
**UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO**
**POSGRADO EN CIENCIAS BIOLÓGICAS**
FACULTAD DE CIENCIAS
BIOLOGÍA EXPERIMENTAL

# ANÁLISIS GLOBAL DE PATRONES PERIÓDICOS EN EL GENOMA

# TESIS

QUE PARA OPTAR POR EL GRADO DE:

**DOCTORA EN CIENCIAS**

PRESENTA:

**DANIELA SOSA PEREDO**

**TUTOR PRINCIPAL DE TESIS:**
**DR. PEDRO EDUARDO MIRAMONTES VIDAL**
**FACULTAD DE CIENCIAS**

**COMITÉ TUTOR:**
**DR. GERMINAL COCHO GIL**
**INSTITUTO DE FÍSICA**
**DR. CARLOS ARTURO II BECERRA BRACHO**
**FACULTAD DE CIENCIAS**

**MÉXICO, D.F. DICIEMBRE, 2013.**


**UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO**
**POSGRADO EN CIENCIAS BIOLÓGICAS**
FACULTAD DE CIENCIAS
BIOLOGÍA EXPERIMENTAL

# ANÁLISIS GLOBAL DE PATRONES PERIÓDICOS EN EL GENOMA

# TESIS

QUE PARA OPTAR POR EL GRADO DE:

**DOCTORA EN CIENCIAS**

PRESENTA:

**DANIELA SOSA PEREDO**

**TUTOR PRINCIPAL DE TESIS:**
**DR. PEDRO EDUARDO MIRAMONTES VIDAL**
**FACULTAD DE CIENCIAS**

**COMITÉ TUTOR:**
**DR. GERMINAL COCHO GIL**
**INSTITUTO DE FÍSICA**
**DR. CARLOS ARTURO II BECERRA BRACHO**
**FACULTAD DE CIENCIAS**

**MÉXICO, D.F. DICIEMBRE, 2013.**

VNIVERSIDAD NACIONAL AVTNOMA DE MEXICO

POSGRADO EN CIENCIAS BIOLÓGICAS
FACULTAD DE CIENCIAS
DIVISIÓN DE ESTUDIOS DE POSGRADO

OFICIO FCIE/DEP/606/13

ASUNTO: Oficio de Jurado

**Dr. Isidro Ávila Martínez**
**Director General de Administración Escolar, UNAM**
**P r e s e n t e**

Me permito informar a usted que en la reunión ordinaria del Comité Académico del Posgrado en Ciencias Biológicas, celebrada el día **9 de septiembre de 2013**, se aprobó el siguiente jurado para el examen de grado de **DOCTORA EN CIENCIAS** del (la) alumno (a) **SOSA PEREDO DANIELA** con número de cuenta **96532669** con la tesis titulada: **"Análisis Global de Patrones Periódicos en el Genoma"**, realizada bajo la dirección del (la**) DR. PEDRO EDUARDO MIRAMONTES VIDAL:**

| | |
|---|---|
| Presidente: | DR. GERMINAL COCHO GIL |
| Vocal: | DR. MARCO ANTONIO JOSÉ VALENZUELA |
| Secretario: | DR. LUIS FELIPE JIMÉNEZ GARCÍA |
| Suplente: | DR. LUIS JOSÉ DELAYE ARREDONDO |
| Suplente | DR. ARTURO CARLOS II BECERRA BRACHO |

Sin otro particular, me es grato enviarle un cordial saludo.

**A t e n t a m e n t e**
"POR MI RAZA HABLARA EL ESPÍRITU"
Cd. Universitaria, D.F. a 8 de noviembre de 2013.

Dra. María del Coro Arizmendi Arriaga
Coordinadora del Programa

FACULTAD DE CIENCIAS
DIVISION DE ESTUDIOS DE POSGRADO

**MCAA/MJFM/ASR/grf***

# Agradecimientos

Al Posgrado en Ciencias Biológicas, UNAM.
Al apoyo recibido por parte del CONACYT
Muy especialmente a mi tutor principal Pedro Miramontes y a mi comité tutoral Germinal Cocho Gil y Arturo Becerra Bracho por su buena dirección y su gran apoyo en este proyecto

# Agradecimientos personales

A Pedro Miramontes por el invaluable impulso que me ha dado, la dedicación, la confianza, el ejemplo de ser una persona con enormes valores y principios, su gran amistad y por todo lo gratificante y enriquecedor que ha sido trabajar con él.

A Marco José Valenzuela, por su gran apoyo y colaboración en este trabajo, su increíble entusiasmo, el valioso tiempo que dedico al proyecto y sobre todo por la confianza en mi trabajo y la amistad que me ha brindado.

Al Lic. en informática Juan Roman Bobadilla por el apoyo técnico en computación.

A Cristina Peredo por todo el esfuerzo que ha hecho para que pudiera realizar este trabajo y muchos logros más en mi vida.

A Victor Mireles, por ser un extraordinario amigo y compañero de trabajo en este proyecto y por su valiosa colaboración en programación y sus estrictos cuestinamientos para lograr resultados objetivos.

A Luis Delaye y Germinal Cocho por seguir orientando el timón de mi carrera y todo el apoyo que me han dado...

A Luis Fernandez, Rosalía Peredo y Engracia Meléndez por su incondicional compañia y cariño

A mi familia Edgar, Vale y Max por convertirse en mi gran motor.

A mis amigas Lorena, Mariana y Natalia, con quienes pude enriquecer el contenido de este trabajo compartiendo muchas pláticas en tantos agradables momentos y quienes me ayudaron a ubicar las prioridades cuando fue necesario, :D.

*Dedicado a*
*mi mamá*

# Índice general

# Índice de figuras

# Resumen

La secuenciación masiva de genomas ha facilitado importantes avances en el estudio del origen y la evolución de los organismos; con el manejo de genomas completos y la identificación de secuencias codificantes, se ha podido constatar que existe una serie de interacciones complejas en el genoma que intervienen en los procesos dinámicos que dan lugar a la gran variabilidad fenotípica de los organismos. En estas interacciones, las regiones no codificantes del genoma, como se plantea en este trabajo, juegan un papel estratégico. Por esta razón, se tuvo como objetivo explorar la estructura del genoma usando herramientas matemáticas para el análisis de periodicidad, con la finalidad de buscar y describir patrones en la arquitectura genómica que pudieran explicar la relevancia biológica y funcional de estas cualidades estructurales. Con las distintas lecturas que se le pueden dar al genoma, describiéndolo con dichas características, fue posible realizar estos análisis. Los primeros resultados de las exploraciones aplicadas en algunos genomas, muestran la relación que puede existir entre la naturaleza biológica de los organismos con algunos patrones encontrados. Posteriormente, se reconocieron elementos modulares y de presencia periódica que permitieron identificar, entre los primates, patrones estructurales muy similares en la distribución periódica de secuencias específicas asociadas a la formación de complejos nucleosómicos. También se encontró que estas secuencias presentan algunas periodicidades no reportadas en los genomas y otras que están asociadas a secuencias altamente repetidas entre las que se encuentran las secuencias *Alu*. Además, este mismo tipo de secuencias asociadas a nucleosomas también mostraron claros patrones de periodicidad en genomas de Arqueas. Finalmente, los resultados de este trabajo dejan suponer que las secuencias repetidas *Alu*, pueden ser consideradas como los elementos básicos

de la estructura más elemental que permite la formación de nucleosomas en eucariontes superiores; además de explicar, con esta relación, su notable éxito de retrotransposición para colonizar los genomas de primates.

# Abstract

The massive sequencing of genomes has triggered important advances in the study of the origin and evolution of organisms. The possibility of handling whole genomes and the identification of coding sequences, reveal a number of complex interactions in the genome involved in dynamic processes that give rise to phenotypic variability of organisms. In these interactions, the non coding regions of the genome should play an important role. The aim of this work is to explore the structure of genomes, using mathematical tools of periodicity analysis in order to find and describe patterns in the genomic architecture that could explain the biological and functional relevance of these structural traits. These analysis could be made with the different readings that can be given to the genome describing it by these structural features. The first results of scans applied to some genomes reflect the relationship that may exist between the biological nature of organisms with some structural patterns detected; also provide an expectation of the reaches that could be achieved with these approaches. Subsequently, modular elements of specific sequences associated with nucleosome positioning were recognized with periodic presence among primates; in fact, a similar structural periodic patterns could be described. It has also been found that these sequences show some unreported periodicities in genomes and others are associated with highly repeated sequences, some of them belonging to *Alu* sequences. Furthermore, these kind of sequences, associated with nucleosomes, was observed in Archaeas showing periodicity patterns. Finally, the results allow to propose that *Alu* repetitive elements may contribute to the most fundamental structure of nucleosome positioning in higher eukaryotes and this relation may explain its remarkable success of retrotransposition to colonize the genomes of primates.

# Introducción

El genoma de los organismos contiene una amplia diversidad de elementos que mantienen y promueven múltiples interacciones en él, con sus productos y con el medio ambiente y cuyos efectos no se pueden acotar a una sencilla relación de gen, proteína y función para explicar el funcionamiento entero o el desarrollo de un organismo. El genoma es una composición heterogénea de nucleótidos, que además de ser un reservorio de información traducible a proteínas está dispuesta y organizada con un orden no aleatorio que se puede compartimentar para su debido funcionamiento.
La diversidad y conservación de las interacciones complejas que existen a distintos niveles entre todos los elementos genéticos, sus productos y el medio ambiente se manifiestan tanto en los distintos destinos fenotípicos a nivel celular, que se pueden observar en organismos eucariontes, como durante el desarrollo de los organismos y los cambios que ocurren en éste a lo largo de su vida. Cualquier variante de estos eventos puede tener efectos en el organismo entero, relevando la importancia de estas interacciones en sus implicaciones evolutivas y en la variabilidad fenotípica de los organismos [66].

Desde la perspectiva que ha permitido tener este trabajo, el genoma podría estudiarse como un conjunto de módulos funcionales, que contendrían tanto elementos traducibles a RNA como secuencias repetidas o no codificantes. Estos módulos, por tanto, forman parte de procesos específicos pero tampoco están completamente aislados o independientes entre sí. Por ejemplo, un conjunto de genes puede operar independientemente de otro conjunto

de genes en un momento durante el desarrollo de un organismo, aunque posiblemente en otro momento estén ligados o interactúen en sincronía con genes de cada conjunto [26].

Sin embargo, esta propuesta de funcionalidad modular del genoma es posible en un contexto más completo de procesos y factores que faciliten su dinamismo y operatividad. En este sentido, como se propone en este trabajo, la cualidad estructural debería facilitar o permitir las interacciones de dichos módulos funcionales y del medio ambiente. Así, bajo distintas lecturas, la composición del genoma, con regiones codificantes y no codificantes, puede ser estudiado también por las características físicas y químicas que le confieren al genoma arquitecturas con la posibilidad de organizarse, proteger, restringir y facilitar interacciones entre unidades codificantes e incluso formar cromosomas.

En esta exploración estructural, un genoma dado se ha podido interpretar como lecturas de un mismo texto con distintos valores semánticos, con lo que se pudieron encontrar diversos y repetidos arreglos que, ahora bajo esta perspectiva, ya no dependen de su traducción a proteínas, sino de los valores químicos y físicos que distinguen y agrupan propiedades estructurales en los nucleótidos y que pueden leerse en un lenguaje binario al asignarles, por ejemplo, dos distintos valores por sus cualidades para formar enlaces de puentes de hidrógeno dobles o triples con la cadena complementaria; o escribir la secuencia asignándole dos valores distintos dependiendo del tamaño o naturaleza química del nucleótido, es decir, si son purinas o pirimidinas. Pero también pueden usarse más valores para caracterizar la secuencia genómica, como los datos de energía libre entre dímeros [4] [35].

Las herramientas matemáticas para el análisis de la periodicidad, como la transformada de Fourier [15] o la función de información mutua [49], se pudieron aplicar al genoma una vez que éste se ha convertido a valores numéricos, como los valores de entalpía entre nucleótidos, valores de energía libre o valores binarios que reflejen las descripciones anteriores.

En diversos estudios para el seguimiento de estas señales estructurales, las herramientas de análisis periódicos evidencian la presencia de osilciones de periodos que revelan los diversos elementos genómicos que comparten cualidades en común; que en muchos casos son muy claramente identificables, como en el caso de las señales de periodo *tres* en regiones codificantes [12] [55]. Pero también, se han encontrado osilciones de otros periodos cuyo sentido biológico es todavía difícl de dilucidar, sobretodo en organismos eucariontes cuyo contenido genómico no es completamente traducido a proteínas, un ejemplo son las oscilaciones periódicas de 500,000 pb encontradas en el cromosoma 21 del humano [30] .

Existen una gran cantidad de secuencias repetidas que llegan a formar más de la mitad del genoma en algunos organismos, como lo es en el humano [5] en donde un tercio del genoma es ocupado por las secuencias repetidas interdispersas. Estas secuencias se componen principalmente de copias degeneradas de elementos transponibles [50] o *transposones*. Estos elementos, son pequeñas secuencias en el genoma que tienen la habilidad de transportarse ellas mismas a otros lugares dentro del mismo genoma [26]. Las secuencias repetidas han sido consideradas, por algunos autores, como la principal unión entre los procesos genéticos y epigenéticos [66] y se cree que algunas secuencias repetidas como minisatélites y microsatélites, podrían ser responsables de eventos evolutivos rápidos [43].

El objetivo de este trabajo se resume en la exploración del uso de algunas herramientas para el análisis de periodicidad que permitieran describir la arquitectura de patrones periódicos en diversos genomas para entender la función o relevancia biológica de estas características estructurales.

Después de una exploración general para encontrar señales periódicas, la principal línea de investigación que se siguió en este trabajo fue la búsqueda de secuencias repetidas y en particular asociadas a la formación de nucleosomas, por ser uno de los mecanismos relevantes de regulación epigenética que promueve el remodelaje de la cromatina para regular físicamente la ex-

presión de genes, además de proteger al genoma de la actividad de *DNAsas* y organizarlo para formar cromosomas metafásicos.

Uno de los resultados más interesantes que se obtuvo en este trabajo, analizando diversos genomas, fue el reconocimiento de peculiares distribuciones de pequeñas secuencias repetidas de diez nucleótidos restringidas a un orden determinado por su naturaleza pirimídica. Estas secuencias fueron descritas previamente para *C.elegans* [57] como secuencias que se agregan y facilitan la asociación a nucleosomas, en esta investigación se encontró que la presencia y distribución de estas pequeñas secuencias en tres grupos de primates, revelan distintos perfiles de periodicidad en los cromosomas de una misma especie, lo que permitió distinguir distintos grupos de cromosomas y observar que entre las especies de primates estos grupos tienen un alto grado de conservación.

También se encontró, según algunos experimentos realizados, que estas pequeñas secuencias están probablemente asociadas a otro tipo de secuencias repetidas y cuya presencia en algunos organismos procariontes, en particular arqueas, nos puede dejar suponer que forman parte de una estructura ancestral para la formación a nucleosomas.

Con el principal objetivo de contextualizar la importancia de los resultados y los argumentos en la discusión, se inicia el trabajo con los antecedentes descritos en el siguiente capítulo y subdivididos en tres secciones. Primero se explica el uso de las herramientas periódicas y su aplicación a secuencias genéticas; la clasificación que se tiene hasta ahora de las secuencias repetidas y sus características de importancia biológica, finalmente, se mencionan los mecanismos generales de regulación epigenética hasta ahora descritos. En el capítulo de Metodología se explicarán las dos herramientas matemáticas usadas en este trabajo: el análisis de Fourier y la función de información mutua. En el capítulo de resultados se irán describiendo los experimentos aplicados con estas herramientas para la exploración del genoma con análisis de Fourier y en la segunda parte la incorporación de la función de información mutua aunada a los experimentos con el análisis de Fourier para secuencias asocia-

das a nucleosomas. Finalmente se presenta en la discusión un resumen de los primeros resultados que muestran los alcances y la interpretación que se pueden obtener con estas herramientas en el estudio del genoma.
De los resultados obtenidos en la segunda parte del trabajo, se discute la relación entre los datos que se obtuvieron con el análisis de información mutua de secuencias específicas asociadas a nucleosomas, su presencia en secuencias repetidas altamente conservadas en primates y su contribución a la conservación de una arquitectura genómica vinculada a este tipo de secuencias. Además se propone que, el encuentro de señales periódicas con estas secuencias asociadas a nucleosomas en arqueas, permite suponer que dichas secuencias pueden ser el origen de la estructura más fundamental para la formación de nucleosomas.

# Capítulo 1

# Antecedentes

## 1.1. Secuencias repetidas

En los organismos eucariontes existen una gran cantidad de secuencias repetidas que no son transcritas a proteínas y que llegan a formar en algunos organismos más de la mitad del genoma. En el humano [5] un tercio del genoma es ocupado por las secuencias repetidas interdispersas. Estas secuencias se componen principalmente de copias degeneradas de elementos transponibles [50], llamados también *transposones*. Estos elementos son secuencias relativamente pequeñas y dispersas en el genoma que tienen la capacidad de transportarse ellas mismas a otros lugares dentro del mismo genoma [26].

El origen de su movilidad radica en la capacidad que poseen para hacer copias de sí mismas. De hecho, se pueden replicar junto con el genoma o independientemente. Además, la transposición puede ocurrir con un RNA intermediario a través de una transcripción en reversa, conocida como retrotransposición, o por una escisión y reintegración del propio DNA, llamada simplemente transposición.

En las secuencias transponibles o *transposones* que no usan un RNA intermediario para replicarse, la inserción a un nuevo sitio específico en el DNA ocurre creando rompimientos escalonados en ambas cadenas, dejando en los

extremos de la cadena cortada algunos nucleótidos sin su cadena complementaria, como se muestra en la figura 1.1; luego de la inserción del transposón esas secuencias serán rellenadas y selladas.

Este tipo de transposición puede ocurrir tanto con mecanismos replicativos como con mecanismos no replicativos; en el primer caso, el transposón se replica y el producto puede insertarse en un sitio nuevo; en el segundo caso, simplemente cambia de sitio dejando un rompimiento en la cadena que puede ser letal a menos de que sea reparado [26]. Figura 1.2.

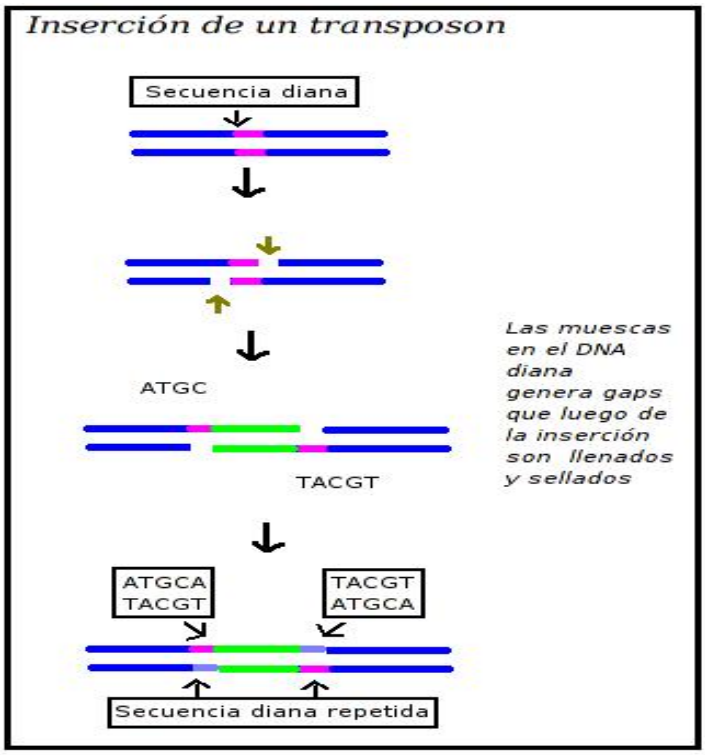


Figura 1.1: La inserción de un *transposón* ocurre tras el rompimiento escalonado en una región específica de la doble cadena de DNA. Los nucleótidos en los extremos de la inserción, que se quedan sin pareja en la cadena complementaria, son compensados al incorporar a sus respectivos nucleótidos complementarios mediante los mecanismos de reparación del DNA.

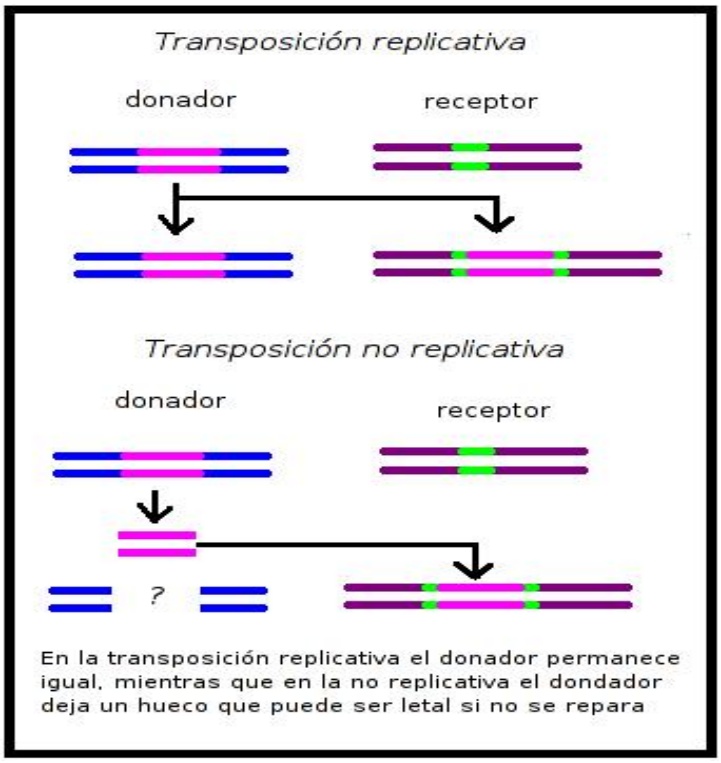


Figura 1.2: Las transposiciones pueden ocurrir al replicarse el *transposón* en su sitio de origen, en donde sólo la copia será incorporada en otra sección del genoma, pero también ocurren al desprenderse de su sitio original en el genoma para incorporarse en otro, lo cual dejará un espacio en el sitio de origen que de no ser reparado puede ser letal.

La gran mayoría de las secuencias repetidas retrotransponibles, si no es que todas, se encuentran clasificadas como elementos tipo retrovirus, que se caracterizan por tener un extremo largo de nucleótidos repetidos y son llamadas secuencias *LTR*, o como elementos nucleares interdispersos, los que pueden ser largos o cortos y por sus siglas en inglés son llamadas secuencias *LINE* y secuencias *SINE*.

En la figura 1.3 se pueden observar las características de estos elementos. Las secuencias *SINE* y *LINE* comparten un sitio de inserción y duplicación

de longitud variable y una terminación *poli-A* o una simple terminación de nucleótidos repetidos; por su parte, los elementos tipo retrovirus tienen las secuencias *LTR* que contienen a las secuencias regulatorias de la transcripción y son reproducidas cada una parcialmente por un proceso complejo de transcripción en reversa, en el cual las secuencias *LTR* encontradas a los extremos son utilizadas como *primers* para el inicio de la transcripción en reversa [26].

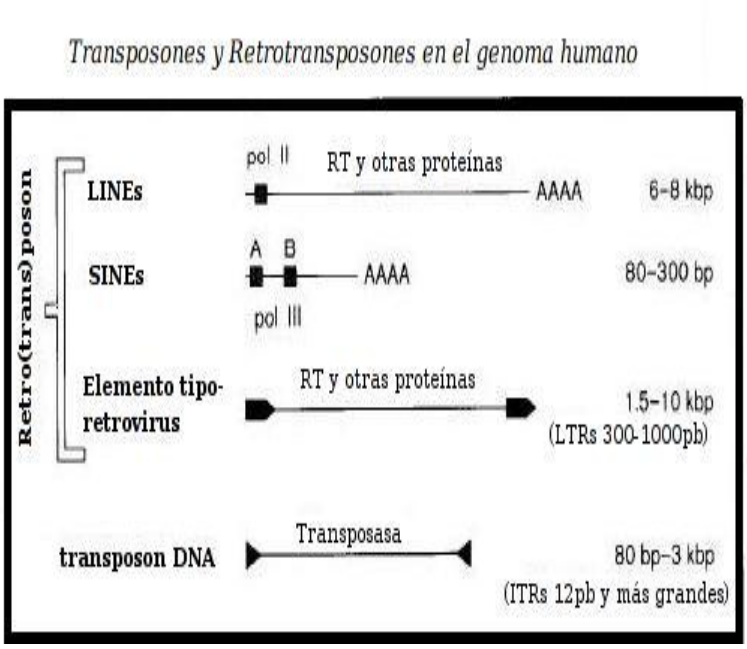


Figura 1.3: Esquema de las estructuras esenciales que permiten reconocer y distinguir a los elementos transponibles y retrotransponibles.

Las secuencias de *retrotransposones*, *transposones* y elementos tipo retrovirus, representan los tres mecanismos generales de amplificación que son responsables de la vasta mayoría de las secuencias interdispersas repetidas en el genoma [50]. Son mecanismos parecidos al que usa un retrovirus pero sin patógenas formaciones individuales. Las secuencias *LINE* y *SINE* forman la fracción más grande de las secuencias repetidas interdispersas en el genoma

humano. En la figura 1.4 se puede ver el porcentaje que ocupa cada una de estas secuencias en el genoma humano [50].

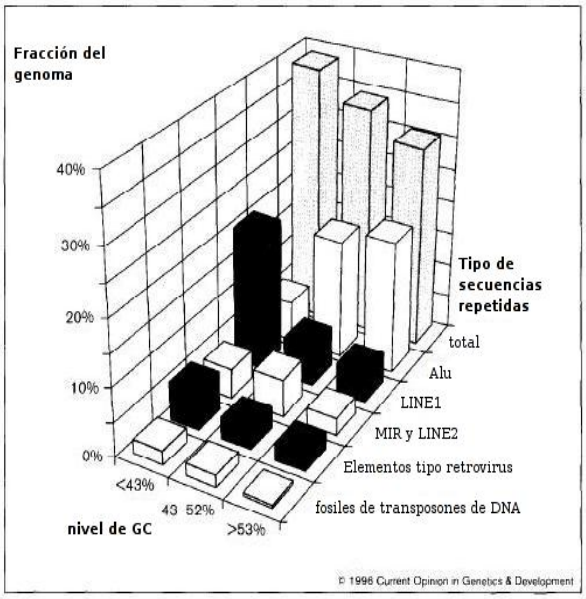


Figura 1.4: Las secuencias repetidas representan un $\sim$40 % de la composición del genoma humano y la mayor parte corresponden a secuencias *Alu* y LINE1 [50].

Las secuencias *SINE*, miden de 100 a 400 pb, y son caracterizadas por tener un promotor interno de la *polimerasa III* [17] que asegura la actividad de transcripción en las nuevas copias. Sin embargo, no tienen secuencias que codifiquen para ninguna proteína, por lo que no son autónomos; las características que comparten con las secuencias *LINE*, como la terminación *poli-A*, sugiere que su movilidad depende de las proteínas provistas por estas secuencias más que por las producidas por los elementos tipo retrovirus [50].

Estas secuencias *SINE* agrupan a las secuencias *Alu*. Se ha calculado que surgieron hace aproximadamente 55 millones de años de una fusión de los 50 y 30 extremos de un gen de RNA 7SL, que codifica para el resto de RNA de la partícula de reconocimiento de señal (*SRP*), en su forma dimérica son

únicos y muy conservadas en los primates [37]. En la figura 1.5 se muestra la divergencia de subfamilias *Alu* en linajes de primates.

Las secuencias *Alu* son las secuencias más representativas de las *SINE*, miden aproximadamente 300pb; están formadas por dos monómeros, uno de los cuales tiene una inserción muy conservada de islas de GpC, a distancias de 31 a 32 bases [2], que podrían funcionar como puertos de regulación epigenética dependientes de metilaciones para nucleosomas [47].

En contraparte, están las secuencias *BI* de roedores. Ambas, *Alu* y *RI*, parecen compartir un origen común con el gen *7SL* RNA (gen que codifica para la partícula de reconocimiento de señal *SRP*, interviene en la unión de la membrana y el ribosoma durante la translocación de proteínas) [50] [41]. El motivo de RNA SRP9/14 de la partícula de reconocimiento de la señal, encontrado también en el RNA 59 del dominio *Alu* [52] [61], es un motivo universalmente conservado en los *SRP* de eucariotas superiores hasta levadura, en archaea y en algunas eubacterias Gram-positivas [52].

El genoma humano, además de contener más de un millón de secuencias *Alu*, lo que corresponde aproximadamente al 10 % del genoma humano, tiene medio millón de secuencias *SINE* llamadas *MIR* (secuencias repetidas interdispersas en mamíferos) ampliamente representadas antes de la radiación de los placentarios; copias de estas secuencias están también en monotremas y marsupiales [50]. Algunas de estas *MIR* tienen extensiones de aproximadamente 3Kb que las revelan también como elementos tipo- *LINE* (*LINE2*). Mucho más relacionados con elementos en genomas de reptiles y anfibios [50].

Las secuencias *LINE* miden de 6 a 8kb, han estado activas en los mamíferos antes de la separación entre linajes de marsupiales y placentarios y tienen copias que forman el 15 % de nuestro genoma. En estos elementos se encuentran secuencias que codifican para reverso-transcriptasas y otras proteínas necesarias para la retrotransposición.

En el humano las secuencias *LINE1* tienen dos marcos de lectura, uno para la unión a proteínas y el otro para la actividad endonucleasa y trans-

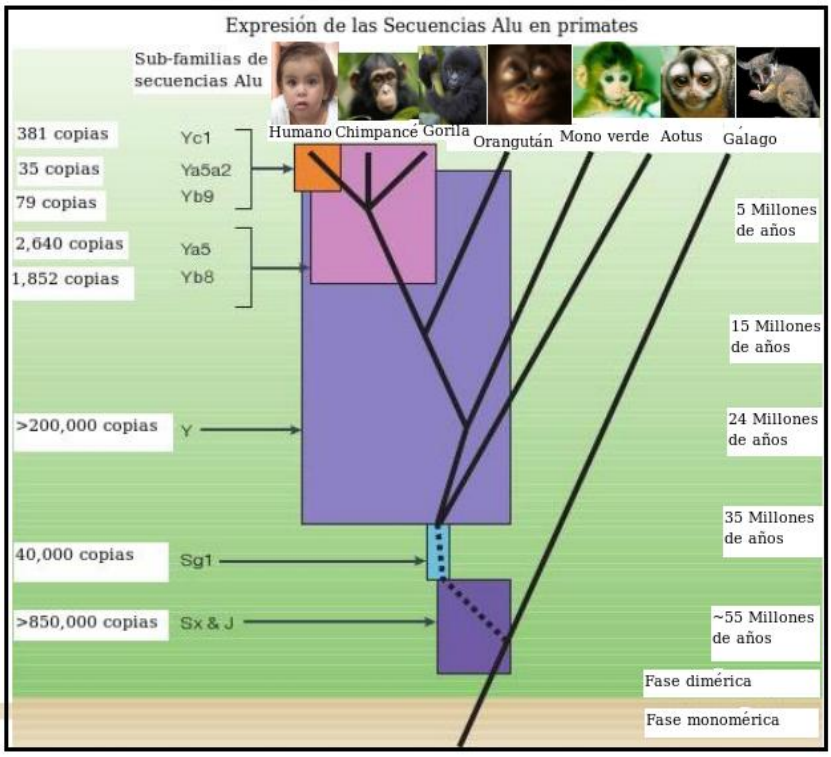


Figura 1.5: La expresión de todas las subfamilias *Alu* con estructuras diméricas, encontradas excusivamente en primates, se reflejan en la divergencia de linajes en primates [2].

criptasa. En algunos elementos tipo-*LINE* se encuentra un promotor interno de la *polimerasa II* que asegura su expresión [50]. Las secuencias *LINE1* pertenecen a un grupo monofilético ampliamente distribuido de *retrotransposones* cuyas copias se caracterizan generalmente por compartir una extremo $5'$ truncado o con deleciones variables [54].

En un modelo descrito para elementos *tipo-LINE* en insectos se sugiere que la reverso transcriptasa reconoce el extremo $3'$ e inicia la transcripción en reversa usando como *primer* o cebador el DNA desnudo [31]. En la figura 1.6 se ejemplifica el mecanismo de retrotransposición con un modelo para secuencias *Alu* (representantes de secuencias *SINE*).

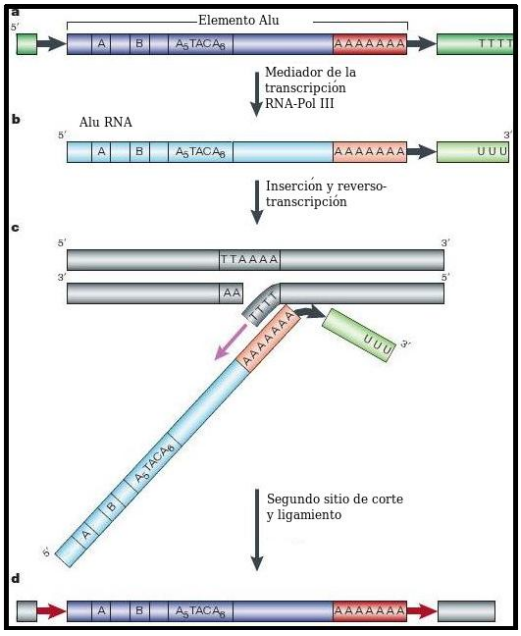


Figura 1.6: A diferencia de los *transposones* de DNA, La transposición para secuencias *LINE* y de secuencias *SINE*, como se reperesenta aquí con una secuencias tipo *Alu*; al igual que los elementos tipos retrovirus, ocurre de manera muy similar al romperse una de las hebras en la cadena de DNA y ante la presencia del RNA específico de la secuencia repetida que funciona como templado para que, mediante una transcripción en reversa, se forme un nuevo segmento de DNA que contenga la secuencia del retrotransposón [31][2].

El mecanismo de transcripción de *LINE* y *SINE* es muy parecido, usan la misma maquinaria pero además de las diferencias en la producción de proteínas propias, las secuencias *LINE* se transponen de forma autónoma o como resultado de una actividad *cis*, en donde las proteínas que producen tienen una mayor afinidad por los transcritos de las secuencias *LINE*, lo que hace más eficiente la movilidad de los transcritos *LINE* sobre los transcritos de las secuencias *SINE*.

Por todas estas características descritas, podemos concluir que la importancia de los elementos transponibles radica en la posibilidad que tienen para promover rearreglos en el genoma directa o indirectamente; ya sea por los eventos propios de auto transposición, que, además de permitir el movimiento de una secuencia a una nueva localización, sirven también como sustrato celular para el sistema de recombinación celular, como se muestra en la figura 1.7, funcionando como portadores de regiones homólogas en las que, al haber dos copias de un transposón en diferentes ubicaciones o incluso en diferentes genomas, pueden proveer sitios para la recombinación recíproca. Tales cambios pueden provocar deleciones, inserciones, inversiones o translocaciones. Por ello, pueden ser la mayor fuente de mutaciones en el genoma [26] y, además de contribuir a la arquitectura del genoma, permiten la emergencia de innovaciones en los distintos linajes de eucariontes [13].

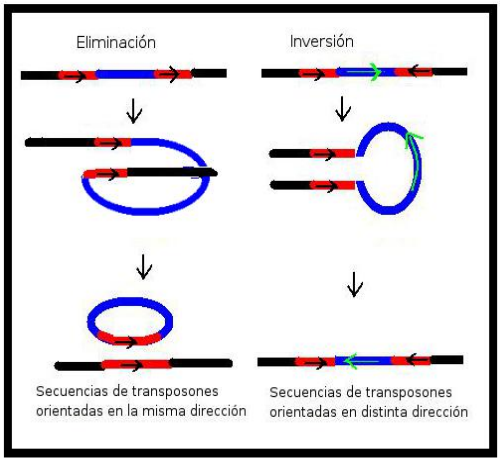


Figura 1.7: Los eventos de inversión y la pérdida de regiones de DNA en el genoma, son algunos de los efectos que pueden ocurrir durante la recombinación tras la inserción de *transposones*

En organismos procariontes también se han descrito numerosas secuencias repetidas dentro de las cuales hay una gran variedad de *transposones*; son consideradas también como representantes de importantes mecanismos evolutivos que permiten a la bacteria adaptarse rápidamente a cambios ambientales [59]. Estos mecanismos involucran la creciente habilidad para transferir genes o realizar variaciones antigénicas y dispersarlas en linajes bacterianos [44] [36]. Los *transposones* más sencillos en procariontes son las *secuencias de inserción*, características también de plásmidos [26]. Las islas de patogenicidad en bacterias son secuencias repetidas muy similares a las *secuencias de inserción*.

## 1.2. Secuencias repetidas en los mecanismos de regulación epigenética

La regulación epigenética, involucra diversos procesos que se reflejan en las distintas expresiones funcionales y fenotípicas tanto a nivel celular, como en el organismo entero; estos mecanismos no resultan solamente de la expresión de la información contenida en el genoma y, uno de los aspectos más interesantes en la regulación de estos procesos, es la facultad de heredar estas cualidades a las células hijas, una vez adquirido un patrón de regulación por estos mecanismos.

La regulación epigenética, por tanto, tiene importantes implicaciones en los procesos ontogenéticos y evolutivos de los organismos. Durante la especialización celular, por ejemplo, se presentan estos mecanismos de regulación; también le dan a los organismos la capacidad de responder a cambios en el medio ambiente y les confiere plasiticidad fenotípica, desde efectos de variegación en plantas [66] hasta cambios morfológicos en animales [8] [64]. Son eventos que finalmente podrían promover divergencia en las especies [62].

Estos procesos de regulación, se pueden entender como sistemas de información epigenética [22]. Uno de estos sistemas ocurre con la retroalimentación al momento de activar un gen cuyo producto influirá en el mantenimiento o silenciamiento de su actividad. Independientemente de cuál sea el primer inductor, el producto podrá permanecer en cantidades suficientes para que el patrón adquirido se exprese en las nuevas células hijas, sin necesidad de que exista el mismo inductor que propició este patrón en la célula madre. Esto significa que en las mismas condiciones no inducidas puede haber dos tipos celulares genéticamente idénticos pero fenotípicamente distintos. El número y el modo de interactuar de este tipo de eventos, en una sola célula, proporciona un potencial evolutivo en donde la selección puede permitir adaptaciones alternativas.

Otro sistema de información epigenética, trascendental en este trabajo,

está relacionado con la estructura de la cromatina, aquí, la conformación del DNA juega un papel trascendental para organizar y regular físicamente la expresión de genes. Las características físico químicas de los nucleótidos le pueden conferir al DNA la capacidad para enrollarse en histonas y formar complejos nucleosómicos, al igual que le permiten la unión a proteínas de transcripción; en este sentido, el tipo de distribución de los nucleótidos o el tipo de secuencias que conforman el genoma, se vuelven un importante objeto de estudio; un ejemplo de ello son las secuencias repetidas en el genoma que, por su capacidad de propagarse en él, pueden proveer también plasticidad fenotípica a los organismos al influir en la organización del genoma e interferir en las interacciones de secuencias codificantes [36], promoviendo marcajes, efectos de variegación y diferenciación celular [66]. También pueden ser causa de daños en regiones codificantes, volviendo inoperantes a genes con inserciones de secuencias repetidas o teniendo efectos negativos en la regulación de la expresión génica normal; como ocurre con la inserción de secuencias *Alu* en genes específicos que manifiestan patologías como cánceres, hemofilia o neurofibromatosis en el humano [2] [32]. Incluso, se ha sugerido que la proliferación de estos elementos en el genoma podría ser la causa del origen de los intrones [18]. Curiosamente estas secuencias también están vinculadas a funciones regulatorias metil dependientes en las que se podrían asociar nucleosomas [47]. De hecho, la unión de secuencias genómicas a nucleosomas está relacionada con la posibilidad que tienen para metilarse, puesto que la metilación facilita la asociación del DNA con las histonas, la reversibilidad de la metilación tiene efectos de remodelaje en la cromatina, lo que permite la expresión genética diferenciada y la duplicación del genoma completo en distintos momentos del ciclo celular.

En el remodelaje de la cromatina participan otros grupos químicos que se asocian a los nucleótidos y a los tallos de las histonas [11], pero los patrones de metilación en el DNA son un principio para la formación de nucleosomas e incluso son suficientes para silenciar completamente regiones del genoma y

tener efectos fenotípicos permanentes que sean distintos entre células genéticamente idénticas. La inducción de regiones metiladas es desconocida pero una vez adquirida esta marca, los patrones de metilación son heredables a las células hijas. Esto es facilitado por la replicación semiconservativa del DNA; en la cual, la hebra que contiene los metilos permanece metilada durante la duplicación y la hebra que no se queda con estos marcajes se comienza a metilar en los mismos sitios al tiempo que se va formando su nueva cadena complementaria.

La actividad de los RNAs pequeños en el genoma, es considerada también como otro sistema de información epigenética [22]. Los RNAs pequeños se conocen como microRNAs (*miRNA*) y RNAs de interferencia(*iRNA*s); su actividad es la misma pero su origen es distinto. Los *iRNAs* se forman de RNAs de doble cadena perfectamente pareados que son procesados por *Dicers* para generar dos cadenas de *sentido* y *antisentido*; al asociarse a complejos protéicos (*RISC*) pueden regular con actividad *cis* y *trans* uniéndose a la secuencia del mRNA complementaria y marcándola para ser degradada por otras moléculas y así evitar su transcripción.

Esta actividad funciona como mecanismo para inhibir secuencias virales y del mismo modo puede controlar la actividad de retrotransposición de secuencias repetidas o *transposones* en el genoma. La formación de *iRNAs* de los RNAs virales puede impedir la expresión de estas secuencias en un segundo virus, aunque muchos virus codifican para proteínas que suprimen a los *iRNAs*.

Los *miRNAs* son transcritos por la *polimerasa II* al igual que un mRNA pero, a diferencia de éstos, son procesados en una sencilla cadena de 19 a 22 nucleótidos que tiene la habilidad de plegarse y formar una imperfecta doble cadena de RNA que es procesada por la endonucleasa *RNAasa III* de la familia de los *Dicers* y *Drosha*. Una de sus cadenas se eliminará y la que se mantiene es cargada en el mismo complejo de silenciamiento de RNA (*RISC*) que usan los *iRNAs*, lo que le permitirá pegarse al mRNA que silenciará. Sólo

tienen actividad *trans*

En resumen, los *iRNAs* actúan en los procesos para inhibir la expresión de secuencias virales; en procesos para eliminar transcritos de secuencias móviles y secuencias repetidas; en procesos para bloquear la síntesis de proteínas a través de los microRNA generados en la misma célula y la supresión de la transcripción mediada por *iRNAs*. También se utiliza esta misma maquinaria cuando se introduce un *iRNAs* experimentalmente para inducir el silenciamiento de algún gen [14].

En animales, los *miRNAs* controlan directamente la expresión de un tercio de genes y tienen en algún momento influencia sobre la expresión casi completa del genoma. Diversos estudios en *Caenorhabditis elegans*, *Drosophila melanogaster* y *Danio rerio* han identificado distintas funciones de *miRNAs* en la coordinación de la proliferación y muerte celular durante el desarrollo embironario, en la resistencia al estrés, en el metabolismo de grasas y en la morfogénesis del cerebro [25]. Muestran también que la actividad de *miRNAs* genera una red que confiere robustez a un fenotipo y los hace funcionar como canalizadores o amortiguadores de variaciones por mutación; este efecto oculta estas variaciones y limita la variedad fenotípica, por lo que la canalización no sólo contribuye potencialmente a la robustés sino también a las innovaciones evolutivas pues, al perderse la canalización de un fenotipo, puede dejar de encubrir estas mutaciones [19]. En plantas es mucho menor el número de genes que son regulados por pequeños RNAs y sin embargo el espectro de acción es muy amplio; están directamente involucrados en aspectos del desarrollo, de respuesta al estrés, y de regulación de su propia formación [33]. Por lo pronto, la importancia y el interés del reconocimiento de secuencias de *miRNAs* también se ha convertido en un tema relevante en los análisis bioinformáticos [25].

Los patrones de regulación epigenética, sin duda, han contribuido a la evolución de los seres vivos; son incluso un prerrequisito en la evolución de organismos complejos. Sin embargo, ni siquiera a nivel celular, se tienen to-

dos los elementos para poder explicar el complejo de inducciones específicas que deben intervenir en la formación de los patrones fenotípicos que se mantienen y se heredan a las células hijas en los diversos linajes celulares que se forman durante el desarrollo embrionario de un organismo. Además, se deben considerar las posibles interacciones entre los sistemas de información epigenética en la formación de estos patrones. Hasta ahora, sólo se han descrito eventos aislados de algunas relaciones entre mecanismos tipo *miRNAs* y el silenciamiento de genes; algunos casos de metilación en genes suprimidos y las respectivas modificaciones en histonas nucleosómicas [3] [34] [60].

## 1.3. Periodicidad en los genomas

La periodicidad se refiere a la presencia regular de un evento y las secuencias repetidas guardan una distribución que no necesariamente es regular a lo largo del genoma. Por esto, se considera que el DNA es una molécula no estacionaria que exhibe diferentes periodicidades en distintas regiones; así, una secuencia repetida podría mostrar un patrón particular de periodicidad solamente en algunas regiones específicas del genoma.

En un análisis de Fourier se pueden observar e inferir, de una manera muy general, las características o longitudes específicas de las secuencias que se encuentran en un genoma que, por tener una presencia recurrente en el genoma, intensifican o aumentan la amplitud de sus frecuencias. Estas observaciones generales, como se sugiere en los resultados de esta investigación, pueden ser indicativas en posteriores análisis para determinar una longitud de ventana que recorra el genoma generando los espectros de frecuencias en función de la secuencia o periódo que se tenga por objeto evaluar; con ello, se puede determinar qué tan representada puede estar la secuencia en una región específica del genoma, o simplemente evaluar su presencia.

Los análisis de espectros de potencia, o análisis de frecuencias, son una de las herramientas ampliamente utilizadas para detectar señales periódicas

en el genoma; con distintas variantes y rastreando distintos patrones o cualidades de la molécula, se ha podido tener una visión general de diversos rasgos que caracterizan a los genomas. Por ejemplo, para un análisis de fluctuaciones de $GC$ a lo largo de un genoma [28] [29] [7] se puede utilizar la transformada de Fourier, al igual que para detectar regiones codificantes en el genoma, puesto que son regiones en donde el espectro de frecuencias debe tener un pico muy discernible en la frecuencia que corresponda al periodo *tres* [55] [58] [65] [43]. La diferencia en estos trabajos tiene que ver con las herramientas computacionales, los criterios para aplicar estos algoritmos y los genomas o secuencias génicas que han permitido tener, tanto un reconocimiento muy general de regiones codificantes, como la identificación un poco más precisa entre los exones e introns de regiones codificantes [9]. Una modificación importante pero más compleja de la transformada de Fourier, llamada transformación de periodicidad cuaterniónica, parece ser muy útil para identificar secuencias repetidas [1].

Otros métodos alternativos para los análisis de frecuencias son, entre otros, el análisis de ondeletas, que son transformaciones que conllevan una traslación y un escalamiento, útiles para señales no estacionarias; las funciones de correlación [47] [46], métodos probabilísticos basados en la teoría de la información, en donde se calcula la entropía del sistema como una medida de información [49]. En base a ello se han hecho análisis de funciones de distancia en algunas partes del genoma para caracterizar las propiedades estadísticas de sus secuencias componentes [39]. La función de información mutua, también basada en las medidas de información o entropía, ha sido utilizada para detectar regiones codificantes [23] y distinguir regiones homogéneas y heterogéneas en el DNA y sus distintos componentes de frecuencia [27]. Incluso, se han combinado distintos análisis de este tipo para detectar con mejor precisión secuencias codificantes [7]; las redes neuronales se han usado también para detectar regiones codificantes, aunque éste no es un análisis basado en la periodicidad. [10].

En las exploraciones para identificar signos periódicos en los genomas, la mayor parte de las señales que sobresalen con una mayor amplitud en determinadas frecuencias, están las relacionadas con secuencias repetidas [42]. No resulta tan claro que las diversas secuencias observadas con estos métodos se encuentren muy conservadas o de manera regular a todo lo largo del genoma como para describirlas con un patrón que las caracterice; sin embargo, como en este trabajo se demostró, parecen existir pequeños arreglos con señales estructurales muy conservados aunque su distribución no es regular. Hasta ahora, la descripción, ubicación y comparación de estas secuencias o elementos genómicos dentro de los organismos y entre organismos nos ha dado pauta para especular más sobre su importancia.

En suma, la composición y distribución de elementos genómicos con características estructurales comunes, permite proponer la conceptualización del genoma como un conjunto de módulos estructurales, en donde el objetivo más deseado y planteado en este trabajo con esta perspectiva, es el poder encontrar la relación y los efectos de estos módulos estructurales y funcionales para que permitan entender no sólo la fotografía de algunos eventos particulares, sino la dinámica genómica que compone y promueve los procesos ontogénicos y evolutivos en los seres vivos.

# Capítulo 2

# Metodología General

En esta sección se explican los dos algoritmos matemáticos de análisis periódico, usados para elaborar programas en lenguaje *phyton*, que permitieron manipular secuencias de DNA y poder detectar periodicidades en diversos genomas usando algunas de las características estructurales de esta molécula descritas en la introducción y especificadas en los resultados.

## 2.1. Análisis de Fourier

Las series de tiempo son una colección numérica de observaciones que ocurren en un orden natural. Este orden natural esta dado por intervalos regulares en los que suceden los eventos. Estas observaciones pueden ser comparables con puntos equitativamente distantes a lo largo de una linea, pero siempre asociados a una sola variable: el tiempo [56].

Cuando más de un fenómeno es observado en cada punto del tiempo o cada observación esta asociada con valores de una serie de variables, se da lugar a múltiples series de tiempo cuyos datos son también llamados series espaciales; la forma más común son observaciones asociadas con un punto en el plano. Ambos tipos de datos son generalizados en simples series de tiempo, cada una requiere una apropiada extensión del método para ser analizada.

Así, uno de los métodos de análisis de series de tiempo es la transformada de Fourier o análisis armónicos de series de tiempo; ésta se resume como la descomposición de series en la suma de sus componentes sinusoidales. Cada serie es representada en los coeficientes de la transformada discreta de Fourier. El análisis de Fourier, en diversos ejemplos, provee una precisa y económica descripción de los datos. Una forma local de análisis armónico conocida como *desmodulación compleja*, puede ser usada para describir oscilaciones que no son suficientemente regulares, como es el caso de las manchas solares, en donde los datos contienen una clara sucesión de picos que ocurren, aproximadamente, cada 11 años. Éstos no son lo suficientemente regulares para ser representados por alguna sinusoide. Los análisis armónicos revelan un pequeño pero persistente componente sinusoide en este tipo de datos. Sin embargo, las oscilaciones en estas series es decir, las series transformadas, pueden describirse en términos sinusoidales por un análisis espectral. Este análisis describe la tendencia de las oscilaciones y no las oscilaciones por sí mismas.

El análisis de Fourier se usó para buscar componentes periódicos con relativo éxito en diversos fenómenos. Sin embargo, para estimar espectros de grandes series de datos, se tuvo que desarrollar una versión del modelo usando un algoritmo que redujo significativamente los esfuerzos computacionales, al que se le conoce como *transformada rápida de Fourier* [6].

La esencia del análisis de Fourier es la representación de conjuntos de datos en términos de funciones sinusoidales. La propiedad más esencial de estas funciones sinusoidales, que las hacen generalmente apropiadas para análisis de series de tiempo, es su comportamiento bajo cambios de escalas de tiempo.

Una sinusoide de frecuencia expresada en radianes o periodo $2\pi/\omega$ se puede escribir:

$$f(t) = R\cos(\omega_t + \varphi)$$

En donde $R$ es la amplitud, $\varphi$ es la fase y si se cambia la variable de tiempo $u = (t - a)/b$, se incorporan tanto los cambios de escala como los de

tiempo.

$$g(u) = f(a + bu) = R\cos(\omega_b u + \varphi + \omega_a) = R\cos(\omega'_u + \varphi')$$

En donde $R' = R, \omega' = \omega$ y $\varphi' = \varphi_a$

Así, la amplitud no cambia, la frecuencia es multiplicada por $b$ y la fase es alterada por la cantidad utilizada en el cambio de tiempo original y la frecuencia de la sinusoide. Dado que el origen del tiempo siempre es arbitrario para la serie de datos, esta simple relación es muy útil.

En particular, ya que la frecuencia no es dependiente ni del origen ni de la escala variable del tiempo, puede considerarse como una cantidad absoluta sin arbitrariedades en su definición. Una característica todavía más útil de una sinusoide es su comportamiento en el muestreo, que es observado como una función de variable continua $t$ en un conjunto de datos equitativamente espaciados; si el intervalo del muestreo es $\triangle$, la sinusoide $R\cos(\omega_1 t + \varphi)$ y $R\cos(\omega_2 t + \varphi)$ son indistinguibles si $\omega 1$ y $\omega 2$ es múltiplo de $2\pi/\ \triangle$. Este fenómeno es conocido como *alisamiento*.

El análisis armónico puede ser difícil de interpretar, aún cuando los datos muestran periodicidades definidas en forma de sucesivos y regulares picos y valles. El problema de analizar datos es el cambio de escala dado que las sinusoidales pueden no coincidir con oscilaciones que crecen en amplitud. La transformada rápida de Fourier es usada de diferentes maneras en programas computacionales. La manera en la que se implementa para cualquier propósito depende del tipo de datos, por ejemplo, si siempre se usara la misma longitud en la serie, entre otras cosas. En general, con pocas consideraciones, la eficiencia del algoritmo de la transformada rápida de Fourier permite hacer un análisis de datos extensivos y analizar diferentes tipos de periodicidades como conicidades (ventanas de hanning).

Por otra parte, la primera gran dicotomía se encuentra en los tipos de datos que pueden ser continuos o discontinuos, pero además, estos datos pueden

ser o no periódicos. Los datos generalmente están en el dominio del tiempo y, mediante la descomposición de este análisis, se obtendrá la misma información, pero en el dominio de frecuencias; de la misma forma, si se tiene el dominio de frecuencias, entonces, por síntesis o la inversa de la Transformada de Fourier, se obtendrá el dominio del tiempo; por lo tanto, es posible pasar de un dominio a otro. Figura 2.1


Figura 2.1: Representación de series periódicas en el dominio de tiempo y su relación inversa en el dominio de frecuencias

En datos aperiódicos y continuos, como podría ser una distribución gaussiana o un decaimiento exponencial, no hay un patrón que se repite; la transformación de Fourier para estos datos se llama simplemente *análisis de Fourier.*

En datos periódicos y continuos, como son ondas sinusoidales u ondas cuadradas, la versión de la transformación de Fourier se llama *series de Fourier.*

Para datos aperiódicos y discretos, o puntos discretos entre el positivo y negativo infinito que no se repiten en forma periódica, este tipo de transformación de Fourier es llamada *transformación discreta del tiempo de Fourier.*

Finalmente, las señales discretas y periódicas son puntos discretos que se repiten en forma periódica entre el infinito positivo y negativo. Para este tipo de datos, la transformación de Fourier algunas veces es llamada *series discretas de Fourier*, pero es más comúnmente llamada *Transformación discreta*

*de Fourier* [51]. Esta última descripción es la más apropiada para estudiar los datos manejados en este trabajo. Adelante se muestran los dos algoritmos usados para la transformada rápida de Fourier con su componente real e imaginario.

$$ReX[k] = \sum_{i=0}^{N-1} X[i]\cos(2\pi kiN)$$

$$ImX[k] = -\sum_{i=0}^{N-1} X[i]\cos(2\pi kiN)$$

La transformada rápida de Fourier se aplicó usando programas desarrollados en lenguaje *phyton* para la búsqueda de patrones periódicos en diversos genomas, utilizando algunas de las cualidades estructurales descritas en la introducción. El análisis se realizó usando ventanas de distintas longitudes que recorrieran los genomas, sin traslape, mostrando todos los espectros de frecuencias por cada ventana analizada, lo cual, permitió reconocer y comparar las distintas señales periódicas encontradas y sus distintas magnitudes en cada genoma o en cada ventana del genoma.

De todos los datos generados, en los resultados, solamente se muestran algunos ejemplos. Para realizar un análisis más específico, se modificó el algoritmo general de la transformada de Fourier con el fin de evaluar exclusivamente la intensidad del periodo *tres* en distintos genomas, el cuál está relacionado con el uso de codones [55] [58] [65] [43]. En el siguiente recuadro se muestra el programa que especifíca el análisis del periodo *tres* en la transformada rápida de Fourier.

```
def dft_n(X,t):
    if t==1:
        return 0
    acc=0;
    acc2=0
    N=len(X) - (len(X) %t);
    T=float(N)/float(t);
    a = 2.0*math.pi*T/float(N)
    RANGO = range(0,N)
    for n in RANGO:
        x=float(n)*a
        acc=acc+X[n]*(math.cos(x))
        acc2= acc2 +X[n]*(math.sin(x))
        acc3= math.sqrt((acc2**2)+(acc**2))

    return acc3;
```

## 2.2. Información Mutua

La información mutua es una medida de la correlación entre variables discretas. Para secuencias simbólicas, la información mutua entre dos símbolos separados por una distancia cualquiera ($k$) se define como la función de información mutua (FIM) [49] y es particularmente útil para analizar las propiedades de correlación en secuencias simbólicas [28]. Esta otra herramienta nos permitió medir la recurrencia de cadenas de bases de diferentes tamaños ($k$) a lo largo de la secuencia, detectando con ello distintas periodicidades.

En un alfabeto definido como: $A = \{a, c, g, t\}$ y en una cadena infinita definida como: $s = (...\alpha_0, \alpha_1, ...)$ en donde $\alpha_i \in A$, $i \in Z$ y los valores se pueden repetir. La FIM entre la cadena $s$ con una cadena idéntica que se desliza $k$ posiciones adelante definida como:

$$I(k, s) = \sum \sum P_{\alpha,\beta}(k, s) \log_2 \frac{P_{\alpha,\beta}(k, s)}{P_\alpha(s) P_\beta(s)}$$

en donde $P_{\alpha,\beta}(k, s)$ es la suma de las probabilidades de tener un símbolo $\alpha$ seguido de un símbolo $\beta$ en la cadena $s$, $k$ posiciones adelante y $P_\alpha(s)$ y $P_\beta(s)$ son las probabilidades de encontrar $\alpha$ o $\beta$ en la cadena $s$. Usando logaritmos en base dos se mide $I(k, s)$ en bits. El número $I(k, s)$ se interpreta como el promedio, sobre todas las posiciones, de la información que puede darse sobre el actual valor de una cierta posición en la cadena y dado el actual valor de esa posición, el de $k$ posiciones adelante.

La información mutua $I(k, s)$ sera casi nula, si y sólo sí, todos los eventos son estadísticamente independientes entre sí, por ejemplo, si la suma de las cuatro probabilidades son factorizables. Entonces la FIM es capaz de detectar cualquier desviación de la independencia estadística. En otras palabras, la función $I(k, s)$ es la medida de información contenida de un evento $\alpha$ respecto a otro evento $\beta$, el cual es localizado $k$ lugares adelante de $\alpha$ (o vice versa) en una serie simbólica. Es importante notar en la ecuación mostrada

arriba que $I(-k, s)$ y que $I(k, s) \geq 0$.

Para este trabajo, se realizaron los cálculos informáticos de la función de información mutua sobre las secuencias de DNA usando medidas consecutivas de desplazamientos ($k$), de 1 a 500, con las que se pudieron obtener distintos arreglos en los perfiles de autocorrelaciones para cada secuencia utilizada, en donde, al graficar $k$ contra la medida de información mutua $I(k, s)$, los picos que se observan en las posiciones $K$ significan que es relativamente fácil conocer el contenido de la cadena $k$ posiciones adelante de una posición cuyo contenido es ya conocido. En los resultados se muestran estas gráficas y las interpretaciones de los patrones periódicos que se pueden deducir de ellas.

# Capítulo 3

# Resultados

## 3.1. Primera parte: Exploración de los espectros periódicos

Los espectros de potencias hechas con la transformada rápida de Fourier en el DNA son un registro específico y particular de cada genoma que muestra la intensidad con la que se presenta un periodo. El periodo, en este caso, se puede interpretar como la longitud de la secuencia de nucleótidos (número de nucleótidos) con determinadas características que se encuentra de manera recurrente a lo largo del genoma. Estas características dependerán de los valores numéricos que se le asignen a la secuencia de DNA.
En el primer experimento se calculó la transformada rápida de Fourier con el registro de señales de enlaces fuertes y débiles de la molécula. Se le asigno el cero (0) a nucleótidos formadores de enlaces débiles (adenina y timina) y el uno (1) a nucleótidos formadores de enlaces fuertes (citocina y guanina). Se observó todo el espectro de frecuencias para distintas muestras de genomas y con el objetivo de poder distinguir periodos que se pudieran perder junto a otros de mayor amplitud, se realizaron experimentos con ventanas de distintos tamaños de ventanas.

En las figuras suplementarias se muestran gráficas de algunas de las ventanas para cada genoma análizado. Las primeras figuras solo ejemplifican tres diferentes tamaños para observar espectros en distintos organismos. cromosoma 1 de *Homo sapiens* partido en ventanas de 20000 (figura 3.25), comosoma 1 *Arabidopsis thaliana* partido en ventanas de 300000, y en el cromosoma 24 de *Leishmania major* en ventanas de 2000 bases. (Figura 3.24)

Con el mismo tamaño de ventanas y para la misma muestra de genomas se realizó el análisis rápido de Fourier, pero usando la distribución de energía libre dependiente de la secuencia de nucleótidos a lo largo de una cadena del DNA; los valores de energía entre nucleótidos contiguos fueron tomados de los datos experimentales de Breslauer [4]. Con estos datos, el valor númerico de la secuencia tendrá 10 posibles valores. Las señales periódicas registradas en algunos casos son distintas y en otros se registran los mismos periodos. No obstante, la mayoría de los periodos observados, tanto con la distribución de enlaces fuertes y débiles *(WS)* como con la distribución de energía libre, están cercanos o son iguales a 3, 6, 11, 50 a 80 y 200. (Figuras 3.26 y 3.27)

### 3.1.1. Análisis del periodo 3

Con los siguientes experimentos se explica que la interpretación de un periodo en el genoma puede representar el tamaño de una señal o la longitud de una secuencia, por lo tanto, puede haber muchas señales del mismo tamaño que se repitan a lo largo de la muestra sin tener necesariamente la misma composición, como es el caso de los codones.

Para calcular el tamaño de las ventanas y la composición de secuencias que se generaron aleatoriamente en los siguientes experimentos, se consideraron las premisas sobre la organización en tripletes o codones de secuencias que codifican para formar proteínas y que, por tanto, restringen el orden de la secuencia; también se consideró el uso de 64 codones para codificar solamente 20 aminoácidos en donde puede haber hasta 6 tripletes que codifiquen para un mismo aminoácido, aunque se sabe que existe un uso preferencial de co-

dones en los organismos [26], esto quiere decir, que se repetirán los tripletes o codones en cualquier secuencia codificante por pequeña que esta sea. También se tomó en cuenta que en un módulo proteico puede haber entre 40 y 100 aa. incluso para una enzima hay de 62 a 2,500 aa.
Las siguientes gráficas son experimentos con secuencias artificiales, creadas con nucleótidos aleatoriamente distribuidos y con codones aleatoriamente distribuidos, variando los tamaños de ventana. Los valores numéricos que se asignaron a las secuencias corresponden a los datos de energía libre. (Figuras 3.1 y 3.2).

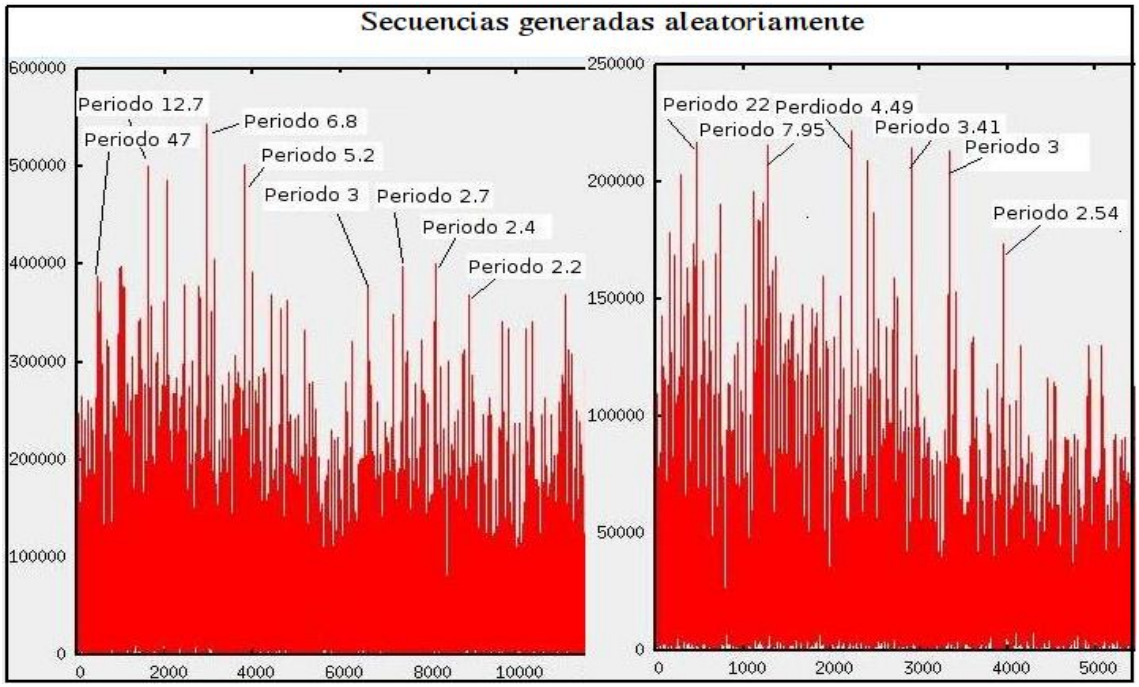


Figura 3.1: Espectro de frecuencias de una secuencia construida con nucleótidos distribuidos aleatoriamente.

Para saber cuántas repeticiones con la misma longitud tiene que haber para que se note la señal periódica, en el espectro de frecuencias y en determinados tamaños de secuencias, se generaron secuencias tanto con nucleótidos

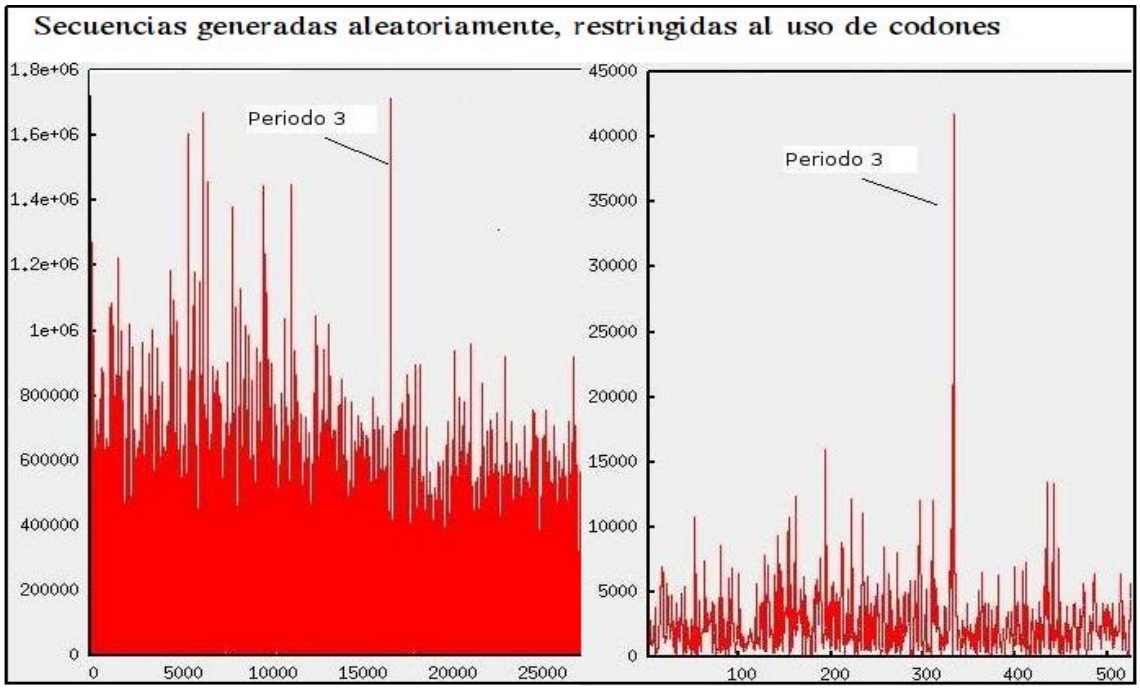


Figura 3.2: Espectro de frecuencias de una secuencia construida con codones distribuidos aleatoriamente usando ventanas de 1000 y 50000 bases.

aleatoriamente distribuidos como con nucleótidos organizados en codones aleatoriamente distribuidos pero con distintas proporciones entre las dos clases de secuencias y variando también el tamaño de la ventana (figuras 3.3 y 3.4). El tamaño más chico de ventana que se usó fue de 1000 bases puesto que es el tamaño aproximado de un gen. En las tres proporciones,1:1,3:1 y 4:1, en las que se pretendió ir reduciendo la proporción de secuencias de codones, siempre se mantuvo la señal de periodo *tres* y esta intensificó su amplitud en las tres proporciones cuando aumentó el tamaño de la ventana. Aunque en la proporción 4:1, se distinguen otras señales de mayor intensidad, la señal para periodo *tres* estuvo muy bien diferenciada.

Para comparar las gráficas de secuencias artificiales restringidas al uso de codones, se analizaron también secuencias codificantes del genoma de *Dro-*

*sophila melanogaster*. (Figura 3.5)

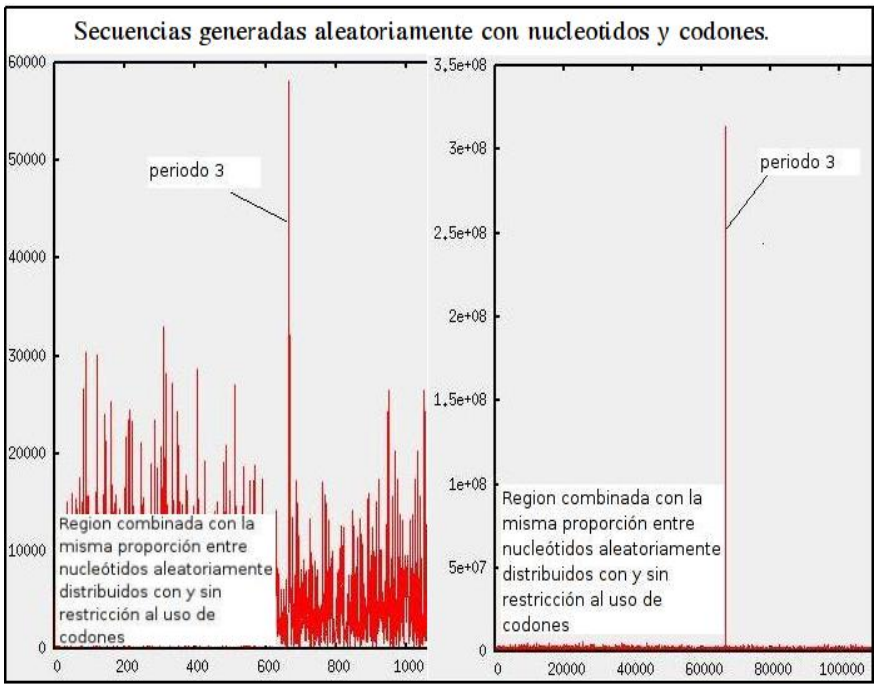


Figura 3.3: Espectro de frecuencias de una secuencia construida con secuencias de nucleótidos y secuencias de codones distribuidos aleatoriamente en la misma proporción, en ventanas de 2000 y 200000 bases.

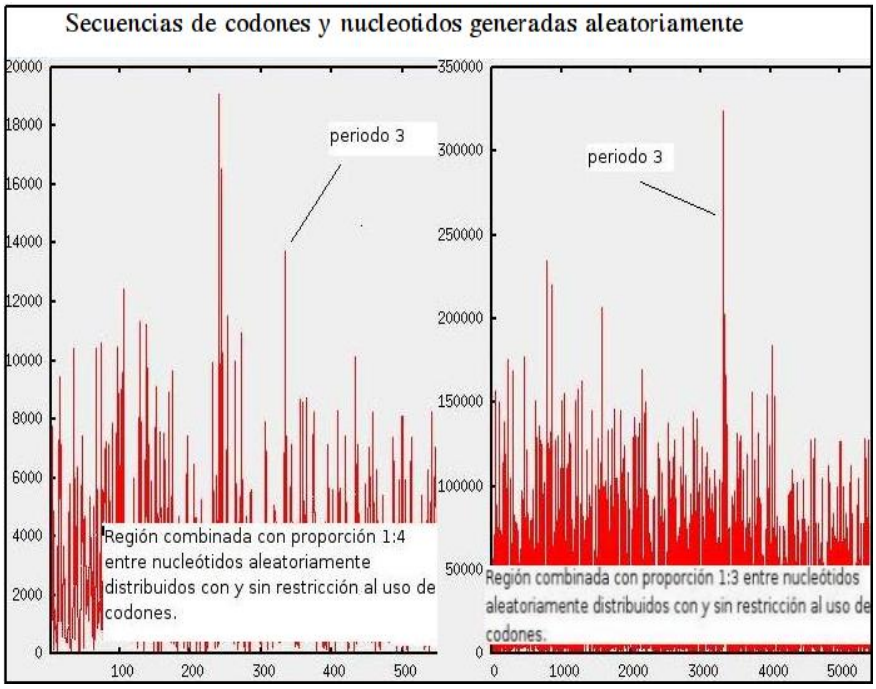


Figura 3.4: Espectro de frecuencias de una secuencia construida tanto con secuencias de nucleótidos, como con secuencias de codones distribuidos aleatoriamente; en proporción 3:1 usando ventanas de 10000 bases y en proporción 4:1 usando ventanas de 1000 bases.

Con el registro de todo el espectro de frecuencias, se puede observar todo el potencial periódico de la muestra y por esto es posible perder fácilmente una señal para un periodo que no tiene mucha presencia. Podría suponerse con estos resultados que, entre más grande sea la muestra, la intensidad de un periodo con un poco más de presencia que el resto, se notará por sobre los demás periodos, como se observó en el caso de la señal para el uso de codones en donde, la suma de secuencias de la misma longitud hicieron más notable la intensidad de ese periodo, o por lo menos estuvo muy bien definido en el espectro de frecuencias, aún, sin que la cantidad de codones presentes en la secuencia estuviera tan representada. Lo que importa, en todo caso,

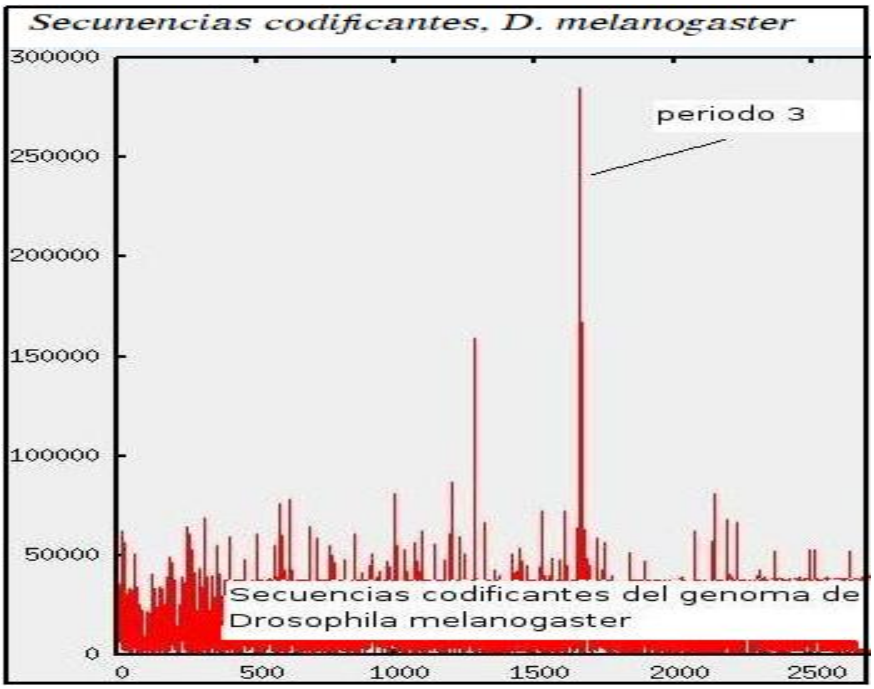


Figura 3.5: Espectro de frecuencias de secuencias codificantes del genoma de *Drosophila melanogaster* usando ventanas de 5000 bases.

es el ruido o la intensidad de otros periodos presentes en la muestra. Esto también puede explicar por qué se encontraron señales de periodo *tres* en el espectro de frecuencias de genomas eucariontes en distintas ventanas escogidas al azar pues bien podrían existir pequeñas regiones codificantes y se notaría la presencia del periodo *tres* aunque con poca intensidad.

Para poder dilucidar mejor sobre el comportamiento o la presencia de un periodo determinado en las muestras, se modificó el análisis de Fourier para registrar solamente el componente de periodo *tres* en ventanas de tamaños modificables en la secuencia. En este registro se pudo evaluar la amplitud de la señal para determinar el registro de un periodo significativamente aceptable. Al calcular la varianza se tuvo una primera evaluación de la diferencia

estándar de las amplitudes para cada ventana con respecto a la media de nuestra muestra. Este valor fue útil para comparar distintas regiones del genoma y estimar con ello las posibles regiones con periodo *tres* significante.

Se sabe que los genomas de organismos procariontes tienen un genoma compuesto mayormente por regiones codificantes y sin intrones por lo que resultaron ser un parámetro apropiado para compararlos con genomas eucariontes. También resultó interesante compararlos con eucariontes de genomas reducidos, típicamente parásitos, que no tienen muchas regiones intergénicas, como el caso de *Encephalitozoon cuniculi*. Enseguida se presentan algunas gráficas para el registro del componente de periodo *tres*. Las primeras dos (figura 3.6 y 3.7) están acompañadas por el espectro entero de frecuencias de la misma región del genoma.

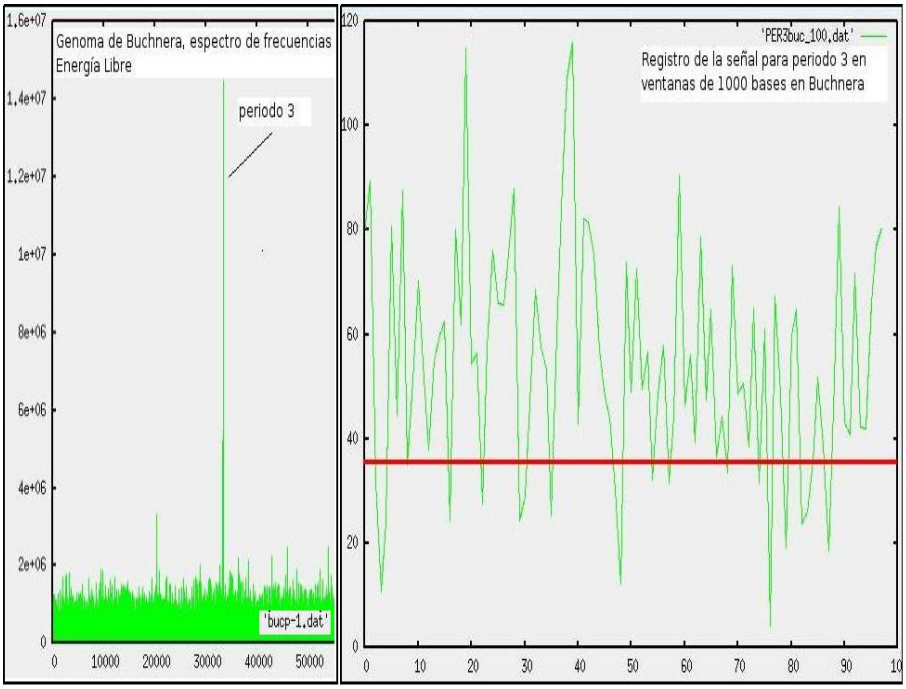


Figura 3.6: Espectro de frecuencias del genoma de *Buchnera aphidícola* y su registro del periodo *tres* en ventanas de 1000 bases.

En general, las secuencias de eucariontes y procariontes tienen un mismo patrón en cuanto al registro del periodo *tres*. Esto indica que por la combinación de regiones codificantes y no codificantes, la media para el registro del periodo *tres* se distingue perfectamente entre ambos tipos celulares, excepto en el genoma *E. cuniculi* que más adelante se describe. Las siguientes gráficas muestran los patrones de periodo *tres* en secuencias aleatorias con uso de codones y sin él. (Figuras 3.8, 3.9 y 3.10)

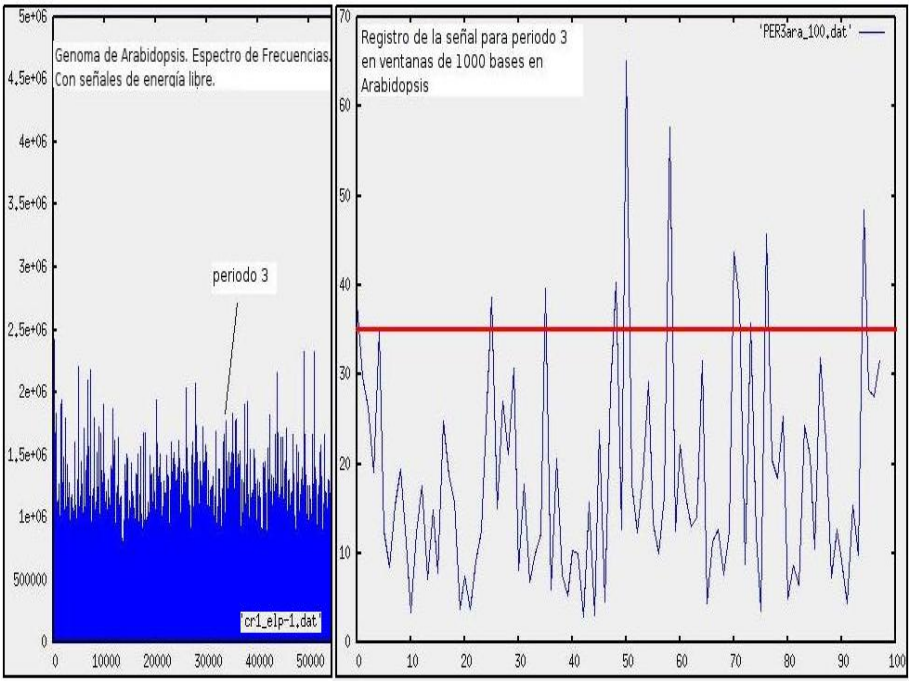


Figura 3.7: Espectro de frecuencias del genoma de *Arabidopsis thaliana* y su registro del periodo *tres* en ventanas de 1000 bases.

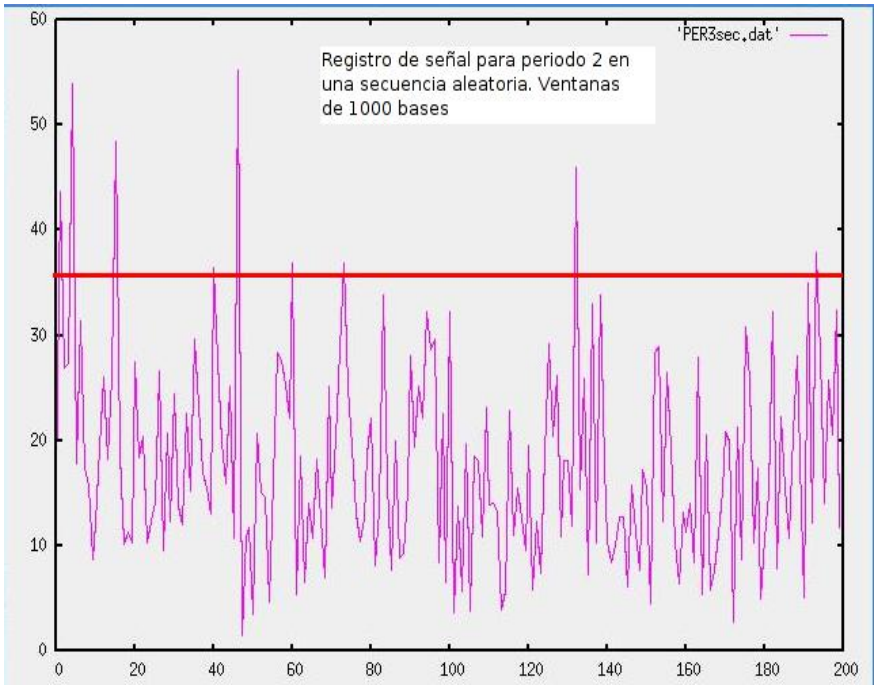


Figura 3.8: Registro del periodo *tres* con la transformada rápida de Fourier para secuencias aleatorias de nucleótidos en ventanas de 1000 bases.

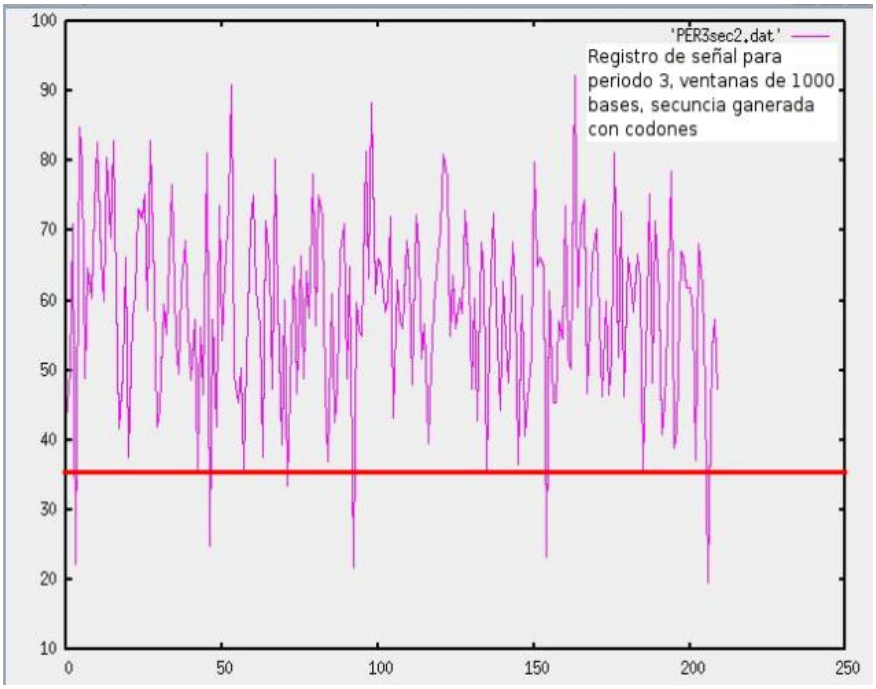


Figura 3.9: Registro del periodo *tres* con la transformada rápida de Fourier para secuencias aleatorias de codones en ventanas de 1000 bases.

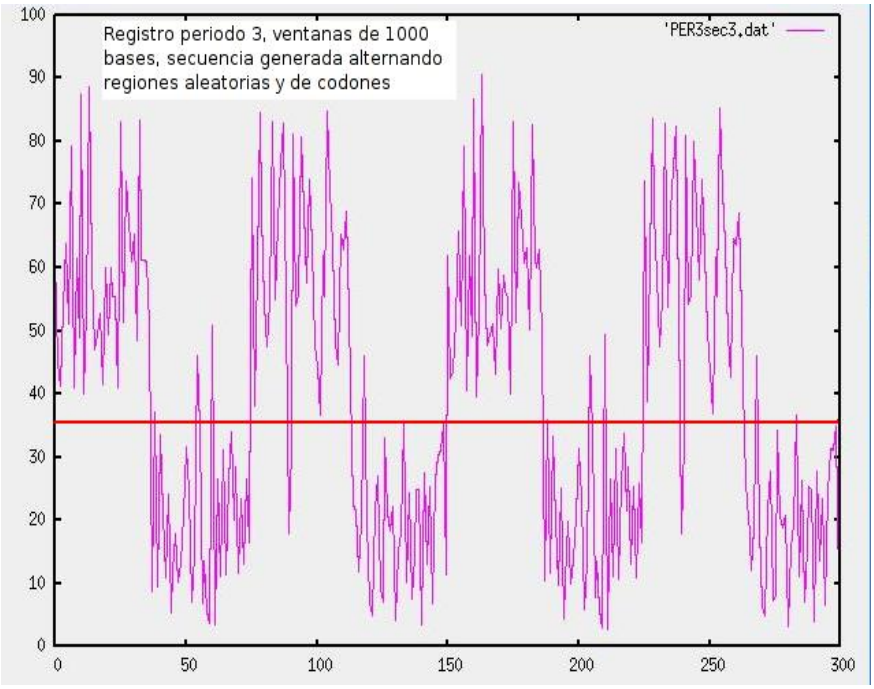


Figura 3.10: Registro del periodo *tres* en intercalado de secuencias aleatorias para nucleótidos y codones en ventanas de 1000 bases.

*E.cuniculi*, es un eucarionte unicelular de microsporidia, la compactación de su genoma ha reducido enormemente, comparado con otros genomas eucariontes, regiones intergénicas o regiones no codificantes. Lo anterior se puede observar en la figura 3.11 que, al parecer, no coincide con los mismos patrones en la amplitud de la frecuencias de periodo *tres* que se observan en el resto de los eucariontes.

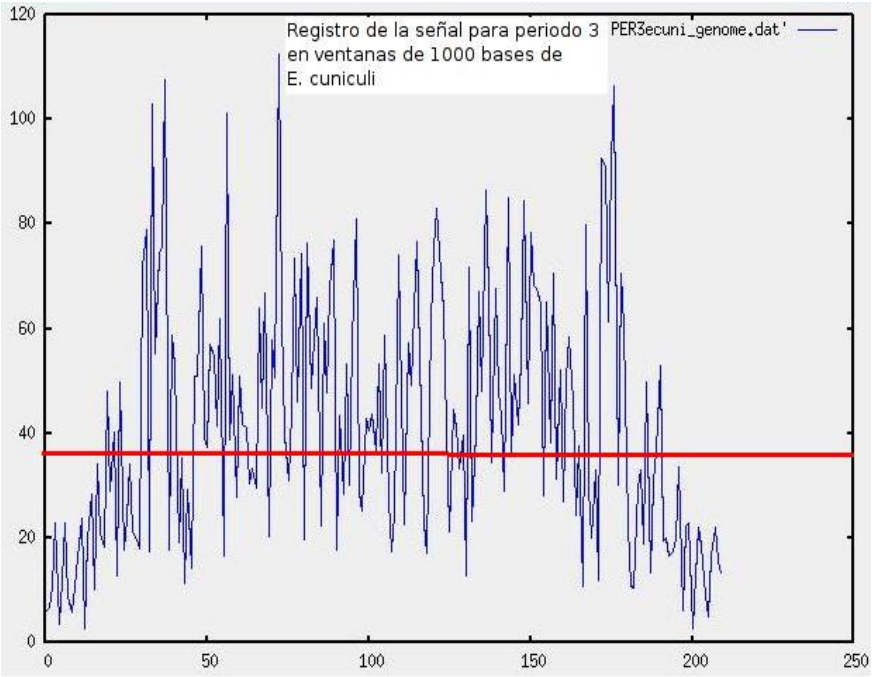


Figura 3.11: Espectro de frecuencias del genoma de *Encephalitozoon cuniculi* en ventanas 1000 bases.

La gráfica anterior muestra el componente de periodo *tres* a lo largo del cromosoma I.

La siguiente Tabla muestra la varianza y el registro de todos los genomas analizados con el mismo método.

| Organismo | Dominio | Parásito | Media | Varianza |
|---|---|---|---|---|
| *H. sapiens* | eucarionte | no | 16.38 | 75.93 |
| *O. sativa* | eucarionte | no | 27.15 | 647.22 |
| *A. thaliana* | eucarionte | no | 18.33 | 156.46 |
| *D. melanogaster* | eucarionte | no | 20.21 | 90.22 |
| *C. elegans* | eucarionte | no | 18.59 | 210.95 |
| *P. falciparum* | eucarionte | si | 29.83 | 600.13 |
| *E. cuniculi* | eucarionte | si | 41.61 | 582.24 |
| *B. aphidicola* | bacteria | no | 54.65 | 498.4 |
| *X. fastidiosa* | bacteria | si | 35.30 | 382.82 |
| *E. coli* | bacteria | si | 41.28 | 541.28 |
| *A. fulgidus* | arquea | no | 55.95 | 696.13 |
| *S. solfataricus* | arquea | no | 35.41 | 262.45 |
| Secuencia1 (aleatoria) | | | 17.98 | 93.33 |
| Secuencia2 (aleatoria) | | | 58.61 | 181.59 |
| Secuencia3 (aleatoria) | | | 39.53 | 555.51 |

Para todos los organismos analizados se graficaron tanto la componente de periodo *tres* en ventanas de 1000 bases, como la relación con el porcentaje de *GC* en cada ventana, resultando que en algunos organismos hay una correlación positiva con el contenido de *GC*. (Figuras 3.12, 3.13 y 3.14)

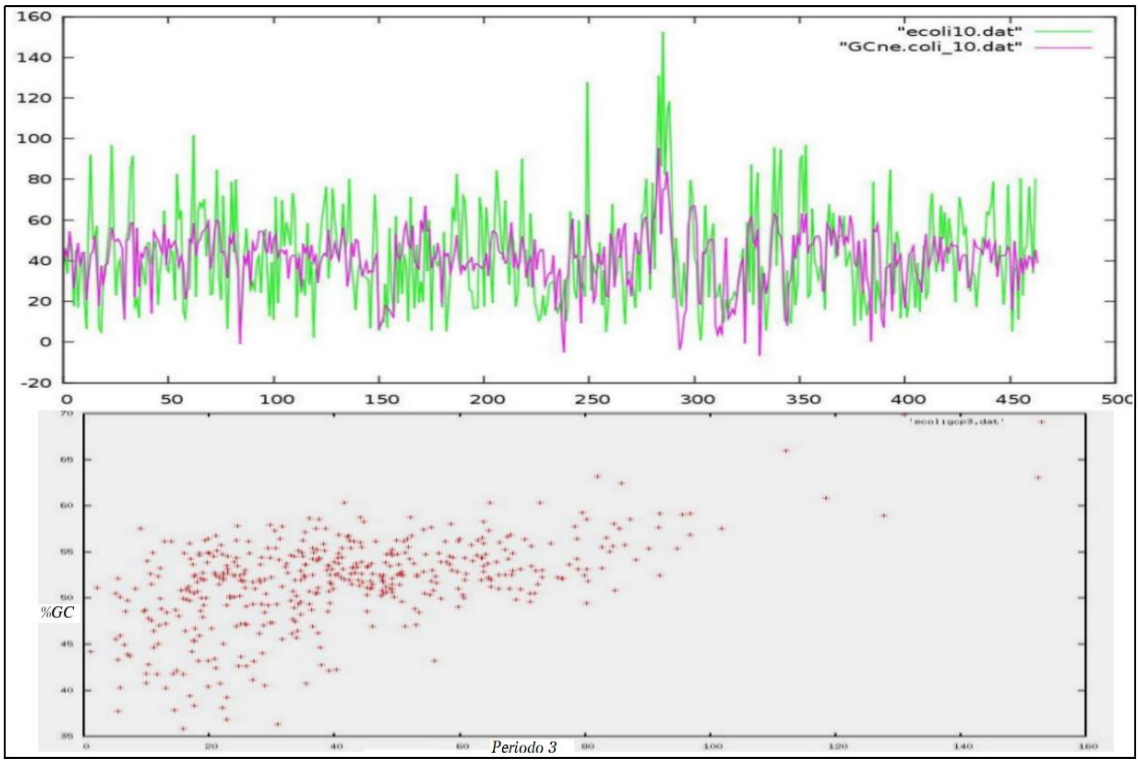


Figura 3.12: Periodo *tres* y porcentaje de *GC* en ventanas de 1000 bases *E. coli*.

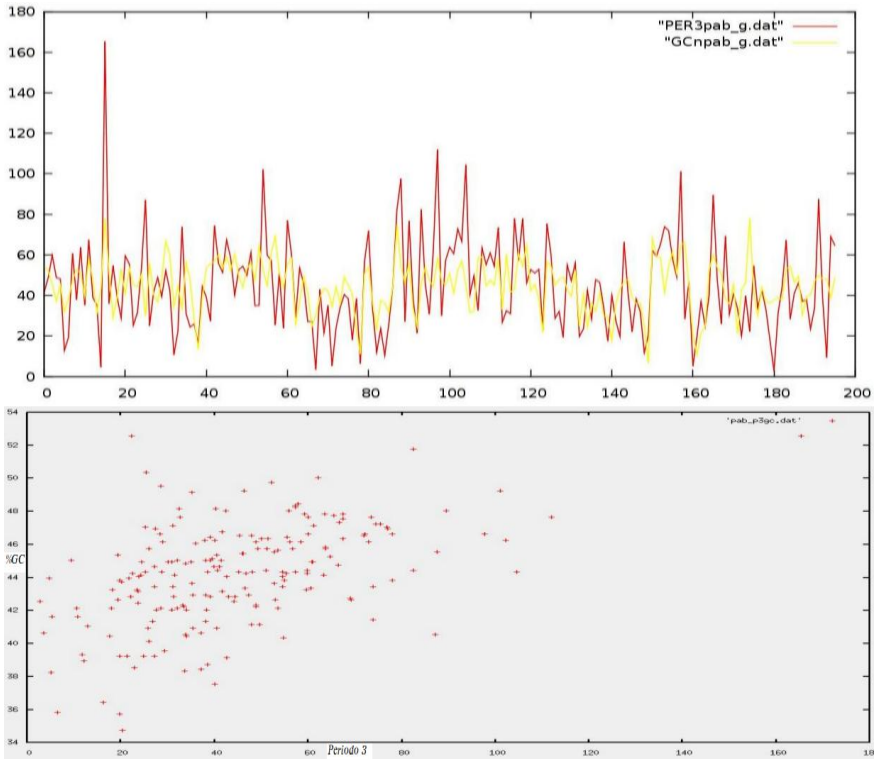


Figura 3.13: Periodo *tres* y porcentaje de *GC* en ventanas de 1000 bases *P. absy*.

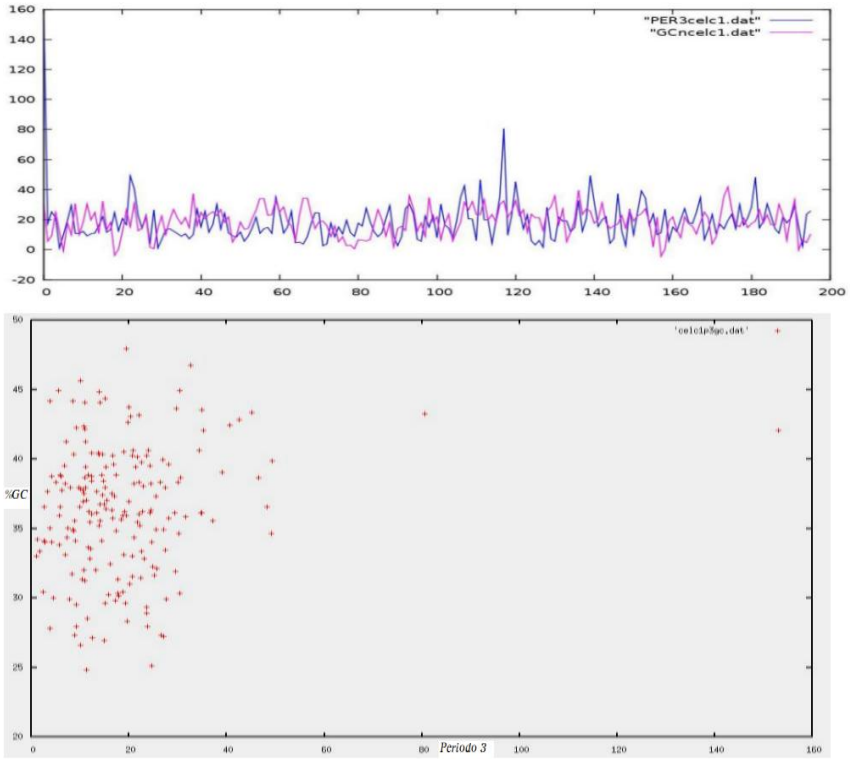


Figura 3.14: Periodo *tres* y porcentaje de *GC* en ventanas de 1000 bases *C.elegans*.

## 3.2. Segunda parte: Exploración de decámeros asociados a nucleosomas

En esta segunda parte, el trabajo se enfocó en reconocer señales que pudieran servir para identificar la formación de nucleosomas. Particularmente se determinó usar una secuencia decamérica que, dispuesta en conjuntos, permite la formación de nucleosomas. Esta secuencia fue estimada en un trabajo de Trifonov *et. al* con secuencias obtenidas experimentalmente del genoma de *C. elegans* [16] y además consensa, en el lenguaje de purinas y pirimidinas (Y/R), otras propuestas de secuencias específicas que se asocian a nucleosomas [57][46].
Se usaron todos los cromosomas de primates, los cromosomas de *C. elegans* y algunos procariontes, para marcar todas los posibles decámeros que, por su naturaleza pirimídica, tuvieran las combinaciones del arreglo decamérico YYYYYRRRRR reportado por Trifonov en *C. elegans* [16]. Para todos los genomas marcados, se hicieron histogramas de densidad y se realizó el análisis de periodicidad usando la función de información mutua y, sólo para algunos cromosomas de humano, se analizó la transformada rápida de Fourier.
Con el fin de evaluar el potencial universal de este decámero para la formación de nucleosomas, se realizó el mismo marcaje en algunos genomas de arqueas y como control, en algunas bacterias. El análisis de información mutua, al igual que el análisis de Fourier, mostró la presencia de las mismas señales periódicas; aunque es un método más sencillo, computacionalmente es más rápido procesar y fue suficiente para este análisis.

El perfil de la función de información mutua, en los cromosomas marcados de primates, mostró una constante regularidad en los picos que corresponden a los periodos 150, 161-172, 190-200, 218, 240, 340-345 y 380, con ligeras variaciones de fase y amplitud que permitieron clasificar a los cromosomas de una misma especie en distintos grupos. Con el apoyo de un análisis estadístico de correlación de Pearson se pudo observar que, incluso, existe una mayor

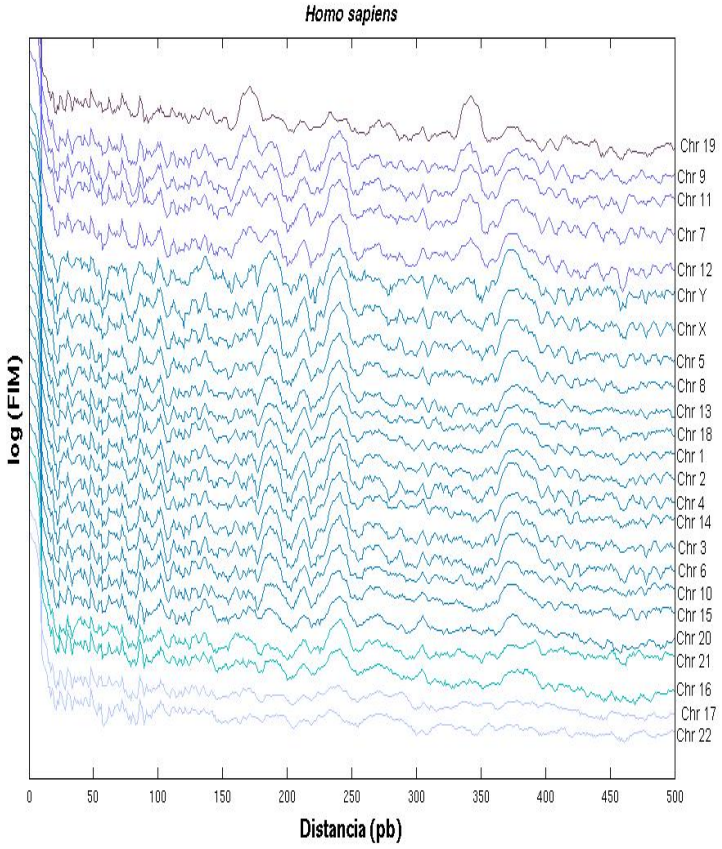


Figura 3.15: Perfiles de la función de información mutua de todos los cromosomas de *Homo sapiens*.

correlación entre algunos grupos de cromosomas entre especies de primates, que en otros grupos de cromosomas dentro de la misma especie. Este hecho es muy notable en el perfil de información mutua del cromosoma 19 de los tres grupos de primates.

En *C.elegans* se observa un perfil de periodicidades mucho más regulares y constantes a lo largo de los cromosomas, aunque los histogramas también registran la presencia de grandes distancias sin este decámero, pero menores

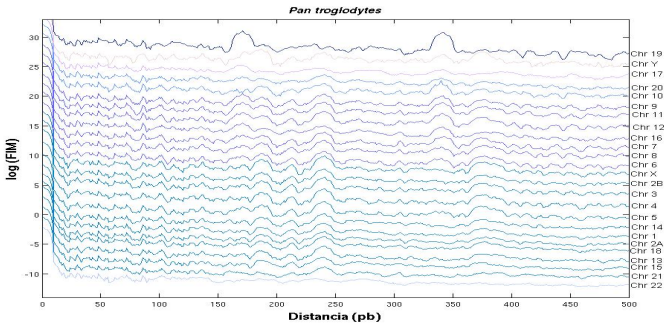


Figura 3.16: Perfiles de la función de información mutua para todos los cromosomas de *Pan troglodytes*.

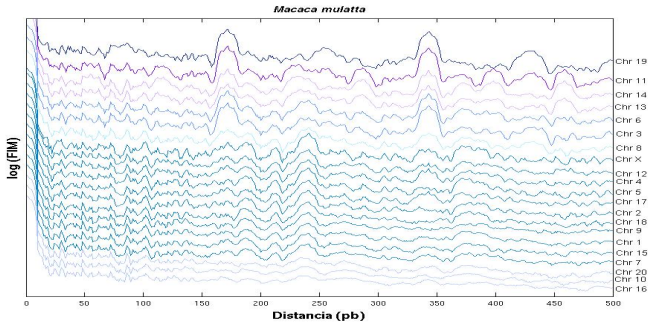


Figura 3.17: Perfiles de la función de información mutua para todos los cromosomas de *Macaca mulatta*.

a las que se encuentran en el humano.

Por la gran variedad de combinaciones de nucleótidos que pueden resultar para formar un decámero, restringido a un orden de cinco consecutivas purinas y cinco consecutivas pirimidinas, no sorprende que se le pueda encontrar en cualquier genoma de manera abundante; sin embargo, la presencia con periodos específicos a lo largo del genoma marca la diferencia entre una distribución simplemente aleatoria y otra que proporciona elementos estructurales en la arquitectura del genoma. Esto es muy evidente en los perfiles

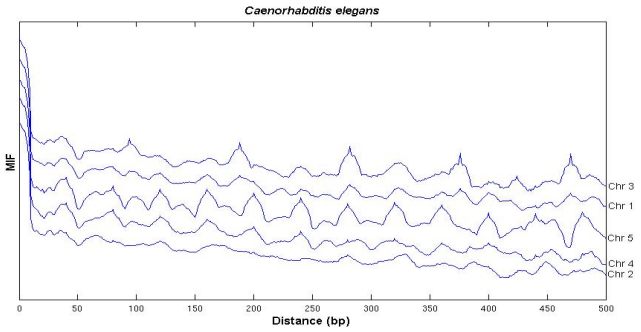


Figura 3.18: Los perfiles de la función de información mutua de todos los cromosomas de *C. elegans* se observan, en general, con osilaciones muy regulares de periodos múltiplos de ∼40 a lo largo de todos los cromosomas.

obtenidos en archaea, que a diferencia de otros procariontes, tienen formación de nucleosomas [40]. En estos organismos, se observan perfiles tanto con señales periódicas de relativa baja amplitud, como casos en donde éstas tienen una muy pronunciada amplitud; tal es el caso de *N. equitans* y *M. jannaschii*. Mientras que en los perfiles de bacterias no se ve prácticamente ningún patrón de periodicidad. Aunque estos notables patrones en archaea no parecen tener similitudes con los perfiles obtenidos en organismos más complejos, la sóla presencia de un patrón puede sugerir que este decámero puede ser un motivo ancestral del genoma que facilita la formación de nucleosomas.

Todos los histogramas de distancias entre decámeros, para cada cromosoma en todos los organismos analizados, resaltaron los mismos picos que se observaron para las respectivas funciones de información mutua. Sin embargo, con los datos obtenidos de los histogramas se pudieron distinguir a detalle las densidades de distancia a lo largo de todo el cromosoma. Con ello se observó que existen regiones con altas densidades del decámero y también regiones muy grandes en donde no se encontró el decámero. De manera muy general, se puede decir con estos resultados que, a lo largo del genoma hay un gran número de decámeros que guardan distancias menores a 500 bases

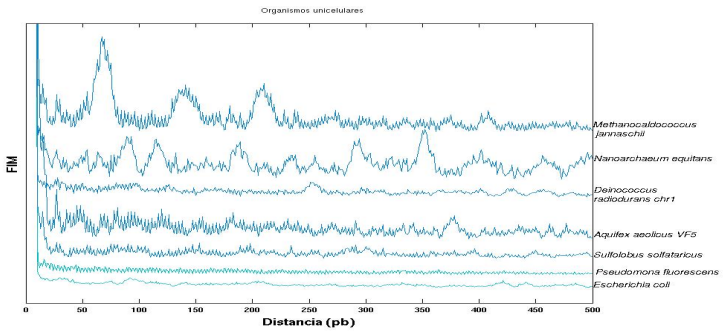


Figura 3.19: Los perfiles de la función de información mutua de organismos unicelulares muestran la presencia de osilciones periodicas en arqueas.

entre ellos y muy pocos que guardan distancias más grandes entre ellos, pero las hay. Por lo tanto, si bien este decámero es una región contundentemente asociada a nucleosomas, la presencia de grandes distancias en las que no se presentan decámeros indica que esta secuencia no debe ser el único motivo que facilite la formación del complejo nucleosómico. En otros análisis de correlaciones paramétricas, que se reportan en el segundo artículo anexado al apéndice se puede ver que, a pesar de su abundancia en el genoma humano, éste decámero mantiene una correlación negativa con la densidad de secuencias repetidas en el genoma, lo cual se puede explicar con la particular distribución y los peculiares hacinamientos reportados en la siguiente subsección de los resultados.

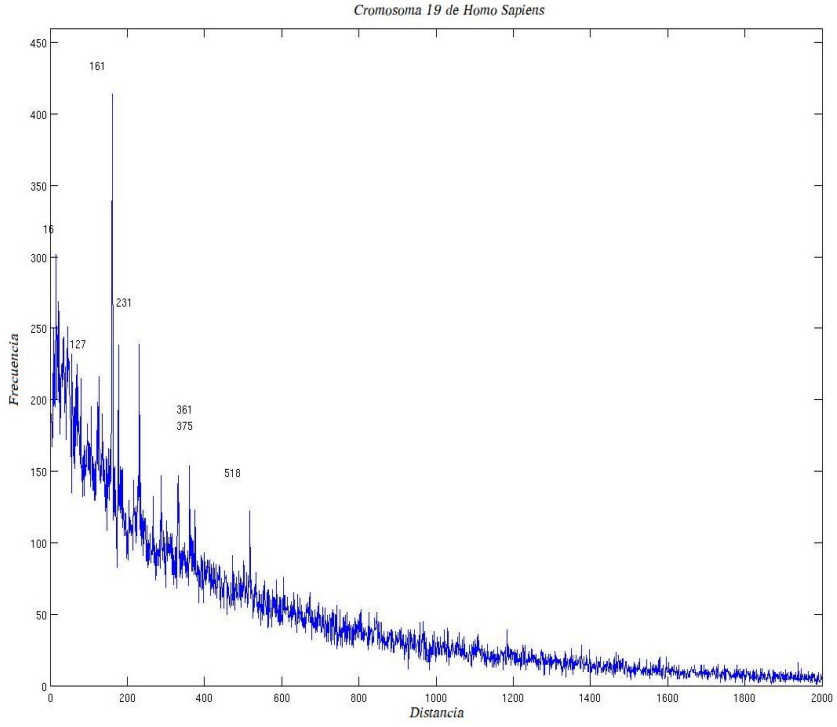


Figura 3.20: Histograma de las distancias del cromosoma 19 del *H. sapiens*.

### 3.2.1. Análisis de arreglos conservados

En los histogramas y perfiles de la función de información mutua en primates fue muy notable la presencia de grandes señales periódicas, algunas previamente reportadas y otras que se pudieron relacionar con secuencias repetidas. (150, 161-172, 190-200, 218, 240, 340-345, 380) Estos resultados se analizan y discuten con mayor detalle en el artículo anexado en el apéndice.

Por otro lado, con algunos análisis parciales con la transformada rápida de Fourier a lo largo de todo el cromosoma 19 del humano y con los datos de las distancias entre decámeros, se observó que sólo en algunas regiones del cromosoma se encontraban densidades del decámero que mantenían distancias entre ellos de alrededor de 80, 161 y 320 bases, de forma intercalada. Estas distancias coinciden con los picos observados tanto en el perfil de la función de información mutua para este cromosoma como en su histograma. (Figura 3.21) Estas densidades del decámero se observan como islas en distintas regiones a lo largo de todos los cromosomas.

En otros casos, se observaron pequeños arreglos o distribuciones del decámero mucho más regulares, en los que, las diferentes distancias que guarda de un decámero con otro, se repiten de manera consecutiva manteniendo el mismo orden a lo largo de pequeñas regiones. (Figura 3.22). En dos casos se encontraron este mismo tipo de arreglos y unas bases adelante, como si fueran espejos, se encontraron arreglos con las mismas distancias pero en orden inverso. (Figura 3.23) La relación de estas estructuras y las periodicidades encontradas con secuencias repetidas, se discuten también en el artículo anexo.

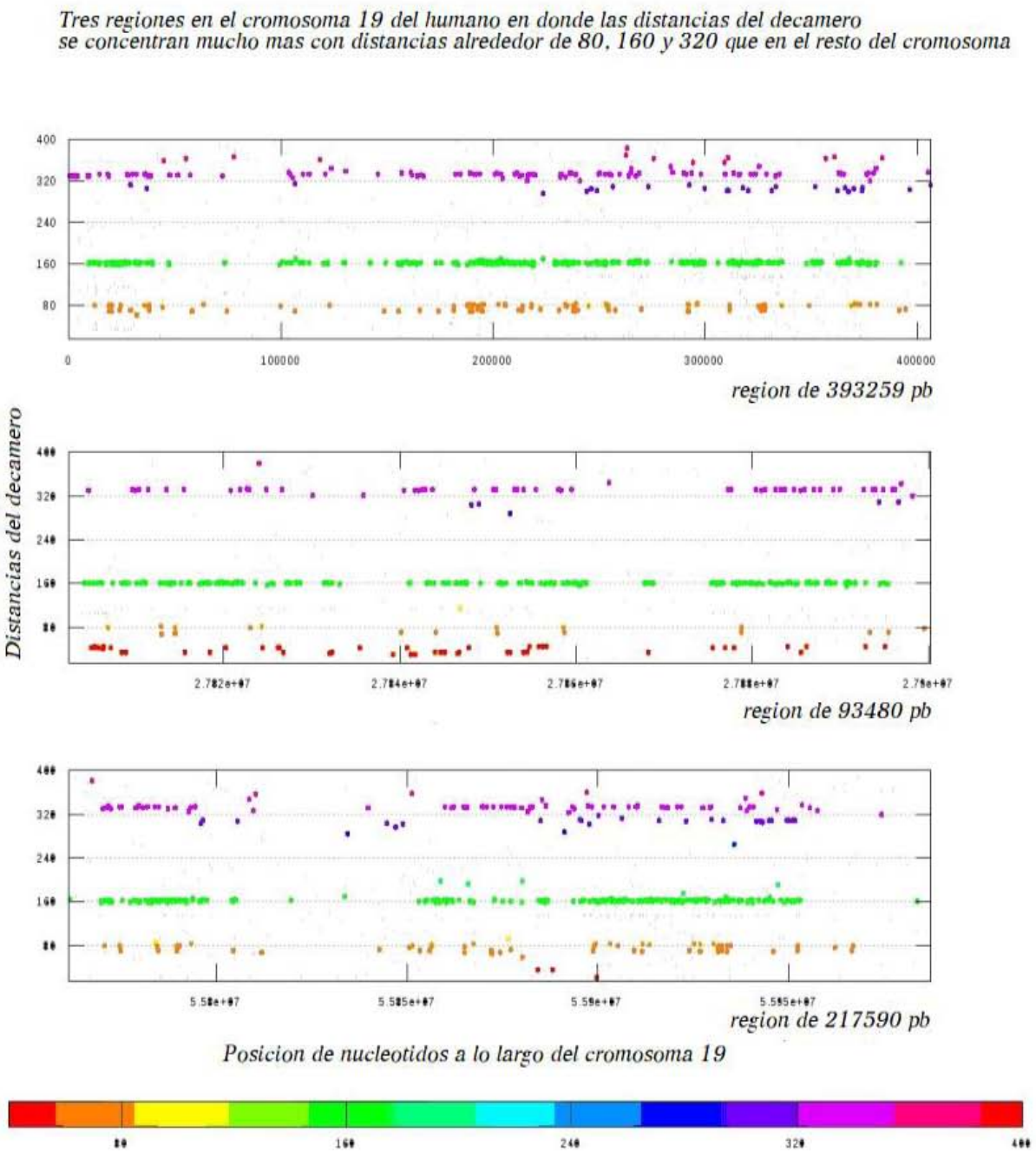


Figura 3.21: Tres de las regiones que se observan como islas en el cromosoma 19 de *H. sapiens*, con una alta densidad de decámeros YYYYYRRRRR que se mantienen separados por distancias de ∼80, ∼161 y ∼320 bases.

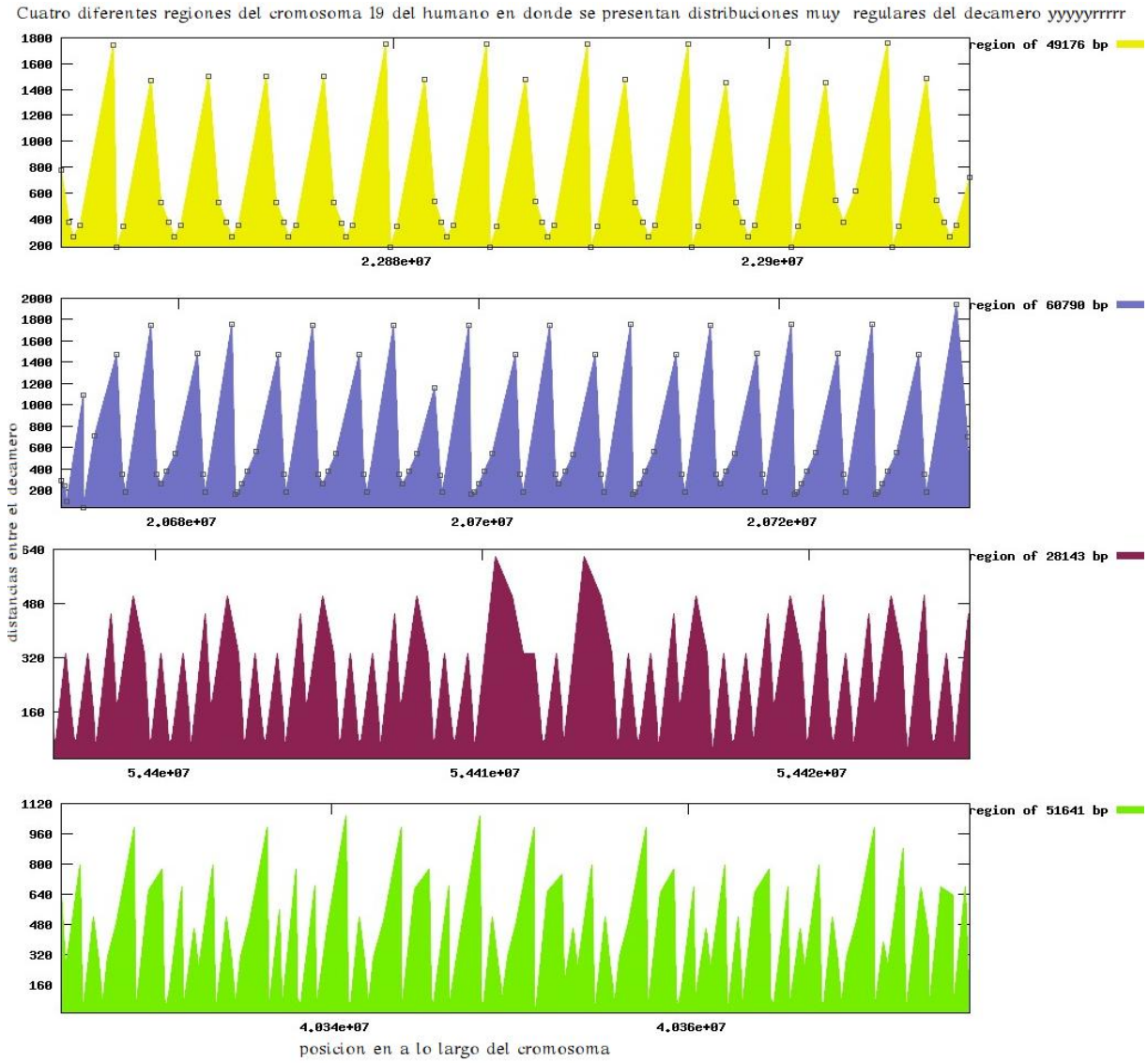


Figura 3.22: Cuatro distintas regiones del cromosoma 19 humano, en donde hay distinas distancias entre los decámeros de tipo YYYYYRRRRR que se repiten de forma periódica.

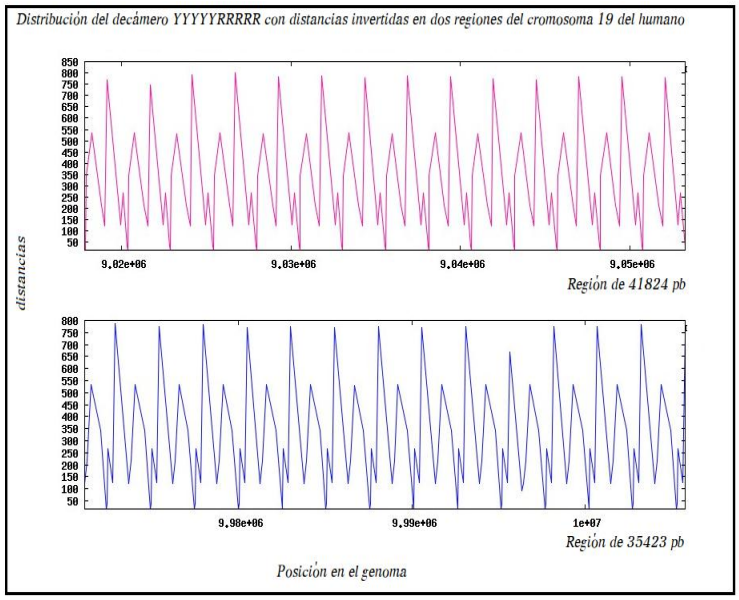


Figura 3.23: Dos regiones del cromosoma 19 del humano. En la primera, se observan diferentes distancias entre los decámeros YYYYYRRRRR que se repiten con un orden de forma periódica, mientras que en la segunda región se observan las mismas distancias entre estos decámeros, de igual manera repetidas de forma periódica, pero con un orden exactamente inverso.

# Discusión y conclusiones

Como conclusión de los primeros análisis de periodicidad en distintos genomas, se observó que los espectros de frecuencias, en ventanas, pueden interpretarse como señales de secuencias con determinadas longitudes o características, presentes de manera recurrente a lo largo de la ventana analizada; la amplitud de estos picos indica qué tan representadas están en esta región.

Sin embargo, también se observo que cualquier señal o secuencia, que corresponda a una determinada frecuencia, se puede perder en presencia de otras más representadas; además, una frecuencia no necesariamente refleja la presencia de una secuencia específica. Un ejemplo de ello es el registro del periodo *tres* sobre representado en regiones codificantes, en donde hay varias secuencias de longitud *tres* repitiéndose en una misma región.

Por otro lado, la distribución de energía libre a lo largo del genoma, según la tabla de Breslauer, deja suponer que los nucleótidos asociados al uso de codones, es decir, las regiones codificantes del genoma, mantienen una distribución constante de energía libre que, a diferencia del resto de las frecuencias visibles en el espectro total de frecuencias, se observa con una amplitud superior de la frecuencia que corresponde al periodo *tres*; es interesante señalar que esto se puede observar aunque existan siete posibles valores de energía libre que dependen de la posición contigua entre nucleótidos a lo largo de la secuencia. Por lo anterior, se puede plantear que el uso de codones mantiene un sesgo mutacional o una restricción inherente a la distribución de la energía libre en la secuencia de nucleótidos.

Además, los análisis de Fourier, registrando sólo el componente de frecuencia $\frac{1}{3}$ con las señales de energía libre, resaltaron patrones distintivos entre genomas eucariontes, eucariontes de genomas reducidos y procariontes. En este análisis, el periodo *tres* se observó sobre amplificado en los organismos cuyo contenido genómico es mayormente codificante, como es el caso de los organismos procariontes u organismos parásitos con genomas reducidos. En contraste, los genomas de eucariontes u organismos que contienen una considerable proporción de material genético, que no se restringe al uso de codones, el periodo *tres* no marca una amplitud tan intensa. Todo lo anterior permitió diferenciar dos rangos de amplitud para la señal de periodo *tres* que facilita distinguir estas dos clases de patrones genómicos relacionados con el periodo *tres* y la naturaleza biológica de los organismos. En algunos organismos se observó una correlación positiva en estos patrones genómicos y el contenido de $GC$, sin embargo, las hipótesis que se pueden plantear con estos resultados pueden ser muy interesantes sólo si se profundizan estos estudios.

De los primeros resultados se puede concluir, también, que el estudio estructural del genoma con estas herramientas tiene muchos alcances y que, tanto la manipulación del algoritmo de la transformada rápida de Fourier, como las características para definir las secuencias analizadas, pueden optimizar el valor informativo del análisis.

Debido a que en estos mismos análisis, usando los valores de energía libre, se detectaron periodos cercanos a 150, 200 y 80 que podrían relacionarse con la distribución de nucleosomas o con los espacios iternucleosómicos, [26] en la segunda parte del trabajo se intenta establecer una relación más clara e informativa entre las posibles señales nucleosómicas observadas y la formación de nucleosomas en el genoma.

Sobre este tema, es importante considerar que existen dos propuestas casi antagónicas respecto a la formación de nucleosomas. Algunos trabajos sugie-

ren que cualquier región del genoma es igualmente asociable a nucleosomas mientras que otros trabajos, por el contrario, describen patrones de secuencias asociadas a nucleosomas que se han obtenido experimentalmente y están sustentados, principalmente, en el hecho de que deben existir regiones distribuidas de tal modo que le confieran al DNA la flexibilidad necesaria para envolver al complejo de histonas [20] [48]. Por otra parte, los estudios sobre la capacidad de remodelaje de la cromatina con la intervención de grupos químicos externos [11], facilitan explicar las cualidades que pudiera tener una secuencia de DNA para permitir la formación de nucleosomas y regular la expresión de genes o la duplicación del genoma.

En los resultados de esta segunda parte se muestran los perfiles de la función de información mutua (FIM), obtenidos en genomas de primates y arquea, marcando a lo largo de los genomas un patrón decamérico de purinas y pirimidinas [57] [46]; el cual se obtuvo experimentalmente del genoma de *C.elegans* por Trifonov *et al.* [57], que además parece consensuar una variedad de secuencias propuestas para formar nucleosomas.[16]

En los genomas de primates se observaron valores periódicos que corresponden a los rangos de distribución que pueden exhibir los nucleosomas en el genoma, además, los perfiles de la función de información mutua muestran tanto picos en periodos ya reportados, tales como las firmas habituales de 31 y 32, 84, 146, 157, 171 y 200 [45] [47] [53], y periodicidades no reportadas y muy conservadas en los genomas analizados de primates tales como 100, 167, 240 y 320.

Tanto en humanos como en el resto de los primates analizados, la mayoría de los cromosomas excepto 19, 22, X e Y, muestran una similitud notable en los periodos que marcan la posición putativa de los nucleosomas obtenidos con este decámero. La regularidad observada en los patrones periódicos, usando este decámero, facilitó describir una clasificación de familias de cromosomas. Esta clasificación esta descrita en el artículo publicado mostrado

en el apéndice.

Sin duda, una de las observaciones más destacables fue que el perfil de la FIM de un cromosoma dado, es más similar al perfil del cromosoma correspondiente entre las especies de primates que al resto de los cromosomas del genoma al que pertenece. El perfil tan conservado de la FIM marcando este decámero, puede reflejar la importancia de este decámero genérico en la arquitectura de los genomas en primates. Las gráficas mostradas en el artículo publicado, demuestran que algunas de las periodicidades encontradas en los genomas de primates con este decámero genérico, están fuertemente asociadas a secuencias altamente repetidas.

En particular, las constantes periodicidades de 31 y 32 , en el caso de los genomas de primates, parecen estar relacionados con las inserciones conservadas de secuencias *Alu*. Estudios recientes muestran que hay islas de *GpC* encontradas en estos elementos *Alu*, que mantienen distancias de 31-32 entre cada una [38], lo que puede explicar su relación con estos periódos. Por otro lado, estas islas de *GpC* también las hacen regiones propensas a los procesos de metilación y desmetilación, lo que significa que son elementos activos en el remodelaje de la cromatina por lo que podrían ser considerados como "nucleosomas epigenéticos"[38][47].

Esto concuerda con el reciente descubrimiento de que el posicionamiento de los nucleosomas parecen estar en fase con los elementos *Alu*, tal como se refleja en los picos 84 pb y 167 pb del espectro de Fourier ya reportados [53] y también encontrados en este trabajo, pero usando el decámero de purinas y pirimidinas, tanto en los primeros análisis de Fourier como con los perfiles de la FIM.

Por otro lado, analizando los histogramas de distribución de este decámero, se lograron detectar regiones con distancias muy regulares del decámero a lo largo de los cromosomas de primates en largos tramos y a distintas distancias en los diferentes cromosomas, así como también se observaron patrones

que reflejan repeticiones invertidas (de simetría inversa). En un seguimiento específico de este análisis para cromosoma 19 del humano, se pudieron observar con claridad regiones en donde este decámero se agrega densamente cada 80, 160 y 320 pb, cifras que coinciden con las periodicidades encontradas en los perfiles de información mutua.

Estos resultados nos permiten concluir que la relevancia biológica de este decámero no sólo radica en su contribución para facilitar la formación de nucleosomas, si no que su relación con la arquitectura de cromosomas y secuencias repetidas, en particular con las secuencias *Alu*, pueden dejarnos suponer que puede estar implicado en las funciones auto regulatorias que caracterizan a estas secuencias [54][41][20] y explicar, por consiguiente, las simetrías invertidas observadas, así como su exitosa propagación en los genomas de primates.

La abundante presencia de estas secuencias, únicas en su forma dimérica dentro de los primates, así como su participación en funciones de regulación de la expresión genética, tales como el *splicing alternativo* cuando está presente en regiones intrónicas de genes[24], realza la importancia que pueden tener en la evolución de los primates [37] y quizá contribuye a explicar las diferencias fenotípicas que existen en este grupo. Esta propuesta se discute con más detalle en el artículo publicado anexado en el apéndice.

Curiosamente, los perfiles de la FIM del decámero genérico en todas las archaeas analizadas, mostraron picos prominentes que muestran una recurrencia periódica de este decámero, lo que sugiere la necesidad de estudios adicionales con el propósito de determinar si pudiera haber existido una selección en el origen de la estructura del nucleosoma, o considerársele como la estructura más fundamental de un nucleosoma, sin los refinamientos evolutivos posteriores conferidos al posicionamiento de nucleosomas en eucariontes superiores. Sin embargo, es la primera vez que se describe esta secuencia putativa para la formación de nucleosomas en arqueas usando la función de información mutua.

Finalmente, con los perfiles de la FIM se pudieron identificar estos motivos decaméricos que facilitan el posicionamiento de nucleosomas asociados a elementos repetitivos en los primates no humanos, humanos y en archaea; pero también se encontraron grandes islas desiertas de este decámero en algunas regiones de los cromosomas; lo anterior junto con el hecho de que los perfiles de la FIM observados en diferentes cromosomas o de diferentes especies a menudo difieren sustancialemnte, implica que este decámero no es la única composición inherente a la formación de nucleosomas.

Los resultados de este trabajo pueden contribuir a la comprensión del origen de las estructuras de nucleosomas en archaeas y al del notable éxito de *Alu*-retrotransposición para colonizar los genomas de primates.

# Figuras suplementarias

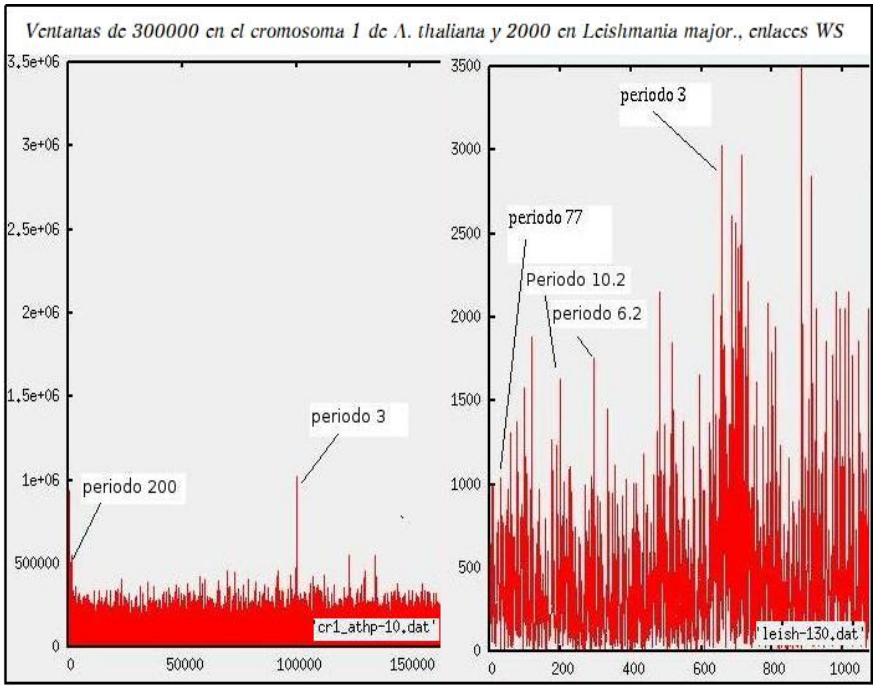


Figura 3.24: Espectro de frecuencias usando la distribución de enlaces *WS* en ventanas de 300000 bases en *Arabidopsis thaliana* y de 2000 bases en *Leishmania major*.

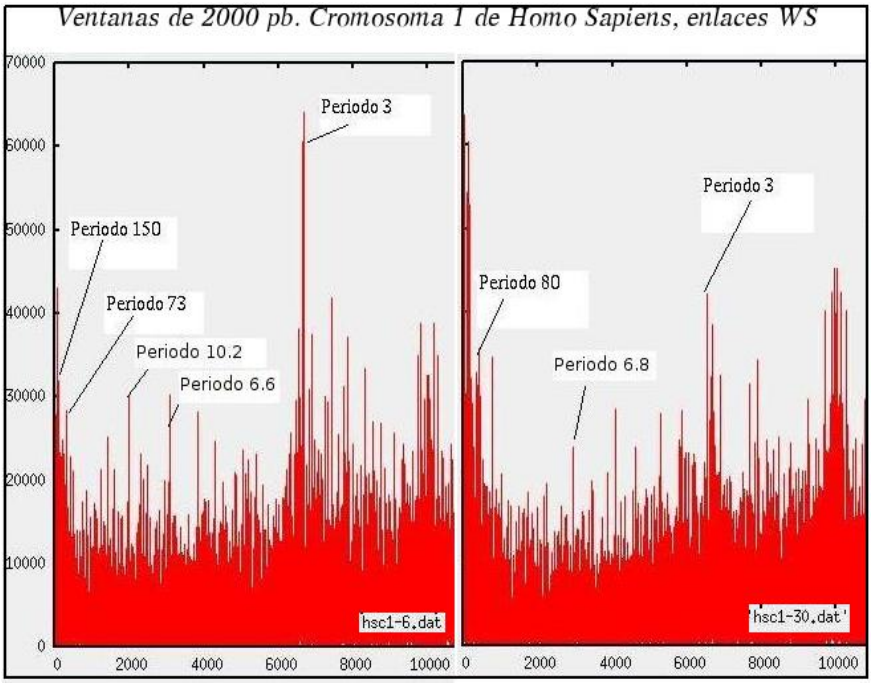


Figura 3.25: Espectro de frecuencias usando la distribución de enlaces *WS* en ventanas de 20000 bases. *Homo sapiens*.

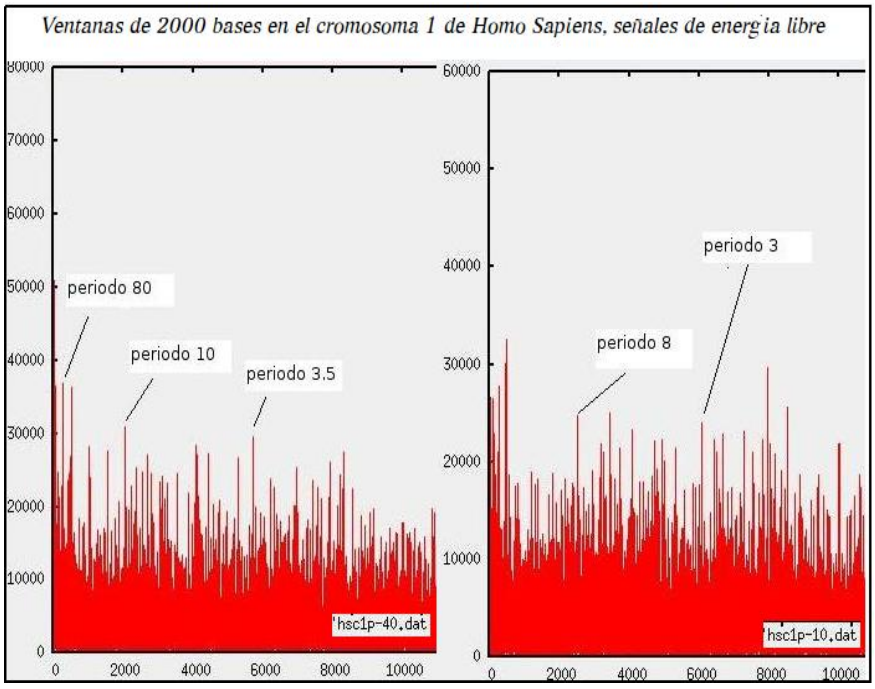


Figura 3.26: Espectro de frecuencias usando la distribución de energía libre en ventana de 20000 bases. *Homo sapiens*.

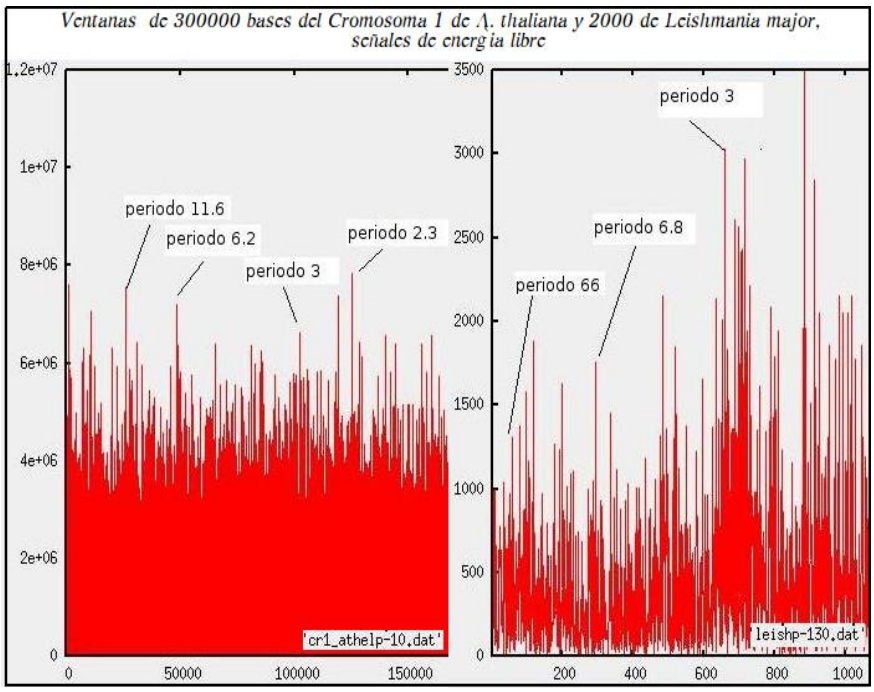


Figura 3.27: Espectro de frecuencias usando la distribución de energía libre en ventanas de 300000 bases en *Arabidopsis thaliana* y 2000 bases en *Leishmania major*.

# Bibliografía


[1] Andrezej K. Brodzik , 2007 Quaternionic periodicity transform: an algebraic solution to the tandem repeats detection problem. Bioinformatics 23(6):694-700

[2] Batzer MA, Deninger PL, 2002 *Alu* repeats and human genome diversity. Nat.Genetics 3:370-380

[3] Bao, N.,Lye,K.W., Barton M.K.,2004. MicroRNA binding sites in Arabidopsis class III HD-ZIP mRNAs are required for methylation of the template chromosome. Dev. Cell 7:653-662

[4] Breslauer K. J., Marky L.A.,1987. Calculating thermodynamic data for transitions of any molecularity from equilibrium melting curves. Biopolymers. 26(9): 1601-1620.

[5] Britten, R.J. Davidson, E.H. 1971. repetitive and nonrepetitive DNA sequences and a speculation on the origins of evolutionary novelty. Q. Rev. Biol. 46 111-133

[6] Cooley, James W.; Tukey, John W. (1965). .ᴬn algorithm for the machine calculation of complex Fourier series". Math. Comput. 19: 297âĂŞ301

[7] Boekhorst Rene, Abnizova Irina, Nehaniv Chrystopher, 2008. Discriminating coding, non-coding and regulatory regions using rescaled range and detrended fluctuation analysis. BioSystems 91(2008):183-194

[8] Dworkin Ian, 2005. Towards a genetic architecture of cryptic genetic variation and genetic assimilation:the contribution of K.G Bateman. Journal of Genetics 84(3):223-226

[9] Eskesen Stephen T, Eskesen Frank N, Kinghorn Brian, Ruvinsky Anatoly,2004 . Periodicity of DNA in Exons. BMC Molecular Biology 5:12

[10] Farber Robert, Lapedes Alan, Sirotkin Karl. 1992. Determination of Eukaryotic Protein Coding Regions Using Neural Networks and Information Thoery. J.Mol.Biol. 226:471-479

[11] Felsenfeld, Groudine, 2003 Controlling the double helix. Nature 421(6921):448-53

[12] Ferreira Paulo J.S G, Neves Antonio J.R., Vera Afreixo, Pinho Armando J., 2006 Exploring Three-Base periodicity for DNA compression and modeling IEEE 5:877-880

[13] Feschotte C, Pritham EJ. 2007 DNA *transposons* and the evolution of eukaryotic genomes. Annu Rev Genet. 41:331-68.

[14] Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC.,1998 Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans.Nature. 391(6669):806-11.

[15] Tolstov Georgi P, Fourier Series ed. 1962

[16] Gabdank I, Barash D, Trifonov EN. Nucleosome DNA bendability matrix (C. elegans). J Biomol Struct Dyn. 26(4):403-11 (2009)

[17] Häsler J, Strub K (2006) *Alu* RNP and *Alu* RNA regulate translation initiation in vitro. Nucleic acids Res. 34(8):2374-2385.

[18] Hickey DA. 1992 Evolutionary dynamics of transposable elements in prokaryotes and eukaryotes. Genetica. 86(1-3):269-74.

[19] Hornstein Eran & Shomron Noam, 2006 Canalization of development by microRNAs.Nature 38:514-519

[20] Ioshikhes IP, Albert I, Zanton SJ, Pugh BF.,2006. Nucleosome positions predicted through comparative genomics.Nat Genet. Oct;38(10):1210-5.

[21] International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860âĂŞ921.

[22] Jablonka E. Lamb M J.,2005. Evolution in four dimensions. The MIT Press, Cambridge, Massachusetts London England.

[23] Korotkov EV, Korotkova MA., 2010 Study of the triplet periodicity phase shifts in genes. J Integr Bioinform. 430:45-49

[24] Kreahling J, Graveley BR, 2004 The origins and implications of Aluternative splicing. Trends Genet. 20:1âĂŞ4.

[25] Krutzfeld Jan, Poy Matthew N, Stoffel Markus,2006. Strategies to determine the biological function of microRNAs. Nature 38:514-510

[26] Lewin Benjamin. 2004, GenesVIII. Pearson prentice Hall

[27] Li W, 1997 The complexity of DNA. J W & Sons,Inc. 3(2):33-37

[28] Li W, 2004. Spectral analysis of guanine and cytosine fluctuations of mouse genomic DNA. Fluctuation and Noise Letters 4(3): L453-L464

[29] Li W. 2007, The complexity of DNA Comlexity, 3: 33-37

[30] Li W. and Miramontes P.2006, Large-scale Oscillation of Structure-Related DNA Sequence Features in Human Chromosome 21,Physical Review E. 74 021912

[31] Luan D D and T H Eickbush.1995, RNA template requirements for target DNA-primed reverse transcription by the R2 retrotransposable element. Mol Cell Biol. 15(7): 3882âĂŞ3891.

[32] Madsen BE, Villesen P, Wiuf C. 2008 Short tandem repeats in human exons: a target for disease mutations.BMC Genomics. 9:410.

[33] Mallory Allison C., Vaucheret Hervé, 2006 Functions of microRNAs and related small RNAs in plants. Nature 38:531-535

[34] Mette MF., Aufstz W., van de Winder J., Matzke M., Matzke A, 2000 Transcriptional gene silencing and promoter methylation triggered by double stranded RNA. EMBOJ 19:1519-1524.

[35] Miramontes P, Medrano L, Cerpa C, Cedergen R, Ferbyre G, Cocho G,(1995) Structural and Thermodynamic properties or DNA uncover different evolutionary histories. J Mol. Evol; **40**(6): 698-704

[36] Guo X., Mrázek J. 2008. Long simple sequence repeats in host-adapted pathogens localize near genes encoding antigens, housekeeping genes, and pseudogenes. J Mol Evol. 67(5):497-509.

[37] Mighell AJ, Markham AF, Robinson PA, 1997, *Alu* sequences. FEBS Lett.417: 1-5.

[38] Ong MS, Richmond TJ, Davey CA (2007) DNA stretching and extreme kinking in the nucleosome core. J. Mol. Biol. 368(4):1067-74.

[39] Orlov YL, Te Boekhorst R, Abnizova II. 2006. Statistical measures of the structure of genomic sequences: entropy, complexity, and position information. J Bioinform Comput Biol. 4(2):523-36.

[40] Pereira SL, Grayling RA, Lurz R, Reeve JN.,1997. Archaeal nucleosomes.Proc Natl Acad Sci U S A.Nov 11;94(23):12633-7.

[41] Quentin Y, 1994. Emergence of master sequences in famiies of retropsons derived from 7sl RNA. Genetica 93:203-215.

[42] Raman Arora, William A, Seatheres, James A Buckew, 2008. Latent Periodicities in Genome Sequences. IEEE Journal of selected topics in signal processing 2(3)

[43] Richard G, KerrestA, Dujon B. 2008 Comparative genomicas and molecular dynamicas of DNA repeats in eukaryotes. Micriobiol. Mol. Biol.Rev. 72(4):686-727.

[44] Rocha, Eduardo P.C , Danachin Antonie , Viari Alain. 1999. Funtional and evolutionary roles of long repeats in prokaryotes. 150(1999):725-733

[45] Sasaki S, Mello CC, Shimada A, Nakatani Y, Hashimoto S, Ogawa M, Matsushima K, Gu SG, Kasahara M, Ahsan B, Sasaki A, Saito T, Suzuki Y, Sugano S, Kohara Y, Takeda H, Fire AZ, Morishita S (2009) Chromatin-associated periodicity in genetic variation downstream of transcriptional start sites. Science 323(5912):401-404.

[46] Salih F, Salih B, Trifonov EN. 2007 Sequence structure of hidden 10.4-base repeat in the nucleosomes of C. elegans. J Biomol Struct Dyn. 26(3):273-82.

[47] Salih F, Salih B, Kogan S, Trifonov EN. 2008 Epigenetic nucleosomes: *Alu* sequences and CG as nucleosome positioning element. J Biomol Struct Dyn. 26(1):9-16.

[48] Segal E, Fondufe-Mittendorf Y, Chen L, ThÃěstrÃűm A, Field Y, Moore IK, Wang JP, Widom J.,2006.A genomic code for nucleosome positioning.Nature.Aug 17;442(7104):772-8.

[49] Shannon Claude E. , 1948. A Mathematical Theory of Communication, Bell System Technical Journal, 27: 379âĂŞ423 y 623âĂŞ656.

[50] Smit Arian FA, 1996. The origin of interspersed repeats in the human genome. Curr Curr Opin Genet Dev. 6(6):743-8.

[51] Steven W. Smith, Ph.D.The Scientist and Engineer's Guide to Digital Signal Processing. Chapter 8.

[52] Strub K, Moss J, Walter P, 1991 Binding sites of the 9- and 14-kilodalton heterodimeric protein subunit of the signal recognition particle (SRP) are contained exclusively in the *Alu* domain of SRP RNA and contain a sequence motif that is conserved in evolution. Mol. Cell. Biol. 11:3949-3959.

[53] Tanaka Y, Yamashita R, Suzuki Y, Nakai K, 2010 Effects of *Alu* Elements in nucleosome positioning in the human genome. BMC Genomics 11:309-318.

[54] Thomas H. Eickbush and Varuni K. Jamburuthugoda,2008.The diversity of *retrotransposons* and the properties of their reverse transcriptases. 134(1-2):221-234

[55] Tiwari Shirh, S.Ramachandran, Alok Bhattacharya, Sudha Bhattacharya, Ramakrishna Ramaswamy. 1997. Prediction of probable genes by Fourier analysis of genomic sequences. Cabios 13(3):263-270

[56] college of LEX. Who is Fourier, 1995. Language Research Foundation.7th printing 2004. USA

[57] Trifonov, 2010 Nucleosome Positioning by Sequence. State of Art and Apparent Final J Biomol Struct Dyn. 27(6):741-6.

[58] Tsonis Anastasios A., Elsner James B.,Tsonis Panagiotis A. 1991. Periodicity in DNA Coding Sequences: Implications in Gene Evolution. J. Theoretical Biol.151(3):3âĂŞ331

[59] van Belkum A, Scherer S, van Alphen L, Verbrugh H.1998 Short-sequence DNA repeats in prokaryotic genomes. Microbiol Mol Biol Rev. 62(2):275-93.

[60] Volpe TA., Kidner C., Hall IM. Teng G., Greawel SI., Martienssen R.A 2002. Regulation of heterocromatic silencing and histone H3 lysine 9 methilation by RNA. Science 297(5588):1833-7

[61] Weichenrieder O, Wild K, Strub K, Cusack S, 2000 Structure and assembly of the *Alu* domain of the mammalian signal recognition particle. Nature 408:167-173.

[62] West-Eberhard Mary Jane, Developmental plasticity and the origin of species differences,2005.PNAS 103(1):6543-6549

[63] Widlund H,R. Kudvalli, P.N. Bengtsson, M., Cao, H., Tullius, T.D., and ubitsa, M. 1999. Nucleosome structural features and intinsic properties of the TATAAACGCC repeat sequences. J.Biol.Chem. 274: 31847-31852

[64] Williamk K.D., Helin A.B.,Posluszny J.,Roberts S.P. and Feder, M. E. 2003. Effect of heat shock, pretreatment and hsp70 copy number on wing development in Drosophila melanogaster.Mol.Ecol.12: 1167-1177

[65] Yin Changchuan, Yau Stephen S.-T.,2007 Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. J. Theoretical Biol.247(2007): 687-694

[66] Zuckerkandl E, Cavalli G.,1992 Combinatorial epigenetics, Junk DNA, and the evolution of complex organisms. Genetica, 86(1-3):269-74

# Apéndice

Hindawi Publishing Corporation
International Journal of Genomics
Volume 2013, Article ID 963956, 13 pages
http://dx.doi.org/10.1155/2013/963956

*Research Article*

# Periodic Distribution of a Putative Nucleosome Positioning Motif in Human, Nonhuman Primates, and Archaea: Mutual Information Analysis

**Daniela Sosa,[1,2] Pedro Miramontes,[1,2] Wentian Li,[3] Víctor Mireles,[1,2] Juan R. Bobadilla,[4] and Marco V. José[4,5]**

[1] *Facultad de Ciencias, Universidad Nacional Autónoma de México, 04510 CP, DF, Mexico*
[2] *Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, 04510 CP, DF, Mexico*
[3] *The Robert S. Boas Center for Genomics and Human Genetics Manhasset, Feinstein Institute for Medical Research, North Shore LIJ Health System, Manhasset, NY, USA*
[4] *Theoretical Biology Group, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, 04510 CP, DF, Mexico*
[5] *Centro Internacional de Ciencias, Cuernavaca, Morelos, Mexico*

Correspondence should be addressed to Marco V. José; marcojose@biomedicas.unam.mx

Received 13 February 2013; Accepted 29 April 2013

Academic Editor: Ancha Baranova



Recently, Trifonov's group proposed a 10-mer DNA motif YYYYYRRRRR as a solution of the long-standing problem of sequence-based nucleosome positioning. To test whether this generic decamer represents a biological meaningful signal, we compare the distribution of this motif in primates and Archaea, which are known to contain nucleosomes, and in Eubacteria, which do not possess nucleosomes. The distribution of the motif is analyzed by the mutual information function (MIF) with a shifted version of itself (MIF profile). We found common features in the patterns of this generic decamer on MIF profiles among primate species, and interestingly we found conspicuous but dissimilar MIF profiles for each Archaea tested. The overall MIF profiles for each chromosome in each primate species also follow a similar pattern. Trifonov's generic decamer may be a highly conserved motif for the nucleosome positioning, but we argue that this is not the only motif. The distribution of this generic decamer exhibits previously unidentified periodicities, which are associated to highly repetitive sequences in the genome. *Alu* repetitive elements contribute to the most fundamental structure of nucleosome positioning in higher Eukaryotes. In some regions of primate chromosomes, the distribution of the decamer shows symmetrical patterns including inverted repeats.

## 1. Introduction

It is generally accepted that the chromatin organization of eukaryotic DNA is strongly governed by a code inherent to the DNA sequence. Modulating the accessibility of individual DNA sequences involves many complex interactions, the most prevalent of which are the interactions between histone octamers and DNA in compacted chromosomes [1, 2]. The condensation of DNA into an ordered chromatin structure allows the cell to solve the topological problems associated with storing huge amount of information of chromosomal DNA within the nucleus. In Eukaryotes and Archaea, DNA is packaged into chromatin in orderly repetitive protein-DNA complexes called nucleosomes. Each nucleosome consists of approximately 146-147 bp of dsDNA wound 1.7-1.8 times around a histone octamer [3–5] to form the basic unit of chromatin structure, the nucleosome. Each octamer is composed of two H3-H4 histone dimers bridged together as a stable tetramer that is flanked by two separate H2A-H2B dimers [6]. Stretches of DNA called linker up to 100 bp, often with an increment of 10 bp, separate adjacent nucleosomes. Multiple nuclear proteins bind to this linker region, some of which may be responsible for the ordered wrapping of strings of nucleosomes into higher-order chromatin structures [7].

Histone proteins condense DNA into complex nucleosome structures both in Eukaryotes and Archaea [2, 8]. Nucleosomes were originally regarded as a distinguishing feature of Eukaryotes prior to identification of histone orthologs in Archaea [9, 10]. The underlying DNA sequence, sometimes called "nucleosome core sequence" or "nucleosome positioning sequence," acts to bias its own packaging in nucleosomes through preferential positioning of histone octamer. It can facilitate DNA wrapping by placing AA dinucleotides along the portion of the DNA helix that faces the histone core complex [11–13]. Thus, DNA sequences that favor nucleosome formation are enriched with AA dinucleotides spaced ~10 bp apart, resulting in a deficiency of TT dinucleotides at the same location and on the strand facing the histone [11–14]. Five to six nucleotides in either direction, where the complementary strand faces the histone core, the trend is reversed (TT enrichment and a deficit of AA). Two main classes of nucleosome positioning sequence (NPS) patterns have been described. In the first class, AA, TT, and other WW dinucleotides (W = A or T) tend to occur together (in phase) in the major groove of DNA closest to the histone octamer surface, while SS dinucleotides (S = G or C) are predominantly positioned in the major groove facing outward. In the second class, AA and TT are structurally separated (AA backbone near the histone octamer and TT backbone further away), but grouped with other RR (where R is purine A or G) and YY (where Y is pyrimidine C or T) dinucleotides. As a result, the RR/YY pattern includes counterphase AA/TT distributions [15].

In the literature, nucleosome positioning is widely regarded as being sequence specific, enabling them with features of regulation of the access of nonhistone proteins to DNA *in vivo* (e.g., [16]). Albeit, the sequence-dependency of nucleosome positioning is still under debate (see, e.g., [16–21]), the fact that histone proteins in Eukaryotes are highly conserved whereas the genome sequences and the positioning sequence motifs seem to be highly divergent among organisms opens an intriguing question.

Both DNA sequence and nucleosome positioning are important factors in gene regulation [22–24]. Accessibility of transcription binding sites crucially depends on the nucleosome positioning [25, 26]. Nucleosomes are distributed in a highly nonrandom fashion around transcription start sites [27, 28]. Replication is dependent on nucleosome positioning [29].

Yet the so-called chromatin code has not been fully determined. This code is a well hidden, weak periodical DNA sequence pattern that is recognized by histone octamers. However, the weak signal is not a problem for the histone octamer. It may select the best bendable segments in random DNA sequences. Additionally, as experimental nucleosome mapping indicates, most of the nucleosomes have only marginal stability [13, 29]. It does not mean, however, that their positions are fully uncertain [30, 31]—as much as 50% contribution may come from sequence itself to determine whether a region is covered by a nucleosome or not [16].

The original assumption that DNA sequence is the major factor in nucleosome positioning was first made as early as 1975 [32] and later in 1984 [33] and confirmed afterwards [34, 35]. However, the exact formulation of the positioning pattern remained elusive. Recently, Trifonov's group has provided a pattern that they claim to be an ultimate solution of the long-standing problem of sequence-based nucleosome positioning [36]. Two basic binary periodical patterns are well established: in purine/pyrimidine alphabet—YRRRRRYYYYYR and in strong/weak alphabet—SWWWWWSSSSSW (S/W). Their merger (shifted by 5 bases) in four-letter alphabet sequence coincides with the first complete matrix of nucleosome DNA bendability [37], which was derived from a large database of nucleosome core DNA sequences generated by micrococcal nuclease (MNase) digestion of *C. elegans* chromatin [38, 39]. The results from the bendability analysis indicate that the sequence CGGAAATTTC, called a CG/AT motif, with CG and AT elements 5 bases apart, is predominant in nucleosome cores at the centers of complementary symmetry of the consensus nucleosome-binding pattern derived from bendability data. A more inclusive, but consistent with all previous proposals, consensus nucleosome positioning pattern observed in *C. elegans* was $(YYYYYRRRRR)_n$. Note that on the reverse complementary strand, the motif is still YYYYYRRRRR (Y/R), but if shifted by 5 bases, it becomes RRRRRYYYYY (R/Y) [40].

The solution was claimed by Trifonov's group to be unique, hence universal, since the physics of DNA bendability should, in principle, be the same for all species [36]. The simple higher occurrence common consensus of the motifs is TTTCCGGAAA, which is identical to their CG/AT motif derived from *C. elegans* nucleosomes [25, 41]. None of other suggested motifs scores better when compared to the rest of the set. Indeed, the experimental data on *C. elegans* were convincingly consistent with the decamer YYYYYRRRRR in regard to its association to nucleosome positioning partly because the motif was derived from the *C. elegans* MNase digestion data. This alone is a good reason to believe that the CG/AT sequence, as well as the more general YYYYYRRRRR motif, is a universal DNA bendability pattern. Another reason is that this motif can be derived from simple DNA deformability considerations, by minimizing unstacking of bases and base pairs caused by DNA bending on the surface of the histone decamer [36].

Analysis of periodicities in 13 fully sequenced eukaryotic genomes [42] showed that weakly periodically positioned TA dinucleotides are detected only in *Saccharomyces cerevisiae*.

The rationale of our work is as follows. If the generic decamer possesses inherent stability properties making it a universal nucleosome positioning sequence throughout Eukarya, we hypothesized that this decamer signal, caused by a regular spacing of nucleosomes, could also be detected in Archaea, whereby vestiges of primitive nucleosome structures could be identified [10, 43], but lacking in Eubacteria where the nucleosome structure does not exist. The goal of this work was to test the universality hypothesis of the putative nucleosome motif YYYYYRRRRR. To this end, we used mutual information function (MIF) profiles of the generic decamer YYYYYRRRRR along the entire genomes of 3 primate species and 4 species of Archaea. We also tested the S/W motif in all organisms. We show that the overall MIF profiles for the Y/R decamer for each chromosome in each

primate species followed a similar periodic pattern, whereas the S/W motif is regular but only in a few chromosomes of primate species. In Archaea species the MIF profiles were different but showed conspicuous periodic features. Hence, with the assumption that an appropriate periodic signal is an indication of the regular spacing of nucleosomes, the Y/R decamer seems to be a highly conserved motif of nucleosome positioning. We used as controls genomes of 3 bacteria, in which there are no nucleosomes, to show that the periodic signal is absent.

On the other hand, the long distance of the regular spacing reflects a low density of the Y/R decamer in these genomes. One implication is that the decamer may not occupy positions at every helix turn, more likely at every nucleosome. Another implication is that other motifs beside this decamer may play a role in the nucleosome positioning.

To further test whether decamer Y/R was able to cast light upon the nucleosome positioning, we generated 10 random sequences of decamers preserving the 5 Ys and 5 Rs content for each chromosome. We found that the random decamers did not present clear-cut patterns in the MIF profiles along chromosomes in contrast to Trifonov's decamer.

Our work is consistent with the assumption that Trifonov's generic decamer is one of the nucleosome positioning motifs in primates and in Archaea, and nucleosomes are regularly spaced. However, this motif was derived by conditioning on CG (or AA, AT, TT) as the flanking dinucleotide with periodicity of 10 (CG-8-CG, or AA-8-AA, etc.), which excludes any nucleosome positioning motifs that do not have these periodicities—10 to start with. There may be other motifs that may be associated to nucleosome positioning. This statement comes from our observation that Trifonov's decamer is not found with the same frequencies along different regions of a given chromosome, different local regions within a gene, or GC-rich versus GC-poor segments, even when the DNA is indeed uniformly supercoiled. Actually, there are long stretches in which the generic decamer is absent.

For comparison purposes and for validation of the use of the MIF, here we also report the same analysis in the five chromosomes of *C. elegans*, for which experimental data are available and certain results are expected [41]. With our approach we found that this motif not only reflects well-known periodicities of the nucleosome positions but also there seems to be other previously unidentified periodicities both in primates and Archaea. We conclude that Trifonov's decamer is not the "one-and-only" universal nucleosome positioning motif. We give evidence that these periodicities are associated with highly repetitive sequences in primate genomes. In particular, we show that the Y/R motif is clearly associated to *Alu* repetitive elements in primate species.

## 2. Materials and Methods

*2.1. Data Sources.* Human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), and macaque (*Macaca mulatta*) complete genome sequences were downloaded from NCBI released, respectively, in March, October, and June of 2006 from ftp://ftp.ncbi.nih.gov/genomes/. In particular, the whole genomes of human, chimpanzee, and rhesus macaque were downloaded from: ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/; ftp://ftp.ncbi.nih.gov/genomes/Pan_troglodytes/, and ftp://ftp.ncbi.nih.gov/genomes/Macaca_mulatta/, respectively.

We selected the following Archaea which were also downloaded from the NCBI website: *Methanocaldococcus jannaschii, Sulfolobus solfataricus, Nanoarchaeum equitans, Archaeoglobus fulgidus,* with the corresponding accession numbers: NC_000909, NC_002754, NC_005213, NC_00917. The selected Eubacteria used as controls are: *Escherichia coli, Pseudomonas fluorescens*, and *Deinococcus radiodurans R1* with accession numbers: NC_000913, NC_004129, and NC_001263.1, respectively.

*2.2. An Overview of the Mutual Information Function.* Initially the mutual information (MI) was used to measure the difference between the average uncertainty in the input of an information channel before and after the outputs were received [44]. The MI is a general measure of correlation between discrete variables, analogous to the Pearson product-moment correlation coefficient for continuous variables. For symbolic sequences, MI between two symbols separated by a distance $k$ is a function of $k$, called mutual information function (MIF) [45]. The MIF is particularly useful for analyzing correlation properties of symbolic sequences [45].

Let us denote by $A = \{a, t, g, c\}$ an alphabet and by $s = (\ldots, a_0, a_1, \ldots)$ an infinite string with $a_i \in A$, $i \in \mathbb{Z}$, where $\mathbb{Z}$ represents the set of all integer numbers and the values of $a_i$ can be repeated. The MIF of the string $s$ and an identical string shifted $k$ positions upstream is defined as

$$I(k, s) = \sum_{\alpha \in A} \sum_{\beta \in A} P_{\alpha,\beta}(k, s) \log_2 \left[ \frac{P_{\alpha,\beta}(k, s)}{P_\alpha(s)\, P_\beta(s)} \right], \quad (1)$$

where $P_{\alpha,\beta}(k, s)$ is the joint probability of having the symbol $\alpha$ followed $k$ sites away by the symbol $\beta$ on the string $s$ and $P_\alpha(s)$ and $P_\beta(s)$ are the marginal probabilities of finding $\alpha$ or $\beta$ in the string $s$. By choosing the logarithm in base 2, $I(k, s)$ is measured in bits. Both the joint probability and the marginal probabilities are estimated throughout the sequence as a global property. The function $I(k, s)$ can be interpreted as the average information over all positions that one can obtain about the actual value of a certain position in the string, given that one knows the actual value of the position $k$-characters away. The mutual information vanishes if, and only if, the events are statistically independent, that is, if all 16 joint probabilities $P_{\alpha,\beta}(k, s)$ factorize. Thus, the MIF is a function capable of detecting any deviation from statistical independence. It must be noted from (1) that $I(k, s) \geq 0$. Computing the MIF for a given sequence using different shifts of magnitude $k$ provides an autocorrelation profile.

*2.3. The MIF Profile.* In this work, we calculated for each given sequence $s$ the contribution made to $I(k, s)$ by the generic decamers YYYYYRRRRR and SWWWWWSSSSSW. For this purpose, we computed $I(k, s)$ of the sequence $s$ and then marked $s$ such that each occurrence of, say, the decamer YYYYYRRRRR appeared in upper case, thereby extending

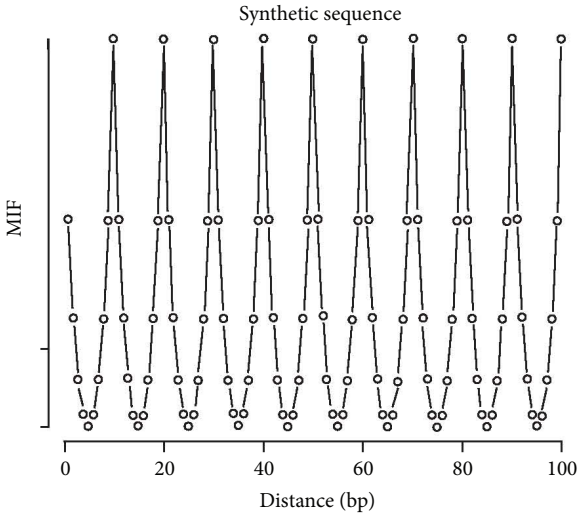


Figure 1: MIF profile of the decamer YYYYYRRRRR from a synthetic sequence. Note the 10-base periodicity.

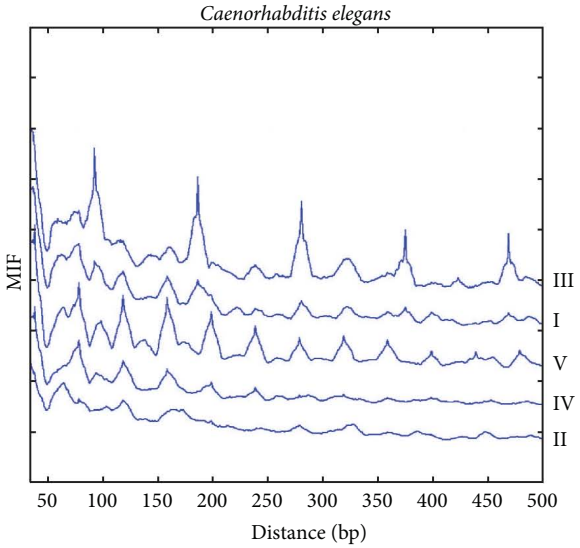


Figure 2: The MIF profiles of the decamer YYYYYRRRRR in 5 *C. elegans* chromosomes.

the alphabet to $A' = \{a, t, g, c, A, T, G, C\}$. If we call this marked sequence $s'$, then the difference $I(k, s') - I(k, s)$ is a measure of how much additional information the decamer YYYYYRRRRR contributes to our prediction of the content of a position in the sequence $k$ spaces away from a position whose information content is already known. This renders a brief description of how much of the correlations of a given chromosome are due specifically to the occurrences of the decamer YYYYYRRRRR. MIF, being similar to the autocorrelation function, is a method to detect periodicity in a sequence. A peak in MIF at spacing $k$ indicates that the decamer prefers a spacing of $k$ bases. In order to test our MIF profile, we generated a synthetic DNA sequence in which the decamer YYYYYRRRRR was placed at regular intervals (Figure 1). Note that the MIF profile clearly exhibits a 10-base periodicity.

For each chromosome the MIF was computed for $k$ between 1 and 500. Besides this excess mutual information between symbol and symbol (base and base $k$-position away), an alternative measure of the decamer-decamer correlation is to convert a DNA sequence to a binary (0/1) sequence: 1 for an appearance of YYYYYRRRRR, 0 otherwise. These two methods lead to equivalent results.

Since tandem repeats of YYYYYRRRRR leads to periodicities at $k = 10, 20, \ldots$, in our MIF and since regular spacing of nucleosomes (e.g., 146 bp plus a 45 linker length corresponds to a spacing of 191 bp) leads to periodicity at, for example, 191, 382, ..., any periodicities at short (<150) and intermediate (>150 and <400) distances in MIF may indirectly confirm the role of YYYYYRRRRR in nucleosome positioning.

This strategy is played out at several levels; we expect to see a periodic presence (absence) of peaks in the MIF profile for genomes known to possess nucleosomes (in those known to have no nucleosomes). We expect to see peaks at both short and intermediate distances. Finally, any observations in contrast to our expectation may lead to new insight; for example, the absence of peaks when expected may point to other nucleosome positioning motifs not included in YYYYYRRRRR; or presence of peaks at unexpected distances may point to other roles of the YYYYYRRRRR motif.

## 3. Results

*3.1. MIF Profiles of C. elegans.* Since the decamer was derived from the *C. elegans* MNase digestion data, we expect periodicities to be present in the MIF profile, either due to the tandem repeats of the decamer or due to the regular spacing of the nucleosomes. The MIF profiles of the decamer on chromosomes I, III, and V, but not on chromosome II or IV, of *C. elegans* display a regular pattern of peaks that appear every 10, 20, 40, and 92–94 bp approximately (Figure 2), and they correspond to distance histograms (not shown). This pattern is even met by chromosome X (not shown). The MIF profile of chromosome V shows regular spacings of multiples of 20 bp (e.g., at 120, 160, 200, 240, 320, 360 400, and 480). Given a decamer, we would have expected that bumps (a lumped region like the top of a mountain different from an acute peak) would have a length of 10 bp but what we observed in both primates and *C. elegans* is that the larger the bumps the more repetitions of the decamer in those regions. The different patterns of the MIF profile observed in *C. elegans* imply that nucleosomes do not favor a universal structure even among chromosomes of the same species.

*3.2. MIF Profiles of Homo Sapiens.* In Figure 3, the MIF profiles of the decamer YYYYYRRRRR, for each human chromosome, are illustrated. These profiles, equivalent to correlation functions, correspond to the distribution of spacings between the generic decamer suggested to be associated to the nucleosome positioning. In general, they show rugged landscapes with several troughs (these spacings are avoided) and peaks (these spacings are preferred). The MIF profiles

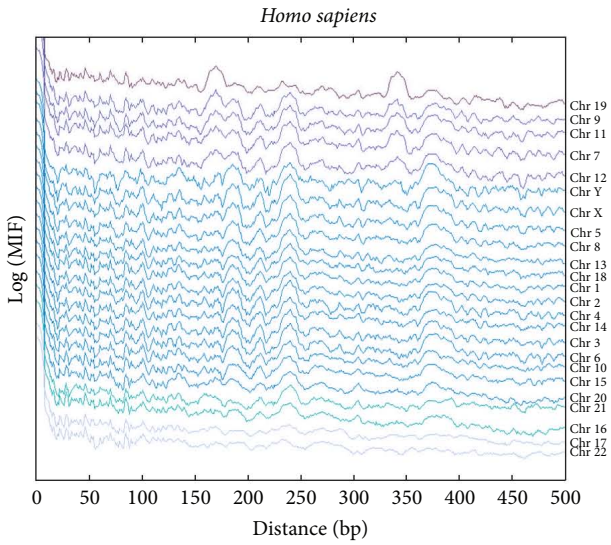


FIGURE 3: The MIF profiles of the decamer YYYYYRRRRR in all *Homo sapiens* chromosomes.

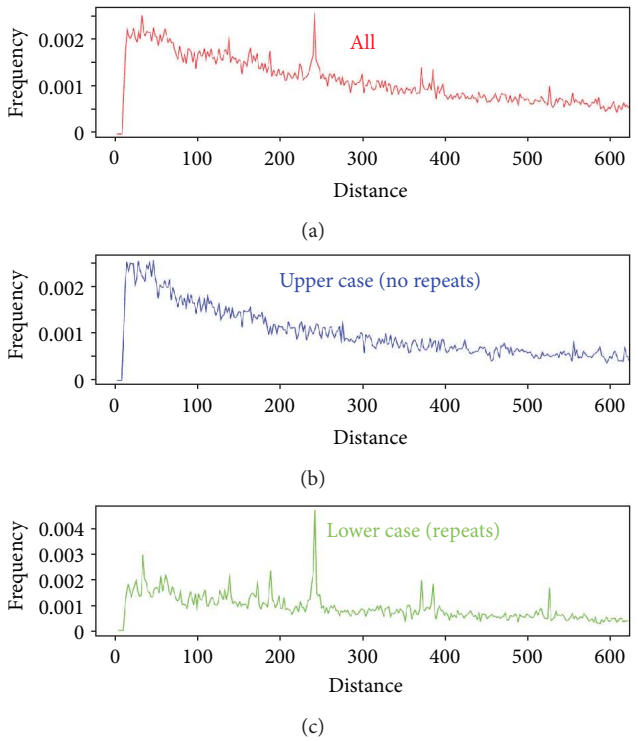


FIGURE 4: (a) Histogram of the spacing of the decamer YYYYYRRRRR in chromosome 21 of *Homo sapiens* (b) Histogram of the intact chromosome without repetitive elements; (c) Histogram of repetitive elements only.

of the decamers on the chromosomes were ordered in order to determine how similar (or different) they are among them. Each MIF profile was shifted upwards by successive integer multiples of 0.5 to facilitate visual inspection. At first glance, there seems to be 3 classes of profiles: class 1 comprises chromosomes 1 to 21 (except 17 and 19) and the sex chromosomes X and Y; class 2 includes chromosomes 17 and 22; and class 3 is represented by chromosome 19. Class 1 can still be subdivided into class 1a (chromosomes 1 to 21 excluding class 1b) and 1b (chromosomes 7, 9, 11, and 12), where the latter displays a bump at around 340 bp and two bumps in the range of 150 to 200 bp. It is widely recognized that the nucleosome has peaks at 80, 146, 165–167, and at around 240 bp [46].

A series of peaks up to −162 bp are clearly found in all chromosomes with the use of the MIF. At 10 bp, all chromosomes do not display a peak but they show a deviation in the falling trend. The observed periodicities occur at 31, 47, 62, 72, 84, 103, 110, 132, 136, and 162; bumps occur in regions 180–195, 225–255, and 365–395; long-range periodicities are found at 212, 240, 306, and 345.

Most chromosomes display a small peak at 165–167 bp, or 190 bp or 218 bp, which may reflect the periodic spacing between nucleosomes. With the exception of chromosomes 17, 19, and 22, all show a bump at around 240 bp due to repetitive elements as we will shortly illustrate (Figure 4). In addition to these peaks or hills, there are others like the ones found at 345 and 380, which might be considered as the spacing between next-nearest-neighbor nucleosomes.

This pattern from MIF is consistent with a direct measurement of histograms of the frequency distribution of spacing between the decamer along each chromosome except for the 10-base periodicity. The periodicities observed in the histogram occur at 10, 16, 20, 42, 55, 79, 93, 127, 146, 161, 178, 215, 230, 268, 287, 330, 360, 378, and 472 (not shown). Note that there is a great density of decamers at distances less than 500 bp, and at the same time there are specific peaks directly related to the more conspicuous ones of the MIF profile differences at these distances (Figure 3).

As close to 50% of the human genome consists of repetitive sequences, and we examine its contributions to the peaks seen in Figure 3. Figure 4 shows the histogram of spacing between the nearest decamer motifs for human chromosome 21 (Figure 4(a)), after masking all repetitive elements (Figure 4(b)), and finally, after masking all the sequence except repetitive elements (Figure 4(c)). A similar behavior is also seen in other human chromosomes (results not shown). Most peaks of intact chromosomes appear also in the histogram of repetitive sequences only. From Figure 4, it can be seen that the histogram of spacing of decamer YYYYYRRRRR for *H. sapiens* chromosome 21 shows peaks that appear in both the whole genome (Figure 4(a)) and in the only repetitive sequences (Figure 4(c)). In particular, this is true for the 240-241 peak. This means that in repetitive sequences there is a great deal of consecutive occurrences of the YYYYYRRRRR decamer spaced 240-241 bp apart. The biological meaning of this observation is still unknown. Due to the large proportion of repetitive sequences in the human genome, its potential function cannot be ignored. Our findings, as well as those in [46, 47], point to a potential role of repetitive elements in the nucleosome positioning. In Supplementary Information S1 available online at http://dx.doi.org/10.1155/2013/963956, we show a table of spacings between YYYYYRRRRR found at highly repetitive sequences in the human genomes.

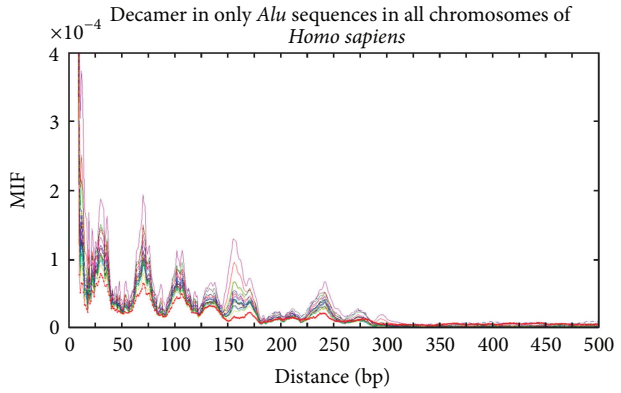


Figure 5: MIF profiles of the R/Y decamer in only *Alu* sequences in all *Homo sapiens* chromosomes.

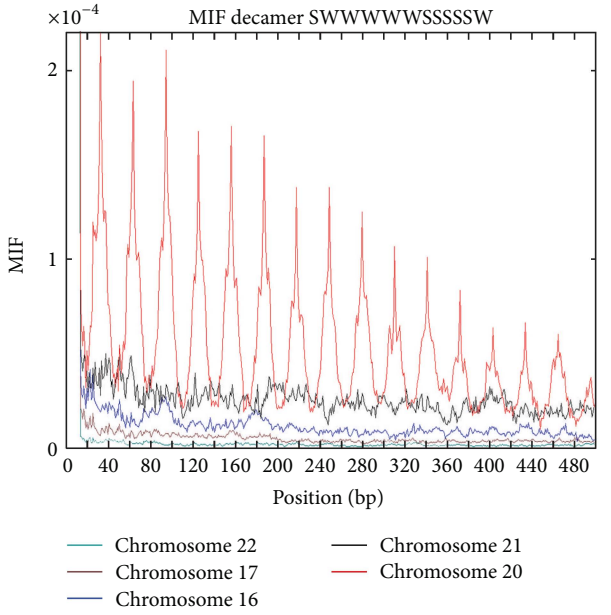


Figure 6: The MIF profiles of the 12-mer SWWWWWSSSSSW in some *Homo sapiens* chromosomes.

A possible relation between nucleosome positioning and one particular type of repetitive sequences, the *Alu* elements, has been suggested before [46]. It was observed that if one ignores the *Alu* repeats, several peaks in the Fourier spectra for AA/TT sequence (1 for AA or TT, 0 otherwise) disappear, but some peaks like the one found at about 165 bp still linger [46]. A similar observation was reported in [36]. Note that here we are analyzing a very different sequence (1 for YYYYYRRRRR, 0 otherwise), and repetitive sequences besides *Alu* are also masked. When only *Alu* sequences are considered, the MIF profiles of the decamer R/Y in all human chromosomes (Figure 5) display the same pattern in all chromosomes, indicating a strong association between the decamer Y/R (and R/Y) with *Alu* sequences. There are pronounced peaks at 32, 62, 110, 134, 160, and at 240 in all human chromosomes. There is a slight departure of this pattern in chromosome X (red curve).

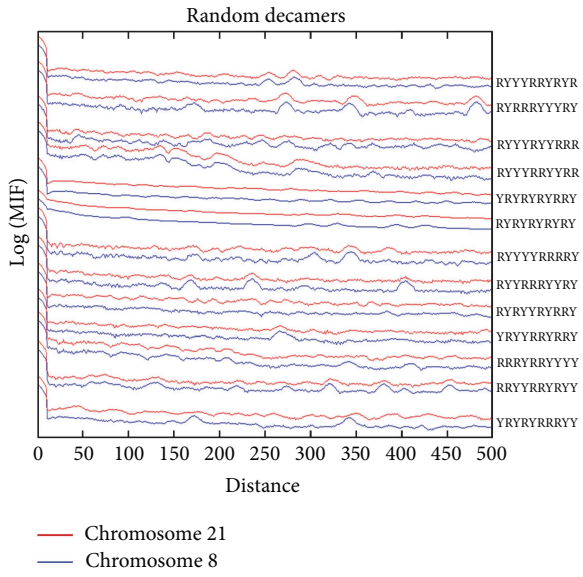


Figure 7: MIF profiles of random decamers with 5 purines and 5 pyrimidines along *Homo sapiens* chromosome 21, in order to compare the meaningful signal of YYYYYRRRRR as a binder nucleosome motif.

When the spacing of more specific decamers of the type YYYYYRRRRR (e.g., CGGAAATTTCCG) is analyzed, the periodic signal weakens considerably. The MIF profiles of the 12-mer SWWWWWSSSSSW in some chromosomes of *H. sapiens* are shown in Figure 6. Note that there is a regular behavior only in chromosome 20 in which there are peaks every 30 bp. A regular behavior is also observed in chromosome 12 whereas the remaining chromosomes exhibit a more irregular and nonuniform pattern.

If one selects at random a given decamer (preserving the number of Ys and Rs), not surprisingly, in most cases no prominent periodic signals are found as it is illustrated for the two chromosomes 21 and 8 of *H. sapiens* in Figure 7. The MIF profiles of the controls were not statistically similar among them (average correlation coefficient $r^2 = 0.56$) as the generic decamer in intact chromosomes do ($r^2 = 0.76$). The average correlation between the actual chromosome 8 with all random controls was $r^2 = 0.42$ whereas the average correlation between chromosome 21 with all random controls was $r^2 = 0.65$. Note that in the intact chromosomes we preserve the YYYYYRRRRR content and in the shuffled control we respect the nucleotide content but we disrupt the YYYYYRRRRR sequence. Therefore, the MIF profiles of the controls were not similar among chromosomes as the generic decamer do in intact DNA sequences.

*3.3. Nonhuman Primate MIF Profiles.* We also calculated the MIF profiles of the decamer on all available chromosomes of *Pan troglodytes* and *Macaca mulatta* (Figures 8 and 9). There is a consistency between MIF profiles of all chromosomes for each primate species, even though subtle differences exist. However, a more striking finding is that when two species are compared, a MIF profile for the decamer on a chromosome of a given species is more similar to that on the

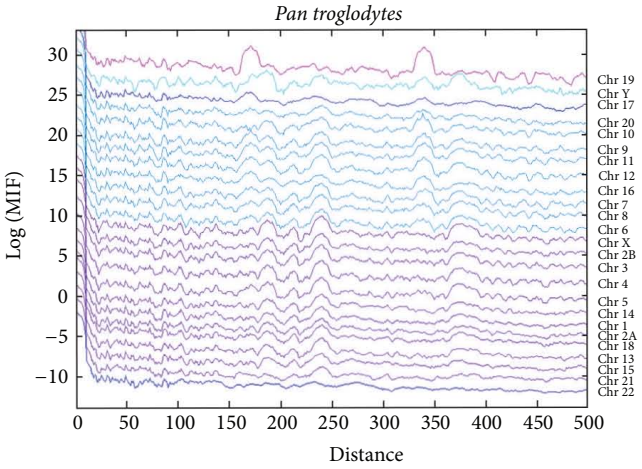


FIGURE 8: The MIF profiles of the decamer YYYYYRRRRR in all *Pan troglodytes* chromosomes.

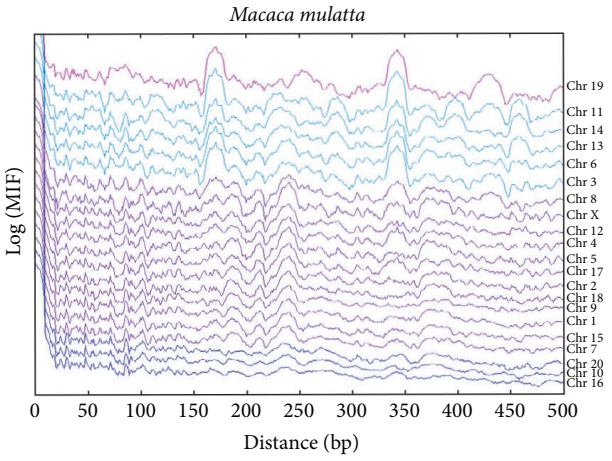


FIGURE 9: The MIF profiles of the decamer YYYYYRRRRR in all *Macaca mulatta* chromosomes.

same chromosome but in the other two species, than to the MIF profile on different chromosomes of the same species.

In general, it is clear that despite the different evolutionary histories of the 3 primate species, there is a common pattern in the MIF profiles of the generic decamer on a given chromosome.

For the same comparative purposes, the MIF profiles of the decamer on the chromosomes of *P. troglodytes* (Figure 8) can also be divided into the same three classes in which the *H. sapiens* chromosomes were divided. In the first class, there are chromosomes 1 to 21 and the sex chromosomes X and Y; in class 2, chromosomes 17 and 22 can even be subdivided given a conspicuous widening similar to a bump in the range of 150 to 175 that is present in chromosome 17. Class 3 is also represented by chromosome 19. But the first class can be subdivided into class 1a (chromosomes 1 to 21 excluding class 1b and 1c), class 1b with the same characterization that is, in human MIF profile (chromosomes 6 to 12, 16 and 22), and class 1c is represented by chromosome Y which has a different

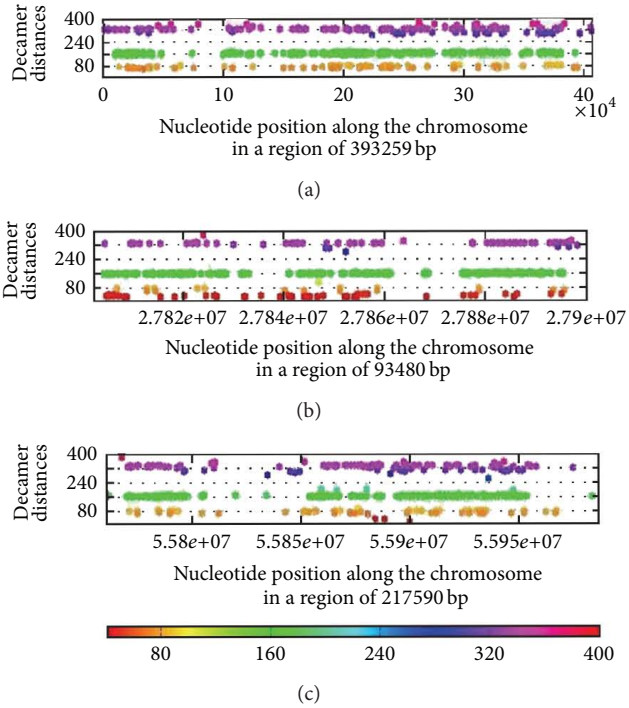


FIGURE 10: The three main regions of human chromosome 19, where distances around 80, 160, and 320 between the generic decamer are more highly concentrated than in the rest of the chromosomes are highlighted. Note that the clusters correspond to the peaks observed in their respective MIF profiles.

kind of bumps in the range of 64–84 bp and in the range of 164–200 bp (Figure 8).

The MIF profiles of the decamer on all chromosomes analyzed in *M. mulatta* (Figure 9) seem to pertain to only two classes. Class 1 can be subdivided by shorter amplitudes in the same bp signals between class 1a (chromosomes 10, 16, and 20) and class 1b (chromosomes 1, 2, 4, 5, 7 to 9, 12, 15, 17, and 18). With a similar profile than the two mentioned subclasses, class 1c presents highly conspicuous peaks at around 172 and 342 bp, and a bump in the range of 274 to 300 bp (chromosomes 3, 6, 11, 13, and 14). In class 1c, the chromosomes 11, 13 and 14 also have a peak at around 460 bp.

Class 2 is represented by chromosome 19 that, in contrast to chromosome 19 for *P. troglodytes* and *H. sapiens,* has a bump in the range of 400 to 450 bp which it shares only with chromosomes 3, 8, and 11, beside the features of its own profiles class (Figure 8). It is important to note that there are several common peaks among human, chimpanzee, and rhesus macaque at 31, 47, 62, 72, 84, 103, 110, 132, 136, and 162; even some bumps are shared among the three species at 180–195, 225–255, and 365–395 and some long-range periodicities at: 212, 240, 306, and 345.

It is important to mention that considering an alphabet of $A = \{A, T, G, C\}$, we calculated the MIF for the 3 species of primates masking all repeats (not shown) and all peaks disappear.

We estimated the similarities of the chromosomes within and between species based upon the cross-correlations of the

MIF profiles of the chromosomes for the 3 primate species. All Pearson's correlation coefficients within chromosomes of a given primate species display values which are in a rough agreement with the classes mentioned above (see correlations in S2). The Pearson correlation coefficients of chromosomes between the 3 primate species in general also support our visual inspection of the previous observations of heterogeneity between chromosomes within species, and uniformity among the same chromosome between species (see S2).

In general, it is clear that despite the subtle differences among chromosomes within species, there is a common pattern in the MIF profiles of the decamer. Given that this decamer is a consensus motif for nucleosome positioning sequence, the hypothesis that the statistical properties of the decamer can be translated to those of the nucleosome positioning can be put forward.

The distribution of the decamer YYYYYRRRRR along each chromosome is not uniform since there are regions in which clusters are crisply recognized whereas there are long stretches lacking this decamer (Figure 10). Since MIF and spacing histograms are averaged over all regions in a chromosome, Figures 3–10 do not show the heterogeneity information in regions deserted of this decamer. Therefore, other decamers or signals associated to the fine structure of the chromosomes cannot be ruled out.

To further examine the issue of heterogeneity, we show examples of physical maps of the location of the generic decamer along a given chromosome. In Figures 10 and 11, the location of the generic decamer along chromosome 19 of *H. sapiens* for different magnification scales is shown. The most striking observation is that the decamer positions are not random but they are not uniformly distributed along the chromosome either. The decamer distribution is clumped in certain regions but there are long stretches in which the decamer is plainly absent. For the remaining human chromosomes, nucleosomes are also more consistently positioned than expected by chance and many are organized in regularly spaced arrays that are enriched near active chromatin. Hence, nucleosome positions are also clearly influenced by DNA sequence. A striking example is an array of regularly spaced nucleosomes created by tandem repetition of sequences with strong nucleosome positioning properties across approximately 35,423 and 41,824 bp of chromosome 19 (Figures 10 and 11). Similar arrays can also be found in other chromosomes.

If we take a look at the distances between consecutive appearances of the decamer, there are regions of the chromosomes in which the decamer appear in a periodic manner. That is, there are two (or actually more) stretches of the chromosome which contain the same number of this decamer, but with the peculiarity that the first and second occurrences of the decamer are spaced by the same distance in both stretches, and so are the second and third, and so on (not shown). Another interesting feature is that there are arrangements of different distances in which the order of distances of the generic decamer can also be encountered in some downstream regions but exactly in reverse order

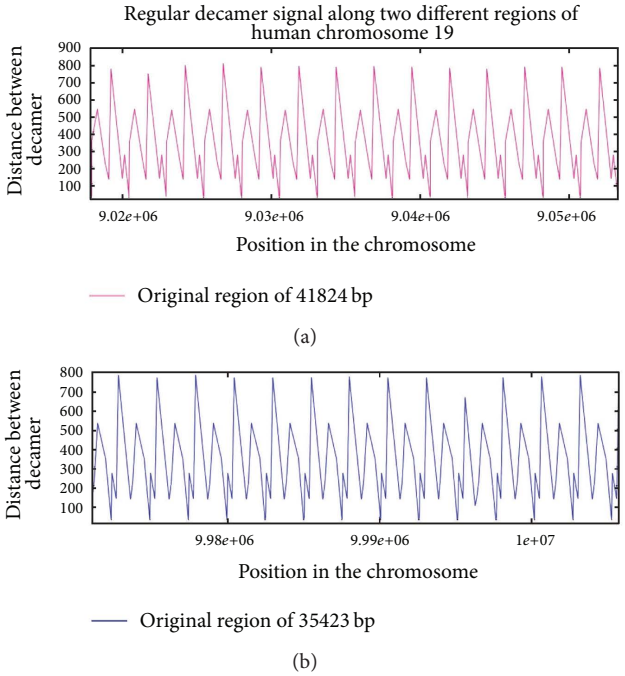


Figure 11: Plot of the distances between the generic decamer (ordinate) along two regions of chromosome 19 (abscissa) of *Homo sapiens*. Note the inverted repeat sequence.

of those distances. In other words, the distribution of the decamer exhibits an inverse symmetry (see Figure 11). It is worth mentioning that in this region there are genes of cadherin, beta-catenin and zinc fingers. This is consistent with a recent finding about rare roughly symmetrically positioned nucleosomes such as the zinc-finger containing protein that showed roughly symmetrically positioned nucleosomes [48].

*3.4. MIF Profiles of Archaea and Eubacteria Species.* We examine the MIF profiles of the decamer for the following Archaea species: *Methanocaldococcus jannaschii*, *Archaeoglobus fulgidus*, *Sulfolobus solfataricus*, and *Nanoarcheum equitans*.

In Figure 12, the MIF profiles of the decamer on several Archaea species are illustrated. It is remarkable to observe that this decamer still exhibits conspicuous periodicities. Similar to the MIF profiles observed in primates, in general, the MIF profiles of the decamer in archean species also manifest rugged landscapes with several troughs and peaks.

In *M. jannaschii,* there are several prominent peaks at around 67, 141, 210, and 408 bp whose magnitudes decrease with distance and they are interspersed throughout high-frequency oscillatory dynamics. The spacing of 141 apparently matches that of a nucleosome core sequence length and that of 67 close to half that length. The spacing of 210 could match the distance between two neighboring nucleosomes, and 375 for next-nearest-neighbor nucleosomes. As various linker sequence length may coexist in different regions in the genome, two nucleosome spacings (375/2 = 187.5 and 210) may not necessarily contradict each other.

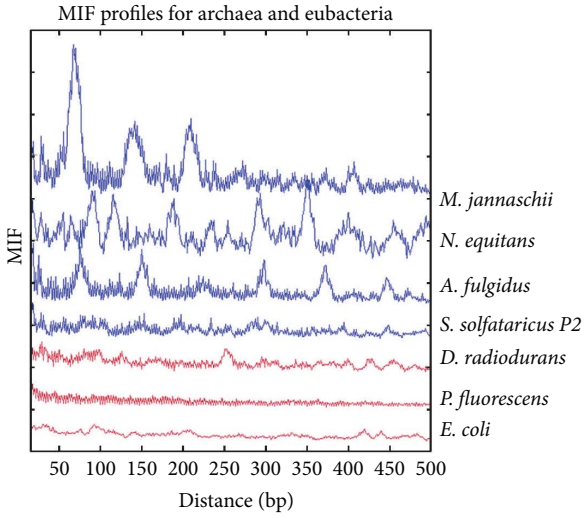


Figure 12: The MIF profiles of the decamer YYYYYRRRRR in some Archaea and Eubacterial genomes.

In *N. equitans,* there are several salient peaks at around 27, 89, 93, 115, 189, 234, 294, 352, 408, 456, and 496 bp with a great variability in their magnitudes, and they are embedded in a high-frequency oscillatory behavior.

Note that in *M. jannaschii* and *N. equitans* there are peaks at ~60 bp and ~85 bp as were found in pLITMUS28 and in *Methanothermus fervidus* [49]. Archaea nucleosomes resemble the structure formed by the $(H3 + H4)_2$ tetramer at the center of the eukaryotic nucleosome. Both structures have a histone tetramer core that recognizes positioning signals, directly contacts ~60 bp, and wraps ~85 bp of DNA alternatively in either a positive or negative toroidal supercoil [49].

In *A. fulgidus,* the MIF profile displays a pattern of high-frequency oscillatory structure that they themselves form jagged bumps, and there are salient peaks at around 75, 150, 300, 375, and 450.

In *S. solfataricus,* the MIF profile is essentially composed by high-frequency oscillatory structure from which jagged bumps are formed with no discernible prominent bumps.

In order to test whether the MIF profiles of Archaea are biologically meaningful, that is, the periodic appearance of the putative nucleosome positioning decamer is due to the repetitive motif within a nucleosome core and the regular spacing of nucleosomes, we also show the MIF of the decamer for several bacteria which are known to be lacking nucleosomes. Three of them (*Escherichia coli, Pseudomonas fluorescens,* and *Deinococcus radiodurans*) are shown in Figure 12. Note that in the corresponding MIF profiles of these three bacteria, the signal is so weak that we cannot ascribe, as expected, that there is a periodicity of the YYYYYRRRRR decamer along the genome. If the decamer is indeed associated with the nucleosome positioning sequence in any species, this is consistent with the absence of nucleosomes in bacterial genomes. We included these bacteria to test whether Archaea cells show evolutionary selection either for or against sequences that favor nucleosome formation. As bacteria do not possess histones, but do show 3 and 10-11 base periodicity due to coding regions, we presumed that *E. coli, P. fluorescence,* and *D. radiodurans* DNA sequences are evolutionarily neutral with respect to nucleosome formation, such that preferred nucleosome forming sequences will occur by chance. These results strongly argue that the Archaean genomes have evolved to favor nucleosome formation.

## 4. Discussion

In this work, we have found that the proposed nucleosome positioning motif YYYYYRRRRR exhibits expected periodicities in primates and Archaea, thus consistent with the hypothesis that it plays a role in nucleosome positioning. In particular, we placed emphasis on the effect of repetitive sequences on the observed periodicities of the motifs R/Y and Y/R, as well as the S/W motif. We succeeded in the detection of the periodical repetition of the DNA patterns in all chromosomes tested despite weak or previously undetected periodicities with other methods. The extraction of the periodical signals in all chromosomes was due to the fact of using both MIF profiles and the generic decamer R/Y and Y/R to document a comprehensive distribution of nucleosome DNA sequences in primate species and even perhaps in Archaea. The MIF profiles display peaks or bumps in places previously recognized, such as the typical signatures at 31-32, 84, 146, 157, 171 and 200 [25, 41, 46]. New periodicities such as 100, 167, 240, and 320 are reported here. We did find the 10-bp in the histograms (not shown) but not in the MIF profiles because it may be unlikely to detect it. The rationale is as follows: there are 10 million copies of YYYYYRRRRR/RRRRRYYYYY in the human genome and if we assume they do not overlap, this would lead to 90 million bases (when they do overlap, the number would still be smaller). For example, for a segment ...YYYYYRRRRRYYYYY... which contains 2 copies of the motif, it covers only 15 bases instead of covering 20 bases; 90 million bases represent 3% of the human genome (if overlap exists, could be 2%), but at least 20% of the human genome is well positioned with nucleosomes. Therefore, there are not enough 10 mers to cover densely within a nucleosome positioning region. This dense packing is what would lead to the periodicity of 10. On the other hand, we can have longer periodicities. Suppose we have this order: beginning-middle (dyad)-end-linker-beginning-next-nucleosome-.... Assume also that this motif tends to sit at the beginning of a nucleosome, then we do not need 20 copies per nucleosome to cover the whole region, only 1-2 copies per nucleosome at the beginning. This density is more consistent with our observations. Then, the regular spacing of nucleosomes would lead to longer (+200) periodicities, but not the 10-base periodicity within a nucleosome. When repetitive elements were masked in whole chromosomes it became evident that the decamer contributes not only to the presence of the nucleosome structure but it also manifests itself as part of highly repetitive sequences (see S1).

With more than one million copies, *Alu* elements are the most abundant repetitive elements in the human genome; they represent ~10% of the genome mass and belong to

the SINE (short interspersed elements) family of repetitive elements [50]. *Alu* elements emerged ~55 million years ago from a fusion of the 50 and 30 ends of a 7SL RNA gene, which encodes the RNA moiety of the signal recognition particle (SRP). Modern *Alu* elements are ~300 bp in length and are classified into subfamilies according to their relative ages [51]. Dimeric *Alu* elements are unique to primates. *Alu* RNAs, transcribed from *Alu* elements, are present in the cytosol of primate cells. *Alu* elements inherited the internal A and B boxes of the RNA polymerase III (Pol III) promoter from the 7SL RNA gene [52]. The typical *Alu* RNA is a dimer of related but nonequivalent arms that are joined by an A-rich linker and followed by a short poly(A) tail [52].

Not surprisingly, the MIF profiles of the shuffled decamers showed no discernible pattern and no rugged landscape. The MIF profiles of the controls were not similar among chromosomes as the generic decamer was. The MIF profiles of the generic decamer in the three primate species exhibited a uniformity between species for the same chromosome, but heterogeneity within species between different chromosomes. The observed regularity of the patterns allowed us to provide families for the distribution of the generic decamer tested.

We selected the three densest regions in which there were clearly clusters of the decamer which appeared every 80, 160, and 320 bp (multiples of 80) of human chromosome 19 (Figure 10). These clusters of the decamer clearly correspond to the peaks of the MIF profile of human chromosome 19 (Figure 3).

The finding of regular periodic patterns of the decamer along primate chromosomes visualized in distance series of long stretches of the different chromosomes, as well as the patterns reflecting inverted repeats (inverse symmetry), discards the possibility that the generic decamer is biologically meaningless. Periodicities naturally arise if the decamer is tandemly repeated, and/or if the nucleosomes are regularly spaced. Inverted symmetry can be caused by the central role of dyad in the nucleosome cores. We think that the probability of finding such arrangements just by chance would be very low. The patterns of the MIF profiles of the five chromosomes of *C. elegans* are not entirely consistent with the regular reported structure of their nucleosomes [41]. Therefore, most nucleosomes in primate genomes are consistently positioned, either because they are forced into positioned arrays by chromatin remodeling or DNA binding proteins, and/or because they adopt favored sequence positions in genomic regions without active binding. Interestingly enough, the MIF profiles of the generic decamer in all Archaea tested showed prominent peaks in an oscillatory background. We propose that this decamer deserves further studies in order to determine if it has been selected since the origin of nucleosome structure.

It has been noted that the RNA motif SRP9/14 binds primarily to the universally conserved core of the *Alu* RNA 59 domain, which forms a U-turn in the context of a tau-junction [53]. This RNA motif is highly conserved in the SRP RNAs from higher eukaryotes to yeast and from Archaea to some Gram-positive Eubacteria [54]. A dimeric *Alu* RNP complex might be important in the origin or propagation of tandemly arranged *Alu* retroposons, as retropositional success was clearly correlated with the emergence of dimeric *Alu* elements during primate evolution [55]. *Alu* elements play an important role in the regulation of gene expression at various levels, such as in alternative splicing when present in intronic regions of genes [56]. The observed MIF profiles from different chromosomes or different species often differ substantially. Therefore, all these patterns cannot be attributable to the origin of nucleosome structures, or nucleosomes sequence preferences. It is likely that many of the peak features may be ascribed to some species-specific or chromosome-specific DNA sequence features, such as *Alu* repeats, but not necessarily limited to them.

What then accounts for the phenotypic differences between nonhuman primates and humans? It stands to reason to propose that part of the difference might be because of species-specific alternative splicing.

We were able to characterize different classes of MIF profiles within each primate species. The outstanding observation is that the MIF profile of a given chromosome is more similar to the corresponding profile among species than within species. The observed peaks of the MIF profiles using the generic decamer in primates are strongly associated with several highly repetitive sequences. This is in agreement with the recent discovery that the positioning of neighboring nucleosomes seems to be in phase with *Alu* elements as reflected by peaks in the Fourier analysis at 84-bp and 167-bp [46]. In this work, we corroborate this result with both Fourier (not shown) and MIF analyses using the decamer.

We have also found that human repetitive sequence densities are mostly negatively correlated with R/Y-based nucleosome-positioning motifs (NPM) and positively correlated with W/S-based motifs [57]. The positive correlation between YYYYYRRRRR/RRRRRYYYYY and repetitive sequence density is intriguing, as it provides an exception to negative correlation between densities of repetitive sequences and that of R/Y-based NPMs. The scatter plot for YYYYYRRRRR/RRRRRYYYYY is particularly interesting; despite the negative trend followed by the majority of the points, there is a minority trend for high repetitive sequence densities and high NPM densities [57]. We believe that it is in this region in which the generic decamer can be found positively associated with *Alu* elements.

Herein, we focused on MIF profiles of the type R/Y and Y/R with several peaks that overlap with repetitive elements. Amongst the most prominent peaks for most chromosomes in primates are at 84, 100, 167, and 240. In fact, in certain chromosomal segments, a well-defined periodic pattern of the decamer within highly repetitive sequences was observed. Appearance of the CG dinucleotide in the nucleosome positioning pattern is rather surprising, considering its generally low occurrence in eukaryotic sequences. However, recent studies suggest that CG dinucleotides play a special role indeed [36]. First, it displays 10.4-base periodicity almost as often as the AA and TT dinucleotides do, in particular in G+C-rich regions [42, 58]. In the *Alu* sequences, the CG element appears at a distance of 31-32 bases from one another [59], suggesting involvement of the sequences in the nucleosomes. Methylation/demethylation of CpG would

modulate the nucleosome stability, so that the CG-containing nucleosome could be considered as "epigenetic nucleosomes" [59]. Most chromosomes, except 19, 22, X, and Y, show a notorious similarity in regard to the putative positioning of the nucleosomes as obtained with our approach. Even among species within the primate family, the latter still holds. Hence, the conserved MIF profile on primates can reflect the importance of these generic decamer into the architecture of primate genomes. In addition, the conserved and peculiar organization of islands into repetitive elements may allow us to consider that this specific decamer could be implicated in the self-regulation functions inherent in these types of sequences.

The finding of peaks in Archaea and its absence in Bacteria may not be a surprising result since it is known that the former contain histones whereas the latter do not. But it is noteworthy that we have detected for the first time via the MIF profiles putative nucleosome signals in Archaea. In addition, there is a prominent presence of the generic decamer in Archaea as it is shown in their corresponding histograms (not shown). To our knowledge, this is the first description of this generic decamer for the nucleosome in this group and it remains to prove that it may be considered a nucleosome without the subsequent evolutionary refinements conferred by the repetitive elements. Hence, repetitive elements turn out to be basic ingredients of the most fundamental structure of nucleosome positioning in higher Eukaryotes.

In summary, putative nucleosome positioning motifs (NPM) associated to repetitive elements in human, nonhuman primates, and Archaea have been identified by means of mutual information profiles (MIF). Trifonov's group suggested a most recent "finale motif" of the long-searched "chromatin code." The biological significance of this decamer motif and its two degenerate parental motifs is examined in primates and Archaea. Common features in the patterns of the generic decamer R/Y on MIF profiles among primate species are found. The distribution of R/Y motif exhibits previously unidentified periodicities, which are associated to highly repetitive sequences in the genome. *Alu* repetitive elements may contribute to the most fundamental structure of nucleosome positioning in higher Eukaryotes. In some regions of primate chromosomes, the distribution of the R/Y decamer shows symmetrical patterns including inverted repeats. We have detected for the first time via the MIF profiles putative nucleosome signals in Archaea. It is clear that the R/Y motif is relevant in the NPM but it is also certain that there must be other relevant motifs besides the Trifonov "finale." Our findings may contribute to the understanding of the origin of nucleosome structures in Archaea and its remarkable success of *Alu* retroposons in colonizing primate genomes.

## Acknowledgments

Marco V. José was financially supported by PAPIIT UNAM, Project IN107112. They thank the Posgrado en Ciencias Biológicas UNAM and the Centro de Ciencias de la Complejidad, UNAM for the server computer support. Tzipe Govezensky offered assistance with S1.

## References

[1] G. Felsenfeld and M. Groudine, "Controlling the double helix," *Nature*, vol. 421, no. 6921, pp. 448–453, 2003.

[2] A. Valouev, S. M. Johnson, S. D. Boyd, C. L. Smith, A. Z. Fire, and A. Sidow, "Determinants of nucleosome organization in primary human cells," *Nature*, vol. 474, no. 7352, pp. 516–522, 2011.

[3] D. E. Sterner and S. L. Berger, "Acetylation of histones and transcription-related factors," *Microbiology and Molecular Biology Reviews*, vol. 64, no. 2, pp. 435–459, 2000.

[4] K. Luger, T. J. Rechsteiner, A. J. Flaus, M. M. Y. Waye, and T. J. Richmond, "Characterization of nucleosome core particles containing histone proteins made in bacteria," *Journal of Molecular Biology*, vol. 272, no. 3, pp. 301–311, 1997.

[5] C. A. Davey, D. F. Sargent, K. Luger, A. W. Maeder, and T. J. Richmond, "Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution," *Journal of Molecular Biology*, vol. 319, no. 5, pp. 1097–1113, 2002.

[6] E. I. Campos and D. Reinberg, "Histones: annotating chromatin," *Annual Review of Genetics*, vol. 43, pp. 559–599, 2009.

[7] R. L. Redner, J. Wang, and J. M. Liu, "Chromatin remodeling and leukemia: new therapeutic paradigms," *Blood*, vol. 94, no. 2, pp. 417–428, 1999.

[8] T. J. Richmond and C. A. Davey, "The structure of DNA in the nucleosome core," *Nature*, vol. 423, no. 6936, pp. 145–150, 2003.

[9] K. Sandman, J. A. Krzycki, B. Dobrinski, B. Lurz, and J. N. Reeve, "HMf, a DNA-binding protein isolated from the hyperthermophilic archaeon *Methanothermus fervidus*, is most closely related to histones," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 87, no. 15, pp. 5788–5791, 1990.

[10] R. L. Fahrner, D. Cascio, J. A. Lake, and A. Slesarev, "An ancestral nuclear protein assembly: crystal structure of the *Methanopyrus kandleri* histone," *Protein Science*, vol. 10, no. 10, pp. 2002–2007, 2001.

[11] I. Ioshikhes, A. Bolshoy, K. Derenshteyn, M. Borodovsky, and E. N. Trifonov, "Nucleosome RNA sequence pattern revealed by multiple alignment of experimentally mapped sequences," *Journal of Molecular Biology*, vol. 262, no. 2, pp. 129–139, 1996.

[12] S. C. Satchwell, H. R. Drew, and A. A. Travers, "Sequence periodicities in chicken nucleosome core DNA," *Journal of Molecular Biology*, vol. 191, no. 4, pp. 659–675, 1986.

[13] A. Valouev, J. Ichikawa, T. Tonthat et al., "A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning," *Genome Research*, vol. 18, no. 7, pp. 1051–1063, 2008.

[14] A. B. Cohanim, Y. Kashi, and E. N. Trifonov, "Yeast nucleosome DNA pattern: deconvolution from genome sequences of *S. cerevisiae*," *Journal of Biomolecular Structure and Dynamics*, vol. 22, no. 6, pp. 687–693, 2005.

[15] I. Ioshikhes, S. Hosid, and B. F. Pugh, "Variety of genomic DNA patterns for nucleosome positioning," *Genome Research*, vol. 21, no. 11, pp. 1863–1871, 2011.

[16] E. Segal, Y. Fondufe-Mittendorf, L. Chen et al., "A genomic code for nucleosome positioning," *Nature*, vol. 442, no. 7104, pp. 772–778, 2006.

[17] Y. Zhang, Z. Moqtaderi, B. P. Rattner et al., "Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions *in vivo*," *Nature Structural and Molecular Biology*, vol. 16, no. 8, pp. 847–852, 2009.

[18] Y. Zhang, Z. Moqtaderi, B. P. Rattner et al., "Evidence against a genomic code for nucleosome positioning," *Nature Structural and Molecular Biology*, vol. 17, no. 8, pp. 920–923, 2010.

[19] N. Kaplan, I. Moore, Y. Fondufe-Mittendorf et al., "Nucleosome sequence preferences influence *in vivo* nucleosome organization," *Nature Structural and Molecular Biology*, vol. 17, no. 8, pp. 918–920, 2010.

[20] A. Travers, "The nature of DNA sequence preferences for nucleosome positioning. Comment on 'Cracking the chromatin code: precise rule of nucleosome positioning' by Trifonov," *Physics of Life Reviews*, vol. 8, no. 1, pp. 53–55, 2011.

[21] E. N. Trifonov, "Cracking the chromatin code: precise rule of nucleosome positioning," *Physics of Life Reviews*, vol. 8, no. 1, pp. 39–50, 2011.

[22] K. Sha, S. G. Gu, L. C. Pantalena-Filho et al., "Distributed probing of chromatin structure *in vivo* reveals pervasive chromatin accessibility for expressed and non-expressed genes during tissue differentiation in *C. elegans*," *BMC Genomics*, vol. 11, no. 1, article 465, 2010.

[23] M. Yaniv and S. C. Elgin, "Chromosomes and expression mechanisms: bringing together the roles of DNA, RNA and proteins," *Current Opinion in Genetics and Development*, vol. 18, no. 2, pp. 107–108, 2008.

[24] L. E. Gracey, Z. Chen, J. M. Maniar et al., "An *in vitro*-identified high-affinity nucleosome-positioning signal is capable of transiently positioning a nucleosome *in vivo*," *Epigenetics and Chromatin*, vol. 3, no. 1, article 13, 2010.

[25] S. Sasaki, C. C. Mello, A. Shimada et al., "Chromatin-associated periodicity in genetic variation downstream of transcriptional start sites," *Science*, vol. 323, no. 5912, pp. 401–404, 2009.

[26] D. Tillo, N. Kaplan, I. K. Moore et al., "High nucleosome occupancy is encoded at human regulatory sequences," *PLoS ONE*, vol. 5, no. 2, Article ID e9129, 2010.

[27] I. Ioshikhes, E. N. Trifonov, and M. Q. Zhang, "Periodical distribution of transcription factor sites in promoter regions and connection with chromatin structure," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 6, pp. 2891–2895, 1999.

[28] R. Sadeh and C. D. Allis, "Genome-wide "re"-modeling of nucleosome positions," *Cell*, vol. 147, pp. 263–266, 2011.

[29] S. Henikoff, "Nucleosome destabilization in the epigenetic regulation of gene expression," *Nature Reviews Genetics*, vol. 9, no. 1, pp. 15–26, 2008.

[30] F. Moreno-Herrero, R. Seidel, S. M. Johnson, A. Fire, and N. H. Dekker, "Structural analysis of hyperperiodic DNA from *Caenorhabditis elegans*," *Nucleic Acids Research*, vol. 34, no. 10, pp. 3057–3066, 2006.

[31] H. R. Widlund, H. Cao, S. Simonsson et al., "Identification and characterization of genomic nucleosome-positioning sequences," *Journal of Molecular Biology*, vol. 267, no. 4, pp. 807–817, 1997.

[32] P. Oudet, M. Gross Bellard, and P. Chambon, "Electron microscopic and biochemical evidence that chromatin structure is a repeating unit," *Cell*, vol. 4, no. 4, pp. 281–300, 1975.

[33] W. Linxweiler and W. Hörz, "Reconstitution of mononucleosomes: characterization of distinct particles that differ in the position of the histone core," *Nucleic Acids Research*, vol. 12, no. 24, pp. 9395–9413, 1984.

[34] H. R. Drew and C. R. Calladine, "Sequence-specific positioning of core histones on an 860 base-pair DNA. Experiment and theory," *Journal of Molecular Biology*, vol. 195, no. 1, pp. 143–173, 1987.

[35] N. Kaplan, I. K. Moore, Y. Fondufe-Mittendorf et al., "The DNA-encoded nucleosome organization of a eukaryotic genome," *Nature*, vol. 458, no. 7236, pp. 362–366, 2009.

[36] E. N. Trifonov, "Nucleosome positioning by sequence, state of the art and apparent finale," *Journal of Biomolecular Structure and Dynamics*, vol. 27, no. 6, pp. 741–746, 2010.

[37] I. Gabdank, D. Barash, and E. N. Trifonov, "Open access article nucleosome DNA bendability matrix (*C. elegans*)," *Journal of Biomolecular Structure and Dynamics*, vol. 26, no. 4, pp. 403–412, 2009.

[38] S. M. Johnson, F. J. Tan, H. L. McCullough, D. P. Riordan, and A. Z. Fire, "Flexibility and constraint in the nucleosome core landscape of *Caenorhabditis elegans* chromatin," *Genome Research*, vol. 16, no. 12, pp. 1505–1516, 2006.

[39] S. G. Gu and A. Fire, "Partitioning the *C. elegans* genome by nucleosome modification, occupancy, and positioning," *Chromosoma*, vol. 119, no. 1, pp. 73–87, 2010.

[40] F. Salih, B. Salih, S. Kogan, and E. N. Trifonov, "Epigenetic nucleosomes: *Alu* sequences and CG as nucleosome positioning element," *Journal of Biomolecular Structure and Dynamics*, vol. 26, no. 1, pp. 9–15, 2008.

[41] F. Salih, B. Salih, and E. N. Trifonov, "Sequence structure of hidden 10.4-base repeat in the nucleosomes of *C. elegans*," *Journal of Biomolecular Structure and Dynamics*, vol. 26, no. 3, pp. 273–281, 2008.

[42] T. Bettecken and E. N. Trifonov, "Repertoires of the nucleosome-positioning dinucleotides," *PLoS ONE*, vol. 4, no. 11, Article ID e7654, 2009.

[43] S. L. Pereira and J. N. Reeve, "Histones and nucleosomes in Archaea and Eukarya: a comparative analysis," *Extremophiles*, vol. 2, no. 3, pp. 141–148, 1998.

[44] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.

[45] W. Li, "Mutual information functions versus correlation functions," *Journal of Statistical Physics*, vol. 60, no. 5-6, pp. 823–837, 1990.

[46] Y. Tanaka, R. Yamashita, Y. Suzuki, and K. Nakai, "Effects of *Alu* elements on global nucleosome positioning in the human genome," *BMC Genomics*, vol. 11, no. 1, article 309, 2010.

[47] D. Holste, I. Grosse, S. Beirer, P. Schieg, and H. Herzel, "Repeats and correlations in human DNA sequences," *Physical Review E*, vol. 67, no. 6, Article ID 061913, pp. 1–7, 2003.

[48] A. Kundaje, S. Kyriazopoulou-Panagiotopoulou, M. Libbrecht et al., "Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements," *Genome Research*, vol. 22, pp. 1735–1747, 2012.

[49] K. A. Bailey, F. Marc, K. Sandman, and J. N. Reeve, "Both DNA and histone fold sequences contribute to archaeal nucleosome stability," *The Journal of Biological Chemistry*, vol. 277, no. 11, pp. 9293–9301, 2002.

[50] E. S. Lander, L. M. Linton, and B. Birren, "International human genome sequencing consortium (2001) Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921.

[51] M. A. Batzer and P. L. Deininger, "Alu repeats and human genomic diversity," *Nature Reviews Genetics*, vol. 3, no. 5, pp. 370–379, 2002.

[52] J. Häsler and K. Strub, "*Alu* RNP and *Alu* RNA regulate translation initiation *in vitro*," *Nucleic Acids Research*, vol. 34, no. 8, pp. 2374–2385, 2006.

[53] O. Weichenrieder, K. Wild, K. Strub, and S. Cusack, "Structure and assembly of the *Alu* domain of the mammalian signal recognition particle," *Nature*, vol. 408, no. 6809, pp. 167–173, 2000.

[54] K. Strub, J. Moss, and P. Walter, "Binding sites of the 9- and 14-kilodalton heterodimeric protein subunit of the signal recognition particle (SRP) are contained exclusively in the Alu domain of SRP RNA and contain a sequence motif that is conserved in evolution," *Molecular and Cellular Biology*, vol. 11, no. 8, pp. 3949–3959, 1991.

[55] A. J. Mighell, A. F. Markham, and P. A. Robinson, "Alu sequences," *FEBS Letters*, vol. 417, no. 1, pp. 1–5, 1997.

[56] J. Kreahling and B. R. Graveley, "The origins and implications of *Aluternative* splicing," *Trends in Genetics*, vol. 20, no. 1, pp. 1–4, 2004.

[57] W. Li, D. Sosa, and M. V. José, "Human repetitive sequence densities are mostly negatively correlated with R/Y-based nucleosome-positioning motifs and positively correlated with W/S-based motifs," *Genomics*, vol. 101, no. 2, pp. 125–133, 2013.

[58] T. Bettecken, Z. M. Frenkel, and E. N. Trifonov, "Human nucleosomes: special role of CG dinucleotides and *Alu*-nucleosomes," *BMC Genomics*, vol. 12, article 273, 2011.

[59] M. S. Ong, T. J. Richmond, and C. A. Davey, "DNA stretching and extreme kinking in the nucleosome core," *Journal of Molecular Biology*, vol. 368, no. 4, pp. 1067–1074, 2007.

Genomics 101 (2013) 125–133

Contents lists available at SciVerse ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

# Human repetitive sequence densities are mostly negatively correlated with R/Y-based nucleosome-positioning motifs and positively correlated with W/S-based motifs

Wentian Li [a,*], Daniela Sosa [b,c], Marco V. Jose [d]

[a] The Robert S. Boas Center for Genomics and Human Genetics, The Feinstein Institute for Medical Research, North Shore LIJ Health System, Manhasset, 350 Community Drive, NY 11030, USA
[b] Facultad de Ciencias, Universidad Nacional Autónoma de México, México 04510 DF, Mexico
[c] Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, México 04510 DF, Mexico
[d] Theoretical Biology Group, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, Apdo Postal 70228, México 04510 DF, Mexico

ARTICLE INFO

*Article history:*
Received 9 July 2012
Accepted 29 October 2012
Available online 5 November 2012

*Keywords:*
Nucleosome positioning
Repetitive sequences
DNA motifs
Wavelet transformation

ABSTRACT

We examined statistical correlations between the frequencies of seven proposed nucleosome positioning motifs and the densities of repetitive sequences in the human genome. For both parametric and non-parametric measures of statistical correlations there is a tendency for repetitive sequence density to be negatively correlated with the density of R/Y-based nucleosome positioning motifs, while being positively correlated with that of W/S-based motifs. These results largely hold even when motifs are examined only within repeat-filtered sequences. The RRRRRYYYYY motif and its 5-base shift YYYYYRRRRR, in particular, is over-represented in the human genome; and its negative correlation is consistently present at different regions and at different length scales. For some other nucleosome positioning motifs, the relationship with repeats can be regional or length scale dependent. Considering the importance of nucleosome formation in epigenetic regulations, these results may provide new insight to the evolution of repetitive sequences.



## 1. Introduction

After double helix, nucleosome provides the next level of physical structure for DNA molecules (the chromatin structure) that play an important role in gene regulation [1–3]. With the chromatin being accessible at the promoter region, sequence is well positioned with nucleosome downstream from the promoter [4,5]. It has been long recognized that some DNA segments have a higher affinity to the nucleosome core histones, perhaps due to their own intrinsic bending, than other segments [6]. This observation led to many proposals of the nucleosome positioning motifs (NPM) (other names are also used, such as nucleosome core sequence pattern, nucleosome positioning code, etc.) which presumably cause certain DNA sequences to be located in the nucleosome core (as versus linker), at specific positions with respect to the central "dyad" region of the two-round wrapping of DNA around histone octamer. These motifs only increase the nucleosome positioning probability, and do not necessarily dictate absolute presence of them (or absolute absence of others) in the nucleosome cores. To cite from ref. [7], "you can position all of the nucleosomes some of the time and some of the nucleosomes all the time, but you can't position all the nucleosomes all of the time".

A major focus of NPM is to examine what sequences are preferred in the major and in the minor groove. This would define a sequence pattern which spans 5 basepair positions. Two types of these spacing-of-5-base motifs were proposed. One is the R/Y-based (R for purine: A or G, Y for pyrimidine: C or T), carving two segments from the … YRNNNRYNNNYR… sequence [8] around the two grooves: YRNNNRY and RYNNNYR (N for any nucleotide base). In this paper, these two patterns are written as the motif [YR-3-RY, RY-3-YR]. The motif YR-3-RY reads: a YR dinucleotide followed by any three bases, then followed by a RY dinucleotide. Another motif is the W/S-based (W for weak: A or T, S for strong: C or G), written as [WW-3-SS, SS-3-WW] [9]. The WW-3-SS is actually a more general motif than the originally observed [AA,TT,TA]NNNGC [10], i.e., either AA, TT, or TA dinucleotide followed by any three bases, then followed by the GC dinucleotide.

One extension of the above two types of short motifs (5 bases spacing or 7-mer or heptamer) is by a tandem repeat of them, leading to a periodicity of ten. For example, a tandem repeat of the W/S-based motif would lead to [WW-8-WW, SS-8-SS]; these two motifs are out of phase by 5 bases. In fact, the [AA,TT]NNNNNNNN[AA,TT] pattern is a main result in ref. [9], though the peak-to-peak distance does not always stay at 10 bases. Trifonov and Sussman uncovered the periodicity of 10.5 bases for dinucleotides [AA,TT], [GG,CC], TA, and TG [11], with the first three belonging to the W/S-type.

The recent genome-scale sequencing of nucleosome core DNA has generated large amount of data and provided fertile ground for testing ideas on NPM [12–16]. In particular, Trifonov's group suggested

* Corresponding author at: The Robert S Boas Center for Genomics and Human Genetics, Feinstein Institute for Medical Research, North Shore LIJ Health System, 350 Community Drive, Manhasset, NY 11030, USA. Fax: +1 516 562 1153.
*E-mail addresses:* wtli2012@gmail.com, wli@nshs.edu (W. Li).

0888-7543/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ygeno.2012.10.005

**Table 1**
Correlation coefficients between NPM densities and repetitive sequence density for human chromosome 20. The densities are calculated from non-overlapping windows. Window sizes are doubled consecutively, starting from 1 kb to 2048 kb (2.048 Mb). The first column is the window size; columns 2-4 are the number of windows, Pearson correlation coefficients and the corresponding *p*-values for testing zero correlation, Spearman correlation coefficient and the corresponding *p*-value; The next three columns are similar for NPM densities calculated from the unique (repeat-filtered) sequence only. (A) [RY-3-YR, RY-3-YR]; (B) [WW-3-SS, SS-3-WW]; (C) WW-8-WW; and (D) [RRRRRYYYYY, YYYYYRRRRR].

| W size (kb) | All seq | | | Unique seq | | |
|---|---|---|---|---|---|---|
| | No. W | Pearson/pv | Spearman/pv | No. W | Pearson/pv | Spearman/pv |
| *A* | | | | | | |
| 2 | 29751 | −0.013/0.03 | −0.019/*E*−3 | 28686 | −0.25/0 | −0.15/0 |
| 4 | 14875 | −0.041/5*E*−7 | −0.045/5*E*−8 | 14692 | −0.23/2*E*−179 | −0.14/4*E*−63 |
| 8 | 7437 | −0.80/4*E*−12 | −0.084/4*E*−13 | 7427 | −0.20/1*E*−69 | −0.15/2*E*−36 |
| 16 | 3718 | −0.13/4*E*−16 | −0.13/2*E*−16 | 3716 | −0.19/1*E*−30 | −0.16/1*E*−21 |
| 32 | 1859 | −0.19/8*E*−17 | −0.18/2*E*−15 | 1859 | −0.21/3*E*−19 | −0.19/8*E*−17 |
| 64 | 929 | −0.27/2*E*−16 | −0.26/7*E*−16 | 929 | −0.26/9*E*−16 | −0.27/1*E*−16 |
| 128 | 464 | −0.33/*E*−13 | −0.33/2*E*−13 | 464 | −0.31/4*E*−12 | −0.34/7*E*−14 |
| 256 | 232 | −0.38/2*E*−9 | −0.40/*E*−10 | 232 | −0.36/2*E*−8 | −0.37/6*E*−9 |
| 512 | 116 | −0.48/7*E*−8 | −0.53/8*E*−10 | 116 | −0.45/3*E*−7 | −0.48/9*E*−8 |
| 1024 | 58 | −0.51/4*E*−5 | −0.60/*E*−6 | 58 | −0.51/5*E*−5 | −0.52/4*E*−5 |
| 2048 | 29 | −0.58/9*E*−4 | −0.75/6*E*−6 | 29 | −0.61/5*E*−4 | −0.70/4*E*−5 |
| *B* | | | | | | |
| 2 | 29751 | 0.057/0 | 0.049/2*E*−17 | 28686 | −0.29/0 | −0.18/5*E*−205 |
| 4 | 14875 | 0.081/0 | 0.060/3*E*−13 | 14692 | −0.25/7*E*−205 | −0.14/4*E*−64 |
| 8 | 7437 | 0.11/0 | 0.070/2*E*−9 | 7427 | −0.16/3*E*−46 | −0.10/2*E*−16 |
| 16 | 3718 | 0.16/0 | 0.091/3*E*−8 | 3716 | −0.043/8*E*−3 | −0.057/5*E*−4 |
| 32 | 1859 | 0.20/0 | 0.095/4*E*−5 | 1859 | 0.013/.6 | −0.052/0.02 |
| 64 | 929 | 0.25/4*E*−15 | 0.11/8*E*−4 | 929 | 0.12/2*E*−4 | −0.046/0.2 |
| 128 | 464 | 0.34/*E*−13 | 0.13/6*E*−3 | 464 | 0.20/1*E*−5 | −0.026/0.6 |
| 256 | 232 | 0.40/2*E*−10 | 0.12/.07 | 232 | 0.27/4*E*−5 | −0.040/0.5 |
| 512 | 116 | 0.45/2*E*−7 | 0.11/0.2 | 116 | 0.33/3*E*−4 | −0.093/0.3 |
| 1024 | 58 | 0.52/2*E*−5 | 0.12/.4 | 58 | 0.40/2*E*−3 | −0.068/0.6 |
| 2048 | 29 | 0.58/*E*−3 | 0.070/.7 | 29 | 0.44/0.02 | −0.25/0.2 |
| *C* | | | | | | |
| 2 | 29751 | 0.25/0 | 0.20/2*E*−258 | 28686 | 0.038/0 | 0.058/1*E*−22 |
| 4 | 14875 | 0.28/0 | 0.21/3*E*−142 | 14692 | 0.11/0 | 0.10/3*E*−37 |
| 8 | 7437 | 0.28/0 | 0.20/2*E*−67 | 7427 | 0.17/0 | 0.13/5*E*−30 |
| 16 | 3718 | 0.27/0 | 0.17/4*E*−25 | 3716 | 0.18/0 | 0.12/3*E*−13 |
| 32 | 1859 | 0.24/0 | 0.13/8*E*−9 | 1859 | 0.17/4*E*−14 | 0.10/2*E*−5 |
| 64 | 929 | 0.21/6*E*−11 | 0.089/.007 | 929 | 0.16/1*E*−6 | 0.069/0.04 |
| 128 | 464 | 0.21/6*E*−6 | 0.060/.2 | 464 | 0.18/6*E*−5 | 0.065/0.2 |
| 256 | 232 | 0.20/.002 | 0.012/0.9 | 232 | 0.19/5*E*−3 | 0.046/0.5 |
| 512 | 116 | 0.16/.08 | −0.10/.3 | 116 | 0.16/0.1 | −0.060/0.5 |
| 1024 | 58 | 0.17/.2 | −0.15/.3 | 58 | 0.16/0.2 | −0.081/0.5 |
| 2048 | 29 | 0.11/0.6 | −0.31/.1 | 29 | 0.085/0.6 | −0.25/0.2 |
| *D* | | | | | | |
| 2 | 29751 | −0.21/4*E*−284 | −0.21/3*E*−289 | 28686 | −0.13/3*E*−113 | −0.23/0 |
| 4 | 14875 | −0.23/2*E*−174 | −0.24/*E*−189 | 14692 | −0.13/2*E*−55 | −0.16/4*E*−83 |
| 8 | 7437 | −0.26/8*E*−115 | −0.28/2*E*−131 | 7427 | −0.10/4*E*−19 | −0.10/1*E*−19 |
| 16 | 3718 | −0.29/*E*−74 | −0.31/9*E*−86 | 3716 | −0.10/5*E*−9 | −0.072/1*E*−5 |
| 32 | 1859 | −0.35/*E*−54 | −0.36/*E*−58 | 1859 | −0.094/5*E*−5 | −0.077/9*E*−4 |
| 64 | 929 | −0.40/4*E*−36 | −0.40/5*E*−36 | 929 | −0.094/4*E*−3 | −0.061/0.06 |
| 128 | 464 | −0.46/3*E*−25 | −0.42/7*E*−21 | 464 | −0.10/0.03 | −0.051/0.3 |
| 256 | 232 | −0.50/9*E*−16 | −0.50/5*E*−16 | 232 | −0.10/0.1 | −0.041/0.5 |
| 512 | 116 | −0.58/*E*−11 | −0.56/4*E*−11 | 116 | −0.094/0.3 | 0.042/0.6 |
| 1024 | 58 | −0.63/*E*−7 | −0.62/4*E*−7 | 58 | −0.068/0.6 | 0.11/0.4 |
| 2048 | 29 | −0.85/.02 | −0.86/.02 | 29 | −0.055/0.8 | 0.15/0.4 |

GRAAATTTYC as a most recent "finale" of the long-searched "chromatin code" [17–19]. This decamer motif and its two degenerate parental motifs, RRRRRYYYYY and SSWWWWWWSS (also the derived ones from tandem repeat followed by shift) are all mergers of the R/Y-based and W/S-based spacing-of-5 motifs mentioned early.

Human genomes are full of repetitive sequences [20] which occupy at least 50% (e.g., [21]) of the genome (it is even suggested that they may occupy as much as 2/3 of the genome [22]). It is natural to ask whether a relationship exists, if any, between NPM and repetitive sequences [23]. In an ongoing work, we examine the effect of repetitive sequences on the observed periodicities of [RRRRRYYYYY, YYYYYRRRRR] (D. Sosa, P. Miramontes, W. Li, V. Mireles, J.R. Bobadilla, M.V. José, unpublished results). Here we analyze the statistical correlations between the density of NPMs and the density of repetitive sequence directly. Obviously, there are only three possible relationships between the two: negative correlation, positive correlation, and no correlation (or statistically insignificant correlations).

The main technical obstacle in answering the posed question is that composition/density of any sequence type/motif may depend on the length scale at which the density is calculated. In a simple form, even base composition may depend on window size such that a [G,C]-rich domain can contain [G,C]-poor subdomains [24]. We will deal with this problem by directly testing correlations at different length scales, as well as by a more systematic approach of wavelet transformation, particularly useful for capturing multiple scales at once. Due to the large number of calculations and tests, we will start by examining one human chromosome (chromosome 20) in more detail. Then these analyses will be extended to the whole genome.

## 2. Results

### 2.1. Human chromosome 20, [RY-3-YR, YR-3-RY] motif

We partition DNA sequence of chromosome 20 into 62,965 non-overlapping 1 kb windows. Windows with less than 90% sequencing rate are discarded, leaving 59,502 windows, or 94.5% of the original number. For each window, densities of various NPMs are calculated, as well as the density of repetitive sequences. These densities at the length scale of 1 kb are the basis for similar calculation at larger length scales. The first NMP we examined is [RY-3-YR, YR-3-RY] [8], whose density in chromosome 20 is 0.067 copies per base if overlapping motif is prohibited (0.10 if overlapping is allowed). Note that YR-3-RY is not only a 5-base shift in a RY-3-YR tandem repeat, but also a reverse complement pattern of RY-3-YR.

#### 2.1.1. YR-3-RY and RY-3-YR density is negatively correlated with the repetitive sequence density

We consider both heptamers YR-3-RY and RY-3-YR, so the definition of the motif is independent from which strand is used, and whether the 5-base move is from major to minor groove or from minor to major groove. At the 1 kb window level, the repetitive sequence density and R/Y-based heptamer density is not significantly correlated (Pearson correlation coefficient (cc) is −0.0016 with $p$-value 0.69, and non-parametric Pearson correlation coefficient −0.0090 with $p$-value 0.029).

At larger window sizes, however, it is increasingly clear that the two are negatively correlated, as summarized in Table 1(A) (left columns) (Note: the notation (e.g.) 5E-7 means $5\times10^{-7}$). We combine the two consecutive windows into one to move to the next length scale, from 1 kb window size to 2 kb, then to 4 kb, etc. The magnitude of the negative correlation, for both Pearson and Spearman correlation, gradually increases. Despite the loss of sample size (number of windows), the statistical significance still increases from $p$-value $\sim10^{-2}$ at 2 kb to $p$-value $\sim10^{-15}$ to $10^{-17}$ at 32 kb to 64 kb. Then for even larger window sizes, the significance is reduced as the number of samples is reduced, though the magnitude of negative correlation increases.

#### 2.1.2. Regional variation of the correlation

Fig. 1(A) shows the scatter plot of number of repetitive sequence bases per kb ($x$-axis) and number of copies of [YR-3-RY, RY-3-YR] motif per kb for chromosome 20. There are 232 points (a point is a 256 kb window) in Fig. 1(A). The points/windows in the left, middle, or right 1/3 of the sequence are labeled by green, blue, and red colors, respectively. Linear regression lines for all points and for the three groups of points are shown. Although all three groups show negative regression slopes with somewhat comparable significance, the third group spans a wider range of repetitive sequence densities (with more windows at low repetitive sequence densities). This indicates a spatial heterogeneity between different chromosome regions.

Both the sign of the correlation and its regional variation along the chromosome has been confirmed by an independent wavelet analysis (see Table S1 of the Supplementary material).

#### 2.1.3. Calculating [YR-3-RY, RY-3-YR] density in repeat-filtered sequence

To further understand the source of the negative correlation, we calculated the [YR-3-RY, RY-3-YR] motif density in the unique sequence,

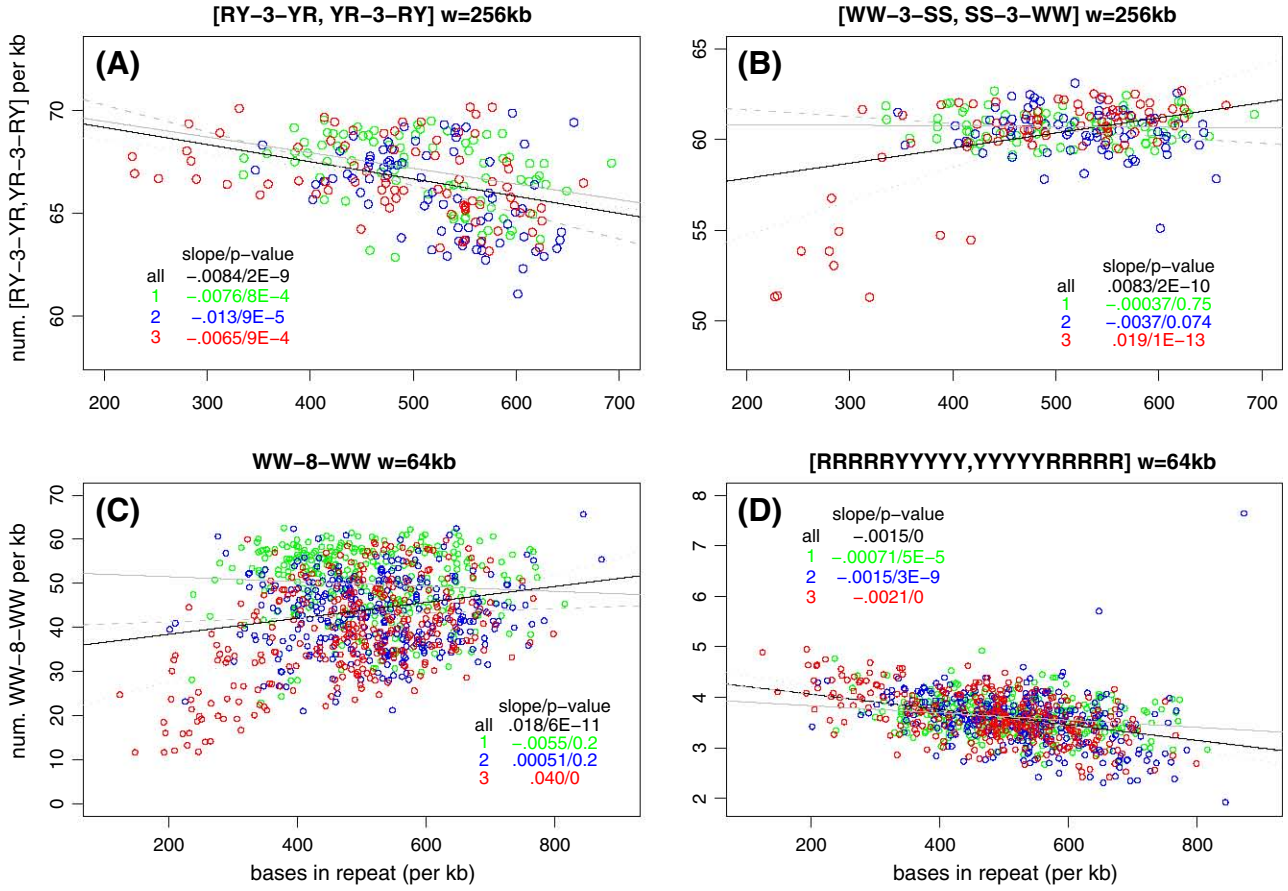


**Fig. 1.** Scatter plot of number of copies of NPMs in non-overlapping windows versus number of bases in repetitive sequence (in the same window) for human chromosome 20. The windows from the left, middle, and right 1/3 of the chromosome are labeled by green, blue, and red colors, respectively. Regression lines for all points, and for points from the three regions are shown. (A) [RY-3-YR, RY-3-YR], window size is 256 kb; (B) [WW-3-SS, SS-3-WW], window size is 256 kb; (C) WW-8-WW, window size is 64 kb; (D) [RRRRRYYYYY, YYYYYRRRRR], window size is 64 kb.

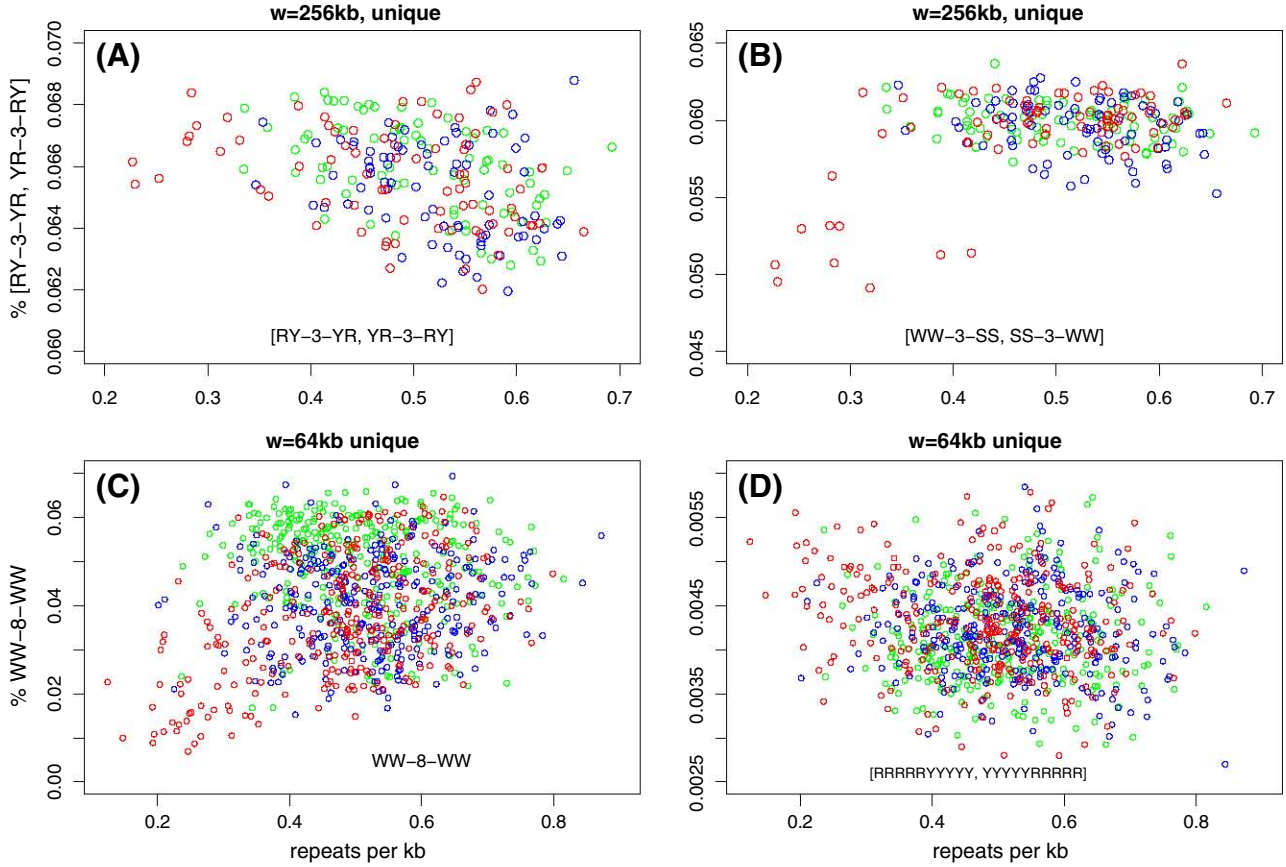


**Fig. 2.** Scatter plot of NPM densities within unique sequences (repeat-filtered sequences) in non-overlapping windows versus repetitive sequence density (in the same window) for human chromosome 20. The windows from the left, middle, and right 1/3 of the chromosome are labeled by green, blue, and red colors, respectively. (A) [RY-3-YR, RY-3-YR], window size is 256 kb; (B) [WW-3-SS, SS-3-WW], window size is 256 kb; (C) WW-8-WW, window size is 64 kb; (D) [RRRRRYYYYY, YYYYYRRRRR], window size is 64 kb.

i.e., in the sequence after the repetitive sequences are filtered/removed. Windows with 100% repetitive sequences are discarded.

Table 1(A) (right columns) shows that the correlation between NPM density and repetitive sequence density remains negative, with comparable magnitude of cc and *p*-values. The scatter plot in Fig. 2(A) shows this trend more directly at the 256 kb window size.

### 2.2. Human chromosome 20, [WW-3-SS, SS-3-WW] motif

The second NPM we examine is the [WW-3-SS, SS-3-WW] [9] whose density in chromosome 20 is 0.06 copies per base if the next motif is at least 7 base away from the current one, but 0.085 if overlapping is allowed. Note that SS-3-WW is not only a 5-base shift of the SS-3-WW tandem repeat, but also a reverse complement pattern of WW-3-SS.

#### 2.2.1. WW-3-SS and SS-3-WW density is positively correlated with the repetitive sequence density

Direct calculation of correlation coefficient, both Pearson's and non-parametric Spearman's, at different window sizes, shows that [WW-3-SS, SS-3-WW] density is positively correlated with the repetitive sequence density (Table 1(B)). The statistical significance is the best (*p*-value is indistinguishable from zero) at smaller window sizes, mainly because there are more samples. However, the magnitude of the correlation coefficient increases with the window size. This simultaneous increase of correlation coefficient and decrease of statistical significance with the increase of length scale has been previously observed in other applications [25].

#### 2.2.2. Regional variation still exists

We show the scatter plot for window size 256 kb in Fig. 1(B). The points from the first, second, and last 1/3 of the chromosome are labeled by green, blue, and red colors, respectively. Linear regression of motif density over repetitive sequence density in the three non-overlapping subsets show that the positive correlation mainly originates from the third subset, which contains windows with low repetitive sequence density (and these low repetitive sequence density windows have low motif density). Similar conclusion by wavelet analysis can be found in Table S2 of the Supplementary material.

#### 2.2.3. Density of [WW-3-SS, SS-3-WW] motif in repeat-filtered sequence is not consistently correlated with the repetitive sequence density

When [WW-3-SS, SS-3-WW] motif is obtained from the unique sequence (repeat-filtered/removed sequence), its density becomes negatively correlated with the repetitive sequence density at smaller window sizes (2 kb–16 kb), as shown in Table 1(B). This reversal from positive to negative correlation at these length scales hints that repetitive sequence itself contains the relevant NPMs. However, at larger window sizes, the correlation is back to positive (though less significant) (Table 1(B)).

The scatter plot in Fig. 2(B) shows that the situation is more complicated. Compared to Fig. 1(B), the points in region-3 are still low-repeat-density and low-NPM-density in unique sequence. However, the flat trend for the remaining points in Fig. 1(B) begin to have a negative trend in Fig. 2(B). A single correlation coefficient value cannot describe the nonlinear relationship between the two densities, and the sign of the correlation may depend on the repetitive sequence density.

### 2.3. Human chromosome 20, WW-8-WW motif

#### 2.3.1. Periodicity-10 of WW dinucleotides is positively correlated with the repetitive sequence density

The results in Table 1(C) shows a very strong positive correlation between WW-8-WW and repetitive sequence density at small window sizes. The corresponding wavelet-based analysis is in Table S3 of the Supplementary material. However, it does not mean lack of heterogeneity. Fig. 1(C) shows that there are more low-repeat-density and low-motif-density windows in the last 1/3 of windows (at window size of 64 kb). Without these windows, the strength of the positive correlation between WW-8-WW and repetitive sequence density would be weaker.

As an [A,T]-rich motif, WW-8-WW is expected to be less common in [G,C]-rich regions. We would like to check whether the repetitive sequences tend to be more [G,C] rich. In chromosome 20, the [A,T]-content in unique sequences is 0.553 which is indeed lower than that in repetitive sequence, 0.564. But this is a very small difference, and its expected effect on WW-8-WW density is only by a ratio of $0.564^4/0.553^4 = 1.08$. This ratio is too small to account for the drop of WW-8-WW density in low-repeat-density regions. The relationship between [G,C]-contents of unique and repetitive sequence at the 100 kb window level was plotted in Fig. 3 of ref. [26], and besides systematic deviation between the two, the [G,C]-contents in the two types of sequences are generally matched.

#### 2.3.2. Positive correlation remains when WW-8-WW motif density is calculated from the unique sequence

When WW-8-WW density is determined from the repeat-filtered sequence, its correlation with the repetitive sequence density remains positive (Table 1(C), Fig. 2(C)). Both the magnitude of cc and $p$-value do not seem to be altered very much.

### 2.4. Human chromosome 20, [RRRRRYYYYY, YYYYYRRRRR] motif

The [RRRRRYYYYY, YYYYYRRRRR] motif is a more recently proposed NPM whose density in chromosome 20 is 0.0035 copies per base. Interestingly, this observed density is much higher than the expected by the random sequence model (see Table S8 of the Supplementary material). Note that the reverse complement of RRRRRYYYYY is itself (i.e., palindromic).

#### 2.4.1. Decamer [RRRRRYYYYY,YYYYYRRRRR] density is negatively correlated with the repetitive sequence density

A tandem repeat of [RRRRRYYYYY, YYYYYRRRRR] contains the [RY-3-YR, YR-3-RY] motif with NNN replaced by [YYY, RRR]. One may consider [RRRRRYYYYY, YYYYYRRRRR] as a longer, but more specific example of [RY-3-YR, YR-3-RY].

The result in Table 1(D) shows a very strong and statistically significant negative correlation between [RRRRRYYYYY, YYYYYRRRRR] and repetitive sequence density, at almost all length scales examined. Fig. 1(D) shows the scatter plot between the two at window size of 64 kb, marked by whether the window is from the first 1/3, middle 1/3, or the last 1/3 of the chromosome. And Fig. S1 shows the spatial fluctuation of both densities along chromosome 20, as well as the position-scale heatmap by the wavelet transformation.

Different from the similar scatter plots in Figs. 1(A–C), the negative correlation in Fig. 1(D) is consistently observed in all regions (there are exceptions, however, such as an outlier, visible in both Fig. 1(D) and Fig. S1, where a very high motif density appears in a high repetitive sequence density window). The negative slopes of linear regression in the three segments have similar magnitude and similar $p$-values. The consistent negative correlation is also observed in a wavelet-based correlation calculation (Table S4 of the Supplementary material).

#### 2.4.2. Negative correlation remains when [RRRRRYYYYY, YYYYYRRRRR] motif density is calculated from the unique sequence

Table 1(D) shows that neither the magnitude nor the $p$-value of correlation between [RRRRRYYYYY, YYYYYRRRRR] density and repetitive sequence density are much affected, when the NPM density is calculated from the repeat-filtered sequences. Fig. 2(D) shows a scatter plot at the 64 kb window size. When it is compared with the similar plot in Fig. 1(D), the correlation is weaker and much less significant.

### 2.5. Human chromosome 20, other NPMs

Besides the four proposed NPMs analyzed so far: [YR-3-RY, RY-3-YR], [WW-3-SS, SS-3-WW], WW-8-WW, [RRRRRYYYYY, YYYYYRRRRR], there are other extensions and/or specific proposed NPMs. For example, an extension of [YR-3-RY, RY-3-YR] from spacing-of-5 to spacing-of-10 leads to the [YR-8-YR, RY-8-RY] motif. Another recent proposal of decamer NPM is [GRAAATTTYC, TTTYCGRAAA] [19]. Besides [RRRRRYYYYY, YYYYYRRRRR], the other parental degenerate of [GRAAATTTYC, TTTYCGRAAA] is [SSWWWWWWSS, WWWSSSSWWW]. All these NPMs are palindromic.

We found, generally speaking, densities of [SSWWWWWWSS, WWWSSSSWWW], [GRAAATTTYC, TTTYCGRAAA], and [YR-8-YR, RY-8-RY] to be positively correlated with the repetitive sequence density (see Tables S5,S6,S7 of the Supplementary material). This summary cannot characterize the whole range of complexity of the correlation analyses, as the results may differ at different length scales, between parametric and non-parametric correlation, and between the magnitude of correlation and statistical significance.

The positive correlation between [YR-8-YR, RY-8-RY] and repetitive sequence density is intriguing, as it provides an exception to negative correlation between densities of repetitive sequences and that of R/Y-base NPMs. However, if the NPM density is calculated only within unique sequences, the correlation with the repetitive sequence density becomes negative (and the correlation is statistically very significant). When we take a close look of the correlation by a scatter plot in Fig. 3 (at window size of 64 kb), the [YR-8-YR, RY-8-RY] density is essentially independent of the repetitive sequence density for windows in the first region. For points in the second region, removing an outlier changes the $cc = 0.005$ ($p$-value $= 2\times10^{-5}$) to $cc = 0.0029$ ($p$-value $= 2\times10^{-3}$). Both are very weak correlations. Only for the last 1/3 of the chromosome is the positive correlation more significant ($cc = 0.0038$, $p$-value $= 7\times10^{-6}$). These observations show that regional heterogeneity may affect the sign of the correlation.

### 2.6. Correlation between repetitive sequence density and proposed nucleosome positioning motifs in other chromosomes

#### 2.6.1. The correlation pattern observed in chromosome 20 is consistently observed in all other chromosomes in the human genome

Calculations carried out on chromosome 20 are extended to all autosomal chromosomes. Table 2 is intended to summarize a large number of results which are all based on consecutively doubling of window sizes from 1 kb to 8.192 Mb. Note that very high percentage of all windows are used (second column in Table 2) in the correlation analysis (the filtering criterion being that 90% of bases within the window are typed), with the only low percentages being in acrocentric chromosomes (13, 14, 15, 21, 22) due to untyped heterochromatin regions.

The [SSWWWWWWSS, WWWSSSSWWW] motif (3 million copies) and the more specific [GRAAATTTYC, TTTYCGRAAA] motif (a low-count of only 37,000 copies) are positively correlated with the repetitive sequence in all chromosomes and almost all window sizes. The [SS-3-WW, WW-3-SS] (190 million copies), WW-8-WW (249 million copies), [RY-8-RY, YR-8-YR] (269 million copies) motifs are positively correlated with repetitive sequence density for most chromosomes, though the correlation could become negative at

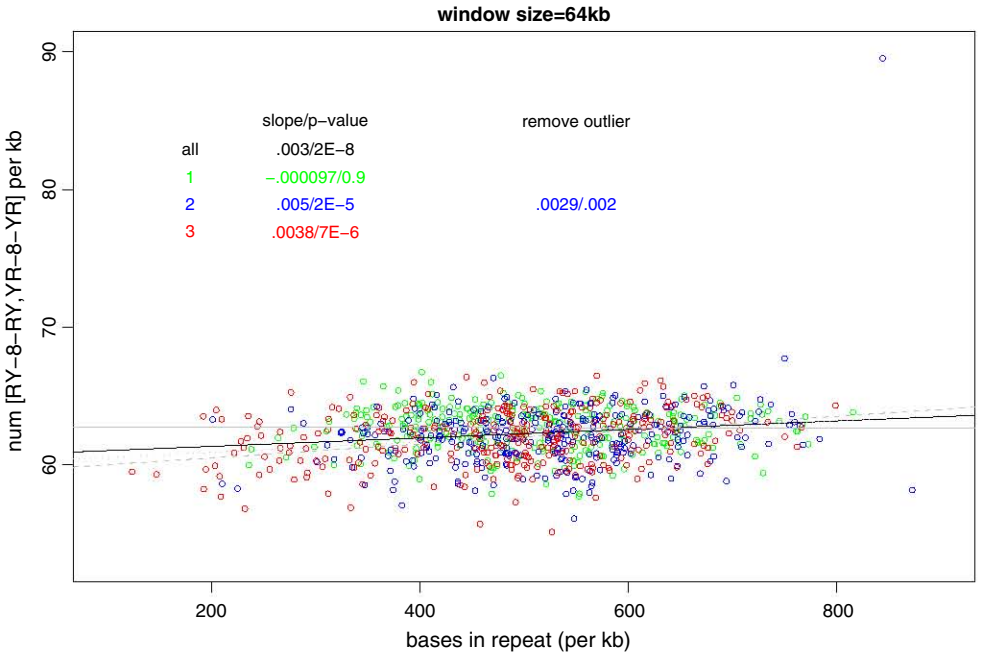


**Fig. 3.** Scatter plot of number of copies of [RY-8-RY, YR-8-YR] versus number of bases in repetitive sequence at window size of 64 kb. Data are from chromosome 20 only.

larger window sizes. These inconsistency may be caused by spatial heterogeneities in the correlation. The [RRRRRYYYYY, YYYYYRRRRR] (9.5 million copies) and [RY-3-YR, YR-3-RY] (274 million copies) are negatively correlated with repetitive sequence density for most chromosomes and for most window sizes, though the correlation may become positive at larger window sizes in some chromosomes.

*2.6.2. Combining all chromosomes into one dataset for correlation analysis*

When windows from all chromosomes are combined to one analysis, due to the increase of sample size, statistical significance for testing zero correlation is expected to improve (smaller *p*-values). Since there is only one single correlation calculation between a NPM and repetitive sequence density, heterogeneity between chromosomes will be a factor. All the signs of correlation obtained in chromosome 20 data are confirmed in the combined genome-wide data (see Tables S9–S15 of the Supplementary material).

Fig. 4 shows the scatter plot of four NPM densities versus repetitive sequence densities at the window size of 256 kb, with the genome-wide data. Comparing Fig. 4 with Fig. 1, the negative correlation with the two R/Y-based NPMs (Figs. 4(A, D)) and positive correlation with the two W/S-based NPMs (Figs. 4(B, C)) are confirmed. The scatter plot for [RRRRRYYYYY, YYYYYRRRRR] is particularly interesting: despite the negative trend followed by the majority of the points, there is a minority trend for high repetitive sequence densities and high NPM densities.

## 3. Discussion

In principle, there could be three types of NPMs using binary symbols: those based on R/Y, W/S, and on M/K (M for amino, C or A, K for keto, G or T). The first two types have been studied in this paper, but not the M/K-based ones. One simple explanation is that

**Table 2**
Correlation between the seven NPMs and repetitive sequence in 22 human autosomal chromosomes. The "+" ("−") mean positive (negative) correlation; "−/+" means negative correlation at smaller window sizes but positive correlation at larger window sizes; "[ns]" means the correlation is not statistically significant (*p*-value > 0.01).

| chromosome | %w used | RRRRRYYYYY | SSWWWWWWSS | GRAAATTTYC | RY-3-YR | SS-3-WW | WW-8-WW | RY-8-RY |
|---|---|---|---|---|---|---|---|---|
| 1 | 90.4 | − | + | + | +/− [ns] | + | + | + |
| 2 | 97.9 | − | + | + | − [ns] | +/− [ns] | + | + |
| 3 | 98.4 | − | + | + | − | + | + | + |
| 4 | 98.2 | − | + | + | −/+[ns] | + | −/+ | −/+ |
| 5 | 98.2 | − | + | + | − | + | + | + |
| 6 | 97.8 | − | + | + | +[ns] | + | + | + |
| 7 | 97.6 | −/+[ns] | + | + | − | + | +/− [ns] | +/− [ns] |
| 8 | 97.6 | −/+ | + | + | − | + | + | +/− [ns] |
| 9 | 85.1 | − | + | + | − [ns]/+[ns] | + | + | + |
| 10 | 96.9 | − | + | + | − | + | + | + |
| 11 | 97.1 | −/+ | + | + | + | + | + | + |
| 12 | 97.5 | −/+ | + | + | − [ns] | + | + | + |
| 13 | 83.0 | − | + | + | − | +/-[ns] | + | −/+ |
| 14 | 82.2 | − | + | + | − [ns]/+[ns] | +/-[ns] | + | + |
| 15 | 79.7 | − | + | + | + | +/− | + | + |
| 16 | 87.3 | − | + | + | − | + | + | + |
| 17 | 95.8 | − | + | + | + | + | + | + |
| 18 | 95.6 | − | + | + | − | +/− [ns] | + | + |
| 19 | 94.4 | − | + | + | − | + | + | + |
| 20 | 94.4 | − | + | + | − | + | + | + |
| 21 | 72.9 | − | + | + | − | +/− [ns] | + | + |
| 22 | 68.0 | − | + | + | − | + | + | + |

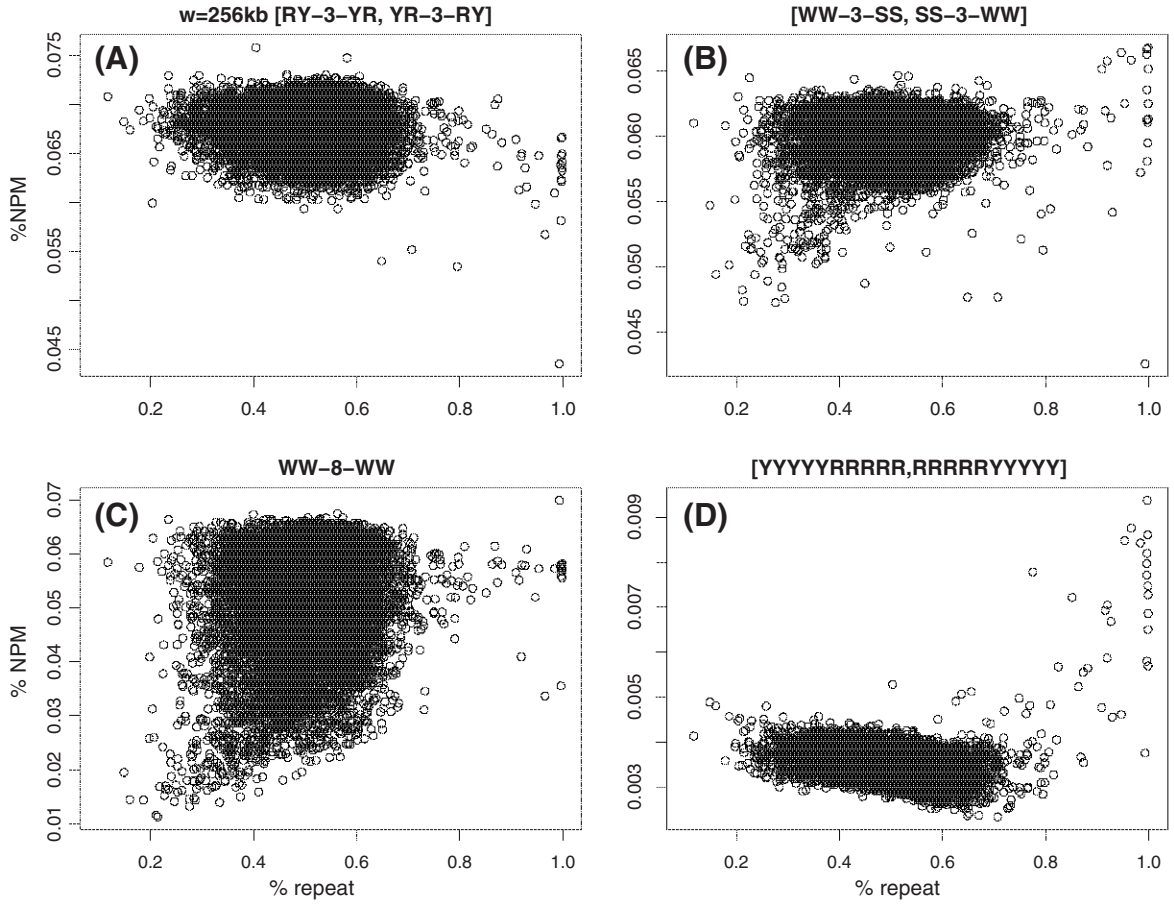


**Fig. 4.** Scatter plot of genome-wide (chromosomes 1-22) NPM densities versus repetitive sequence density at the level of 256 kb windows. (A) [RY-3-YR, RY-3-YR]; (B) [WW-3-SS, SS-3-WW]; (C) WW-8-WW; and (D) [RRRRRYYYYY, YYYYYRRRRR].

none such motif has been proposed. W/S-based motifs are closely related to the thermodynamic stability of the double helix DNA molecule. R/Y-based motifs, due to the difference of nucleotide sizes (R is larger than Y) and the limitation of physical space, are intrinsically related to the bending and rigidity of the DNA molecule. If a M/K-based motif exists, it could be either due to interactions between keto bases (G or T) and the histones, or related to the fact that keto bases have two alternative forms of the structure (keto vs. enol form).

The total number of copies of a NPM in the genome provides important information on how much this NPM may contribute to the nucleosome positioning. It is estimated that 20% of the human genome [15], around 600 Mb, or even more [27], are occupied by nucleosome with stable positioning. In order for repeating [RRRRRYYYYY, YYYYYRRRRR] motif to cover the 600 Mb region, 60 to 120 million copies of them are needed. When only 9 million copies (or 3% of the genome) are actually observed (Table S8), several consequences can be expected.

One is that it is less likely to observe the periodicity-10 signal as there would not be enough copies of the motif to repeat tandemly (D. Sosa, P. Miramontes, W. Li, V. Mireles, J.R. Bobadilla, M.V. José, unpublished results). Another consequence of lower number of copies of a NPM is that instead of a densely packing of the NPM in nucleosome regions, we may only need a few NPM per nucleosome, while other factors contribute to the positioning. For example, it is suggested in ref. [28] that barriers near a gene's promoter region may help the positioning of nucleosomes. It raises the question of the importance of not only a particular proposed NMP in nucleosome positioning, but also of roles played by DNA sequence in general.

We are not aware of previous studies on the correlation between NPMs and repetitive sequences. In ref. [29], an experimentally obtained nucleosome signal is plotted within and around (up to 1 kb) the *Alu* element. The goal of this experimental study is very different from ours as it is centered around the *Alu* element and within a much smaller length scale (the peak-to-trough distance is 200 bp). Even if we know where the nucleosome signal is located within an *Alu* element, we still do not know whether the presence of *Alu* sequence increases the nucleosome positioning probability in that region, though *Alu* elements were claimed to confer nucleosome positioning *in vitro* [30].

Besides treating *Alu* as a subgroup, there are also subgroups within *Alu*. There are roughly 40 different *Alu* sequences such as *Alu*Jb, *Alu*Sx, *Alu*Y, *Alu*Sx1, each of which with more than 100,000 copies. The *Alu*Y sequence is in a relatively younger group [31]. There are also human-specific branches of *Alu*, i.e., Yc1, Ya5a2, Yb9 [32], with much lower frequencies. Preliminary analyses show that most of our results between NPM densities and repetitive sequence densities hold true for *Alu* or *Alu*Y densities also. However, the correlation with densities of W/S-based NPMs may become negative.

In ref. [33], the autocorrelation function of CG dinucleotide is calculated for the original and the repeat-masked sequence. Peaks at distances of 31 and 62 bps disappear in the repeat-masked sequence, but at the same time, new peaks at distances 10 and 21 appear. In that paper, any peak at a multiple of 10 bp is considered to be a nucleosome positioning signal, then such signal is present in both *Alu* and non-repetitive sequences. The authors of ref. [33] suggested that *Alu* elements might play a role of “anchor” for nucleosomes, which is reminiscent of the barrier idea in ref. [15]. If both positive and negative correlations exist between NPMs and repetitive sequence density, it indicates nucleosome positioning information can be enriched either in repetitive sequences or in unique sequences.

Whether more repetitive sequences in a genome increase or decrease the probability for nucleosome positioning may provide insight on the evolution of repetitive sequences [34]. Most repetitive sequences are transposable elements caused by at least three mechanisms [35] and are particularly abundant in sexual organisms [36]. As a major force in expanding the higher organisms' genome including the human's [37], it must have an effect on the genome function [38–44]. But most of the focus concerning impact of repetitive sequences is on the genomic instability introduced, genetic innovation accompanied by the extra DNA sequences [41], and gene expression or regulatory networks [45]. Discussion on repetitive sequences' impact via nucleosome formation was mostly in promoter region [46,47].

The results in this paper hint that repetitive sequences can also have subtle and complicated impact to nucleosome-forming potential by either increasing or decreasing NPM density in repetitive sequence regions. This effect can be small, and may be detectable only in local regions with extreme densities of repetitive sequence.

Detection of sequence signal or statistical correlation between any two sequence measures can always be more complex than the apparent calculations. First, stratifying sequence data by controlling other quantities can much weaken a signal or a correlation. For example, the periodicity of R-7-R [48] or WW-8-NWW motifs [49] around a CpG dinucleotide can be absent if the CpG dinucleotide is located in a [G,C]-rich and unmethylated CpG island. Second, when multiple sequence measures are pairwisely correlated, the cause-effect relationship between these measures intrinsically affect conditional correlation result [50]. The idea that repetitive sequences have relevant evolutionary impact on higher organisms only under certain conditions was discussed in ref. [51]. Whether the correlation between NPM and repetitive sequence densities discussed here can disappear by conditionaling on other sequence measures worth future studies.

## 4. Materials and methods

### 4.1. Human DNA sequence data

The GRCh37/hg19 (Feb. 2009) version of the human genome sequence is downloaded from UCSC's genome browser (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/). Repeat sequences are marked as lowercase in the file as versus the uppercase letters for unique sequences. For specific repetitive sequence family, we use the rmsk.txt file from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/ which lists the starting and ending positions of 5.298 millions occurrence of more than 1300 different types of repetitive sequences.

### 4.2. Motif countings

For some NMPs, whether overlapping motifs are counted as more than one copy or not will change the counting value. For example, WW-3-SS may overlap with another WW-3-SS shifted by one base position, while we may consider both to contribute to one NPM. The countings for the NMP density calculation in Table S8, however, are all obtained by shifting one position.

### 4.3. Motif density in unique sequences

For a window with $N$ bases (e.g. $N=1000$), $N_{\text{not}}$, $N_{\text{rep}}$, $N_{\text{uniq}}$ are the number of bases that are not sequenced, part of a repetitive sequence, or not part of the repetitive sequence (thus part of the unique sequence), and $N=N_{\text{not}}+N_{\text{rep}}+N_{\text{uniq}}$. During the quality control stage, windows with sequencing rate ($N_{\text{not}}/\text{N}\le 0.9$) are discarded, so for almost all windows used, $N_{\text{not}}=0$. Denote $n$ and $n_{\text{uniq}}$ as the number of copies of a NPM in the window and in the unique sequence within the window, respectively, then $n/(N_{\text{rep}}+N_{\text{uniq}})$ is the NPM density, and $n_{\text{uniq}}/N_{\text{uniq}}$ is the NPM density in the unique sequence.

### 4.4. Statistical methods

Pearson's and Spearman's statistical correlation and the corresponding tests were carried out by the *cor.test* function in *R* (http://www.r-project.org/), with the option *method* = "*pearson*" (default) or *method* = "*spearman*". Spearman's correlation is simply a Pearson's correlation by replacing the raw data with its ranking values. Kendall's correlation coefficient is, like Spearman's correlation, another non-parametric measure of correlation, defined as $(\#\text{ concordant pairs}-\#\text{ discordant pairs})/(n(n-1)/2)$, and can be calculated by the above *R* function with *method* = "*kendall*".

### 4.5. Wavelet analysis

Wavelet transformation [52] provides an alternative way in dealing with correlation analysis at different length scales. We adopt the *R* routines *plot.pair.wavelet* used in ref. [25] with the Haar wavelet basis, which requires the installation of two *R* packages: *Rwave* [53] and *wavethresh* [54].

## Acknowledgments

We thank Pedro Miramontes, Jan Freudenberg, Victor Mireles for discussions, and W.L. acknowledges the support from the Robert S Boas Center for Genomics and Human Genetics. MVJ acknowledges support from PAPIIT UNAM, project IN107112.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2012.10.005.

## References

[1] A. Wolffe, Chromatin: Structure and Function, 3rd edition Academic Press, 1999.
[2] C.L. Woodcock, R.P. Ghosh, Chromatin higher-order structure and dynamics, Cold Spring Harb. Perspect. Biol. 2 (2010) a000596.
[3] G. Li, D. Reinberg, Chromatin higher-order structures and gene regulation, Curr. Opin. Genet. Dev. 21 (2011) 175–186.
[4] A. Kundaje, S. Kyriazopoulou-Panagiotopoulou, M. Libbrecht, C.L. Smith, D. Raha, E.E. Winters, S.M. Johnson, M. Snyder, S. Batzoglou, A. Sidow, Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements, Genome Res. 22 (2012) 1735–1747.
[5] R.E. Thurman, et al., The accessible chromatin landscape of the human genome, Nature 489 (2012) 75–82.
[6] C.R. Calladine, H. Drew, B. Luisi, A. Travers, Understanding DNA: The Molecule and How it Works, 3rd edition Academic Press, 2004.
[7] T.C. Bishop, Chromatin in 1, 2 and 3 dimensions. Comment on 'Cracking the chromatin code: precise rule of nucleosome positioning' by E.N. Trifonov, Phys. Life Rev. 8 (2011) 56–58.
[8] V.B. Zhurkin, Specific alignment of nucleosomes on DNA correlates with periodic distribution of purine–pyrimidine and pyrimidine–purine dimers, FEBS Lett. 158 (1983) 293–297.
[9] S.C. Satchwell, H.R. Drew, A.A. Travers, Sequence periodicities in chicken nucleosome core DNA, J. Mol. Biol. 191 (1986) 659–675.
[10] X. Wang, G.O. Bryant, M. Floer, D. Spagna, M. Ptashne, An effect of DNA sequence on nucleosome occupancy and removal, Nat. Struct. Mol. Biol. 18 (2011) 507–509.
[11] E.N. Trifonov, J.L. Sussman, The pitch of chromatin DNA is reflected in its nucleotide sequence, Proc. Natl. Acad. Sci. 77 (1980) 3816–3820.
[12] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I.K. Moore, J.Z. Wang, J. Widom, A genomic code for nucleosome positioning, Nature 442 (2006) 772–778.
[13] S.M. Johnson, F.J. Tan, H.L. McCullough, D.P. Riordan, A.Z. Fire, Flexibility and constraint in the nucleosome core landscape of *Caenorhabditis elegans* chromatin, Genome Res. 16 (2006) 1505–1516.
[14] A. Valouev, J. Ichikawam, T. Tonthat, J. Stuart, S. Ranade, H. Packham, K. Zeng, J.A. Malek, G. Costa, K. McKernan, A. Sidow, A. Fire, S.M. Johnson, A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning, Genome Res. 18 (2008) 1051–1063.
[15] A. Valouev, S.M. Johnson, S.D. Boyd, C.L. Smith, A.Z. Fire, A. Sidow, Determinants of nucleosome organization in primary human cells, Nature 474 (2011) 516–520.

[16] Z. Zhang, C.H. Wippo, M. Wal, E. Ward, P. Korber, B.F. Pugh, A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome, Science 332 (2011) 977–980.
[17] I. Gabdank, D. Barash, E.N. Trifonov, Nucleosome DNA bendability matrix (*C. elegans*), J. Biomol. Struct. Dyn. 26 (2009) 403–411.
[18] E.N. Trifonov, Nucleosome positioning by sequence, state of the art and apparent finale, J. Biomol. Struct. Dyn. 27 (2010) 741–746.
[19] E.N. Trifonov, Cracking the chromatin code: precise rule of nucleosome positioning, Phys. Life Rev. 8 (2011) 39–50.
[20] R.J. Britten, DE Kohne, Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms, Science 161 (1968) 529–540.
[21] W. Li, On parameters of the human genome, J. Theol. Biol. 288 (2011) 92–104.
[22] A.P.J. de Koning, W. Gu, T.A. Castoe, M.A. Batzer, D.D. Pollock, Repetitive elements may comprise over two-thirds of the human genome, PLoS Genet. 7 (2011) e1002384.
[23] H. Takata, K. Maeshima, Irregular folding of nucleosomes in the cell. Comment on 'Cracking the chromatin code: precise rule of nucleosome positioning' by Edward N. Trifonov, Phys. Life Rev. 8 (2011) 51–52.
[24] W. Li, Are isochore sequences homogeneous? Gene 300 (2002) 129–139.
[25] C.C.A. Spencer, P. Deloukas, S. Hunt, J. Mullikin, S. Myers, B. Silverman, P. Donnelly, D. Bentley, G. McVean, The influence of recombination on human genetic diversity, PLoS Genet. 2 (2006) e148.
[26] J. Paces, R. Zíka, V. Paces, A. Pavlícek, O. Clay, G. Bernardi, Representing GC variation along eukaryotic chromosomes, Gene 333 (2004) 135–141.
[27] D.J. Gaffney, G. McVicker, Y. Fondufe-Mittendorf, J. Widom, Y. Gilad, J.K. Pritchard, Most nucleosomes in the human genome are consistently positioned, The Biology of Genome (May 8-12, 2012, Cold Spring Harbor Laboratory). http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE36979.
[28] Y. Zhang, Z. Moqtaderi, B.P. Rattner, G. Euskirchen, M. Snyder, J.T. Kadonaga, X.S. Liu, K. Struhl, Intrinsic histone–DNA interactions are not the major determinant of nucleosome positions in vivo, Nat. Struct. Mol. Biol. 16 (2009) 847–852.
[29] Y. Tanaka, R. Yamashita, Y. Suzuki, K. Nakai, Effects of Alu elements on global nucleosome positioning in the human genome, BMC Genomics 11 (2010) 309.
[30] E.W. Englander, B.H. Howard, Nucleosome positioning by human Alu elements in chromatin, J. Biol. Chem. 270 (1995) 10091–10096.
[31] M. Costantini, F. Auletta, G. Bernardi, The distribution of 'new' and 'old' Alu sequences in the human genome: the solution of a 'mystery', Mol. Biol. Evol. 29 (2012) 421–427.
[32] M.A. Batzer, P.L. Deininger, Alu repeats and human genome diversity, Nat. Rev. Genet. 3 (2002) 370–379.
[33] T. Bettecken, Z.M. Frenkel, E.N. Trifonov, Human nucleosomes: special role of CG dinucleotides and Alu-nucleosomes, BMC Genomics 12 (2011) 273.
[34] J. Jurka, V.V. Kapitonov, O. Kohany, M.V. Jurka, Repetitive sequences in complex genomes: structure and evolution, Ann. Rev. Genomics Hum. Genet. 8 (2007) 241–259.
[35] A.F.A. Smit, The origin of interspersed repeats in the human genome, Curr. Opin. Genet. Dev. 6 (1996) 743–748.
[36] D.A. Hickey, Evolutionary dynamics of transposable elements in prokaryotes and eukaryotes, Genetica 86 (1992) 269–274.
[37] H.H. Kazazian Jr., Mobile elements: drivers of genome evolution, Science 303 (2004) 1626–1632.
[38] M. Syvanen, The evolutionary implications of mobile genetic elements, Annu. Rev. Genet. 18 (1984) 271–293.
[39] D.J. Finnegan, Eukaryotic transposable elements and genome evolution, Trends Genet. 5 (1989) 103–107.
[40] C. Feschotte, E.J. Pritham, DNA transposons and the evolution of eukaryotic genomes, Annu. Rev. Genet. 41 (2007) 331–368.
[41] R. Cordaux, M.A. Batzer, The impact of retrotransposons on human genome evolution, Nat. Rev. Genet. 10 (2009) 691–703.
[42] C. Biémont, A brief history of the status of transposable elements: from junk DNA to major players in evolution, Genetics 186 (2010) 1085–1093.
[43] J.A. Shapiro, Mobile DNA and evolution in the 21st century, Mob. DNA 1 (2010) 4.
[44] A. Hua-Van, A. le Rouiz, T.S. Boutin, J. Filée, P. Capy, The struggle for life of the genome's selfish architects, Biol. Direct 6 (2011) 19.
[45] C. Feschotte, Transposable elements and the evolution of regulatory networks, Nat. Rev. Genet. 9 (2008) 397–405.
[46] A. Huda, L. Mariño-Ramirez, D. Landsman, I.K. Jordan, Repetitive DNA elements, nucleosome binding and human gene expression, Gene 436 (2009) 12–22.
[47] A. Huda, Epigenetic Regulation of the Human Genome by Transposable Elements, Ph.D Thesis, Georgia Institute of Technology, 2010.
[48] O. Clay, W. Schaffner, K. Matsuo, Periodicity of eight nucleotides in purine distribution around human genomic CpG dinucleotides, Somat. Cell Mol. Genet. 21 (1995) 91–98.
[49] A. Tanay, A.H. O'Donnell, M. Damelin, T.H. Bestor, Hyperconserved CpG domains underlie polycomb-binding sites, Proc. Natl. Acad. Sci. 104 (2007) 5521–5526.
[50] J. Freudenberg, M. Wang, Y. Yang, W. Li, Partial correlation analysis indicates causal relationships between GC-content, exon density and recombination rate variation in the human genome, BMC Bioinform. 10 (Suppl. 1) (2009) S66.
[51] E. Zuckerkandl, G. Cavalli, Combinatorial epigenetics, "junk DNA", and the evolution of complex organisms, Gene 390 (2007) 232–242.
[52] D.B. Percival, A.T. Walden, Wavelet Methods for Time Series Analysis, Cambridge University Press, 2005.
[53] R. Carmona, W.L. Hwang, B. Torresani, Practical Time–Frequency Analysis: Gabor and Wavelet Transformations With an Implementation in S, Academic Press, 1998.
[54] G. Nason, Wavelet Methods in Statistics with R, Springer, 2008.