = ''
   x = next(record_iterator)
   dna = x.seq
   r = range(len(dna))
   #print(len(dna))
   for i in r:
      if dna[i] in letras:
        seqWS = seqWS + to_binarydWS(dna[i])
   seqWS = " ".join(seqWS[i:i+2] for i in range(0, len(seqWS), 2))
   d0WS = seqWS.count('0')
   d1WS = seqWS.count('1')
   d00WS = seqWS.count('00')
   d11WS = seqWS.count('11')
   d01WS = seqWS.count('01')
   d10WS = seqWS.count('10')
   dws = (1.0*d00WS*d11WS-d10WS*d01WS)/(d0WS*d1WS)
   #print(len(seqWS), 'long bin')
   #print(dws, 'dWS')
   print(dws)
```

```
#Con el paquete de Biopython hacemos un interador de
#secuencias de nuestro archivo fasta
from Bio import SeqIO
record_iterator = SeqIO.parse("ecoli.fasta", "fasta")
#obtenemos el numero de secuencias en nuestro archivo fasta
records = list(SeqIO.parse("ecoli.fasta", "fasta"))
#el numero de registros sera la longitud en que trabaje la funcion ParseSec
recrange = range(len(records))
for i in recrange:
   ParseSec(record_iterator)
```

Programa para hacer las gráficas de dispersión de valores de IDH

```
#Se abren los archivos .txt de los distintos valores de IDH de todas las secuencias.

#Para alus
AludYR= data.matrix(read.table("GenomaAlusYR.txt"))
AludWS = data.matrix(read.table("GenomaAlusWS.txt"))
AludMK = data.matrix(read.table("GenomaAlusMK.txt"))

#Para ORF H.sapiens
sapiensdYR = data.matrix(read.table("dYRORF2.txt"))
sapiensdWS = data.matrix(read.table("dWSORF2.txt"))
sapiensdMK = data.matrix(read.table("dMKORF.txt"))

#Para CDS
```

```
sapiensCdYR = data.matrix(read.table("dYRcoding.txt"))
sapiensCdWS = data.matrix(read.table("dWScoding.txt"))
sapiensCdMK = data.matrix(read.table("dMKcoding.txt"))

#Para coli
colidYR = data.matrix(read.table ("dYRcoli.txt"))
colidWS = data.matrix(read.table("dWScoli.txt"))
colidMK = data.matrix(read.table("dMKcoli.txt"))

#Para graficas 3D

library(rgl)

#ORFsapiens vs Alu
plot3d(AludYR, AludWS, AludMK,
xlab="dYR", ylab="dWS", zlab="dMK")
points3d(sapiensdYR, sapiensdWS, sapiensdMK, col="red")

#ORFsapiens vs CDS
plot3d(sapiensdYR, sapiensdWS, sapiensdMK, col="red",
xlab="dYR", ylab="dWS", zlab="dMK")
points3d(sapiensCdYR, sapiensCdWS, sapiensCdMK, col="blueviolet")

#ORFsapiens vs ORFcoli
plot3d(sapiensdYR, sapiensdWS, sapiensdMK, col="red",
xlab="dYR", ylab="dWS", zlab="dMK")
points3d(colidYR, colidWS, colidMK, col="cyan")
```

Programa escrito en R para la prueba de Welch

```
#Para realizar la prueba Welch
#Se abren los archivos .txt de los distintos valores de IDH de todas las secuencias.

#Alus
AludYR= data.matrix(read.table("GenomaAlusYR.txt"))
AludWS = data.matrix(read.table("GenomaAlusWS.txt"))
AludMK = data.matrix(read.table("GenomaAlusMK.txt"))

#Para ORF H.sapiens
sapiensdYR = data.matrix(read.table("dYRORF2.txt"))
sapiensdWS = data.matrix(read.table("dWSORF2.txt"))
sapiensdMK = data.matrix(read.table("dMKORF.txt"))

#Para CDS
sapiensCdYR = data.matrix(read.table("dYRcoding.txt"))
sapiensCdWS = data.matrix(read.table("dWScoding.txt"))
sapiensCdMK = data.matrix(read.table("dMKcoding.txt"))

#Para coli
colidYR = data.matrix(read.table ("dYRcoli.txt"))
colidWS = data.matrix(read.table("dWScoli.txt"))
colidMK = data.matrix(read.table("dMKcoli.txt"))

#Hacemos una matrix de todas las secuencias
Alus = matrix()
```

```
Alus = cbind(AludYR, AludWS, AludMK)
ORFsapiens = matrix()
ORFsapiens = cbind(sapiensdYR, sapiensdWS, sapiensdMK)
Coding = matrix()
Coding = cbind(sapiensCdYR, sapiensCdWS, sapiensCdMK)
ORFcoli = matrix()
ORFcoli = cbind(colidYR, colidWS, colidMK)

#Prueba de Welch
t.test(ORFsapiens, Coding)
t.test(ORFsapiens, Alus)
t.test(ORFsapiens, sAlus)
t.test(ORFsapiens, ORFcoli)
```

Apéndice 2

Parámetros utilizados en la aplicación Table Browser del UCSC Genome Browser

Las secuencias Alu de *H. sapiens* se obtuvieron de la base de datos del UCSC Genome Browser utilizando la aplicación Table Browser, los parámetros utilizados para descargar las secuencias de todo el genoma son los siguientes.

Tabla 3 : Parámetros utilizados para descargar secuencias Alu de la aplicación Table Browser	
Clade	Mammal
Genome	Human
Assembly	Dec. 2013 (GRCh38/hg38)
Group	Repeats
Track	RepeatMasker
table	rmsk
Region	genome
Filter	edit
Output format	sequence
file type returned	gzip compressed

Dentro del parámetro Filter se editaron los siguientes parámetros (Tabla 4)

Tabla 4 : Parámetros editados en el parámetro Filter de la aplicación Table Browser	
repClass	SINE
repFamily	Alu

Bibliografía

Abu El-Soud, W. Characterization of histone modifying enzymes and their substrates in maize (2007)

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., y Walter, P. *Molecular Biology of the Cell. New York: Garland Science; 2002.* 5ª edición. New York: Garland Science (2002)

Appel, T. *The Cuvier-Geoffroy Debate. French Biology in the Decades before Darwin.* Oxford University Press, Nueva York, Estados Unidos (1987)

Birdsell, J.A. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Molecular biology and evolution* **19**(7):1181–1197 (2002)

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., y Thurman, R.E. Identification and analysis of functional elements in 1 % of the human genome by the ENCODE pilot project. *Nature* **447**(7146):799–816 (2007)

Blattner, F., Plunkett G, I., Bloch, C., Perna, N., Burland, V., Riley, M., Collado-Vides, J., Glasner, J., Rode, C., Mayhew, G., Gregor, J., Davis, N., Kirkpatrick, H., Goeden, M., Rose, D., Mau, B., y Shao, Y. The Complete Genome Sequence of Escherichia coli K-12. *Science* **277**(September):1453–1462 (1997)

Breslauer, K.J., Frank Ronald Blöcker, H., y Marky, L.A. Predicting DNA duplex stability from the base sequence. *Proceedings of the National Academy of Sciences* **83**(11):3746–3750 (1986)

Brown, T. *Introduction to genetics: a molecular approach.* Garland Science (2011)

Calladine, C.R., Drew, H.R., Luisi, B.F., y Travers, A.A. *Understanding DNA. The Molecule and How It Works.* Tercera edición. Elsevier Academic Press, Chennai, India (2004)

CCSB, Center of Cancer Systems Biology. Human ORFeome (Consultado en Diciembre 2014). `http://horfdb.dfci.harvard.edu/hv7/index.php?page=home`

Clamp, M., Fry, B., Kamal, M., Xie, X., Cuff, J., Lin, M.F., Kellis, M., Lindblad-Toh, K., y Lander, E.S. Distinguishing protein-coding and noncoding genes in the human genome. *Proceedings of the National Academy of Sciences* **104**(49):19428–19433 (2007)

Crick, F.H. *et al.* The origin of the genetic code. *Journal of molecular biology* **38**(3):367–379 (1968)

Darwin, C. *On the origins of species by means of natural selection.* London: Murray (1859)

Deininger, P. Alu elements: know the SINEs. *Genome Biol* **12**(12):236 (2011)

Doolittle, W.F. y Sapienza, C. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**(5757):601–3 (1980)

Dridi, S. Alu mobile elements: from junk DNA to genomic gems. *Scientifica* **2012** (2012)

Duret, L. y Arndt, P.F. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet* **4**(5):e1000071 (2008)

Elgar, G. y Vavouri, T. Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends in genetics* **24**(7):344–352 (2008)

ENCODE Project Consortium. The ENCODE (ENCyclopedia of DNA elements) project. *Science* **306**(5696):636–640 (2004)

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., y Fitzgerald, S. Ensembl 2014. *Nucleic acids research* pág. gkt1196 (2013)

Foerstner, K.U., Von Mering, C., Hooper, S.D., y Bork, P. Environments shape the nucleotide composition of genomes. *EMBO reports* **6**(12):1208–1213 (2005)

Franklin, R.E. y Gosling, R. The structure of sodium thymonucleate fibres. I. The influence of water content. *Acta Crystallographica* **6**(8-9):673–677 (1953)

Futuyma, D.J. *Evolution. Sunderland, MA*, tomo 4. USA: Sinauer Associates. Gouy MAM, Guindon S, Gascuel O (2010). SeaView version (2005)

García-Sancho, M. *Biology, computing, and the history of molecular sequencing: from proteins to DNA, 1945-2000.* Palgrave Macmillan (2012)

Gould, S.J. Darwinism and the Expansion of Evolutionary Theory. *Sciencie* **216**(4544):380–387 (1982)

Gould, S.J. *The Structure of Evolutionary Theory.* The Belknap Press of Harvard University Press, Cambridge, Massachusetts (2002)

Gould, S.J. y Lewontin, R.C. The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proceedings of the Royal Society of London B: Biological Sciences* **205**(1161):581–598 (1979)

Gu, Z., Wang, H., Nekrutenko, A., y Li, W.H. Densities, length proportions, and other distributional features of repetitive sequences in the human genome estimated from 430 megabases of genomic sequence. *Gene* **259**(1):81–88 (2000)

Hartl, D.L. y Ruvolo, M. *Genetics*. Jones & Bartlett Publishers (2011)

Huxley, J. *Evolution. The Modern Synthesis*. London: George Alien & Unwin Ltd. (1942)

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**(7011):931–945 (2004)

Kapitonov, V. y Jurkal, J. The age of Alu subfamilies. *Journal of molecular evolution* **42**(1):59–65 (1996)

Karolchik, D., Hinrichs, A., Furey, T., Roskin, K., Sugnet, C., Haussler, D., y Kent, W. The UCSC Table Browser data retrieval tool. (Consultado en Diciembre 2014). `http://genome.ucsc.edu/cgi-bin/hgTables`

Kimura, M. y Ota, T. On some principles governing molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America* **71**(7):2848–2852 (1974)

Klug, W.S., Cummings, M.R., Spencer, C.A., y Ward, S.M. *Concepts of genetics*. 10ª edición. Pearson Education (2012)

Koo, H.S., Wu, H.M., y Crothers, D.M. DNA bending at adenine· thymine tracts. *Nature* **320**(6062):501–506 (1986)

Koonin, E.V. Darwinian evolution in the light of genomics. *Nucleic Acids Research* **37**(4):1011–1034 (2009)

Krebs, J.E., Goldstein, E.S., y Kilpatrick, S.T. *Lewin's genes X*. Jones & Bartlett Publishers (2009)

Kumar, S., Kingsley, C., y Distefano, J.K. Genome Mapping and Genomics in Human and Non-Human Primates págs. 7–31 (2015)

Lesk, A. *Introduction to bioinformatics*. Oxford University Press (2013)

Luijsterburg, M.S., White, M.F., van Driel, R., y Dame, R.T. The major architects of chromatin: architectural proteins in bacteria, archaea and eukaryotes. *Critical reviews in biochemistry and molecular biology* **43**(6):393–418 (2008)

Lynch, M. y Conery, J.S. The origins of genome complexity. *Science (New York, N.Y.)* **302**(2003):1401–1404 (2003)

Manly, B. Randomization and Monte Carlo methods in biology. *NewYork: Chapman HallManlyRandomization and Monte Carlo methods in biology* (1991)

Miramontes, P., Medrano, L., Cerpa, C., Cedergren, R., Ferbeyre, G., y Cocho, G. Structural and thermodynamic properties of DNA uncover different evolutionary histories. *Journal of molecular evolution* **40**(6):698–704 (1995)

Morris, K.V. *Non-coding RNAs and epigenetic regulation of gene expression: Drivers of natural selection*. Horizon Scientific Press (2012)

Nature Education. Gene expression Scitable Nature Education (Consultado 2015-28-08). `http://www.nature.com/scitable/topicpage/gene-expression-14121669`

Nelson, D.L., Lehninger, A.L., y Cox, M.M. *Lehninger principles of biochemistry*. Macmillan (2008)

Ohta, T. The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics* **23**:265–286 (1992)

Palazzo, A.F. y Gregory, T.R. The Case for Junk DNA. *PLoS Genetics* **10**(5) (2014)

Pennisi, E. Working the (gene count) numbers: finally, a firm answer? *Science* **316**(5828):1113–1113 (2007)

Pigliucci, M.M. y Müller, G.B. *Evolution-the extended synthesis*. MIT Press (2010)

Roy-Engel, A.M., Salem, A.H., Oyeniran, O.O., Deininger, L., Hedges, D.J., Kilroy, G.E., Batzer, M.A., y Deininger, P.L. Active Alu element ."A-tails": Size does matter. *Genome Research* **12**(9):1333–1344 (2002)

Satchwell, S.C., Drew, H.R., y Travers, A.A. Sequence periodicities in chicken nucleosome core DNA. *Journal of molecular biology* **191**(4):659–675 (1986)

Spencer, C.C., Deloukas, P., Hunt, S., Mullikin, J., Myers, S., Silverman, B., Donnelly, P., Bentley, D., y McVean, G. The influence of recombination on human genetic diversity. *PLoS Genetics* (2006)

Sperling, A.S., Jeong, K.S., Kitada, T., y Grunstein, M. Topoisomerase II binds nucleosome-free DNA and acts redundantly with topoisomerase I to enhance recruitment of RNA Pol II in budding yeast. *Proceedings of the National Academy of Sciences* **108**(31):12693–12698 (2011)

Struhl, K. y Segal, E. Determinants of nucleosome positioning. *Nature structural & molecular biology* **20**(3):267–273 (2013)

Teves, S.S. y Henikoff, S. Transcription-generated torsional stress destabilizes nucleosomes. *Nature structural & molecular biology* **21**(1):88–94 (2014)

Travers, A. The evolution of the genetic code revisited. *Origins of Life and Evolution of Biospheres* **36**(5-6):549–555 (2006)

Travers, A. y Muskhelishvili, G. DNA structure and function. *FEBS Journal* (2015)

Watson, J.D. y Crick, F.H. Molecular structure of nucleic acids. *Nature* **171**(4356):737–738 (1953)

Welch, B.L. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika* págs. 28–35 (1947)

WikimediaCommons. A-DNA, B-DNA and Z-DNA (2015). http://commons.wikimedia.org/wiki/File:A-DNA,_B-DNA_and_Z-DNA.png#mediaviewer/File:A-DNA,_B-DNA_and_Z-DNA.png

Woese, C.R. y Fox, G.E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences* **74**(11):5088–5090 (1977)

Wright, S. Evolution in Mendelian populations. *Genetics* **16**(2):97 (1931)

Wright, S. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *Proc. 6th Int. Cong. Genetics* **1**(356-366) (1932)

Wright, S. *Evolution and the Genetics of Populations: A Treatise in Four Volumes: Vol. 4: Variability Within and Among Natural Populations*. University of Chicago Press (1978)

Xing, K. y He, X. Mutation bias, rather than binding preference, underlies the nucleosome-associated G+ C % variation in eukaryotes. *Genome biology and evolution* **7**(4):1033–1038 (2015)

Xu, H. y Morrical, S.W. Protein Motifs for DNA Binding. *eLS* (2010)

Zhao, J., Bacolla, A., Wang, G., y Vasquez, K.M. Non-B DNA structure-induced genetic instability and evolution. *Cellular and Molecular Life Sciences* **67**(1):43–62 (2010)

Zhurkin, V.B. Specific alignment of nucleosomes on DNA correlates with periodic distribution of purine—pyrimidine and pyrimidine—purine dimers. *Febs Letters* **158**(2):293–297 (1983)

Zuckerkandl, E. On the molecular evolutionary clock. *Journal of Molecular Evolution* **26**(1-2):34–46 (1987)

ZUCKERKANDL, E. Y PAULING, L. Evolutionary divergence and convergence in proteins. *Evolving genes and proteins* **97**:97–166 (1965)