
# UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

DOCTORADO EN CIENCIAS BIOMÉDICAS
INSTITUTO DE INVESTIGACIONES BIOMÉDICAS

## "PROPIEDADES ESTADÍSTICAS DE SECUENCIAS GENÓMICAS: UN ENFOQUE ESPECTRAL"

TESIS
QUE PARA OPTAR POR EL GRADO DE:
DOCTORA EN CIENCIAS BIOMÉDICAS

PRESENTA:
MARÍA FERNANDA HIGAREDA HOYOS

DIRECTOR DE TESIS:
RAFAEL ALBERTO MÉNDEZ SÁNCHEZ
INSTITUTO DE CIENCIAS FÍSICAS, UNAM

COMITÉ TUTOR:

LUIS ALBERTO MENDOZA SIERRA
INSTITUTO DE INVESTIGACIONES BIOMÉDICAS, UNAM

OTTO GEIGER
CENTRO DE CIENCIAS GENÓMICAS, UNAM

MÉXICO, D.F. SEPTIEMBRE 2013

Este trabajo fue realizado bajo la tutoría del Dr. Rafael A. Méndez Sánchez, del programa de doctorado en Ciencias Biomédicas del Instituto de Investigaciones Biomédicas y del Instituto de Ciencias Físicas de la Universidad Nacional Autónoma de México (UNAM).

Durante la realización de esta tesis María Fernanda Higareda Hoyos recibió una beca de doctorado por CONACyT del periodo de febrero de 2005 a enero de 2010. Beca complemento de PAPIIT DGAPA-UNAM (IN111308) de febrero de 2010 a julio de 2010. Parte del trabajo de esta tesis de investigación ha sido financiado por el Consejo Nacional de Ciencia y Tecnología de México (CONACyT 79613) y los proyectos PAPIIT DGAPA-UNAM (IN111308 e IN111311).

**Dedicado a:**
*Mi familia.*

# Agradecimientos

A mis padres por el apoyo incondicional que me brindaron durante el transcurso de esta etapa.

A mis hermanos Armando y César, por su compañía, preocupación y ayuda en momentos díficiles en mi vida.

A mi abuela por quererme y ayudarme en todo momento, incliuda a Luz Dary.

Al Dr. Rafael Méndez por el apoyo y la confianza que me brindó para la realización de esta tesis y poderla finalizar.

A los Drs. Luis Mendoza y Otto Geiger de quien tuve su apoyo en todo momento.

A mis sinodales, los Drs. Miguel A. Cevallos, Germinal Cocho, Alma Lara Sagahón y Enrique Merino Pérez que se dieron a la tarea de revisar este trabajo de tesis, por aportar su valiosa opinión para mejorarla notablemente.

A mis amigas Karla, Ivette, Lugui, Sheilla, Pamela y Ane por sus consejos, sobre todo a Tellez por leer y comentar los artículos conmigo.

A la Universidad Nacional Autónoma de México, al Instituto de Investigaciones Biomédicas y al Instituto de Ciencias Físicas por haberme abierto las puertas y darme el privilegio y orgullo de pertenecer a esta casa de estudios y haber estudiado en ella un doctorado.

A los profesores que me formaron como profesionista, por su dedicación y empeño al compartir su conocimiento.

# Resumen

El análisis estadístico de una secuencia genómica completa, usualmente asume la homogeneidad de la densidad de nucleótidos a travez del genoma. Esta hipótesis se ha probado que es incorrecto en diferentes organismos y sólo puede ser correcto localmente. Para evitar dar un valor constante, a esta propiedad variable, se propone el uso de la estadística espectral, la cual caracteriza la densidad de nucleótidos como una función de la posición a lo largo del genoma. Mostramos que la densidad acumulada de bases en genomas bacterianos puede ser separada en una propiedad promedio (secular) mas una parte fluctuante. Esta separación es, en cierto sentido, equivalente al análisis de fluctuaciones libres de tendencia (DFA por sus siglas en inglés). De acuerdo a la descripción cualitativa de su parte secular los genomas bacterianos pueden ser divididos en dos grupos lineal o lineal por partes. Estos dos grupos de genomas muestran diferentes propiedades en la distribución de espaciamiento de los nucleótidos. Con el fin de analizar estadísticamente genomas que tiene una densidad variable de nucleótidos es necesario el uso del desdoblamiento, i.e., obtener la separación entre la parte secular y la parte fluctuante. El desdoblamiento permite una comparación adecuada de las propiedades estadísticas de diversa secuencias genómicas. Con esta metodología fueron analizadas algunas secuencias de la proteina $\beta$-miosina tanto en la *Drosophila melanogaster* como en el humano. También fueron estudiados cuatro géneros de bacterias *Burkholderia, Bacillus, Clostridium* y *Corynebacterium*. La distribución de primeros vecinos es muy similar para especies con el mismo genero pero es muy diferente para especies con diferente genero. Mostramos que esta diferencia es atribuida a la diferencia del uso de codon en bacterias. Se muestra que la distribución de primeros vecinos entre regiones codificantes que contienen introns y las otras que no contienen introns, son las mismas cuando usamos el desdoblamiento. La distribución obtenida es muy similar al azarosa (Poissoniana o exponencial).

Con el fin de hacer un modelo que explique los resultados anteriores construimos un modelo de máxima entropía para la distribución de fluctuaciones de distancias entre pares de bases. Como punto inicial usamos la distribución de Poisson con una repulsión de corto alcance. Dentro de este modelo, se obtiene la distribución de primeros vecinos con $n$ bases intermedias. Para validar este modelo se analizaron muchas secuencias y se mostró que las propiedades fluctuantes son universales.

# Abstract

Statistical analysis of a complete gnomic sequence usually assumes a homogeneous nucleotide density along the genome. This assumption has been proved incorrect for several organisms since the nucleotide density is only locally homogeneous. To avoid giving a constant value, to this variable property, we propose the use of spectral statistics which characterizes the density of nucleotides as a function of its position along the genome. We show that the cumulative density of bases in bacterial genomes can be separated into an average (or secular) plus a fluctuating part. This separation is, in some sense, equivalent to detrended fluctuation analysis (DFA). According to the qualitative description of their secular part bacterial genomes can be divided in two groups linear and piecewise linear. These two groups of genomes show different properties when their nucleotide spacing distribution is studied. In order to analyze statistically genomes having a variable nucleotide density it is necessary the use of the unfolding, i.e., to get a separation between the secular part and the fluctuations. The unfolding allows an adequate comparison of the statistical properties of diverse gnomic sequences.

With this methodology were analyzed some gnomic sequences of *Drosophila melanogaster* and that of the human $\beta$-licina. Also four bacterial genuses were analyzed *Burkholderia, Bacillus, Clostridium* and *Corynebacterium*. Interestingly, the nearest neighbor spacing distributions are very similar for species within the same genus but they are very different for species from different genera. This difference can be attributed to the difference in the codon usage. We show that the nearest neighbor distribution of distances between bases pairs of some intron-less and intron-containing coding regions are the same when the *unfolding* is applied. The form of the distribution obtained is quite similar to that of a random, (Poissonian or exponential), sequence.

Based on the maximum entropy approach, a model for the fluctuations

of distances between pairs of bases in DNA is obtained. As a starting point, a Poissonian distribution with a short-range repulsion is used. Within this model, the nearest-neighbor distribution with $n$ bases in between is obtained. To test the validity of this model we analyze several sequences and show that the fluctuating properties are universal.

# Índice general

# Capítulo 1

# Introducción

En los últimos años la disponibilidad de secuencias se ha incrementado enormemente y una buena parte de ellas se encuentran disponibles en la red (v.g. GenBank). Esto hace muy pertinente su caracterización estadística y la aplicación de técnicas matemáticas, físicas para arrojar respuestas biológicas. Es por ello que el estudio de las propiedades estadísticas de secuencias genéticas ha permitido la obtención de una gran cantidad de información del DNA como son, predicción de genes, estructura, orígenes de replicación, etc. De echo, existe una larga tradición en el análisis estadístico de secuencias de DNA y siguen apareciendo nuevos desarrollos. Entre los que se han utilizado se encuentran las caminatas aleatorias, el análisis de series de distancias, el estudio de correlaciones y autocorrelaciones de diversas medidas como el porcentaje de guanina-citocina (G+C) o los espectros de potencia. Quizá el punto más importante es que, hay varios ejemplos donde se ha utilizado la estadística para obtener propiedades biológicas. Por ejemplo, usando el sesgo[1] se localizó el origen de replicación en eucariontes [1] y procariontes [2]. En otro caso se utilizaron las ondeletas[2] para localizar y orientar los genes [3]. También, utilizando el método de agrupamiento, se localizan genes [4], y con este mismo método se construye árboles filogenéticos [5]. Utilizando una medida alternativa llamada superinformación se puede distinguir las regiones codificantes de las no codificantes [6].

En este proyecto, con un enfoque similar a lo anteriormente esbozado, analizamos la secuencias genómicas con una herramienta conocida como es-

[1] Del inglés *GC skew*
[2] Del inglés *wavelets*

tadística espectral. Esta metodología fue inicialmente desarrollada para estudiar las propiedades de los niveles de energía del núcleo atómico [7, 8] y fue después que se aplicó a varias áreas de la física, desde sistemas microscópicos hasta macroscópicos [9]. Más recientemente hay interés en utilizar esta técnica en áreas fuera de la física, incluyendo la biología. Particularmente hay varios trabajos que aplican la estadística espectral a las propiedades de los electroencefalogramas [10, 11], en la expresión de genes con microarreglos [12] y en las propiedades estadísticas entre las distancias de nucleótidos [13, 14, 15], que incluye un par de trabajos nuestros [16, 17]. Paralelamente a la estadística espectral se ha desarrollado una teoría matemática, la teoría de matrices aleatorias[3] que predice teóricamente las cantidades que se estudian en la estadística espectral. Los resultados de las matrices aleatorias se obtienen usando diversos métodos matemáticos como la máxima entropía, entre otros.

En esta tesis la estadística espectral se aplica al análisis de las secuencias genómicas, como se detalla a continuación.

Iniciamos en el capítulo siguiente donde se hará una revisión de los antecedentes de la biología de genomas, desde el descubrimiento del DNA (ácido desoxi ribonucléico), hasta la características del genoma completo de organismos eucariontes y procariontes, su composición, estructura, etc. También se explicarán los métodos estadísticos en secuencias genómicas como correlaciones, caminatas al azar, análisis de fluctuaciones libres de tendencia[4], etc. También se hará un resumen de los resultados biológicos arrojados como son: evolución de secuencias, localización de genes y de otras estructuras, etc.

El capítulo 3 se explican brevemente la hipótesis y los objetivos tanto generales como particulares, el capítulo 4 es el más importante ya que ahí se dan las contribuciones de los análisis estadísticos a la biología y será dividido de la siguiente forma:

La sección 4.1 se divide en dos partes:

- En la subsección 4.1.1 se analizan algunas regiones codificantes y no

[3]Una matriz aleatoria es una arreglo de $n \times n$ donde sus elementos son asignados al azar, de una distribución de probabilidad dada.

[4]Del inglés *detrended fluctuation analysis*

codificantes de secuencias de la mosca y del humano. Aquí mostramos que la densidad acumulada de bases en las secuencias genómicas está compuesta por una parte suave o secular mas una parte fluctuante. Luego mostramos que la propiedad secular es una propiedad característica de cada secuencia, es decir, es una propiedad biológica. También mostramos que la distribución de distancias a primeros vecinos, una cantidad fluctuante, es universal. Esto quiere decir que no depende si la secuencia es codificante o no. Estos resultados corresponden al primer artículo que publicamos en Physica A [16].

- En la subsección 4.1.2 se analiza una gran variedad de bacterias de diferentes grupos y diferentes características como por ejemplo el contenido de GC, su tamaño, etc. Se observa que éstas se pueden agrupar por la densidad acumulada de bases (CDoB)[5] en dos grupos, uno con una CDoB lineal y otra con una CDoB lineal por partes. La conclusión es que, para las bacterias que pertenecen a este último grupo, muchas propiedades estadísticas se tienen que analizar por partes ya que de otra manera se tiene el riesgo de dar una suma de diferentes propiedades estadísticas. Se tomaron cuatro generos de bacterias que son: *Burkholderia, Clostridium, Bacillus y Corynebacterium*. Se muestra que la distancia a primeros vecinos es muy similar para todas las especies del mismo género. Esta distribución cambia al cambiar de género. Por varios experimentos numéricos hemos mostrado que la explicación de este cambio es debido al uso de codón. Este resultado también lo hemos publicado en la revista Physica A [17].

Para finalizar el capítulo 4, en la sección 4.2 se construye un modelo de máxima entropía que predice teóricamente las distribuciones a primeros vecinos. También se presentan las distribuciones para las distancias cuando hay $n$ nucleótidos intermedios.

En los capítulos 5 y 6, que corresponden a la discusión y a las conclusiones, se discuten los puntos mas importantes de la tesis: utilizando la estadística espectral se pueden comparar diferentes tipos de genomas no importando su tipo. Analizando la distribución de fluctuaciones de diferentes bacterias encontramos una marca o firma genómica a nivel de género y esto es debido al uso de codon.

[5]Del inglés *cumulative density of bases*

# Capítulo 2

# Antecedentes

## 2.1. Biología del DNA

El DNA es un polímero grande con una estructura lineal de residuos de azúcar y fostatos alternados. El azúcar en las moléculas de DNA es 2-desoxirribosa, un monosacáridos de cinco carbonos el cual se une a una base nitrogenada dando origen a un nucleósido. La adición de un fosfato a un nucleósido da lugar a la subunidad del DNA denominado nucleótido.

Las bases nitrogenadas del DNA se dividen en dos clases, las purinas y las pirimidinas [18]. Las purinas, adenina (A) y guanina (G), están conformadas por dos anillos heterocíclicos engranados [19] pero tienen grupos diferentes unidos a ellas [18]. La timina (T) y la citosina (C) forman las pirimidinas cuya estructura solo posee un anillo heterocíclico. Todas las bases nitrogenadas se unen al monosacárido por medio de enlaces glucosídicos. Las purinas se unen al C1 del azúcar mediante un enlace N-glucosídico a través del N9 del anillo mientras que las pirimidinas lo hacen a través del N1 [19].

Las bases nitrogenadas se encuentran hacia el interior de la hélice por su carácter hidrofóbico mientras que el esqueleto azúcar-fosfato se encuentra hacia el exterior. La adenina puede formar enlaces de hidrógeno sólo con la timina mientras que la citocina sólo puede hacerlo con la guanina. A esto se le llamó apareamiento de bases y las bases apareadas se dicen que son complementarias [19]. La conformación de las bases es indispensable para estructurar la doble hélice de DNA, su complementariedad y la formación de los enlaces entre éstas, son la característica fundamental de la hélice doble y contribuyen a la estabilidad termodinámica de ésta [18].

Los primeros estudios del DNA, con difracción de rayos X, permitieron comprobar que existen dos tipos de estructuras, las formas B y A del DNA. La forma B, que se observa en condiciones de humedad elevada, corresponde con mucha precisión al promedio de la estructura del DNA en situación fisiológica [18], es dextrógira y está conformada por 10 pb por giro. La forma A es igualmente dextrógira y aparece en condiciones de humedad escasa [19], tiene 11 pb por vuelta, su surco mayor es más angosto y mucho más profundo que el de la forma B en el cual su surco menor es más ancho y menos profundo [18].

Actualmente se sabe que el DNA también puede adquirir estructuras helicoidales levógiras denominadas Z-DNA constituidas por 12 pb en cada giro [19]. Esto es debido al enlace glucosídico que conecta la base con la posición 1 de la 2-desoxirribosa. Este enlace puede estar en una de las dos conformaciones llamadas "sin" y "anti". En la hélice levógira, la unidad repetida fundamental suele ser un dinucleótido de purina-pirimidina, con el enlace glucosídico en la conformación anti en los residuos de pirimidina y "sin" en los de purina. Esta última, es la causa del estado levógiro de la hélice. El cambio a la posición "sin" en los residuos de purina anti-sin alternantes le imparten a la cadena de DNA levógiro un aspecto de zigzag que lo diferencia de las hélices dextrógiras [18].

### 2.1.1. Estructura del cromosoma

Las bacterias tienen organizado su genoma en dos compartimentos: los plásmidos, en los cuales nos enfocaremos por el momento, y el cromosoma. Los plásmidos son elementos extracromosomales del DNA bicatenario con replicación autónoma. Tradicionalmente se les ha asociado con fenotipos que, bajo ciertas circunstancias, le confiere ventajas selectivas a la bacteria que los posee [20]. El uso indiscriminado de los antibióticos, después de la segunda guerra mundial propició la proliferación de cepas bacterianas mutantes resistentes a estos fármacos. En la década de los cincuentas y principios de los setentas se descubrió que algunos de los plásmidos son, en gran parte, los responsables de la resistencia a los antibióticos. Poco después, se demostró que las cepas resistentes pueden transferir algunos de los plásmidos responsables de cada propiedad a cepas sensibles por medio de la conjugación, es decir por el proceso mediante el cual las bacterias adquieren información genética de otras mediante un contacto célula-célula. Los plásmidos también codifican una enorme variedad de funciones que incrementan la capacidad de las bacterias para colonizar nuevos nichos, y/o para adaptarse a ambientes poco

favorables. Algunos ejemplos son: la resistencia a metales pesados, las adhesinas (proteínas que les permiten adherirse a superficies para una óptima colonización), la resistencia a radiaciones ionizantes, la degradación de los hidrocarburos, la fijación de nitrógeno, etc. [20].

Puesto que una de las cualidades que caracterizan a los plásmidos es su permanencia en las células, uno pensaría que su transmisión está confinada a sus células hijas, es decir, en la transferencia vertical. Sin embargo, por medio de la conjugación algunos plásmidos se pueden transferir de una célula a otra sin importar la especie, fenómeno que se conoce como transferencia horizontal. También se ha demostrado que la movilidad de los plásmidos contribuye de manera importante a la evolución de las especies bacterianas, debido a que pueden conferir una adaptación instantánea a ambientes con cambios constantes [21]. Los plásmidos tienen características propias, es decir, que pueden variar en su forma, en su tamaño y en su número de copias. Aunque la mayoría de los plásmidos, son circulares, también existen muchos plásmidos lineales reportados en géneros tan diversos como *Borrelia, Streptomyces, Thiobacillus, Nocardia, Rhodococcus, Escherichia*, etc [20]. En cuanto a su tamaño podemos encontrar ejemplos, desde pequeños, de sólo unas pocas kilobases, hasta muy grandes como los megaplásmidos de *Sinorhizobium meliloti* de 1.7 y 1.4 Mb (Honeycutt et al. 1993, Margolin y Long , 1993). Algunos incluso llegan a ser del tamao del genoma completo de algunas bacterias, como por ejemplo *Haemophilus influenzae* de 1.8 MB [22], o el de *Mycoplasma genitallium* de 0.58 MB [23]. Otra característica particular de los plásmidos es sin duda su número de copias. Cada plásmido, en condiciones normales de crecimiento controla su número de copias, que puede ser bajo, mediano o alto. Esto se debe a un estricto control en sus sistemas de replicación.

El cromosoma de bacterias, por otra parte, posee pequeñas proteínas básicas (HU y FIS) que pueden compactar el DNA en varios niveles [24]. Muchas especies de arqueas tiene histonas que las ayuda a compactar su cromosoma.

El cromosoma eucariótico, a diferencia del cromosoma de las bacterias, contiene una sola molécula de DNA que se extiende de un extremo a otro del cromosoma a través del centrómero. En la cromatina de la interfase de la divisón celular con sus características de 100 Åde diámetro, este DNA se mantiene en una configuración superenrollada en hélice de modo negativo mediante una acomodación de histonas, proteínas que ayudan a compactar el DNA. La cromatina relajada, es decir la cromatina que está abierta, de la interfase tiene una estructura llamada nucleosoma, que es una estructura

elipsoidal que contiene un centro núcleo muy conservado que consiste en un segmento de DNA de 146 pb que envuelve a un octámero de histonas. Cada octámero consta de dos moléculas de cada una de las histonas H2a, H2b, H3 y H4. La cromatina se condensa aun m´s, durante la mitosis y la meiosis, posiblemente mediante un segundo nivel de superenrrollamiento en fibras de cromatina de 300 Å de diámetro. Las fibras de cromatina de 300 Å enrolladas en forma de hélice o dobladas en la metafase se enrollan en espiral y a su vez en estructuras altamente condensadas visibles en el microscopio óptico.

Los cromosomas eucarióticos contienen secuencias de DNA repetitivo, generalmente localizado en la heterocromatina, ubicada en sitios específicos de los cromosomas, como son las regiones centroméricas o teloméricas.

Durante la mitosis los cromosomas tiene una forma y tamaño característicos, cada uno tiene una región condensada llamada centrómero, que confiere la apariencia general de cada cromosoma. Dependiendo de la posición del centrómero, los brazos tienen longitudes relativamente distintas. Los cromosomas se clasifican en metacéntricos, submetacéntricos, acrocéntricos o telocéntricos de acuerdo con la localización del centrómero y se denomina brazo p, el brazo más largo se encuentra debajo y se denomina brazo q [24].

### 2.1.2. Telómero

Los telómeros son estructuras superenrrolladas que se encuentran en los extremos de los cromosomas eucariotas. Por ello se le han denominado capuchas cromosómicas [25, 26]. El DNA telomérico humano tiene entre 5 y 15 Kb de secuencias repetidas en tándem (TTAGGG) en forma de doble cadena con excepción de la parte final del extremo 3 donde de 100 a 200 nucleótidos son de cadena sencilla. Griffith y De Lange [27] descubrieron mediante microscopía electrónica, unas estructuras en forma de asa, llamadas ciclos o "loops", situadas en la parte final del telómero. Las formación de estas estructuras se debe a una secuencia, rica en guaninas, que permite el plegamiento de las cadenas de complementariedad y a la proteína TRF2 que las mantiene estables. Esto hace posible la formación de dos lazos, uno mayor denominado "t-loop", y otro más pequeño conocido como "d-loop" formada por la inserción de la cadena telomérica sencilla, con una región complementaria de nucleótidos que se encuentran en el DNA telomérico de doble cadena.

El telómero tiene asociadas varias proteínas a lo largo de sus cadenas, esta asociación constituye una estructura única de cromatina no nucleosomal llamada telosoma. Las proteínas TRF1 y TRF2 regulan de manera negativa,

| Proteína | Función |
|---|---|
| TRF1 | Está unida a la doble cadena de repetidos teloméricos TTAGGG, regula negativamente la longitud del telómero. Está presente en el t-loop y promueve el apareamiento de las cadenas teloméricas [27, 28] |
| TRF2 | Está unida a la doble cadena de repetidos teloméricos TTAGGG, regula negativamente la longitud del telómero. Está involucrada en la formación de los loops. La pérdida de su función ocasiona la fusión de los extremos cromosómicos y dispara la senescencia y apoptosis medidas por p53 y ATM [29, 30, 31, 32] |
| POT1 | Está unido a la cadena simple de repetidos TTAGGG y es necesaria para la protección del telómero [33] |
| TIN2 | Está unida a la proteína TRF1, actúa como regulador negativo de la longitud del telómero, reforzando la actuación de TRF1 en el apareamiento de los repetidos telomericos [34] |
| TANK1 (tankyrasa 1) | Inactiva a TRf1 por poli-ADP-ribosilación, provoca el alargamiento del telómero. [35] |
| TANK2 (tankyrasa 2) | Está unida a la proteína TRF1 y su función no está determinada [36] |
| hRAP1 | Está unida a la proteína TRF2. Actúa como regulador positivo del telómero [37] |

situándose en la doble cadena el tamaño del telómero. La proteína Pot1 actúa como cubierta protectora del telómero y se encuentra en el extremo de la cadena sencilla [38].

Hasta el momento se conoce un mayor número de proteínas asociadas al telómero con diferentes funciones se enlistan a continuación.

Los telómeros llevan a cabo las siguientes funciones:

- Protección: evitan la pérdida de genes por degradación o recombinación.
- Estabilización: evitan que los extremos de los cromosomas formen estructuras alteradas o que fusionen extremo con extremo.
- Aseguran la replicación completa de los cromosomas, sin la pérdida de nucleótidos del extremo 5 de cada hebra de DNA.

- Sirven como reloj mitótico que regula el número de replicaciones celulares de los distintos tipos de células [26, 39]

En células somáticas los telómeros disminuyen progresivamente su longitud en cada ciclo de división celular, perdiendo de 50 a 100 pb, a causa de la incapacidad de la DNA polimerasa para sintetizar el telómero. Después de un determinado número de divisiones celulares, el telómero se reduce a un punto crítico. Este acortamiento pudiera eliminar genes indispensables para la vida o silenciar genes cercanos por un fenómeno conocido como efecto de la posición del telómero [40]. En el punto crítico de acortamiento, se manda una señal de paro a la división celular, dando origen a la senescencia celular. Una vez que las células alcanzan la senescencia, éstas son llevadas a la apoptosis [25, 26, 39].

Muchos cromosomas, como los de hongos y algas verdes, son de 1mm o menos de longitud. Los cromosomas de animales y plantas, por otra parte, tienen más de 30mm de longitud. En términos de números de cromosomas *Myrmecia pilosuna* tiene uno y *Ophioglossum reticulatum* tiene 1260 cromosomas diploide.

### 2.1.3. Contenido de genes

Sólo una pequeña fracción de genomas eucariontes está compuesto de regiones codificantes (ej. aprox. 2 % en el genoma humano). En arqueas y bacterias las regiones codificantes están entre el 85 y 95 %. En general hay una correlación entre el tamaño del genoma y el número de genes. El genoma de bacterias se ha encontrado que contiene de 480 genes, como es el caso de *Mycoplasma genitalium*, hasta 8317 en *Bradyrhizobium japonicum*. Esta variación en el número de genes puede reflejar las diferencias en el estilo de vida, un contenido bajo de genes se encuentra en parásitos especializados y simbiontes. Un alto contenido de genes se observa en especies con un ciclo de vida complicado. En Arqueas va en un rango de 563 en *Nanoarchaeum equitans* a 4540 en *Methanosarcina acetivorans*.

### 2.1.4. Composición de pares de bases

En el caso de las bacterias hay una gran variación en el contenido de G+C desde el 22.4 % en *Wiglesworthia glossinidia* a 72 % en *Streptomyces coelicolor*. El contenido de GC esta correlacionado con el tamaño de las bacterias y

puede servir para clasificarlas. Muchas bacterias tiene un alto contenido de AT cerca del término de replicación [41].

El genoma de eucariontes difiere substancialmente en el porcentaje de nucle´tidos consistentes de G+C en el cromosoma. Bernardi y colaboradores han identificado 5 familias de regiones llamadas isocoras en el genoma humano. Cada una de estas regiones se caracteriza por el contenido de G+C. Las regiones de bajo contenido de G+C llamadas L1 y L2, son pobres en genes. Por otra parte regiones con alto contenido en G+C son llamadas H1, H2 y H3 con alto contenido en genes. Otros vertebrados y plantas también tienen la estructura de isocoras. El porcentaje más bajo del genoma es la isocora H3 que es la contiene los genes esenciales del organismo o *housekeeping gene*. El genoma humano contiene aproximadamente un 4.5 % de H3.

Otra interesante marca en los genomas son la densidad de islas CpG. La notación CpG se refiere al que el nucleótido guanina sigue inmediatamente la citosina en la cadena de DNA: 5.CG..3. La p en CpG se refiere la unión fosfodiester que conectan a los nucleótidos adyacentes en la cadena y distingue de las uniones de hidrógeno entre C y G en la cadena complementaria. Las islas CpG son regiones de DNA de 100 nucleótidos de longitud, con un alto contenido en G+C. Cerca de 30000 islas de CpG fueron detectadas en el genoma humano y estas islas correlacionan con la densidad de genes. Otros estudios muestran que la mitad de genes del humano y el ratón son asociados con islas CpG; estas marcas son importantes en la detección de genes [42].

También se ha investigado la distribución de las islas CpG en vertebrados. Se encontró que los genomas de vertebrados de sangre caliente (hombre, ratón y gallina) están caracterizados por abundantes islas de CpG. Esto está en contraste con los genomas de vertebrados de sangre fría (2 reptiles, un anfibio y dos peces) que son escasos o ausentes en islas CpG [43].

El genoma de los vertebrados puede dividirse en dos, el genoma *core* y el genoma *desert*. El genoma core está formado por H2 y H3, con aproximadamente el 12 del genoma total contiene el 50 % de los genes, la densidad de genes es alta, el CpG es alto, la metilación es alta, los intrones son cortos y la cromatina es abierta. El genoma desert está formado de L1, L2 y H1, tiene aproximadamente el 83 % del genoma, tiene el 50 % de los genes, la densidad de los genes es baja, el CpG es bajo, los intrones son largos y la cromatina es cerrada.

### 2.1.5. Secuencias repetidas

En genomas bacterianos se han clasificado diferentes categorías con variaciones considerables en el tamaño (pocos pares a kb), composición (directos o invertidos) y localización (agrupados o dispersos). Algunos ejemplos incluyen varias familias de repeticiones palindrómicas [44], repeticiones en tándem y genes duplicados. También se ha encontrado que los elementos repetidos tienen un papel importante en la compactación del DNA en arqueas.

El genoma de eucariontes tiene un gran porcentaje de secuencias repetidas: el conjunto de repeticiones en tándem de tipo satélite comprende un total de 250Mb del genoma humano. Son secuencias de entre 5 y varios nucleótidos que se repiten en tándem miles de veces generando regiones repetidas con tamaños que oscilan entre 100kb hasta varias megabases.

Las secuencias satélites tienen una riqueza en nucleótidos A+T superior a la media del genoma y en consecuencia son menos densas.

Hay principalmente 6 tipos de repeticiones de DNA satélite en el humano:

1. Satélite 1: es una secuencia básica de 42 nucleótidos y está situado en los centrómeros de los cromosomas 3 y 4.

2. Satélite 2: la secuencia básica es ATTCCATTCG. Está presente en las proximidades de los centrómeros de los cromosomas 2 y 10.

3. Satélite 3: la secuencia básica es ATTCC. Está presente en los cromosomas 9 e Y.

4. Satélite alfa: secuencia básica de 171 nucleótidos. Forman parte del DNA de los centrómeros cromosómicos.

5. Satélite beta: es una secuencia básica de 68 nucleótidos y aparece en torno al centrómero.

6. Satélite gamma: es una secuencia básica de 220 nucleótidos.

#### Minisatélites

Están compuestas por una unidad básica de secuencia de 6 a 25 nucleótidos que se repiten en tándem generando secuencias de entre 100 y 20000 pares de bases. Se estima que el genoma humano contiene unos 30000 minisatélites.

Diversos estudios han relacionado los minisatélites con procesos de regulación de la expresión génica, como el control del nivel de transcripción.

Aproximadamente el 90 % de los minisatélites, en el genoma humano, se sitúan en los telómeros de los cromosomas. La secuencia básica de seis nucleótidos TTAGGG se repite miles de veces en tándem generando regiones entre 5 kb y 20 kb que conforman los telómeros.

### Microsatélites

Están compuestos por secuencias básicas de 2-4 nucleótidos, cuya repetición en tándem origina frecuentemente secuencias menores a 150 nucleótidos. Algunos ejemplos son los dinucleótidos CA y el trinucleótido CAG. Se estima que el genoma humano contiene unos 200000 microsatélites que, al contrario de los minisatélites, se distribuyen homogéneamente.

### DNA repetido disperso

Son secuencias de DNA que se repiten de modo disperso por todo el genoma constituyendo el 45 % del genoma humano. Los elementos cuantitativamente más importantes son los LINEs y SINEs, que se distinguen por el tamaño de la unidad repetida.

### SINE

Elementos nucleares dispersos cortos. *short intersprersed nuclear elements*. Son secuencias cortas, generalmente de unos cientos de bases que aparecen repetidas miles de veces en el genoma humano. Los elementos Alu son secuencias de 250-280 nucleótidos presentes en 1500000 copias dispersas en el genoma humano. Estructuralmente son dímeros casi idénticos, excepto que la segunda unidad contiene un inserto de 32 nucleótidos, siendo mayor que la primera. En cuanto a su secuencia tiene una considerable riqueza en G+C (56 %).

### LINE

Elementos nucleares dispersos largos. *long interspersed nuclear elements*. Constituyen el 20 % del genoma humano. La familia de mayor importancia cuantitativa es LINE-1 o L1 que es una secuencia de 6kb repetidas unas 800000 veces, de modo disperso, por todo el genoma.

Los elementos LINE completos son codificantes y codifican dos proteínas:

1. Proteína de unión a ARN.
2. Enzima con actividad retrotranscriptasa y endonucleasa.

Diversos estudios han mostrado que las secuencias LINE pueden tener importancia en la regulación de la expresión génica, habiéndose comprobado que los genes próximos a LINE presentan un nivel de expresión inferior. Esto es especialmente relevante porque aproximadamente el 80 % del genoma humano contiene algún elemento L1 en sus intrones. Su riqueza en G+C es del 42 %.

### 2.1.6. Recombinación *hotspots* y secuencias repetidas

La recombinación entre extensiones homólogas, pero no alélicas, de DNA (familias de genes, duplicaciones y elementos repetidos) es una fuente importante de mutaciones (y mutaciones patológicas).

En humanos se han identificado motivos cortos de DNA que determinan la localización de 40 % de los hotspots de recombinación (*cross-over*) meióticos y que están significativamente enriquecidos en los puntos de quiebre (*breakpoints*) de los síndromes de recombinación alélica no homóloga (NAHR). La forma más penetrante del motivo ocurre dentro de una familia de elementos repetidos inactiva (THE). El motivo también tiene una actividad recombinogénica fuerte en familias de elementos activos incluyendo Alu y LINE2. En humanos se ha evidenciado que en algunos casos la recombinación puede ser promovida en ciertos tipos de DNA repetido (repeticiones de DNA). En la mayoría de los eucariontes, y también los humanos, los hotspots ocurren cada 50  100 kb y producen un evento de recombinación cada 1300 meiosis aproximadamente [45]. Se ha demostrado también que la NAHR puede ser una fuente mayor de mutaciones patogénicas en los genomas eucariontes [46]. En Drosophila el DNA repetido, incluyendo los elementos transposables (ETs) y las repeticiones cortas asociadas con telómeros y centrómeros típicamente ocurre en las regiones de recombinación reducida [47, 48].

Los esfuerzos recientes para mapear la estructura de la variación de la tasa de recombinación a escala fina en humanos han mostrado un escenario más complejo con la presencia de elevadas tasas de recombinación en algunos elementos repetidos [45]. De hecho, hay evidencia de que en humanos la

recombinación puede en algunos casos, incluso ser promovida en ciertos tipos de repeticiones de DNA y se ha especulado acerca de los procesos evolutivos que han conducido a esta situación [46]. Hay análisis que han demostrado influencias sistemáticas en las posiciones de los hotspots de recombinación. Esto incluye una tendencia a agruparse junto a regiones promotoras pero, activamente evitan regiones transcripcionales [46]. Las comparaciones entre los patrones estimados de esperma y de datos de variación sugieren que los hotspots coinciden típicamente en hembras y machos [49].

Se han realizado estudios que han identificado 2 secuencias en los hots spots: el 7mero CCTCCCT y el 9mero CCCCACCCC como hotspots enriquecidos. Trabajos subsecuentes con el 7mero extendieron el motivo a un 13mero degenerado, CCNCCNTNNCCNC, al que en adelante se hará referencia como el motivo 13mero. La actividad del motivo 13mero degenerado es influida por su contexto así como por las bases presentes en los sitios degenerados.

El "motivo 13mero núcleo" CCTCCCTNNCCAC, en el contexto de un elemento repetido THE1B conduce a un hotspot detectable el 70 % de las veces [45]. La presencia de del motivo degenerado, para los elementos THE1, conduce a un aumento de 10 veces en el promedio de la tasa de recombinación, comparado con el modesto aumento del doble o triple en los elementos L2 y tres familias activas Alu (AluY, AluSg y AlusX).

La presencia del motivo hotspot en estos contextos tan penetrantes determinan aproximadamente el 10 % de todos los hotspots. En contraste, la presencia del motivo núcleo en DNA únicamente conduce a un hotspot solamente el 10 % de las veces y explica sólo el 1.3 % de todos los hotspots [45]. También se han encontrado polimorfismos en los hotspots estudiados. Entre éstos algunas mutaciones que co-segregan con la actividad del hotspot e interrumpen, ya sea el motivo 13mero o un subconjunto del núcleo del motivo 13mero [50, 51]. Todos estos resultados evidencian que el motivo 13meroes un determinante esencial de la actividad en 40 % de los hotspots. Además de los casos de recombinación alélica, se ha sugerido que el motivo 13mero también tienen un papel importante en los síndromes de NAHR recurrente y otras formas de inestabilidad genómica. En la NAHR los eventos de recombinación entre secuencias parálogas producen duplicación, pérdida u otros rearreglos complejos que interrumpen la actividad génica. Existen 6 tipos comunes de desórdenes de NAHR que revelan la presencia del motivo 13mero en repeticiones de copias cortas [45]. El motivo 13mero también ocurre en los *breakpoints* de la deleción común del DNA mitocondrial, una

mutación somática recurrente de 5 kb asociada con síndromes específicos, pero que también sucede con el envejecimiento [45]. Estos hallazgos establecen una relación entre los eventos programados de formación de rupturas de doble cadena (DSB) y la recombinación (conversión de genes y recombinación) en la meiosis. Frecuentemente cuando no son programados producen hipermutabilidad y patogenia. En cada meiosis ocurren en promedio decenas de DSBs en DNA repetido y el 10 % de los hotspots serán producidos por combinaciones de motivos altamente penetrantes de DNA repetido.

Se han descrito varios desórdenes genómicos por eventos de deleción y duplicación mediados por repeticiones y NAHR que producen fenotipos deletéreos [52, 53]. Los *breakpoints* de duplicación segmental y de regiones variables de regiones variables poco comunes de *copy-number* (particularmente deleciones) son enriquecidos por los elementos repetidos. Esto implica que de los elementos de repetición corta que conforman el 40 % del genoma humano, son raramente suficientes para generar NAHR meiótica que produzcan fenotipos patogénicos por sí mismos. En humanos se han identificado cerca de 500 eventos de deleción específica que puedan ser atribuidos a la ilegítima recombinación entre elementos Alu [54]. De manera similar NAHR, entre elementos LINE1, han sido propuestos como responsables de 55 deleciones específicas en humanos [4]. Es importante resaltar que los rearreglos génicos (duplicación, inversión, deleción) solamente ocurrirán por NAHR; los DSBs son procesados como eventos de la recombinación. En contraste, los eventos de conversión génica producirán cambios en la repetición con poca o ninguna consecuencia funcional a una tasa mucho mayor. Se han utilizado métodos de genética de poblaciones para la detección de la tasa de recombinación y conversión génica entre secuencias alélicas [46]. Aplicando estos métodos indican que los eventos de conversión no alélica tienen un papel importante en la historia de las secuencias repetidas (THE, Alu) [46]. La evidencia combinada sugiere que NAHR entre los miembros de las familias de DNA repetido disperso y arreglado en tándem, podría representar una fuente substancial de mutación en el genoma humano, quizá incluso a nivel comparable con la mutación puntual. Las evidencias sugieren que la NAHR ocurre entre miembros de familias de repeticiones cortas, que muestran una identidad de secuencia típicamente en el rango de 80-95 %, y que no necesariamente están contenidas en regiones de alta identidad de secuencia. A partir de esto surge la pregunta de por qué la maquinaria de recombinación no ha evolucionado hacia evitar la iniciacin de DSBs dentro de las secuencias repetidas?

Surgen tres posibles respuestas:

1. Puede existir la selección de NAHR entre las repeticiones pero no hay un mecanismo alternativo disponible. Esta explicación parece poco probable dado que muchos, quizás la mayoría de los hotspots en humanos (y otros eucariontes) no ocurre en DNA repetido.

2. Puede haber cierto beneficio selectivo para el DNA repetido de una asociación con los hotspots, por ejemplo, al permitirle propagarse por el genoma. Sin embargo esta hipótesis parece poco probable ya que el contexto de motivo/repetición más recombinogénico se dá en los elementos inactivos THE1; Otros elementos (v.g. LINE1) muestran fuerte supresión local de recombinación a pesar de ser activas.

3. La asociación de los hotspots con el DNA repetido es producida por otros factores y aunque genera una carga mutacional, ésta es suficientemente pequeña. Por ello la selección por un mecanismo alternativo es débil.

¿Pero qué proceso evolutivo podría "conducir" a la maquinaria de recombinación a una en la que una fracción sustancial de DSBs ocurran en DNA repetido? Una posible pista es la sorprendentemente rápida evolución vista en los hotspots de los primates. Entre humanos y chimpancés parece que comparten poco o nada de hotspots de recombinación [55, 56]. Es bien sabido que los hotspots de recombinación contienen un inherente impulso autodestructivo que puede promover la rápida evolución del sistema. Específicamente el polimorfismo que actúa en cis dentro de un hotspot puede influir la tendencia de una cromátide a experimentar DSBs. Esto ya se ha visto en machos heterócigos para mutaciones que abolen el motivo corto en el hotspot DNA2 [51]. En esos casos el alelo más "caliente" será el que típicamente experimente la formación de DSBs y será reparado por, consecuentemente convertido a el alelo "frío" [57, 51, 58, 59]. Esto impulsa a una dirección lejos del alelo caliente a nivel poblacional.

Este fenómeno condujo a la "paradoja del hotspot" que pregunta: ¿Por qué este impulso no ha eliminado todos los hotspots? [57]. Una solución a esta paradoja es que mientras el número de hotspots decrece, hay una ventaja selectiva en aumento producida por la falla al alcanzar la disyunción meiótica apropiada para un cambio de la maquinaria a reconocer un nuevo

motivo primario y activar un nuevo conjunto de hotspots [46]. Así los motivos hotspot deben ser comunes, propagados por todo el genoma, no tener alguna otra función (v. g. sitios de unión a factores de transcripción) y, por lo menos en base a los datos obtenidos en humanos, eludir genes [46]. Se sugiere que existe fuerte presión selectiva en la maquinaria de recombinación para cambiar que puede producir mecanismos para formación de hotspots en el DNA repetido [46].

### 2.1.7. Secuencias palindrómicas

Recientemente se ha descubierto que varias translocaciones constitucionales, especialmente aquéllas que involucran al cromosoma 22, se llevan a cabo utilizando secuencias palindrómicas en 22q11 y en el cromosoma hermano. El análisis en los fragmentos de empalme de translocación muestra que el punto de ruptura de dichas translocaciones, mediadas por palíndromas, se localiza en el centro de repeticiones palindrómicas ricas en AT (PATRRs). La identificación de estas translocaciones mediadas por PATRR sugiere una ruta universal de rearreglo burdo del genoma humano. Las ocurrencias de las translocaciones mediadas por PATRR pueden ser detectadas por PCR en muestras de esperma normal, pero no de células somáticas. El polimorfismo de varias PATRR influye en su propensión a adoptar una estructura secundaria que en a su vez afecta la frecuencia de la translocación de novo. Se propone que las PATRRs forman una estructura secundaria inestable que conduce a rupturas en el centro de éstas. Las rupturas de la doble hebra parecen ser seguida por una vía de reparación, que une los extremos finales, produciendo finalmente una translocación. Anteriormente se pensaba que la mayoría de los rearreglos genómicos se formaban aleatoriamente, pero los datos más recientes indican que no es el caso. La región 22q11 es un hotspot para reareglos cromosómicos no aleatorios. Las deleciones, duplicaciones y translocaciones en 22q11 ocurren en más de 1/3000-4000 nacimientos [60]. Los rearreglos en están relacionados con defectos congénitos del desarrollo. Por otra parte la mayoría de las deleciones intersticiales y duplicaciones en 22q11 son atribuidas a la presencia de una estructura genética específica de cromosomas de repetición de bajas copias (LCR). La recombinación desigual entre segmentos de DNA de secuencia similar contenidos dentro de los LCRs o recombinación homóloga no alélica puede producir dichas duplicaciones y deleciones cromosomales [61] . La mayoría de las translocaciones constitucionales que involucran a 22q11 comparten el mismo punto de rup-

tura LCR 22q11.2 localizado dentro de LCR-B. Estos puntos de ruptura y translocación del cromosoma 22 se localizan en el centro de una secuencia palindrómica rica en AT. Los puntos de ruptura del cromosoma hermano (11q23, 8q24.1, 17q11.2, 1p21.1 y 4q35.1) también yacen en el centro de una secuencia palindrómica cuya longitud abarca varios cientos de pares de bases. Se ha invocado un mecanismo mediado por palíndromas para la generación de esas translocaciones constitucionales [62, 63, 64, 65, 66, 67, 68]. Por lo tanto se ha sugerido la translocación mediada por palíndromas como una de las vías universales para los rearreglos cromosomales humanos.

El análisis detallado de la configuración genómica de los cromosomas 11 y 22 ha sido difícil porque la secuencia tiene una configuración altamente inestable, representando un hotspot para recombinación y deleción en bacterias, levaduras y mamíferos [69, 70, 71].

Se utilizó una estrategia bastante permisiva de reacción en cadena de la polimerasa PCR, secuenciación por RNA polimerasa y clonación de células de *E. coli* deficientes de recombinación para lograr el genotipo PATRR [72]. Con estos métodos se pudieron elucidar tres regiones de *breakpoints* asociadas a translocación en las que se identificaron estructuras palindrómicas con secuencias altamente ricas en AT. La secuencia palindrómica que contiene dos motivos consensos cabeza a cabeza adyacentes el uno al otro, es teóricamente capaz de formar una estructura de horquilla de cadena sencilla o una estructura cruciforme de doble cadena. Se dice que la inestabilidad de las secuencias palindrómicas es mediada primariamente por su propensión a formar estructuras secundarias [71, 72, 73]. Incluso se ha observado directamente la extrusión cruciforme del plásmido PATRR usando microscopía de fuerza atómica [72]. Es probable que la formación de dicha estructura secundaria inusual pudiera inducir inestabilidad genómica induciendo translocaciones. A pesar de que hay numerosas secuencias palindrómicas en el genoma humano [74], hasta la fecha sólo se han identificado pocas translocaciones mediadas por palíndromas. De hecho, se ha demostrado anteriormente [71] que PATRR11 manifiesta polimorfismos en tamaño como un resultado de la deleción dentro de PATR y que éste influye la frecuencia de la translocación de novo en esperma [75] (la síntesis de novo se utiliza para medir la frecuencia de translocación). Existiendo una relación dependiente del tamaño: a mayor longitud, más síntesis de novo y a menos longitud menor síntesis de novo. El tamaño y la simetría de las PATRRs también refleja una propensión a la formación de estructuras secundarias, que es probablemente un rasgo que aumenta la frecuencia de recombinación. Esto también se ha observado por

microscopía de fuerza atómica [76]. La identificación de una translocacin mediada por palíndromas que no involucra a 22q11 proporciona evidencia de que las translocación mediada por palíndromas representa un mecanismo general de rearreglo. Estos hallazgos respaldan la hipótesis de que la translocación mediada por PATRR representa una vía universal para los rearreglos cromosomales burdos. A partir de estos resultados se forma la hipótesis de que las PATRRs forman estructuras secundarias, que a su vez inducen translocaciones durante la gametogénesis; especialmente la espermatogénesis.

Las PATRRs deberían ser capaces de formar estructuras de horquilla dentro de las regiones de DNA de hebra sencilla en la cadena rezagada durante la replicación del DNA [71, 77, 78, 79, 80]. Si la translocación ocurre durante la meiosis la estructura cruciforme de 4 lados es análoga al empalme de Holliday en la recombinación genética homóloga [71]. Si la translocación se lleva a cabo después de la meiosis, durante la etapa de espermátide, la acumulación de DNA súper enrollado forma una estructura cruciforme que probablemente extrudirá [81, 82, 83]. Esta conformación inusual de DNA podría actuar como blanco potencial de nucleasas específicas de estructura y contribuir a las translocaciones mediadas por palíndromas.

En humanos las anormalidades cromosómicas frecuentemente muestran una inclinación dependiente del género de origen parental. Las anormalidades estructurales ocurren predominantemente en la línea germinal paternal. Esto se debe al gran número de divisiones celulares que ocurren durante la espermatogénesis. De ocurrir las translocaciones por errores de replicación se esperaría una relación positiva entre la edad paternal y las translocaciones de novo. Sin embargo, la frecuencia de las t(11:22) parece no aumentar con el aumento en la edad paternal. Esto indica que estas translocaciones no ocurren vía un mecanismo dependiente de replicación premeiótico. Algunas translocaciones constitucionales recurrentes ocurren vía NAHR durante la recombinación genética. Adicionalmente, los *breakpoints* de la translocación t(11:22) no coinciden con *hotspots* de translocación, por lo que parece que las vías de recombinación meióticas no están involucradas en el mecanismo de translocación. Otros factores pueden influir en las translocaciones mediadas por PATRR. Se ha mostrado que la organización espacial de los cromosomas no es aleatoria y es un factor que contribuye en la formación de translocaciones somáticas específicas. La proximidad espacial de los cromosomas 11 y 22 durante la meiosis puede jugar un papel en la generación de la translocación recurrente; por hibridación fluorescente in situ FISH se ha demostrado que la proximidad de 11q23 y 22q11 es mayor que la de regiones control.

### 2.1.8. Metilación y epigenética

La complejidad del DNA no sólo se resume en sus cuatro bases nitrogenadas, sino que también estas mismas bases pueden presentarse en el genoma de manera modificada por medio de mecanismos epigenéticos, es decir no genéticos.

La metilación del DNA es un fenómeno epigenético clave. Los fenómenos epigenéticos se definen como aquellos cambios hereditarios que repercuten en la función de un gen que no pueden ser explicados por la secuencia del DNA [84]. Esta es encontrada en sitios específicos del DNA: mientras que en algunas especies ocurre en la posición N6 de los residuos de adenina, en otras ocurre en la posición N4 y en el carbono 5 de los residuos de citosina. Esta última es mejor conocida como 5-metilcitosina ($m^5C$). La metilación del DNA se presente en diferentes formas a lo largo de casi todos los organismos. La $m^5C$ es la modificación que se presenta de manera común en eucariontes, es decir, del 1 al 8 % de las citosinas de las células de eucariontes superiores (incluyendo mamíferos) están metiladas [85, 86, 87]. Las $m^5C$ se localizan en el surco mayor del DNA de tal forma que no interfiere con el apareamiento de bases según el modelo de Watson y Crick [88]. Esta modificacin ocurre preferencialmente en el dinucleótido C-G, donde la citosina metilada es seguida en su lado 3 por un residuo de guanina; la metilación tiene como blanco preferencial lo sitios CpG aunque también se observa, con menor frecuencia, en citosinas no-CpG. Estructuralmente la metilación sucede de manera bimodal o palindrómica dando la siguiente estructura:

5 mCpG 3,

3 GpCm 5.

Este tipo de estructura es conocida como completamente metilada. Cuando sólo una hebra de DNA está metilada se conoce como parcialmente metilada o hemimetilada, estructura común en la replicación del DNA, donde el mantenimiento de la metilación depende de que ocurra con el DNA hemimetilado. Cuando se presenta la hemimetilación, el sitio no metilado de la hebra complementaria será restaurado a la condicin completamente metilada, y ésta será a su vez transmitida por medio de la hebra hija en la siguiente división. Este proceso es estable y en raras ocasiones sucede un error que altere el patrón de la metilación. La densidad de grupos $m^5C$ es muy variada a lo largo de los fila existiendo ejemplos extremos de patrones de metilación, como ocurre en algunos nematodos, como *Caenorhabditis elegans* cuyo genoma carece por completo de grupos $m^5C$ [89]. En *Drosophila melanogaster* se ob-

serva una densidad de metilación muy baja o casi nula. Esto a diferencia de los genomas de la mayoría de los invertebrados que poseen niveles moderadamente altos de $m^5$CpG concentrados en dominios largos de DNA metilado, separados por dominios de longitud equivalente no metilados. En el extremo opuesto están los vertebrados, que poseen la densidad más alta de $m^5$C en el reino animal. Dicha variedad de densidades en el reino animal, subraya la posibilidad de que la distribución a lo largo del genoma refleja una diferente función para el sistema de metilación [86, 89].

### 2.1.9. Función de la metilación de DNA

La funciones biológicas de la metilación del DNA son fundamentalmente diferentes entre eucariontes y procariontes. En las bacterias, la metilación juega un papel central en la restricción del hospedero a los fagos de DNA. En eucariontes no existe evidencia convincente de que éste sea un mecanismo de restricción de virus. En bacterias la metilación ocurre principalmente en residuos de adeninas en varios taxa. También se sabe que es importante durante la reparación del mal apareamiento de hebras; este fenómeno no ocurre en eucariontes, como tampoco la metilación de adeninas [90]. En eucariontes superiores, en especial en mamíferos, la metilación tiene varias funciones. Una de las más importantes es la represión de la expresión génica. Un evento asociado a la represión, y sin lugar a dudas una de las características más relevantes del patrón de metilación de vertebrados, son las Islas CpG. Como habíamos visto anteriormente las islas CpG son regiones ricas en G+C ($> 55\,\%$) que contienen una gran cantidad de sitios CpG, las cuales, normalmente se localizan en las terminaciones 5 de los promotores de muchos genes humanos [91]. Actualmente se tienen detectadas alrededor de 29 000 islas CpG en el genoma humano, localizándose en el 76 % de los promotores, y tienen una longitud de 0.4 a 3 Kb [90, 91]. Generalmente las islas de los promotores de los genes se encuentran desmetiladas en los estadios tempranos de desarrollo, en todos los tipos de tejidos. Incluso se ha visto que estas siguen desmetiladas aún cuando su gen asociado está apagado. Pero cuando estas islas se encuentran metiladas se convierten en represores génicos, y el gen que tiene asociado queda establemente silenciado durante el desarrollo. Sin embargo, una fracción significativa de las islas CpG en seres humanos es propensa a una metilación progresiva de ciertos tejidos durante el envejecimiento, lo cual sucede también en algunas células cancerosas y también en las líneas celulares inmortalizadas. Este fenómeno sucede de manera natural

pero a un ritmo muy lento, y está asociado a la metilación de novo que se comentará más adelante [91]. El patrón de metilación es removido en los espermatocitos y durante las primeras fases de la embriogénesis, posiblemente para evitar el silenciamiento de genes importantes durante el desarrollo y mantener así un genoma totipotencial, cuya diferenciación en parte está regulada por el patrón de metilación que se va formando por acción de las DNA metiltransferasas de novo. El patrón de metilación es diferente según el tipo celular, y la conservación de éste es llevada a cabo por la metiltransferasa de mantenimiento, pero dicha conservación no es perfecta y a lo largo de los años el patrón de metilación es susceptible a cambios, lo cual se ha propuesto como una posible causa de la transformación maligna [90, 92, 93]. La represión génica de las islas CpG está asociada a otra de las funciones de la metilación del DNA que es la impronta génica. El estado de metilación cambia en los promotores que son sujetos a un control tejido específico. Si está el gen metilado su transcripción se inhibe, y viceversa. Ambas, la metilación y desmetilación del DNA, ocurren durante la embriogénesis fundamentalmente. Por lo que todas las diferencias al´licas son perdidas en el desarrollo de las células germinales primordiales en el embrión, independiente del sexo, los patrones previos de metilación son borrados. Dicho patrón se restablece en estadios tardíos de la embriogénesis e impuesto diferencialmente según el sexo. Este patrón y expresión diferencial, dependiente de si el gen es heredado por el padre o la madre, se conoce como impronta génica y establece una diferencia en el comportamiento de los alelos heredados de cada uno de los progenitores; es decir, por medio de una metilación monoalélica en $^{5}$mCpG se inactiva en la expresión de un alelo.

Existe evidencia de que los niveles de metilación del DNA están conectados a una amplia organización de la cromatina, donde el DNA no metilado usualmente da lugar a la eucromatina, que es la región codificante del cromosoma o activa para la expresión génica. Por otra parte, las regiones altamente metiladas dan lugar a otra conformación llamada heterocromatina, la cual es inactiva y altamente condensada, pero no de manera permanente, ya que cuenta con mecanismos de relajación para poder replicarse. La evidencia actual sugiere que esto sucede no por la metilación *per se*, sino más bien por una gama de proteínas que se unen a los sitios $^{5}$mCpG, proteínas conocida como proteínas de unión a 5-metilcitosina (MBD). Dichas interacciones contribuirán a la inhibición de la transcripción de promotores metilados. La metilación es un importante portador de información epigenética, sin embargo, el hecho de que ocurra en m$^{5}$C es paradójico, ya que la citosina se encuentra

metilada en el carbono 5 le confiere la desventaja crucial de ser más propensa a mutaciones, esto se debe a que la desaminación de esta base modificada resulta en una transición mutante de mC− >T. Mientras que la desaminación de la citosina no metilada resulta en la transición a uracilo, el cual es reconocido y reparado eficientemente por la vía uracil-deglicosilasa [94]. Lo anterior da como resultado que los sitios $m^5C$ sean sitios calientes (hot spots) para mutaciones, que por sí solos representan el 30 % de las mutaciones puntuales en la línea germinal y de las mutaciones somáticas adquiridas.

### 2.1.10. Cromatina

La eucromatina es una forma de cromatina que se condensa durante la interfase, se encuentra presente la histona 4 (H4) acetilada, mientras que la heterocromatina es definida como regiones de cromatina que permanecen condensadas y densamente teñidas a lo largo del ciclo celular [95]. Existen dos tipos de heterocromatina: la hetecromatina constitutiva y la facultativa. La primera se encuentra en regiones como la que rodea a los centrómeros, que es llamada hetecromatina pericentromérica, la telomérica y en la región de organización nucleolar. Se sabe que este tipo de cromatina se duplica tardíamente durante la fase S, y que forma regiones enriquecidas con secuencias repetitivas, donde la presencia de la H4 acetilada está muy reducida [86]. A diferencia de la heterocromatina constitutiva, la facultativa es regulada en el desarrollo. Fue definida por Brown en 1966, quien observó que los cromosomas funcionales de *Sciara, Coccidos*, y otros insectos, eran heterocromatinizados para la diferenciación y desarrollo de los machos. En el caso de los mamíferos, la heterocromatina facultativa es observada en la inactivación del cromosoma X (Xi) en hembras; parece que la función de este proceso va hacia la regulación de la dosis de productos de genes del cromosoma X. El Xi tiene muchas similitudes con la heterocromatina constitutiva, es replicado tardíamente en la fase S, retiene una heterocromatina condensada durante la interfase, llamada cuerpo de Barr, y la H4 está hipoacetilada [86]. La heterocromatina tiene varios componentes, el primero en detectarse fue la proteína asociada a la heterocromatina (Hp1), la cual está altamente conservada en los organismos. En mamíferos se han encontrado tres proteínas Hp1. Se cree que estas proteínas están involucradas en el silenciamiento de genes, se encuentra a lo largo del genoma y están conservadas desde la levaduras hasta los seres humanos. Otras modificaciones de la heterocromatina es la metilación de la H3 en la lisinas 9 (meH3K9), la cuál es una de las modificaciones epi-

genéticas más robustas, y parece que es permanente en la naturaleza, debido a que se encuentra en todos los organismos. La heterocromatina pericentromérica es un tipo de ejemplo de heterocromatina constitutiva. Esta posee funciones potencialmente importantes para la segregación cromosómica y la expresión de genes [95]. La heterocromatina pericentromérica está constituida por varias regiones repetidas ricas en A+T, incluyendo elementos de DNA transponibles, y repeticiones satélites, estos satélites en el caso de los ratones son llamados satélites mayores, en el caso de los seres humanos son los satélites 2 y 3 [96]. Debido a su yuxtaposición a la región centromerica, no es de sorprenderse que la organización de las regiones pericentroméricas sea vital para asegurar la correcta segregación cromosómica, y el mantenimiento de la estabilidad genómica. Estas regiones han sido implicadas en el silenciamiento de genes que ocurre cuando genes eucromatínicos se localizan en la heterocromatina adyacente como resultado de rearreglos y transposición; este fenómeno fue originalmente reportado en *D. melanogaster* y se le llamó variegación posición-efecto, actualmente se implicado este fenómeno a la HP1.

## 2.2. Análisis estadístico de secuencias genómicas

Una vez que fueron secuenciados los primeros genomas, se iniciaron estudios sobre la estructura de los mismos. Estos trabajos, básicamente de carácter estadístico, [97, 98, 99] se enfocaron en especial a la investigación de las propiedades elementales de las secuencias, tales como la frecuencia de aparición de los nucleótidos y de ciertos dímeros, en particular el contenido de citosina y guanina. En la tabla 2.1 se muestra el número de publicaciones que estudian correlaciones en el DNA, que han aparecido en las diferentes décadas; una gráfica del número acumulado de artículos se da en la Fig. 2.1. Estos estudios estadísticos dieron paso a modelos para explicar la composición y evolución de los genomas.

También se han hecho modelos que tratan de explicar la evolución de genomas como por ejemplo el trabajo de Wentian Li donde propuso en 1989 un modelo que se le conoce como duplicación-mutación (o expansión-modificación). Este modelo permite la duplicación de nucleótidos y la mutación aleatoria de algunos de los duplicados a través de las generaciones y es

| Periodo | Num. de Artículos publicados |
|---|---|
| 50's | 3 |
| 60's | 4 |
| 70's | 9 |
| 80's | 70 |
| 1990-1995 | 156 |
| 1995-2000 | 394 |
| 2001-2005 | 815 |
| 2006-2011 | 541 |

Cuadro 2.1: Número de publicaciones que han aparecido sobre correlaciones en DNA; los datos fueron tomados de la pág. del Prof. Wentian Li, http://www.nslij-genetics.org/dnacorr/

de los primeros en genómica evolutiva que predecía la existencia de correlaciones a largo alcance en las secuencias de DNA. De este modelo hablaremos brevemente a continuación.

## 2.3. Modelo de duplicación-mutación

Sea $\Omega$ el conjunto de todas las cadenas binarias infinitas. Esto es, si $x \in \Omega$ entonces

$$x = \cdots a_{-2}a_{-1}a_0a_1a_2 \cdots \tag{2.1}$$

donde $a_i \in 0, 1$ con $i = 0, \pm 1, \pm 2, \ldots$ Ahora sea $\vartheta : \Omega \to \Omega$ una transformación que toma una cadena $x$ de $\Omega$ en otra cadena $\tilde{x} = \vartheta(x)$ que también está en $\Omega$. Cada dígito binario de $x$ evoluciona por medio de $\vartheta$ de la siguiente manera,

$$0 \to \begin{cases} 1 & \text{con probabilidad} \quad p \\ 00 & \text{con probabilidad} \quad 1-p, \end{cases} \tag{2.2}$$

y

$$1 \to \begin{cases} 0 & \text{con probabilidad} \quad p \\ 11 & \text{con probabilidad} \quad 1-p. \end{cases} \tag{2.3}$$

La transformación anterior expande la longitud de la cadena en cada posición con probabilidad $1 - p$ y la modifica con una probabilidad $p$. El proceso de expansión tiende a aumentar la correlación entre los dígitos binarios, mientras

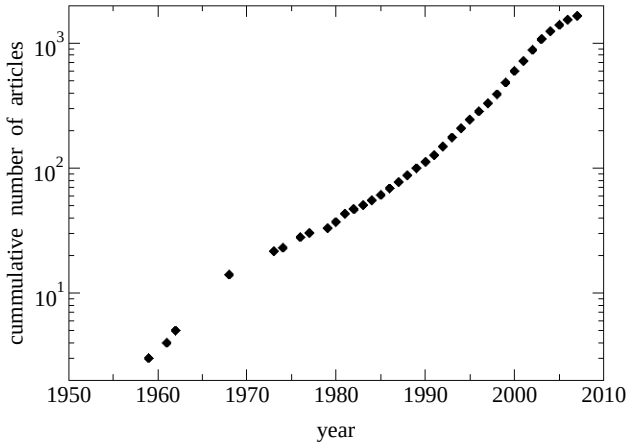



Figura 2.1: Número de artículos de correlaciones de DNA producidos desde los 50's hasta la actualidad.

que el de la mutación, de un dígito binario en otro, tiende a destruirla. Si se hacen algunas aproximaciones la función de autocorrelación se comporta como una ley de potencia con exponente $\varepsilon$. El Prof. Li encontró una relación analítica entre el exponente

$$\varepsilon = -\frac{\log\left[\sum_i \lambda\left(\frac{d_0}{k \pm i}, d_0\right)\right]}{\log(k)}, \tag{2.4}$$

y la probabilidad de mutación $p$ antes descrita. Aquí $d_0$ es la distancia más probable a la que dos elementos de la cadena se separan, $k$ es el factor de dilatación promedio de toda la cadena y $\lambda(d_1, d_2)$ es el valor de la matriz de transición de estados de la distancia $d_1$ a la distancia $d_2$. Este modelo capta los ingredientes básicos de la dinámica mutacional en las secuencias de DNA.

El Dr. Li fue el primero en demostrar la existencias de estas correlaciones de largo alcance en secuencias reales de DNA. Un poco después el grupo de Boston encontró correlaciones similares con otras metodologías (ver abajo). Ya que la duplicación genera redundancia y la correlación local aumenta mientras que la mutación tiende a borrarla. Sin embargo esta correlación local se propaga a grandes distancias, sobretodo si la probabilidad de duplicación es grande. El tipo de autocorrelación que se encuentra en todas las secuencias del DNA hasta ahora analizadas es de un tipo especial que se llama ruido rosa (o ruido $1/f$) en contraste con el llamado ruido blanco o aleatorio en el que no

existe relación alguna entre los nucleótidos de una secuencia de DNA. El ruido $1/f$ tiene propiedades también de autosemejanza, esto es, existe la misma estructura de las secuencias a distintas escalas de observación (fractal). Se ha encontrando este ruido en todos los análisis de genomas completos de organismos desde las bacterias, levadura y plantas, hasta el hombre [100, 101]

A principios de la década de los 80's, Benoit Mandelbrot [102] mostró de que diversas estructuras encontradas en la naturaleza, como las nubes, los continentes o los árboles, presentan una geometría compleja a la que denominó *fractal*, porque se podría caracterizar por una dimensión no entera (fraccionaria). Una característica importante de las estructuras con geometría fractal es que presentan autosimilaridad, es decir, es posible identificar al mismo patrón en diferentes escalas (por ejemplo la coliflor). Una consecuencia de este escalamiento es la presencia de correlaciones a largo alcance. En el caso concreto de los fenómenos biológicos de naturaleza fractal, existe un escalamiento que sigue una ley de potencias y que pueda detectarse por al menos dos órdenes de magnitud [103].

Una función $f(x)$, sigue una ley de potencia cuando la variable independiente, $x$, tiene un exponente, es decir, se encuentra elevado a una potencia, por ejemplo; $y = Ax^{\alpha}$ donde la potencia $\alpha$ puede ser cualquier número y A es una constante. Cuando $\alpha > 0$, la función crece, mientras que cuando $\alpha < 0$, la función decae. En este último caso se habla de una ley de potencias inversa. Usualmente las funciones que siguen leyes de potencia se grafican en escala logarítmica donde puede observarse una relación lineal entre las dos variables. La pendiente de la recta obtenida corresponde al exponente $\alpha$ de la función graficada. Los fenómenos que presentan una dinámica de leyes de potencia también exhiben autosimilaridad y, por lo tanto, una dinámica fractal con correlaciones a largo alcance [104]. Dado que, a partir de secuencias de DNA pueden generarse series de distancias, es factible determinar la presencia de correlaciones a largo alcance en estas secuencias.

## 2.4. Modelo de caminata aleatoria

Partiendo del modelo de caminata aleatoria, Peng, *et al.* han aplicado diversas técnicas para analizar la presencia de correlaciones de largo alcance en secuencias de DNA [105, 106, 107, 108]

En el modelo de caminata aleatoria se asume la presencia de un caminador teórico que recorre una cadena de DNA. El caminador comienza en la posición

$n(n=0)=0$ y da un paso hacia arriba $[u(n)=+1]$ cada vez que encuentra una pirimidina [1], y un paso hacia abajo $[u(n)=-1]$, cada vez que encuentra una purina. De esta manera la cadena original de DNA se convierte en una secuencia de -1 y 1. Usualmente se gráfica la caminata acumulada $y(n)$ contra la posición $n$ en la cadena. Uno de los resultados más importantes de la caminata aleatoria es el hecho de encontrar regiones relativamente largas, que son ricas en cierto tipo de nucleótido (por ejemplo purinas), por lo que la distribución, a lo largo de la cadena de DNA no es uniforme [104]. Dada esta característica se dice que las secuencias de DNA presentan *mosaicos* [109] El primer acercamiento que utilizó este grupo, fue la determinación de la raíz de las fluctuaciones medias cuadradas, $F(n)$ alrededor del promedio de desplazamiento, definida como:

$$F(n)=\sqrt{\overline{(\Delta y(n))^2}-\left(\overline{\Delta y(n)}\right)^2} \tag{2.5}$$

donde $\Delta y(n)=y(n_0+n)-y(n_0)$ y las barras sobre las variables indican la media aritmética para todas las $n$ posiciones en la cadena genómica.

Se entiende por fluctuaciones, a las variaciones a nivel microscópico que se presentan en los sistemas macroscópicos. Así las fluctuaciones en física son equivalentes al concepto de varianza estadística en la biología. En el caso de $F(n)$, se puede presentar dos posibles escenarios: (a) para procesos estocásticos o sólo con correlaciones locales, $F(n)\approx n^{\frac{1}{2}}$; y (b) para procesos con correlaciones libres de escala (correlaciones a largo alcance), las fluctuaciones se describen por la ley de potencias: $F(n)\approx n^{\alpha}$, con $\alpha\neq 1/2$. A través del método conocido como *min-max*, Peng *et al* [106], encontraron correlaciones a largo alcance en regiones no codificantes, en contraste con las codificantes, donde $\alpha\approx 0{,}5$. Estos resultados, han sido ampliamente discutidos e inclusive, algunos autores señalan que no existe diferencia entre la dinámica de regiones codificantes y la de no codificantes [110]. Una de las críticas principales a estos resultados, hace referencia al hecho de que se presenta heterogeneidad en las series de distancias obtenidas por la caminata aleatoria del DNA. Esto quiere es decir que se presentas secuencias relativamente grandes ricas en ya sea purinas o pirimidinas. Debido a este fenómeno de mosaico, métodos como el análisis de función de autocorrelación y la $F(n)$ pierden validez ya que se basan en medidas que cambian a lo largo de la serie de distancias [109].

[1] Pirimidina = Timina y Citosina; Purina = Adenina y Guanina

Posteriormente Peng *et al.*, mejoraron su técnica para evitar el efecto mosaico del DNA, mediante la eliminación de la tendencia de las fluctuaciones sobre los diferentes tamaños de ventanas [106]. A esta técnica se le llama análisis de fluctuaciones libre de tendencias, DFA por sus siglas en inglés. En una serie de tiempo, las fluctuaciones crecen al aumentar la longitud de la ventana donde se calculan éstas. Este hecho nos permite asumir una relación de leyes de potencia entre los factores de escalamiento: $M_y = M_x^{\alpha}$ donde $M_x$ y $M_y$ son factores de escalamiento y $\alpha$ es el exponente de escalamiento. El escalamiento en la variable independiente (usualmente el tiempo) puede calcularse por:

$$M_x = \frac{n_2}{n_1} \tag{2.6}$$

donde $n_1$ y $n_2$ son dos longitudes diferentes de ventana. Una manera de calcular el escalamiento en la variable dependiente es mediante:

$$M_y = \frac{s_2}{s_1} \tag{2.7}$$

donde $s_2$ y $s_1$ son las desviaciones estándar de las distribuciones de los correspondientes histogramas. Con lo anterior, el exponente de escalamiento $\alpha$, puede calcularse por:

$$\alpha = \frac{\ln M_y}{\ln M_x} = \frac{\ln s_2 - \ln s_1}{\ln n_2 - \ln n_1} \tag{2.8}$$

que es la pendiente obtenida al aplicar un modelo de regresión lineal sobre los logaritmos de los datos originales.

El DFA ha sido aplicado para estudiar diversos fenómenos, en el caso de la biología, los trabajos más importantes han sido sobre secuencias de DNA y sobre ritmos cardiacos [104]. En el caso de DNA, los resultados de Peng *et al*, fueron consistentes con sus resultados anteriores, es decir, las correlaciones a largo alcance sólo fueron detectadas para las secuencias no codificantes [106]. En el caso de secuencias codificantes, se detectó un cambio en la pendiente con una $\alpha$=0.51 para la primera parte de la curva que es equivalente a un proceso aleatorio.

En contraste con estos resultados, otros autores han encontrado correlaciones de largo alcance en secuencias codificantes. Esta diferencia se debe fundamentalmente, a modificaciones en la metodología para generar las series de distancia. En particular, Voss [110] y Sousa Vieira [111] calcularon el

espectro de potencias provenientes de series no binarias, mientras que Mohanty y Narayana Rao obtuvieron los momentos factoriales de series que representaban el exceso de purinas sobre pirimidinas [13].

El conocimiento de la estructura cromosómica en los organismos eucariontes estimuló la investigación de correlación entre nucleótidos separados a diferentes distancias en una misma cadena. Esto con el fin de entender la complicada estructura en la cual el DNA se compacta en diferentes niveles para formar los cromosomas, lo que hace que nucleótidos separados a distancias grandes sobre la cadena lineal puedan estar próximos en la estructura cromosómica. En otras palabras se planteó que debían existir correlaciones entre diferentes nucleótidos situados a grandes distancias.

La primera referencia que apareció en la literatura en torno a las correlaciones de largo alcance en el DNA [112], se avocó a describir éstas a partir de las funciones de autocorrelación de los diferentes nucleótidos. Se consideraron las secuencias de los mismos como un texto sólo con cuatro letras (A, T, G, C), pretendiendo obtener una medida numérica de la autocorrelación.

Se han reportado [106, 113, 114] la existencia de al menos siete maneras de codificar las cadenas de nucleótidos, por ejemplo los enlaces entre los mismos, el tipo de las base (purina o pirimidina), etc. Por ejemplo si aplicamos la codificación

$$\begin{aligned} A, T &\rightarrow 0, \\ G, C &\rightarrow 1, \end{aligned}$$

en una secuencia como la siguiente:

$$\cdots ATGCTTAGCCGATTGG \cdots, \tag{2.9}$$

obtenemos la cadena de números:

$$\cdots 0\ 0\ 1\ 1\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ 1\ 1 \cdots. \tag{2.10}$$

Sea $x(t)$ una serie de tiempo y $y(t) = x(t+\tau)$ la misma serie de tiempo, con un retardo de $\tau$ posiciones con respecto a $x(t)$. La función de autocorrelación (ACF) queda definida de la siguiente manera:

$$c(\tau) = \frac{1}{N} \sum_{i=1}^{N} \tilde{x}(i)\tilde{y}(i), \tag{2.11}$$

donde $\tilde{x}(i) = x(i) - \langle x(i) \rangle$, con $\langle x(i) \rangle$ es el promedio de $x(i)$ con ecuaciones equivalentes para $y(i)$. La cantidad de la ecuación anterior mide la correlación que existe entre $x(t)$ y $x(t + \tau)$. En equivalencia con el coeficiente de correlación de Pearson [115], un valor de ACF = 1 indica una autocorrelación positiva del 100, un valor de ACF = -1 indica una autocorrelación negativa del 100 y un valor de ACF = 0 indica una independencia entre los datos de la serie, y por tanto se pude concluir que la serie sigue un comportamiento de ruido blanco en donde no se puede predecir un evento respecto al otro.

Mas recientemente Arneodo y colaboradores [3], analizaron los contenidos de regiones ricas en GC (isocoras) y el sesgo de A/T y G/C en varios cromosomas humanos. Ellos mostraron que los máximos del sesgo son marcadores de replicación de DNA. Se encontró que el contenido de GC y el numero relativo de A/T y G/C oscilan en periodos de 110+20kb y 400+40kb. En otras palabras el contenido de GC y el sesgo esencialmente se repitan cada 110kb y 400kb. Los ritmos lentos en el contenido de GC refleja la estructura natural de las fibras de cromatina, el cual es doblado dentro de bucles del orden de 100kb en el cromosoma. Sin embargo el hecho de que estos ritmos son similares a los encontrados en el sesgo sugiere que la expresión de genes toma lugar en una sincronización múltiple, lo cual provee una extraordinaria guía en la organización espacial y orientación de genes (ver Fig. 2.2).

Una vez que la secuencia del DNA humano estuvo disponible, Mansilla [116] calculó la función de información mutua de los diferentes cromosomas humanos. Es notable el hecho de que todas las funciones de información mutua de los diferentes cromosomas tengan la misma forma. Esto sugiere que, a pesar de las diferencias marcadas en las funciones, tamaño y estructura de los mismos, la correlación entre las bases tienen los mismos patrones es decir, sugiere un mecanismo único de evolución estructural de todos los cromosomas.

En otro trabajo (I. Grosse, *et al* 2006) [117] se analizó la autocorrrelación de todos los cromosomas humanos y de todos los cromosomas del arroz con diferentes métodos de mapeos binario. Se encontró que la correlación de patrones parecida a una ley de potencia es diferente dependiendo del método de que se use pero son idnticos en toda las especie de todos los cromosomas.

Con respecto a las leyes de potencia (L. Matignetti and M. Caselle, 2007) [118], analizaron secuencias de exones que se encuentran en el RNA mensajero pero que no son codificantes tanto en ratones como en humanos. Ellos mostraron que todas estas secuencias tienen un distribución de leyes de potencia dando una explicación teórica y evolutiva. Estas secuencias tienen

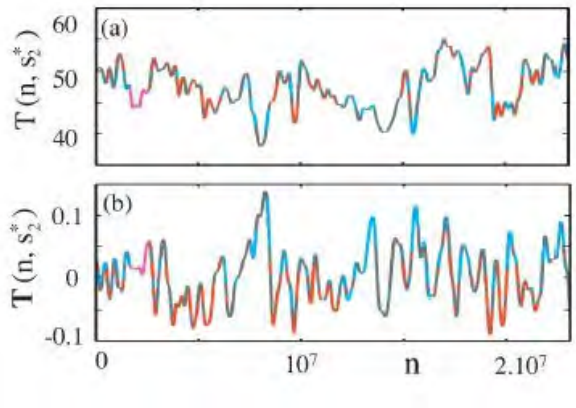



Figura 2.2: Oscilaciones observadas en un fragmento del cromosoma 22 del humano. a) Contenido de GC y b) $x = (A-T)/(A+T)+(G-C)/(G+C)$. Las porciones rojas y azules corresponden a la localización de los genes en la cadena lder y rezagada de las cadenas de DNA. La localización de la inmunoglobulina está en la porción rosa. Tomado de la Ref. [3]

una estructura, como la longitud, estructura secundaria y la presencia de tripletes AUG, que tienen un efecto en la eficiencia de la traslación y por lo tanto son preservados en la evolución.

Un trabajo que separa la funcionalidad del genoma en humanos haciendo análisis de correlaciones de largo alcance es el de Oliver y colaboradores [119]. El genoma humano muestra una estructura composicional común con dos características de escala, una larga correspondiente a la isocoras y otra pequeña o mediana para elementos genómicos. Para detectar estas correlaciones de largo alcance se utilizó la metodología de DFA el cual provee un parámetro cuantitativo que es el exponente $\alpha$ de escalamiento que representa las propiedades de la correlación de la señal. Sobre este mismo tema ¿podemos relacionar las regiones codificantes de las no codificantes en los genomas con correlaciones? E. Stanley, et al, 1995 [120] trabajaron con secuencias codificantes y no codificantes de eucariontes con metodologías como DFA y transformadas de Fourier. Ellos encontraron que las regiones codificantes prácticamente no tiene correlaciones en el rango de 10kb a 100kb en contraste con las regiones no codificantes en donde si se encuentran correlaciones. Anteriormente Peng et al. [106], encontró correlaciones de largo alcance y leyes de potencia de nucleótidos en secuencias de DNA pero utilizando el método

de caminatas al azar, encontrando que las regiones no codificantes que son los intrones y las secuencias intergénicas muestran correlaciones de largo alcance mientras que las secuencias codificantes no muestran estas correlaciones. Estos mismos resultados fueron observados pero con otra metodología diferente que es el Análisis de Fourier estándar por Li y Kaneko [100].

Todos estos trabajos son importantes porque podemos inferir cual es la estructura del DNA y además podemos sacar conclusiones evolutivas de los genomas. Nuestro grupo de investigación estudia el comportamiento de los genomas utilizando la metodología utilizada por los físicos desde los años 50 para el estudio de los núcleos atómicos llamada estadística espectral [7, 8, 9]. De hecho, se han empezado a dar varias aplicaciones de estas técnicas en la biología [10, 11, 12]. Para poder aplicarla es necesario separar las propiedades seculares o el promedio y las propiedades fluctuantes, esto de acuerdo con Voss [110] y Soussa-Vieira [111] . De hecho, se espera, que la propiedad secular es distintiva de cada organismo, mientras que las propiedades fluctuantes sean características generales de la mayoría de los genomas.

La parte secular ha sido estudiada extensivamente por José y colaboradores [121, 122]. Por ejemplo mediante la determinación de la densidad acumulada de los 64 tripletes a lo largo del cromosoma de *Borrelia burgdorferi* se revelaron altas correlaciones en frecuencias de monopletes, dupletes y tripletes entre las cadenas opuestas llamadas replicoras, la cadena lider y rezagada (en la mitad del cromosoma). Usando una secuencia permutada de bases se demostró que las distribuciones acumuladas no son una consecuencia de su composición de bases. Esta propiedad ocurre en las mitades de los cromosomas bacterianos y se ha denotado como reflejo de una simetría bilateral inversa, que al parecer es un tipo universal de simetría en los cromosomas de eubacterias [122]. Otro ejemplo de como varían las propiedades seculares lo dimos en la Ref. [16] en donde graficamos el número de bases acumuladas para algunas cadenas genómicas (Ver. Fig. 2.3). Es claro en esa figura que las densidades no son constantes ya que presentan cambios de pendiente y esto depende de las características individuales de cada organismo.

La distribución de las fluctuaciones de distancias de primeros vecinos del mismo nucleótido en algunas cadenas es casi una exponecial [16]. Esto quiere decir que la probabilidad de encontrar ese nucleótido es más alta en distancias cortas y esta probabilidad decrece exponencialmente mientras la distancia aumenta. Un estudio en el mismo sentido, pero con otra cantidad matemática, fue hecho anteriormente por Mohanty y Narayana Rao [13] quienes mostraron que la varianza de número (también llamada rigidez espectral)

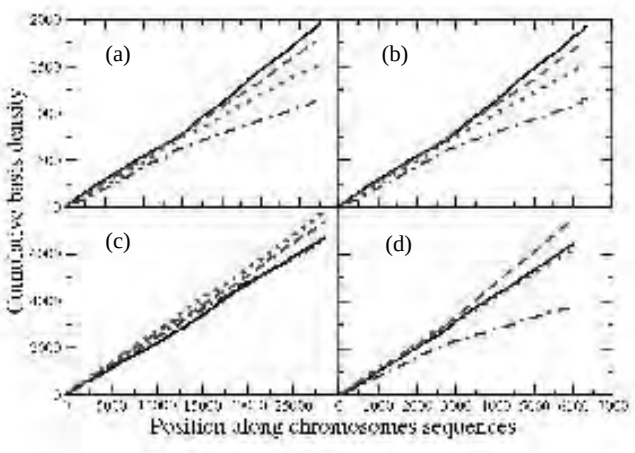



Figura 2.3: Densidad acumulada de bases: La primera columna corresponde a regiones no codificantes y la segunda a codificantes, la primera fila corresponde a *D. melanogaster* y la segunda al humano. La línea continua (negra) es adenina (A), la línea punto-linea (marron) es timina (T), la línea punteada (rojo) es citosina (C) y la línea interrumpida (verde) es guanina (G). Tomado de la Ref [16].

crece casi linealmente a distancias menores de aproximadamente 1kb. La rigidez espectral es equivalente a la $\Delta_3$ de Dyson o a la DFA. Así los resultados de Mohanty y Narayana Rao y los de Higareda *et al* [16] son consistentes con una distribución casi exponencial para las distancias a primeros vecinos la cual está dada por

$$p(x) = e^{-x}. \tag{2.12}$$

Esta ecuación predice como es el comportamiento estadístico de los nucleótidos a primeros vecinos en las secuencia de la mayoría de los organismos (Ver Fig. 2.4). Para esta distribución la rigidez espectral $\Sigma^2$ es lineal con la distancia $L$ entre bases

$$\Sigma^2(L) = L. \tag{2.13}$$

La raíz cuadrada de esta cantidad fue calculada numéricamente por Mohanty y Narayana Rao para las mismas secuencias (Ver Fig. 2.5). En esta figura la raíz cuadrada de la Ec. (2.13) fue superpuesta en rojo.

Afreixo y colaboradores también trabajan con la distancia de internucleótidos y sugieren que esta distancia es capaz de capturar una información esencial acerca de los genomas y que es posible construir árboles filogenéticos [15]. Utilizando esta misma metodología, pero con distancias de

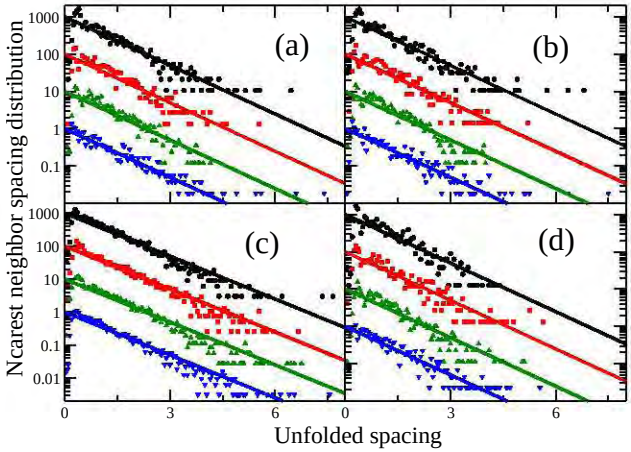



Figura 2.4: Distribución de primeros vecinos de las secuencias genómicas descritas en la Fig 2.3. Tomado de la Ref. [16]

nucleótidos no similares, también se obtiene una firma genómica que distingue las características de cada especie [14].

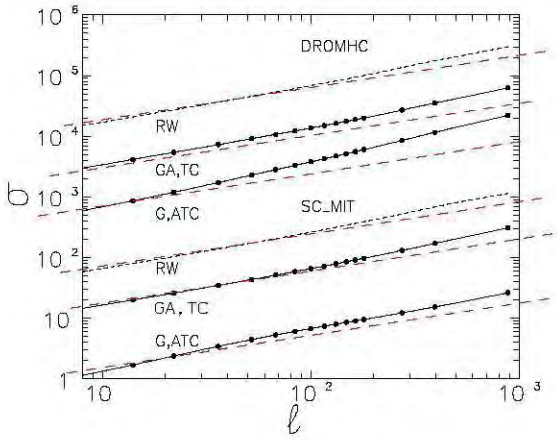



Figura 2.5: Raíz cuadrada de la varianza de número de algunas secuencias genómicas. Adaptada de la Ref [13]. Las líneas discontinuas corresponden al resultado Poissoniano $\sigma = \sqrt{L}$.

# Capítulo 3

# Hipótesis y Objetivos

**Hipótesis:**

El cambio de comportamiento entre generos en la distribución de distancias entre bases a primeros vecinos en secuencias bacterianas, podría depender del uso de codon, es decir, cuando la distribución de distancias cambia de un género de bacteria a otro género es debido a que las bacterias tienen un sesgo alto para utilizar un codón específico para formar los aminoácidos.

**Objetivos:**

Analizar las secuencias de diferentes organismos, específicamente estudiar la distribución de distancias a primeros vecinos de los cuatro nucleótidos (A, T, G y C) con la metodología de estadística espectral, separando su promedio de las fluctuaciones.

**Objetivos particulares:**

1. Analizar la densidad acumulada de bases de las secuencias genómicas separando la parte promedio de las fluctuaciones.
2. Analizar el comportamiento de la distribución de primeros vecinos tanto de eucariontes como de procariontes los cuales tienen diferentes características genómicas y poder compararlos entre sí.

3. Comparar secuencias bacterianas del mismo y diferente genero, tanto en el promedio y como en sus fluctuaciones.

4. Hacer un modelo que ayude a entender el comportamiento de la distribución de distancias a primeros vecinos para cualquier organismo

# Capítulo 4

# Resultados

Los resultados obtenidos en esta tesis doctoral son detallados en la sección 4.1, artículos publicados y en la sección 4.2, resultados adicionales. En la sección 4.2 se muestra la construcción del modelo matemático para poder explicar y compara con los resultados numéricos obtenidos de la secuencias genómicas. Los artículos contiene una discusión de los resultados, la cual es complementada en la discusión de esta tesis.

## 4.1. Artículos publicados

### 4.1.1. Nearest neighbor spacing distribution of bases in some coding and non-coding DNA sequences

Utilizando la estadística espectral en secuencias que contienen intrones y otras que no lo contiene, encontramos que la distribución de distancia entre nucleótidos es la misma. La forma de la distribución es similar al azar.

Available online at www.sciencedirect.com
ScienceDirect
Physica A 372 (2006) 368–373
PHYSICA A
www.elsevier.com/locate/physa

# Nearest-neighbor spacing distribution of basis in some intron-less and intron-containing DNA sequences

M.F. Higareda[a], H. Hernández-Saldaña[b,c,*], R.A. Méndez-Sánchez[d]

[a]*Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, Apdo. Postal 70228, Ciudad Universitaria, C.P. 04510 México D.F., México*
[b]*Universidad Autónoma Metropolitana-Azcapotzalco, México D.F., México*
[c]*Instituto de Física, Universidad Nacional Autónoma de México, México D.F., México*
[d]*Centro de Ciencias Físicas, Universidad Nacional Autónoma de México, A.P. 48-3, 62210 Cuernavaca, Morelos, México*

Available online 5 September 2006

**Abstract**

We show that the nearest neighbor distribution of distances between basis pairs of some intron-less and intron-containing coding regions are the same when a procedure, called *unfolding*, is applied. Such a procedure consists in separating the secular variations from the oscillatory terms. The form of the distribution obtained is quite similar to that of a random, i.e., Poissonian, sequence. This is done for the HUMBMYH7CD, DROMYONMA, HUMBMYH7 and DROMHC sequences. The first two correspond to highly coding regions while the last two correspond to non-coding regions. We also show that the distributions before the unfolding procedure depend on the secular part but, after the unfolding procedure we obtain an striking result: all distributions are similar to each other. The result becomes independent of the content of introns or the species we have chosen. This is in contradiction with the results obtained with the detrended fluctuation analysis in which the correlations yield different results for intron-less and intron-containing regions.


*Keywords:* Spectral statistics; Genomic sequences; Statistical analysis

## 1. Introduction

In recent times DNA of new species have been sequenced, opening new questions in biological physics. Large data of genomic sequences are now available and ready to be analyzed. In particular, the study of their statistical properties is of broad interest. Up to now, several works addressed the correlations in coding and non-coding regions [1,2]. Many statistical measures have been introduced since those works appeared: the detrended fluctuation analysis [3], diffusion entropy analysis [4], factorial moment analysis [5], wavelet analysis [6], among others.

Despite several efforts (see Ref. [5] and references therein), the question about the statistical properties between intron-less and intron-containing coding regions is still open. Here we use two complementary

*Corresponding author. Instituto de Física, Universidad Nacional Autónoma de México, México D.F., México.
*E-mail address:* hugo@fisica.unam.mx (H. Hernández-Saldaña).

0378-4371/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2006.08.043

Table 1
Some statistical properties of the sequences analyzed here

| Code | Size (basis) | Introns (%) | Basis-content (%) | | | | $\langle s \rangle$ (basis) | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | A | T | C | G | A | T | C | G |
| HUMBMYH7CD | 5999 | 0 | 26.7 | 16.0 | 26.0 | 31.3 | 3.7 | 6.3 | 3.8 | 3.2 |
| DROMYONMA | 6338 | 0 | 24.1 | 14.5 | 18.7 | 21.7 | 3.3 | 5.4 | 4.2 | 3.6 |
| HUMBMYH7 | 28438 | 73 | 23.6 | 23.0 | 27.4 | 26.0 | 4.2 | 4.4 | 3.6 | 3.9 |
| DROMHC | 22663 | 72 | 30.6 | 18.4 | 23.6 | 27.5 | 3.3 | 5.5 | 4.3 | 3.7 |

A, T, C, and G correspond to adenine, thymine, cytosine and guanine, respectively. $\langle s \rangle$ stands for the mean level spacing of each basis and is calculated as an arithmetic average.

techniques to study the "short-range" statistical properties of DNA sequences. Our work considers, at this stage, only monoplets, but studies on duplets and triples are in progress [12]. First we use a generalized detrended fluctuation analysis, commonly called *unfolding* [7–9], to separate the secular trend and the non-secular properties of the sequences. Then, the nearest neighbor distribution is obtained and analyzed.

In the next section we define the cumulative basis density and the mathematical procedure called *unfolding*. The cumulative basis density is obtained for: two intron-less sequences (0% of introns), HUMBMYH7CD and DROMYONMA, corresponding to the Human $\beta$-cardiac myosin heavy chain (MHC) and *Drosophila melanogaster* MHC, respectively; and two intron-containing coding sequences, HUMBMYH7 and DROMHC that correspond to the Human $\beta$-cardiac MHC and to the *D. melanogaster* MHC, respectively. The sequences were taken mainly from GenBank [10] and were analyzed with the detrended fluctuation analysis [2]. The number of introns of each sequence is given in Table 1. In Section 3 the nearest neighbor distribution is obtained yielding a universal feature. Examples of the non-sense results obtained without a proper unfolding procedure are also given there. A brief conclusion follows.

## 2. Unfolding procedure

Following Ref. [5] we define the density of basis as

$$\rho^{\beta}(x) = \sum_{i=1}^{N_{\beta}} \delta(x - x_i^{\beta}), \tag{1}$$

where $\{x_i^{\beta}\}_{i=1}^{N_{\beta}}$ is the ordered set of $N_{\beta}$ positions of the basis. Here $\beta$ stands for adenine (A), thymine (T), cytosine (C), or guanine (G) in DNA. In the last equation $\delta(x)$ is the Dirac $\delta$-function. The cumulative basis density or staircase function is defined as

$$N^{\beta}(x) = \sum_{i=1}^{N_{\beta}} \Theta(x - x_i^{\beta}), \tag{2}$$

where $\Theta(x)$ is the Heaviside function. This function is plotted in Fig. 1 for the sequences HUMBMYH7CD, DROMYONMA, HUMBMYH7 and DROMHC. Note that this quantity corresponds to the inverse function of the cumulative position plots of Ref. [11].

Following Refs. [7–9], the cumulative basis density can be expressed as

$$\mathrm{N}^{\beta}(x) = \langle \mathrm{N}^{\beta}(x) \rangle + \mathrm{N}_{osc}^{\beta}(x), \tag{3}$$

where $\langle \mathrm{N}^{\beta}(x) \rangle$ is a mean or secular part and $\mathrm{N}_{\mathrm{osc}}^{\beta}(x)$ is the oscillating one. The brackets in the secular part refer to average in position. Note that the secular part is a function of $x$ and is not necessarily smooth. In quantum physical systems, for instance, $\langle \mathrm{N}^{\beta}(x) \rangle$ contains the periodic orbit information, i.e., the classical one [9]. As seen in Fig. 1, cumulative densities of different basis and/or sequences have different secular parts. Some look almost linear while other appear linear by parts. It is not clear for us the biological origin of the abrupt change

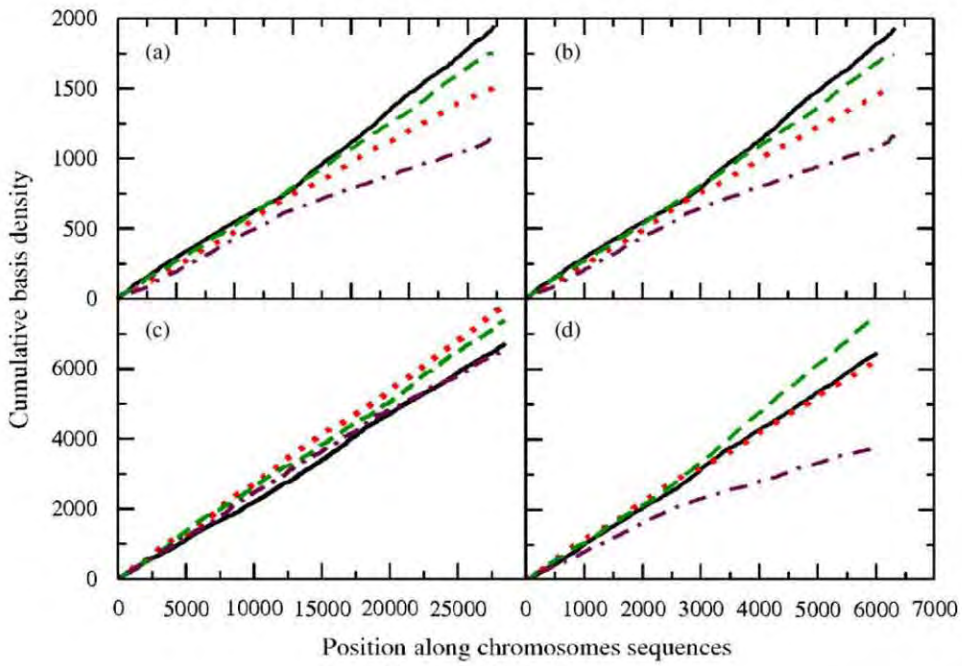



Fig. 1. (Color online) Cumulative basis density of the sequences: (a) DROMYONMA, (b) DROMHC, (c) HUMBMYH7CD and (d) HUMBMYH7. The first column corresponds to non-coding region and the second one to coding one. The first row corresponds to *D. melanogaster* and the second one to human. The solid (black) line corresponds to adenine (A), the dash-dotted (maroon) line to thymine (T), the dotted (red) line to cytosine (C) and the dashed line (green) to guanine (G).

in the slope of $\langle \mathrm{N}^{\beta}(x)\rangle$. However, for bacterial genomes, similar changes in density corresponds to the replication origin [11,13]. Analytical expressions for the secular part are unknown up to now, hence we will use some of the numerical methods commonly used in spectral statistics. For instance, we will use a polynomial fitting. The degree of the polynomial is chosen, as usual, as the minimum in which the fluctuations does not change anymore. Note that this unfolding procedure is more general that the detrended fluctuation analysis, since the latter uses only a linear fitting. A systematical study and a comparison with a different unfolding procedure (window analysis) will be given elsewhere [12]. In the sequences of Table 1 we tested fittings with polynomials from 8th until 18th degree. We also used windows from 100 to 1000 distances. Within this range of parameters (polynomial degree and window) the results we present are stationary, i.e., the results do not change.

The transformation that allows the comparison of the fluctuations for systems with different secular parts is the unfolding. It works as follows. Lets define a new unfolded sequence $\{y_i^{\beta}\}_{i=1}^{N_{\beta}}$, where

$$y_i^{\beta} = \langle \mathrm{N}^{\beta}(x_i^{\beta})\rangle. \tag{4}$$

With this transformation the new sequence has a uniform density and mean level spacing equal to unity, i.e., $\langle y_{i+1}^{\beta} - y_i^{\beta}\rangle = 1$.

## 3. Nearest neighbor spacing distribution

The nearest neighbor spacing distribution, $p(s)$, is the observable most commonly used to study the "short-range" fluctuations in the density of basis. Here $s_i^{\beta} = y_{i+1}^{\beta} - y_i^{\beta}$. The quotes mean that the short range is defined in units of the mean level spacing $\langle s^{\beta}\rangle$. Then, for a very diluted basis in a sequence, $p(s)$ refers to very long distances in the complete sequence and vice versa. The distribution $p(s)$ then yields the probability density to have two neighboring levels $y_i^{\beta}$ and $y_{i+1}^{\beta}$ at a spacing $s$. With the unfolding procedure the function $p(s)$ and its first moment are normalized to unity. The mean level spacing $\langle s\rangle$ on basis, without unfolding, is given in

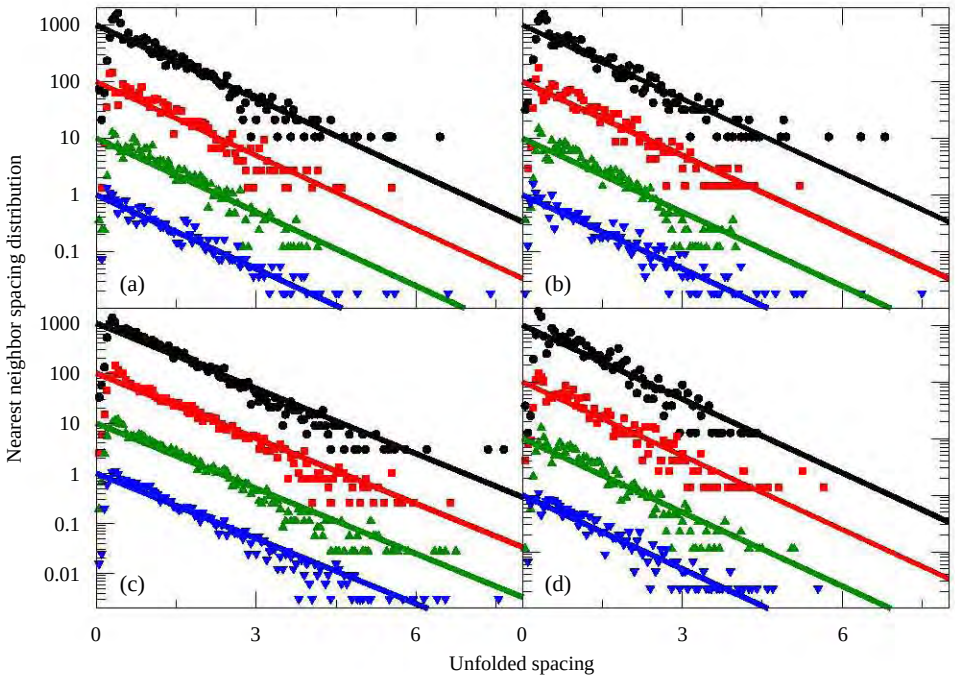



Fig. 2. (Color online) Nearest neighbor spacing distribution, $p(s)$, for the same sequences of Fig. 1 and following the same order. The circles (black) correspond to the $p(s) \times 1000$ of adenine, the squares (red) to the $p(s) \times 100$ of cytosine, the triangles up to the $p(s) \times 10$ of guanine and the triangles down (blue) to the $p(s)$ of thymine. The solid lines correspond to the Poissonian result $p(s) = \exp(-x)$ times the corresponding scaling factor.

Table 1. Note that this is not a well defined quantity for some sequences like thymine in DROMYONMA (dash-dotted in Fig. 1), since it has two different values and hence, the unfolding procedure is required.

In Fig. 2 we show the obtained nearest neighbor spacing distribution $p(s)$ for the different basis for the different sequences. As seen in this figure, they are very similar to the uncorrelated Poissonian prediction

$$p(s) = \exp(-s). \tag{5}$$

The main differences between the results of the genomic sequences and the Poissonian prediction appear at both, very short and very long distances. The explanation at short distances is trivial: they are related to the physical size of the basis which avoid distances smaller than the physical size of a basis. This generates a "repulsion hole". The repulsion hole is not due to the unfolding procedure but this procedure makes the hole much more evident. The result we get in (2) is opposite to the well-known result given in Ref. [3]. This means that, at least in the sequences we used, it is impossible to differentiate coding and non-coding regions with statistical analysis. Biological information is present only in the secular part but not in the fluctuations.

Now we will show some results without the unfolding procedure. In Fig. 3 we show the histograms we obtained for thymine in the four sequences. They are no normalized nor unfolded. As seen, some of them, Fig. 3(a) and (c), are very similar to the Poissonian prediction while other, Fig. 3(b) and (d), are quite different. This behavior give the impression that statistical distribution for non-coding and coding regions are different. However this wrong conclusion is due to an artifact because in Fig. 2 they have clearly the same distribution. The explanation of this behavior is very simple: since the density is not uniform and it depends on the position along the sequences, the distances between basis have different statistical weights. Missing or incorrect rectification (unfolding), gives place to an incorrect measurement of the fluctuations and to possible wrong conclusions. Then, the unfolding procedure eliminates artifacts due to a wrong average. To find the correct way to unfold a sequence is a very delicate and difficult task [9]. However, it is the way to obtain a measure of the fluctuations which allow us a comparison between systems having different trends.

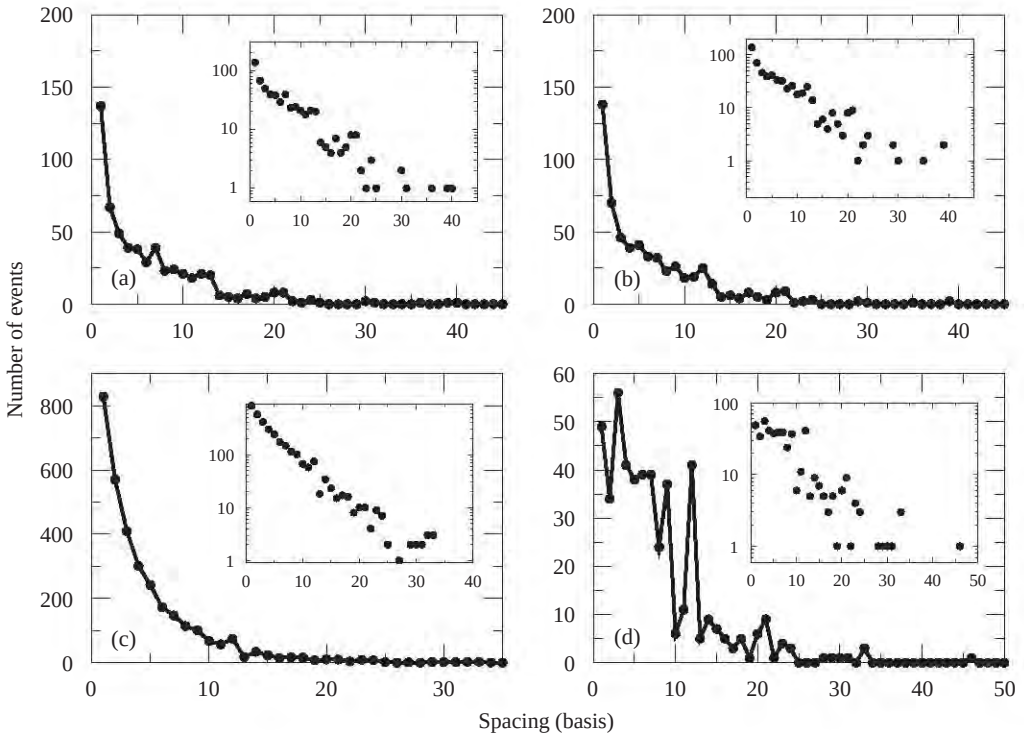



Fig. 3. Nearest neighbor spacing histograms for thymine in the same order as the previous figure. No normalization nor unfolding was performed. Note that the fluctuation at small spacings are large in case (a) and (b) and huge in case (d). An exponential fitting in case (c) is correct however it is non-sense in case (d). The insets show the same plots with the $Y$-axis in logarithmic scale.

## 4. Conclusion

We introduced the cumulative basis density as well as the unfolding procedure to analyze distances between basis in genomic sequences. We have shown that the nearest neighbor spacing distribution is very close to the Poissonian distribution. The main differences between them have been found in short distances where the physical size of each basis generates a ''repulsion hole''. The results we have got for the nearest neighbor spacing distribution are quite different that those available in the literature for other quantities like correlations. We obtained the same result, almost Poissonian, for two intron-less and two intron-containing coding sequences. This suggest that a proper unfolding procedure should be implemented before any statistical analysis of genomic sequences. Artifacts due to a wrong average could appear. Our results indicate strongly that statistical measures in genomic sequences without proper unfolding could give artifacts and probably several results in the literature on the area are not correct.

## Acknowledgements

It is a pleasure for us to deeply thank to MV José and the theoretical biology group at IIB-UNAM for their inestimable support and great help. We wish to thank C. Abreu-Goodberg for helpful discussions and E. Várguez-Villanueva for useful comments. This work was supported by DGAPA-UNAM IN118805. MFH thanks CONACyT by a Ph.D. grant. With this paper we wish to celebrate with Professor Alberto Robledo his 60th anniversary and to have the opportunity to do physics at mesolatitudes. Alberto, not only happy birthday but that all non-liquid works be fluid.

## References

[1] W. Li, W. Kaneko, Europhys. Lett. 17 (1992) 655.
[2] C.K. Peng, et al., Nature 356 (1992) 168.

[3] C.K. Peng, et al., Phys. Rev. E 49 (1994) 1685.
[4] N. Scafetta, V. Latora, P. Grigolini, Phys. Lett. A 299 (2002) 565.
[5] A.K. Mohanty, A.V.S.S.N. Rao, Phys. Rev. Lett. 84 (2000) 1832.
[6] A. Arneodo, E. Bacry, P.V. Graves, J.F. Muzy, Phys. Rev. Lett. 74 (1995) 3293.
[7] M.L. Metha, Random Matrices, second ed., Academic, New York, 1991.
[8] T.A. Brody, et al., Rev. Mod. Phys. 53 (1981) 385.
[9] T. Guhr, A. Müller-Groeling, H.A. Weidenmüller, Phys. Rep. 299 (1998) 189.
[10] ⟨http://www.ncbi.nlm.nih.gov/Genbank/index.html⟩.
[11] J. Sánchez, M.V. José, Biochem. Biophys. Res. Comm. 299 (2002) 126.
[12] M.F. Higareda, H. Hernández-Saldaña, R.A. Méndez-Sánchez, to be published.
[13] M.V. José, T. Govezensky, J.R. Bobadilla, Physica A 351 (2005) 477.

### 4.1.2. Evidence of codon usage in the nearest neighbor spacing distribution of bases in bacterial genomes

Cuando separamos las secuencias de bacterias en la parte secular y la parte fluctuante después del *unfolding*, la parte secular agrupa a las bacterias en dos grupos: lineal y lineal por partes. Por este motivo este método es importante porque se tiene características estadísticas diferentes a través del genoma.
Con esta misma metodología 4 géneros bacterianos fueron comparados. Interesantemente la distribución de las fluctuaciones agrupa las especies del mismo género. Esta diferencia se atribuye al uso de codon en bacterias.

Physica A 391 (2012) 1255–1269

ents lists available at SciVerse ScienceDirect

Physica A

al homepage: www.elsevier.com/locate/physa

# in the nearest neighbor spacing distribution of es

Mendoza [b], R.A. Méndez-Sánchez [a,*]

*tónoma de México, A. P. 48-3, 62210, Cuernavaca, Morelos, Mexico*
*Nacional Autónoma de México, A. P. 70228, Ciudad Universitaria, C.P. 04510 México D.F., Mexico*
*Autónoma de México, A. P. 565-A, 62251, Cuernavaca, Mor., Mexico*

A B S T R A C T

Statistical analysis of whole genomic sequences usually assumes a homogeneous nucleotide density throughout the genome, an assumption that has been proved incorrect for several organisms since the nucleotide density is only locally homogeneous. To avoid giving a single numerical value to this variable property, we propose the use of spectral statistics, which characterizes the density of nucleotides as a function of its position in the genome. We show that the cumulative density of bases in bacterial genomes can be separated into an average (or secular) plus a fluctuating part. Bacterial genomes can be divided into two groups according to the qualitative description of their secular part: linear and piecewise linear. These two groups of genomes show different properties when their nucleotide spacing distribution is studied. In order to analyze genomes having a variable nucleotide density, statistically, the use of unfolding is necessary, i.e., to get a separation between the secular part and the fluctuations. The unfolding allows an adequate comparison with the statistical properties of other genomes. With this methodology, four genomes were analyzed *Burkholderia, Bacillus, Clostridium* and *Corynebacterium*. Interestingly, the nearest neighbor spacing distributions or detrended distance distributions are very similar for species within the same genus but they are very different for species from different genera. This difference can be attributed to the difference in the codon usage.



genomic sequences, a fact that has enabled the development of a wide range of ctional regions in nucleotide sequences [1]. Even though some of these statistical non-biological phenomena, these techniques have been applied for the analysis of n walks provide a measurement of the correlation between nucleotides over long $-C)/(G+C)$ skew was used to locate the origin of replication in prokaryotes [3] as been used in the genes localization and orientation [5]. And finally, correlations have been used to reconstruct phylogenetic trees [6].

*tics* has been widely used in the study of several complex systems, ranging from s to the study of electroencephalograms. This methodology is used to compare gainst some theoretical predictions [7–9]. Briefly, spectral statistics requires the

x: +52 5556227775.
dez-Sánchez).

separation of data series into a secular (i.e. a local average) and a fluctuating part. This technique, as we mentioned, has been applied in the study of the electroencephalogram correlation matrices to distinguish between resting and stimulated people [10,11], and in the analysis of microarray data for the discovery of functional gene modules [12]. Regarding nucleotide sequences analysis, spectral statistics has been used to study the statistical properties of intron-less and intron-containing genomic sequences [13].

Two common quantities analyzed in spectral statistics are the number variance or spectral rigidity and the nearest neighbor spacing distribution, which we will call here detrended distance distribution. On the one hand, using the number variance, Ref. [14], a long range correlation in the DNA sequences was found that was previously detected by other methods [2,15,16]. On the other hand, in Ref. [13], the detrended distance distributions of genomic sequences between similar nucleotides were studied, finding an almost exponential distribution for short distances.

In this paper we analyze a set of bacterial genomes using spectral statistics. Specifically, we first obtain the cumulative density of bases (CDoB), and then we separate the distribution into its secular and fluctuating parts with a procedure called *unfolding or detrending*. Results show that bacterial genomes can be classified into two groups: one containing genomes with a linear nucleotide density, and another with a piecewise-linear density. Here we show that bacteria within a genus, but not among different genera, present similar detrended distance distributions when analyzing the same nucleotide. Finally, we present evidence suggesting that these differences in the detrended distance distribution can be explained in part by codon bias.

## 2. Methods

### 2.1. Genomic sequences

We analyzed the full genomic sequences of 132 bacterial strains (Tables 1–3), spanning all phylogenetic groups. All sequences were taken from the NCBI-GenBank database. As representative examples of our analysis, the results obtained for *B. phymatum* and *C. tetani* are shown.

### 2.2. Cumulative density of bases and the distributions of spacings and detrended distances

In general, any mathematical quantity defined for a genomic sequence can be represented as the sum of a smooth part plus a fluctuating part. In the Appendix, we give an example of how the methodology described in this section is applied to a simple sequence. Following Ref. [13], the cumulative density of bases (CDoB) or staircase function is defined as

$$\mathrm{N}^{\beta}(x) = \sum_{i=1}^{N_\beta} \Theta(x - x_i^{\beta}), \tag{1}$$

where $\{x_i^{\beta}\}_{i=1}^{N_\beta}$ is the set of $N_\beta$ sorted positions of the bases and $\Theta$ is the Heaviside step function. The index $\beta$ represents adenine, thymine, cytosine or guanine (A, T, C or G) in the DNA sequence.

The unfolding procedure consists of two parts. In the first step, the cumulative density of bases $\mathrm{N}^{\beta}(x)$ is separated in a smooth or secular part $\langle \mathrm{N}^{\beta}(x) \rangle$ plus a fluctuating one $\mathrm{N}^{\beta}_{\mathrm{fluc}}(x)$ as follows

$$\mathrm{N}^{\beta}(x) = \langle \mathrm{N}^{\beta}(x) \rangle + \mathrm{N}^{\beta}_{\mathrm{fluc}}(x). \tag{2}$$

The smooth part can be calculated using average in position commonly known as the window average. The size of the window is taken as the minimum value for which the fluctuations are stationary. The window length is fixed to 800 bases. When the average properties do not change, this average corresponds to an average over the full range. In the second step of the procedure, a "detrended sequence" $y_i^{\beta}$ is defined from the positions along the sequence, through:

$$y_i^{\beta} = \langle \mathrm{N}^{\beta}(x_i^{\beta}) \rangle. \tag{3}$$

With this transformation the new sequence has a uniform density which means an average level spacing approximately equal to unity, *i.e.* $\langle s_i^{\beta} \rangle = 1$ where the detrended distance $d_i^{\beta}$ is defined through $d_i^{\beta} = y_{i+1}^{\beta} - y_i^{\beta}$. The difference between $d_i^{\beta}$ and $s_i^{\beta} = x_{i+1}^{\beta} - x_i^{\beta}$ is that $\langle d_i^{\beta} \rangle \approx 1$, while $\langle s_i^{\beta} \rangle = f^{\beta}(x)$, where $f^{\beta}(x)$ is in general a nonlinear function of the position $x$, i.e., $\langle d_i^{\beta} \rangle = \langle d_i^{\beta} \rangle(x)$. This means that the sequence $y_i^{\beta}$ has non-integer values. Although these new positions do not have a biological meaning, they are useful since only the fluctuations remain, no matter what the trend is. An important characteristic of the resulting unfolded position of nucleotides is that it allows the comparison of fluctuations between genomes. The detrended distance distribution is exactly the distribution of $d_i^{\beta}$.

To apply the transformation of Eq. (3), an expression for the secular part $\langle \mathrm{N}^{\beta}(x) \rangle$ is required. Up until now, analytical expressions for the average of the genomic CDoB are unknown. Therefore we use the two common numerical methods in spectral statistics to calculate its value. One such numerical method consists of measuring the spacing $\langle s_i^{\beta} \rangle$ in a window and divides all spacings by $\langle s_i^{\beta} \rangle$. A second method, which is used in this study, among with the window average described above,

**Table 1**
List of bacterial genomes classified with two slopes in the averaged cumulative bases density.

| Name | Ref. seq. | Group | Size | %GC |
|---|---|---|---|---|
| *Acidobacterium capsulatum* ATCC 51196 | NC 012483 | Acidobacteria | 4.1 | 60.5 |
| Candidatus *Koribacter versatilis* Ellin345 | NC 008009 | Acidobacteria | 5.7 | 58.4 |
| *Arthrobacter aurescens* TC1 | NC 008711 | Actinobacteria | 4.6 | 62.4 |
| *Clavibacter michiganensis* subsp. michiganensis NCPPB 382 | NC 009480 | Actinobacteria | 3.3 | 72.5 |
| *Bifidobacterium adolescentis* ATCC 15703 | NC 008018 | Actinobacteria | 2.1 | 59.2 |
| *Mycobacterium abscessus* ATCC 19977 | NC 010397 | Actinobacteria | 5.02 | 64.1 |
| *Rhodococcus erythropolis* PR4 | NC 012490 | Actinobacteria | 6.5 | 62.3 |
| *Rubrobacter xylanophilus* DSM 9941 | NC 008148 | Actinobacteria | 3.22 | 70.5 |
| *Salinispora arenicola* CNS-205 CNS205 | NC 009953 | Actinobacteria | 5.8 | 69.5 |
| *Streptomyces coelicolor* A3(2) | NC 003888 | Actinobacteria | 8.66 | 72.0 |
| *Aquifex aeolicus* VF5 | NC 012438 | Aquificae | 1.59 | 43.3 |
| *Hydrogenobaculum sp.* Y04AAS1 | NC 11126 | Aquificae | 1.6 | 34.8 |
| *Sulfurihydrogenibium azorense* Az-Fu1 | NC 000918 | Aquificae | 1.6 | 32.8 |
| *Persephonella marina* EX-H1 | NC 012440 | Aquificae | 1.95 | 37.1 |
| *Bacteroides fragilis* YCH46 | NC 006347 | Bacteroidetes/chlorobi | 5.31 | 43.2 |
| Candidatus *Amoebophilus asiaticus* 5a2 | NC 010830 | Bacteroidetes/chlorobi | 1.9 | 35.0 |
| *Salinibacter ruber* DSM 13855 | NC 007677 | Bacteroidetes/chlorobi | 3.59 | 66.1 |
| *Chloroflexus aggregans* DSM 9485 | NC 011831 | Chloroflexi | 4.7 | 56.4 |
| *Roseiflexus sp.* RS-1 | NC 009523 | Chloroflexi | 5.8 | 60.4 |
| *Cyanothece sp.* ATCC 51142 | NC 010546 | Cyanobacteria | 5.43 | 37.9 |
| *Synechococcus elongatus* PCC 6301 | NC 006576 | Cyanobacteria | 2.7 | 55.5 |
| *Trichodesmium erythraeum* IMS101 | NC 008312 | Cyanobacteria | 7.75 | 34.1 |
| *Deinococcus geothermalis* DSM 11300 DSM11300 | NC 008025 | Deinococcus/thermus | 3.28 | 66.5 |
| *Thermus thermophilus* HB8 | NC 066461 | Deinococcus/thermus | 2.13 | 69.4 |
| *Clostridium acetobutylicum* ATCC 824 | NC 003030 | Firmicutes | 3.94 | 30.9 |
| *Mycoplasma genitalium* G37 | NC 000908 | Firmicutes | 0.58 | 31.7 |
| *Ureaplasma parvum* serovar 3 str. ATCC 700970 | NC 002162 | Firmicutes | 0.75 | 25.5 |
| *Rhodopirellula baltica* SH 1 | NC 005027 | Planctomycetes | 7.1 | 55.4 |
| *Acidovorax ebreus* TPSY | NC 011992 | Proteobacteria | 3.8 | 66.8 |
| *Bordetella avium* 197N | NC 010645 | Proteobacteria | 3.73 | 61.6 |
| *Haemophilus influenzae* 86-028NP | NC 007146 | Proteobacteria | 1.9 | 38.2 |
| *Helicobacter pylori* P12 | NC 011498 | Proteobacteria | 1.67 | 38.8 |
| *Brucella abortus* S19 | NC 010742 | Proteobacteria | 2.12 | 57.2 |
| *Agrobacterium tumefaciens* str. C58 | NC 003062 | Proteobacteria | 5.65 | 59.0 |
| *Neisseria gonorrhoeae* FA 1090 | NC 002946 | Proteobacteria | 2.15 | 52.7 |
| *Citrobacter koseri* ATCC BAA-895 | NC 009792 | Proteobacteria | 4.72 | 53.8 |
| *Anaeromyxobacter dehalogenans* 2CP-C | NC 007760 | Proteobacteria | 5.0 | 74.9 |
| *Bradyrhizobium japonicum* USDA 110 | NC 004463 | Proteobacteria | 9.1 | 64.1 |
| *Desulfatibacillum alkenivorans* AK-01 | NC 011768 | Proteobacteria | 6.5 | 54.5 |
| *Ralstonia eutropha* JMP134 | NC 007347 | Proteobacteria | 7.26 | 64.4 |
| *Caulobacter crescentus* CB15 | NC 002696 | Proteobacteria | 4.0 | 67.2 |
| *Erwinia tasmaniensis* Et1/99 | NC 010694 | Proteobacteria | 3.9 | 53.4 |
| *Thauera sp.* MZ1T | NC 011662 | Proteobacteria | 4.5 | 68.3 |
| *Variovorax paradoxus* S110 | NC 012791 | Proteobacteria | 6.7 | 67.5 |
| *Gluconacetobacter diazotrophicus* PAl 5 | NC 011365 | Proteobacteria | 3.9 | 66.3 |
| *Syntrophobacter fumaroxidans* MPOB | NC 008554 | Proteobacteria | 5.0 | 59.9 |
| *Rhodobacter sphaeroides* ATCC 17025 | NC 009428 | Proteobacteria | 4.61 | 68.8 |
| *Rhodopseudomonas palustris* BisA53 | NC 008435 | Proteobacteria | 5.5 | 64.4 |
| *Rickettsia rickettsii* str. Iowa | NC 010263 | Proteobacteria | 1.3 | 32.4 |
| *Pseudomonas aeruginosa* PA7 | NC 009656 | Proteobacteria | 6.6 | 66.4 |
| *Wolbachia* endosymbiont of *Drosophila melanogaster* | NC 002978 | Proteobacteria | 1.26 | 35.2 |
| *Salmonella enterica* subsp. arizonae serovar 62:z4,z23:– RSK2980 | NC 010067 | Proteobacteria | 4.6 | 51.4 |
| *Shigella sonnei* Ss046 | NC 007384 | Proteobacteria | 5.03 | 50.8 |
| *Thioalkalivibrio sp.* HL-EbGR7 | NC 011901 | Proteobacteria | 3.46 | 65.1 |
| *Xanthomonas axonopodis* pv. citri str. 306 | NC 003919 | Proteobacteria | 5.27 | 64.7 |

uses a numerical expression for the secular part. As a result, instead of using an analytical expression of the secular part, we use a polynomial fitting taking the degree $n$ of the polynomial as the minimum in which the fluctuations are stationary:

$$\langle \mathrm{N}^{\beta}(x) \rangle \approx a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n. \tag{4}$$

Here $a_0, a_1, \ldots, a_n$, are the constants of the fitting. We use the degree of the polynomial $n = 8$ since, in this case, the fluctuations are stationary and the zone close to the origin of replication can also be analyzed within the same framework. Fortunately the constants tend to zero as $n$ increases. The resulting unfolding procedure is thus equivalent to the detrended fluctuation analysis [17]. However, while the latter method uses a linear fitting, unfolding uses the average of the cumulative density of bases. The obtention of $\langle \mathrm{N}^{\beta}(x) \rangle$ is the same as any regression analysis. We should stress that unfolding detrends the distances conserving the fluctuations.

**Table 2**
List of bacterial genomes classified with one slope in the averaged cumulative bases density.

| Name | Ref. seq. | Group | Size | %GC |
|---|---|---|---|---|
| *Acidimicrobium ferrooxidans* DSM 10331 | NC 013124 | Actinobacteria | 4.4 | 68.3 |
| *Atopobium parvulum* DSM 20469 | NC 013203 | Actinobacteria | 2.1 | 45.7 |
| *Corynebacterium aurimucosum* ATCC 700975 | NC 012590 | Actinobacteria | 2.8 | 60.6 |
| *Porphyromonas gingivalis* W83 | NC 002950 | Bacteroidetes/chlorobi | 2.34 | 48.3 |
| *Chlorobium limicola* DSM 245 | NC 010803 | Bacteroidetes/chlorobi | 2.76 | 51.3 |
| *Cytophaga hutchinsonii* ATCC 33406 | NC 008255 | Bacteroidetes/chlorobi | 4.43 | 38.8 |
| *Chlamydia trachomatis* 434/Bu | NC 010287 | Chlamydiae/verrucomicrobia | 1.03 | 41.3 |
| *Chlamydophila pneumoniae* TW-183 | NC 005043 | Chlamydiae/verrucomicrobia | 1.22 | 40.6 |
| *Dehalococcoides ethenogenes* 195 | NC 002936 | Chloroflexi | 1.46 | 48.9 |
| *Thermomicrobium roseum* DSM 5159 DSM5159 | NC 0119559 | Chloroflexi | 2.0 | 64.3 |
| *Prochlorococcus marinus* str. MIT 9211 MIT9211 | NC 009976 | Cyanobacteria | 1.7 | 38.0 |
| *Desulfotomaculum acetoxidans* DSM 771 | NC 013216 | Firmicutes | 4.54 | 41.4 |
| *Listeria innocua* Clip11262 | NC 003212 | Firmicutes | 3.0 | 37.4 |
| *Lactobacillus casei* ATCC 334 | NC 008526 | Firmicutes | 2.9 | 46.6 |
| *Geobacillus thermodenitrificans* NG80-2 | NC 009328 | Firmicutes | 3.55 | 48.9 |
| *Streptococcus agalactiae* A909 | NC 007432 | Firmicutes | 2.13 | 35.6 |
| *Staphylococcus aureus* subsp. aureus COL | NC 002951 | Firmicutes | 2.8 | 32.8 |
| *Thermoanaerobacter tengcongensis* MB4 | NC 003869 | Firmicutes | 2.69 | 37.6 |
| *Rhodopirellula baltica* SH 1 | NC 005027 | Planctomycetes | 7.14 | 55.4 |
| *Acetobacter pasteurianus* IFO 3283-01 | NC 013209 | Proteobacteria | 2.9 | 53.0 |
| *Acinetobacter baumannii* AB0057 | NC 011586 | Proteobacteria | 4.05 | 39.2 |
| *Arcobacter butzleri* RM4018 | NC 009850 | Proteobacteria | 2.34 | 27.0 |
| *Actinobacillus pleuropneumoniae* L20 | NC 009053 | Proteobacteria | 2.27 | 41.3 |
| *Bartonella bacilliformis* KC583 | NC 008783 | Proteobacteria | 1.4 | 38.2 |
| *Buchnera aphidicola* str. Cc (Cinara cedri) Cc | NC 008513 | Proteobacteria | 0.42 | 20.2 |
| *Campylobacter jejuni* RM1221 | NC 003912 | Proteobacteria | 1.77 | 30.3 |
| *Francisella tularensis* subsp. tularensis FSC198 | NC 008245 | Proteobacteria | 1.9 | 32.3 |
| *Nautilia profundicola* AmH | NC 012115 | Proteobacteria | 1.7 | 33.5 |
| *Anaplasma marginale* str. Florida | NC 012026 | Proteobacteria | 1.2 | 49.8 |
| *Nitrosomonas europaea* ATCC 19718 | NC 004757 | Proteobacteria | 2.81 | 50.7 |
| *Coxiella burnetii* CbuG_Q212 | NC 011527 | Proteobacteria | 2 | 42.6 |
| *Polynucleobacter necessarius* subsp. asymbioticus QLW-P1DMWA-1 | NC 009379 | Proteobacteria | 2.2 | 44.8 |
| *Edwardsiella ictaluri* 93-146 | NC 012779 | Proteobacteria | 3.8 | 57.0 |
| *Desulfovibrio desulfuricans* subsp. desulfuricans str. ATCC 27774 | NC 011883 | Proteobacteria | 2.9 | 58.1 |
| *Francisella novicida* U112 | NC 008601 | Proteobacteria | 1.9 | 32.5 |
| *Ehrlichia canis* str. Jake | NC 011365 | Proteobacteria | 1.3 | 29.0 |
| *Kangiella koreensis* DSM 16069 | NC 013166 | Proteobacteria | 2.9 | 43.7 |
| *Pelobacter carbinolicus* DSM 2380 | NC 007498 | Proteobacteria | 3.7 | 55.1 |
| *Helicobacter pylori* 26695 | NC 000915 | Proteobacteria | 1.7 | 38.9 |
| *Legionella pneumophila* str. Corby | NC 009494 | Proteobacteria | 3.6 | 38.5 |
| *Granulibacter bethesdensis* CGDNIH1 | NC 008343 | Proteobacteria | 2.7 | 59.1 |
| *Marinomonas sp.* MWYL1 | NC 009654 | Proteobacteria | 5.1 | 42.6 |
| *Psychrobacter arcticus* 273-4 | NC 007204 | Proteobacteria | 2.7 | 42.8 |
| *Shewanella denitrificans* OS217 OS-217 | NC 007954 | Proteobacteria | 4.55 | 45.1 |
| *Vibrio cholerae* M66-2 | NC 012578 | Proteobacteria | 3.9 | 47.6 |
| *Leptospira biflexa* serovar Patoc strain Patoc 1 (Ames) | NC 010842 | Spirochetes | 3.95 | 38.9 |
| *Treponema denticola* ATCC 35405 | NC 002967 | Spirochetes | 2.8 | 37.9 |

### 2.3. Codon usage

To test the role of codon bias in the generation of different base distributions, we created artificial genomic sequences where the most frequent codons in the real genomes were substituted by the least frequent synonymous codons and vice versa. The resulting artificial genomes, hereafter referred to as *modified sequences*, codify for the same set of proteins as the real genomes, but with a totally different codon bias. These modified sequences were created from the whole set of bacteria given in Table 3. The codon usage tables were taken from the Codon Usage Database [18]. As an example, consider valine in *C. tetani*. Valine is codified by the codons GUU, GUC, GUA, and GUG with frequencies 18.2, 2.3, 21.7, and 4.7, respectively. Then, we interchanged the least frequently used codon (GUC) and the most frequently used codon (GUA), i.e., we interchanged GTA for GTC in the modified sequence.

In a second experiment, a stronger change was performed. We replace all codons that codify a given amino acid in a certain species by the most frequent codon of a second species. To clarify this point, consider valine in both *C. botulinum* and *B. cepacia*. The codon usage of valine for *B. cepacia* is GUU (0.30%), GUC (3.40%), GUA (0.22%) and GUG (3.83%). The codon usage of *C. botulinum* for the valine is GUU (2.34%), GUC (0.15%), GUA (3.13%) and GUG (0.58%). Then, in this second experiment the codons GTT, GTC and GTA of *Clostridium botulinum* are replaced by GTG to obtain a modified *C. botulinum* sequence.

**Table 3**
List of species of bacterial used in this work.

| Name | Ref. seq. | Group | Size | %GC |
|---|---|---|---|---|
| *Burkholderia cepacia* AMMD | NC 008391.1 | $\beta$-proteobacteria | 8.0 | 65.0 |
| *Burkholderia pseudomallei* | NC 009076 | $\beta$-proteobacteria | 7.31 | 68.0 |
| *Burkholderia thailandensis* E264 | NC 007651 | $\beta$-proteobacteria | 6.71 | 67.6 |
| *Burkholderia phymatum* STM815 | NC 010622 | $\beta$-proteobacteria | 8.7 | 62.3 |
| *Burkholderia xenovorans* LB400 | NC 007951 | $\beta$-proteobacteria | 9.8 | 62.6 |
| *Bacillus amyloliquefaciens* FZB42 | NC 009725 | Firmicutes | 3.9 | 46.5 |
| *Bacillus clausii* KSM K16 | NC 006582 | Firmicutes | 4.3 | 44.8 |
| *Bacillus pumilus* SAFR-032 | NC 009848 | Firmicutes | 3.7 | 41.3 |
| *Bacillus thuringiensis* Al Hakam | NC 008600 | Firmicutes | 5.36 | 35.4 |
| *Bacillus weihenstephanensis* KBAB4 | NC 010184 | Firmicutes | 5.91 | 35.5 |
| *Clostridium acetobutylicum* | NC 003030 | Firmicutes | 4.09 | 30.9 |
| *Clostridium beijerinckii* NCIMB 8052 | NC 009617 | Firmicutes | 6.0 | 29.9 |
| *Clostridium phytofermentans* ISDg | NC 010001 | Firmicutes | 4.8 | 35.3 |
| *Clostridium botulinum* A | NC 009495 | Firmicutes | 3.9 | 28.2 |
| *Clostridium tetani* E88 | NC 004557 | Firmicutes | 2.87 | 28.6 |
| *Corynebacterium diphtheriae* | NC 002935 | Actinobacteria | 2.83 | 60.6 |
| *Corynebacterium efficiens* YS-314 | NC 004369 | Actinobacteria | 3.22 | 63.0 |
| *Corynebacterium jeikeium* K411 | NC 007164 | Actinobacteria | 2.51 | 61.4 |
| *Corynebacterium glutamicum* ATCC 13032 Bielefeld | NC 006958 | Actinobacteria | 3.3 | 53.8 |
| *Corynebacterium urealyticum* DSM 7109 | NC 010545 | Actinobacteria | 2.4 | 64.2 |

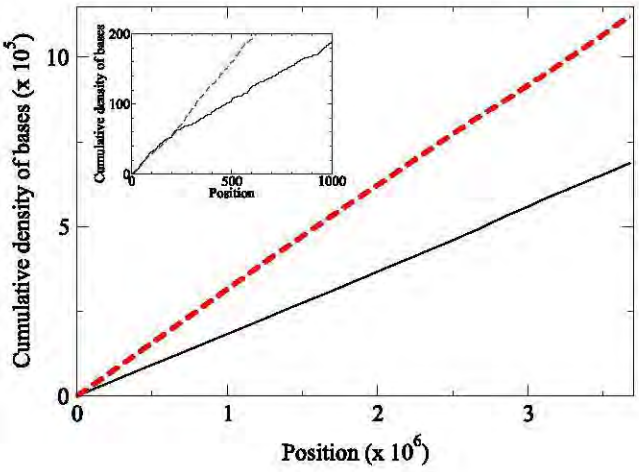



**Fig. 1.** (Color online) Cumulative density of bases in *B. phymatum*. The continuous (black) line corresponds to adenine, and the dashed line (red) corresponds to guanine. Thymine has a similar behavior to adenine. The same happens with cytosine and guanine. In the inset, a zoom showing the fluctuations is given.

To measure the similarity between distributions, we followed the suggestion by Vegelius et al. [19] and used the proportional similarity (PS). PS between two distributions $U$ and $V$ is defined as

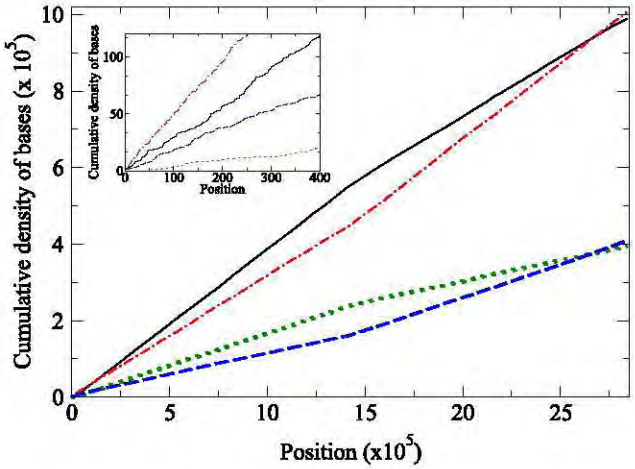
$$\mathrm{PS} = \sum_{i=1}^{C} \min(f_{U_i}, f_{V_i}), \tag{5}$$

where $f_{U_i}$ and $f_{V_i}$, $i = 1, \ldots, C$ are the relative frequencies of $U$ and $V$ in each one of the $C$ categories. Higher (lower) numbers indicate that the distributions are more (less) similar to the other. The proportional similarity ranges from 1 for equal distributions to 0 for distributions with no similarity.

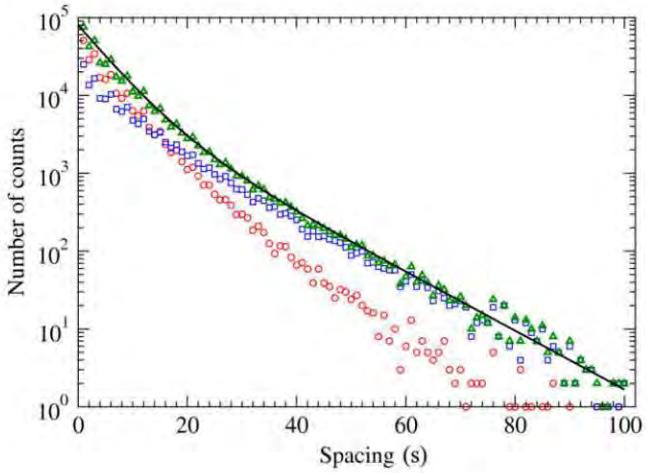
## 3. Results and discussion

The cumulative density of bases was analyzed for the set of 112 bacterial genomes shown in Tables 1 and 2. As explained in the methodology, each genomic sequence is decomposed in two parts, the average and the fluctuations. On the one hand, the average of the CDoB shows one of these two possible behaviors: (a) a linear trend or (b) a piecewise-linear trend, composed of two consecutive lines with different slopes. The fluctuations, on the other hand, show an almost decaying exponential distribution with peaks at short distances. Then, short distances are favored instead of large distances.

Fig. 1 shows the statistical analysis of the *B. phymatum* genome, which is a representative case where the CDoB has a linear trend. Note that the slope of the CDoB of guanine is higher than the slope of the CDoB of adenine. This is consistent with the genome composition since *B. phymatum* is GC rich. Note the difference between Figs. 1 and 2 showing the analysis




ı *C. tetani*. The continuous (black) line corresponds to adenine, the dash-dotted line (red) corresponds to
ıanine, and the dashed (blue) line corresponds to cytosine. In the inset, a zoom showing the fluctuations




ne for the whole genome (green triangles) of *C. tetani*. Red circles correspond to the same analysis, but
ıe (before the change in slope observed in Fig. 2). Blue squares correspond to the analysis performed on
he line corresponds to $f(d) = 70\,000 * \exp(-d/5) + 10\,000 * \exp(-d/11.5)$.

ase with a piecewise-linear CDoB. We can observe from Fig. 2 that the sequence is
behavior is not uniform along the genomic sequence. In adenine and guanine, the
ıigher density than in the second half of the sequence but with different contents.
ıine again with different contents. The abrupt change in the slope in Fig. 2 is located
0,21]. This result was consistently found for all the bacteria with piecewise-linear

·linear CDoB, there are different statistical properties before and after the change
f distances between nucleotides. Note in Fig. 3 that the distribution of distances is
he region before the CDoB bend, and (iii) the region after the same bend. It can be
ɔility to find a guanine close to another is very high and that it is possible to find
the order of ≈80 nucleotides. It is also possible to see that the distribution of the
on of the complete sequence at distances below ≈10 bases, while the second part
ɔlete sequence at distances beyond ≈40 bases. Similar results were found for all
trends. In all these cases, the distribution of distances between nucleotides in the
'o different distributions, one before and one after the origin of replication. Each of
ıential functions, which added are the distribution of the whole genome sequence.
the distribution of distances $d$ between nucleotides for all the genome is very close
$) * \exp(-d/11.5)$, where the first term reflects the behavior of the distribution at
ɜ the behavior at long distances (circles and squares in Fig. 3, respectively). Bacteria

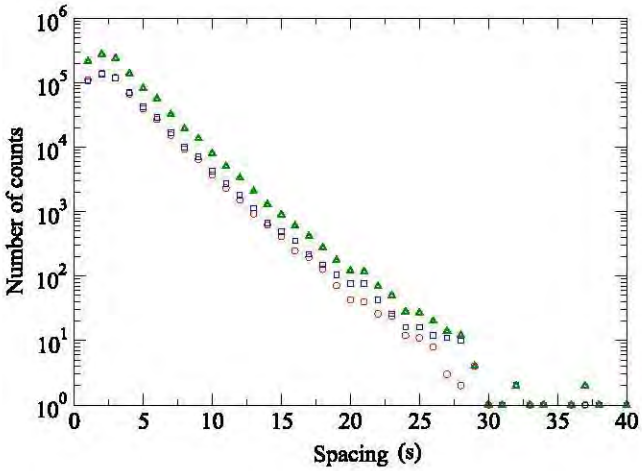



**Fig. 4.** (Color online) Spacing distribution of guanine (not normalized) for *B. phymatum*. The triangles up (green) correspond to the complete sequence, the circles (red) correspond to the first part of the sequence, the squares (blue) correspond to the second part of the sequence.

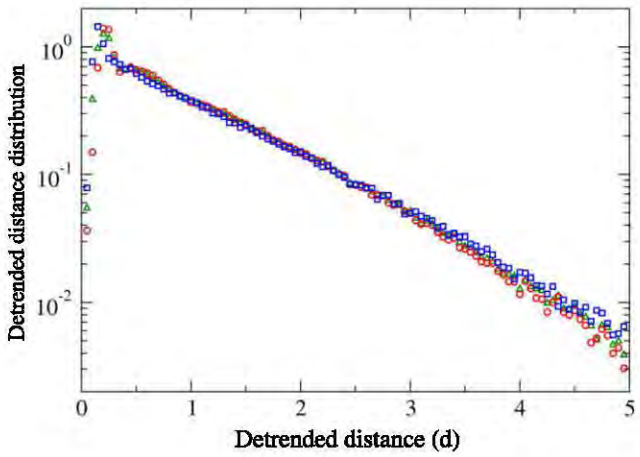



**Fig. 5.** (Color online) Detrended distance distribution of guanine for *C. tetani* for the unfolded sequence. The triangles up (green) correspond to the complete sequence, the circles (red) correspond to the first part of the sequence, the squares (blue) correspond to the second part of the sequence.

with a linear CDoB, on the other hand, have a constant average density of bases throughout the genome. A representative case is shown in Fig. 4, where the histogram of the distances between guanines in the genome of *B. phymatum* is shown. Note that the distribution of the two halves of the genome are almost identical. Again one can see that the probability to find guanines separated by other nucleotides is very high and that it is possible to find two guanines separated up to ≈25 nucleotides.

We have shown that several genomes show statistical properties that vary along the nucleotide sequence, making the comparison of distribution trends among species of bacteria difficult. Because of this characteristic, the application of the unfolding procedure (see Methods) becomes relevant, which eliminates the average slopes and allows a reliable comparison of the fluctuations in the distributions of bases. We applied the unfolding procedure to all the bacterial genomes with piecewise linear trends (Table 1). As a representative result, Fig. 5 shows the detrended distance distribution of *C. tetani* after the unfolding procedure was applied. As can be seen, the difference between the histograms of the regions separated by the origin of replication (Fig. 3) does not appear in the detrended distance distribution of Fig. 5. Another result of the unfolding procedure, is the change of scale (compare the horizontal axes in Figs. 3 and 5). In Fig. 5, an unfolded spacing of 1 corresponds to a true average distance of 5.6 between guanines. Furthermore, the peak at ≈0.18 corresponds to the minimal distance between bases. The biological description of the detrended distance distribution of Fig. 5 is similar to the one of Fig. 3.

When the unfolding procedure is applied to bacterial genomes with linear CDoB, the results conserve the identity observed in all regions of the genome. Fig. 6 shows the result of applying the unfolding procedure to the genome of *B. phymatum*. In this case, an unfolded spacing of 1 corresponds to a true average distance of 3.2. The first and second peaks at ≈0.33 and ≈0.66 correspond to the peaks at true distances 1 and 2 of Fig. 4, respectively. The biological description of the detrended distance distribution of Fig. 6 is similar to the one of Fig. 4.

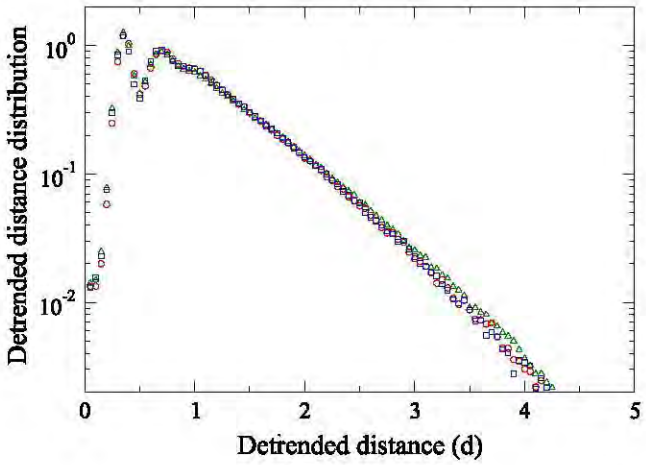



**Fig. 6.** (Color online) Detrended distance distribution of guanine for *B. phymatum* for the unfolded sequence. The triangles up (green) correspond to the complete sequence, the circles (red) correspond to the first part of the sequence, the squares (blue) correspond to the second part of the sequence.

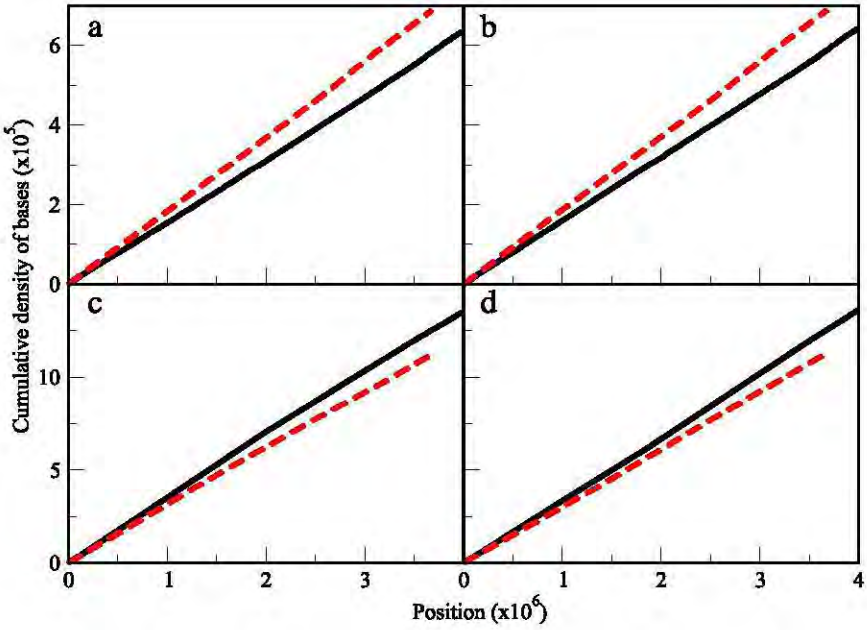



**Fig. 7.** (Color online) Cumulative density of bases for some species of *Burkholderia*. The continuous (black) line corresponds to *B. pseudomallei* and the dashed (red) line to *B. phymatum*: (a) adenine, (b) thymine, (c) guanine, and (d) cytosine.

### 3.1. Codon usage

In what follows, using the same methodology, we will show that different species of the same genus can have different CDoB but a similar detrended distance distribution. Also, we will show that the detrended distance distribution is different for species of different genera and we give evidence that this difference comes from the codon usage. This similarity or non-similarity between distributions will be quantified using the proportional similarity PS. In this example, several species of the *Burkholderia, Bacillus, Clostridium* and *Corynebacterium* genera are analyzed. Here, we will only show figures for *Burkholderia* and *Clostridium*.

In Figs. 7 and 8, the staircase function is plotted for some species of the genera *Burkholderia* and *Clostridium*. The list of the species analyzed is given in Table 3, and the detrended distance distribution obtained for the species of *Clostridium* used in Fig. 8 is given in Fig. 9. As can be seen, the distributions of adenine for all species of *Clostridium* are very similar to each other, although they do not overlap. The proportional similarity between the adenine distributions is PS = 0.86. The same happens for the distributions of thymine (PS = 0.86), guanine (PS = 0.96) and cytosine (PS = 0.97) of *Clostridium*. This distribution is very close to an exponential distribution. In Fig. 10, the detrended distance distribution are plotted for the same species of *Burkholderia* as those of Fig. 7. As can be seen, the adenine distribution is very similar (PS = 0.96) for all species of *Burkholderia* used. Something similar happens for thymine (PS = 0.96), for guanine (PS = 0.89) and cytosine

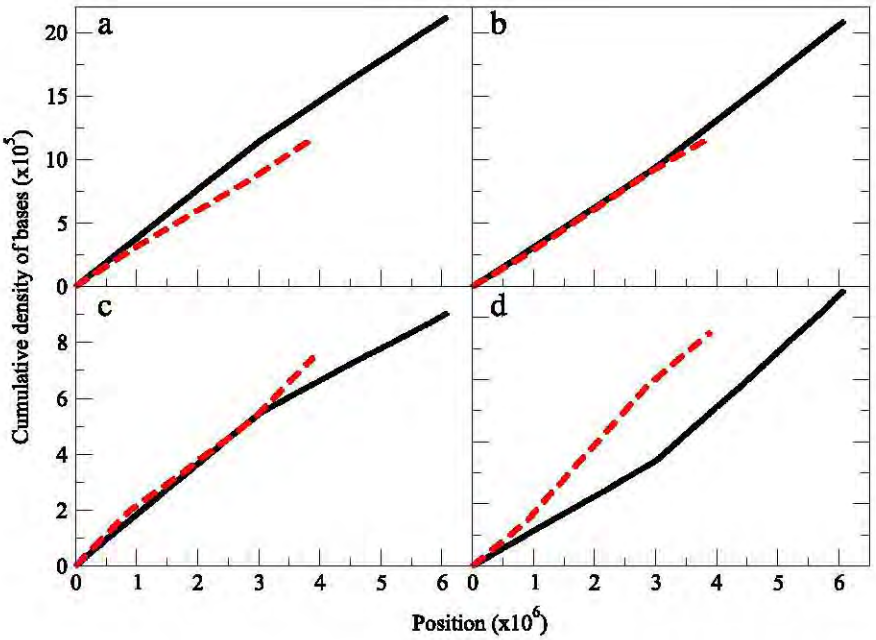



**Fig. 8.** (Color online) Cumulative density of bases for *Clostridium beijerinckii* (continuous black line) and for *C. tetani* (dashed red line); (a) adenine, (b) thymine, (c) guanine, and (d) cytosine.

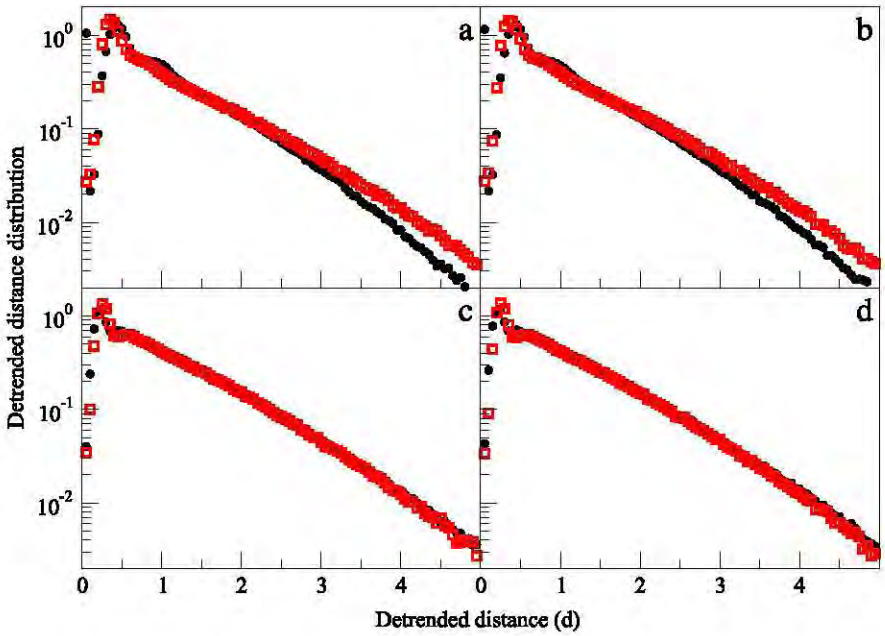



**Fig. 9.** (Color online) Detrended distance distribution for the same sequences of Fig. 8 and following the same order. The circles (black) correspond to *C. beijerinckii* and squares (red) to *C. tetani*.

(PS = 0.89); there are some small differences close to $s = 0$, i.e., different species have a different value of $p(s \approx 0)$. The rest of the distribution is quite similar for all species within the same genus.

Now we will show that species of a different genus have different detrended distance distributions. In Fig. 11, we show the comparison of the detrended distance distribution for two species belonging to different genera, *B. cepacia* and *C. botulinum*. As can be seen, the differences appear at short distances as well as in the slope of the distribution at large distances. On the other case, *B. cepacia* (circles) has three peaks at short distances in guanine and cytosine, it also decays faster than *C. botulinum* which has only one peak at short distances (Fig. 11(c) and (d)). In adenine and thymine, there are two peaks at short distances in *C. botulinum* while in *B. cepacia*, there only appears one peak (Fig. 11(a) and (b)). The proportional similarity between those distributions is 0.81, 0.81, 0.71 and 0.71 for adenine, thymine, guanine and cytosine, respectively. This means that the distributions of *Clostridium* and *Burkholderia* are different ≈30%.

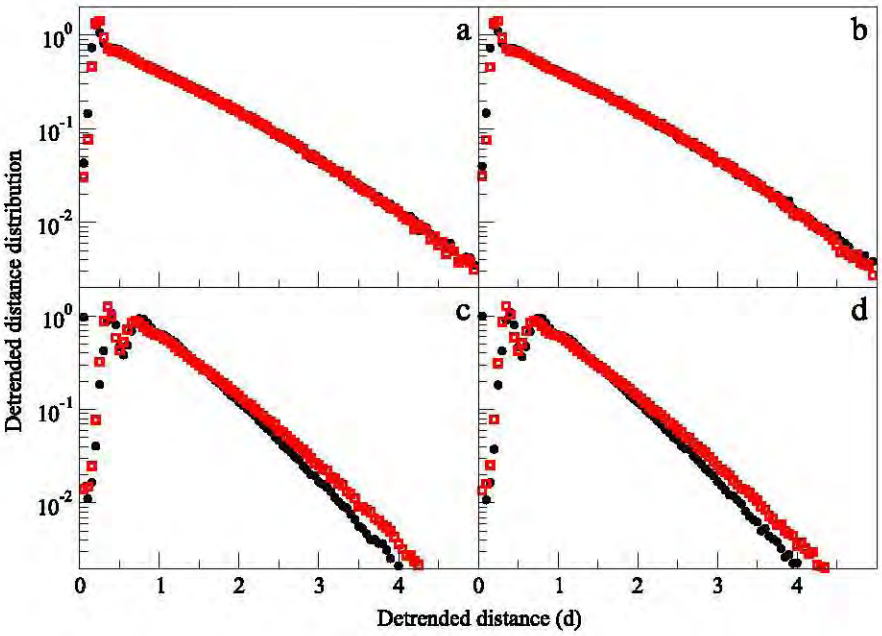



**Fig. 10.** (Color online) Detrended distance distribution for the same sequences of Fig. 7 and following the same order. The circles (black) correspond to *B. pseudomallei* and the squares (red) to *B. phymatum*.

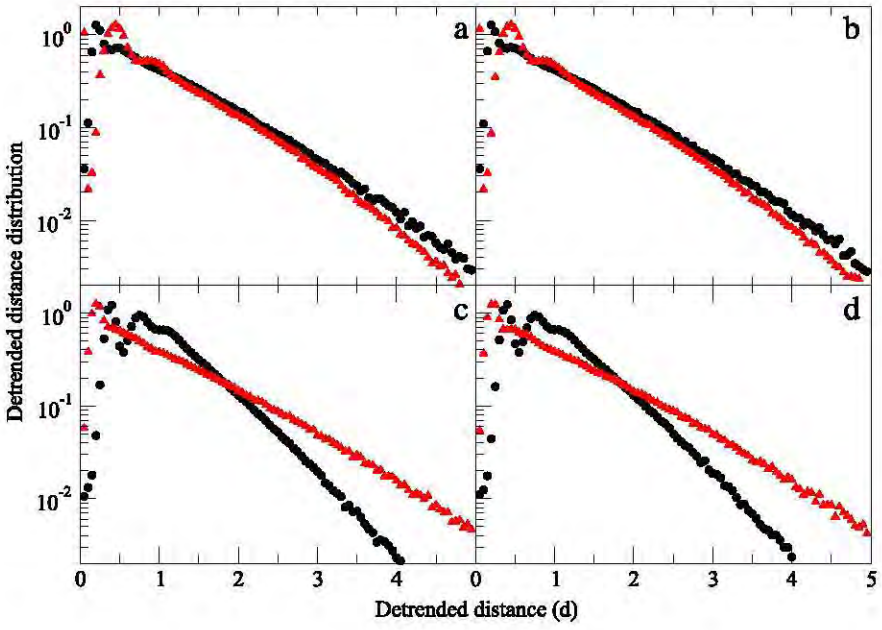



**Fig. 11.** (Color online) Comparison of the detrended distance distribution for two species belonging to different genera: (a) adenine, (b) thymine, (c) guanine, and (d) cytosine. The circles (black) correspond to *B. cepacia* and the triangles (red) to *C. botulinum*.

It is well known that prokaryotes present a higher codon usage bias than eukaryotes [22]. We will now give evidence that this codon bias is to a large extent responsible for species within the same genus having the same distribution, as well as of having different distributions among genera.

To assess the effect of codon bias on the distribution of bases, we compared the detrended distance distribution in the unfolded sequences with the modified sequences in which the most frequently used codon is changed by the least frequently used codon (see methods) obtained from them. To see the difference, we first focused on the peaks of the distributions by locating their maxima. There is an inverse proportionality between the peak heights at short distances and the slope of the distribution at long distances since the distributions are normalized. When there are high peaks at short distances, the slope at large distances decrease, if the peaks at short distances are low, the slopes have high values. The peaks are more sensitive than the slopes. For this reason, we show the results for the peaks instead of the slopes. Nevertheless, the proportional

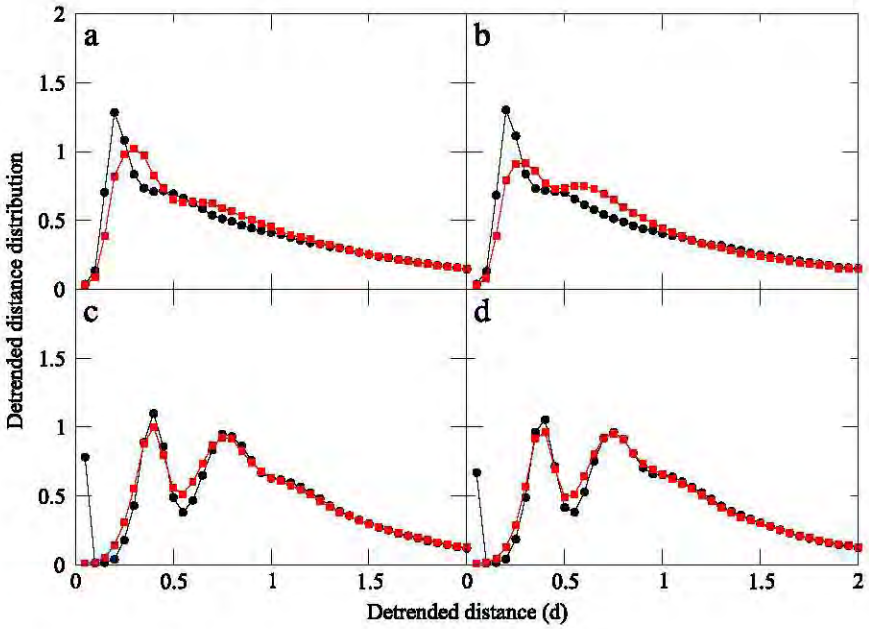



**Fig. 12.** (Color online) Detrended distance distribution of *Burkholderia thailandensis*: (a) adenine, (b) thymine, (c) guanine, and (d) cytosine. The circles (black) correspond to the original sequence and the squares (red) to the modified sequence with silent mutations changing from most to least used codon and vice versa (see text).

similarity measures both the peaks and slopes at the same time: one can see how the PS changes when the original and the modified sequences are compared.

We first show the results obtained from the genus *Burkholderia*. It is important to mention that studied species of this genus (Table 3) have the same codon usage (Not shown), hence the results shown hereby are valid for the whole genus. For the distribution of adenine, the original genomic sequence (Fig. 12(a), circles) has a peak located at $s = 0.224$, while in the modified sequence (Fig. 12(a), squares) the peak is located now at $s = 0.303$. In the distribution of thymine, there is one peak at $s = 0.225$ in the original sequence (see Fig. 12(b)), but in the modified sequence there are two peaks, one located at $s = 0.291$ and another at $s = 0.569$. Notably, the distribution of guanine and cytosine are very similar in the original and modified sequences. PS between the distributions of the original and modified sequences are 0.93, 0.93, 0.94 and 0.95 for adenine, thymine, guanine and cytosine, respectively. In other words, the distribution changes by 7%, 7%, 6%, and 5%, for adenine, thymine, guanine and cytosine, respectively.

Now, the results for the genus *Clostridium* are presented. Here, again, the studied species of this genus (Table 3) have the same codon usage (Not shown), hence the results are valid for the whole genus. Regarding the distribution of guanine, the original sequence shows two peaks at $s = 0.239$ and $s = 0.506$ (Fig. 13(c)), while the modified sequence presents only a peak at $s = 0.293$. In the distribution of cytosine, both the original and the modified sequence show two peaks (Fig. 13(d)). In the original DNA chain they are located at $s = 0.238$ and $s = 0.509$, but in the modified DNA chain they are located at $s = 0.291$ and $s = 0.837$. Back in *Clostridium*, the distributions of adenine and thymine are similar in the original and modified sequences. The proportional similarity in this case is 0.97, 0.97, 0.94 and 0.89 for adenine, thymine, guanine and cytosine, respectively.

The interchange of the less frequent codon by the most frequent and vice versa has only a limited effect on the detrended distance distributions of Figs. 12 and 13 since only less than 10% of the sequence is changed. In the new experiment, for a given amino acid we find the most frequent codon used by *B. cepacia*. Then, in the *C. botulinum* sequence, all codons that yield the same amino acid are replaced by the most frequently used in *B. cepacia*. This procedure was repeated for all amino acids. *Clostridium* is a bacterium that has a low GC content and *Burkholderia* is high in GC. Therefore, changing the codon usage in *Clostridium* by *Burkholderia*, the GC content increases. Fig. 14 shows the detrended distance distribution for two species, *C. botulinum* (black) and *B. cepacia* (red). The third distribution (blue) corresponds to a modified *C. botulinum* with the codon usage of *B. cepacia*.

The strongest changes in the distribution are observed in guanine Fig. 14(c) and cytosine Fig. 14(d). In cytosine, the peaks at short distances and slopes of *Clostridium* and *Burkholderia* are very different. The distribution of the modified *Clostridium* sequence (blue) has a slope that is more like *Burkholderia* (red). This change also occurs in the peaks at short distances. The behavior of guanine in the modified *Clostridium* sequence is similar to that of cytosine. We also analyzed the proportional similarity between the distributions of *Burkholderia* and *Clostridium* and compared the results with those of *Burkholderia* and the modified *Clostridium*. The distributions of the modified *Clostridium* are closer to those of *Burkholderia*. On the one hand, the proportional similarity between the detrended distances distribution of *Burkholderia* and *Clostridium* is 0.71 and 0.72 for guanine and cytosine respectively. On the other hand, the proportional similarity between the detrended distance

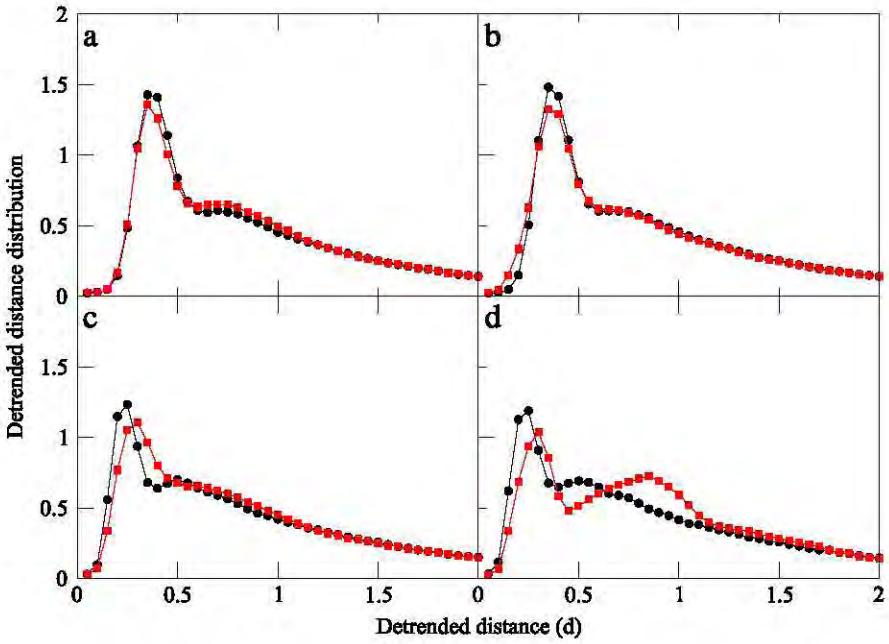



**Fig. 13.** (Color online) Detrended distance distribution of *C. phytofermentans*: (a) adenine, (b) thymine, (c) guanine, and (d) cytosine. The circles (black) correspond to the original sequence and the squares (red) to the modified sequence with silent mutations changing from most to least used codon and vice versa (see text).

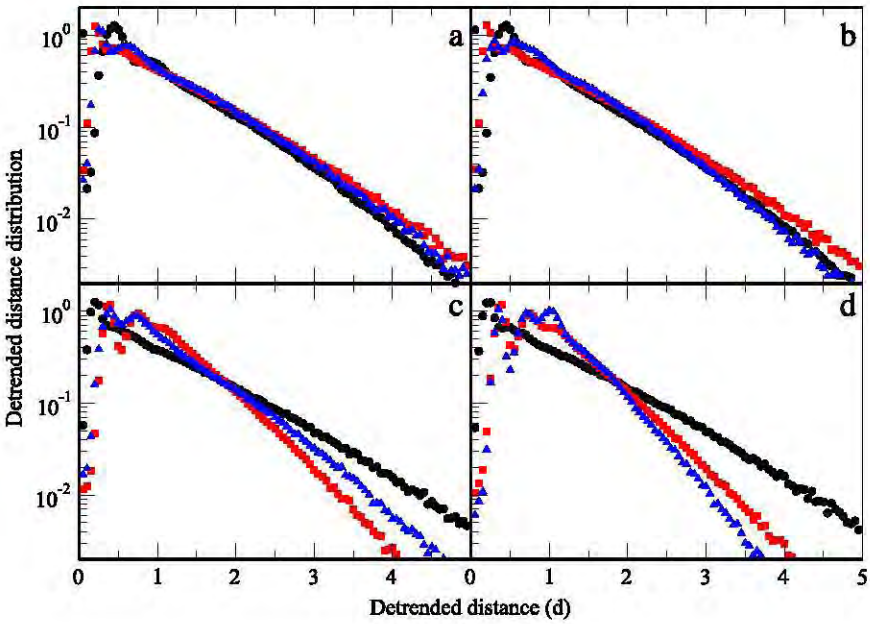



**Fig. 14.** (Color online) Comparison of the detrended distance distribution for two species belonging to different genera: (a) adenine, (b) thymine, (c) guanine, and (d) cytosine. The circles (black) correspond to *C. botulinum* and the squares (red) to *B. cepacia*. The triangles (blue) correspond to *C. botulinum* with the codon usage of *B. cepacia*.

distribution of *Burkholderia* and the modified *Clostridium* increased to 0.91 and 0.89 for guanine and cytosine, respectively. The proportional similarity increased by almost 20%.

## 4. Conclusions

We presented an analysis of the nearest neighbor spacing distribution or detrended distance distribution of nucleotides on several bacteria using *spectral statistics*. Working with the unfolding methodology, we showed that the cumulative density of bases can be separated into a secular (or average) and a fluctuating part. The secular part is always specific for the

bacterium under study. We also observed two types of trends: linear and piecewise linear, but we were unable to find correlations between the type of distribution and the phylogeny of the bacteria studied. However, the nearest-neighbor spacing distributions can group different species of the same genus despite their CDoB changes.

One implication of our results is that statistical analysis of genomes in bacteria should take into account the cumulative density of bases (CDoB), and the reason is that this density is not homogeneous (stationary) for all bacteria. Also, for species with piecewise linear distributions, the total distribution of distances results from the sum of distributions in the different regions of the genome; specifically, it was found that the statistical analysis of the detrended distance distribution of full genomes without unfolding can be interpreted as the addition of two *almost* exponential distributions.

We showed several results obtained for bacteria of the genera *Clostridium* and *Burkholderia*, as we found that they are representative cases of the kind of results obtained by the application of our methodology. Importantly, we showed that the detrended distance distribution is similar for species of the same genus even though their average CDoB is different. This was possible only because we unfolded the genomic sequences, showing that while their secular parts are different, their fluctuating parts are similar. This is a result that might have been overlooked with the use of other types of statistical analysis. Finally, we showed that the detrended distance turned out to be different when comparing bacteria of different genera. This difference, however, can be attributed in a large extent to the difference in codon usage in distinct genera. A more drastic experiment was done: the change of the codon usage of one bacterium (*Clostridium*) to the other (*Burkholderia*) was also performed. A change of the spacing distribution was shown: it is more similar to that of *Burkholderia*. This yields strong evidence that codon usage totally contributes to the form of the detrended distance distribution.

The methodology presented in this work allows a comparison between genomes coming from very different organisms. It can be applied to eukaryotes or prokaryotes and since it uses a function instead of a constant for the average, it can also be applied to bacterial plasmids in which the cumulative density of bases changes notably. This methodology could be functional also in the *Arabidopsis thaliana* and *Homo sapiens* genomes.

## Acknowledgments

M.F.H. was supported during the Ph.D. program (Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México) by a scholarship from the Consejo Nacional de Ciencia y Tecnología (México)). This work was supported by the PAPIIT DGAPA-UNAM under project IN111308. We are indebted to H. Hernández-Saldaña, X.A. Méndez-Báez and Alison Stewart for useful comments and discussions. We wish to thank the referees who helped us to improve the quality of this paper enormously.

## Appendix. Example of the unfolding procedure with a constructed example

To illustrate the implementation of unfolding, we give an example with a constructed sequence. In the first step, the positions $x_i^A$ of the adenine bases are given in a quadratic way:

$$\{x_i^A\} = \{i^2\} = \{1, 4, 9, 16, 25, 36, 49\}.$$

As the second step, some fluctuations are assigned to these positions by adding or subtracting one position at random. The random sequence obtained is $\{+1, +1, -1, -1, -1, -1, +1\}$. The positions with both a quadratic trend (secular part) and the random fluctuations are given by

$$\{x_i^A\} = \{2, 5, 8, 15, 24, 35, 50\}.$$

The empty positions are then filled at random with the nucleotides T, G and C. The resulting sequence is:

GAGTAGTACCTTCGACGGTGCCTAGTGGCGTGGGATCCTTGGTCTGTCCA.

From this sequence, one can see the positions of the T, G, and C nucleotides,

$$\{x_i^T\} = \{4, 7, 11, 12, 19, 23, 26, 31, 36, 39, 40, 43, 45, 47\},$$

$$\{x_i^G\} = \{1, 3, 6, 14, 17, 18, 20, 25, 27, 28, 30, 32, 33, 34, 41, 42, 46\},$$

$$\{x_i^C\} = \{9, 10, 13, 16, 21, 22, 29, 37, 38, 44, 48, 49\}.$$

The CDoB for this sequence is given in Fig. 15. It is easy to see from the figure that the average of the adenine CDoB is given by $\langle N^A(x)\rangle = \sqrt{x} - 1/2$. With this average we define the detrended (unfolded) positions $y_i$ as

$$y_i^A = \langle N^A(x_i)\rangle = \sqrt{x_i} - 1/2. \tag{A.1}$$

In this example we get, to one digit, the detrended positions and distances are,

$$\{y_i^A\} = \{0.9, 1.7, 2.3, 3.4, 4.4, 5.4, 6.6\}$$

and

$$\{d_i^A\} = \{0.8, 0.6, 1, 1, 1, 1.2\},$$

respectively.

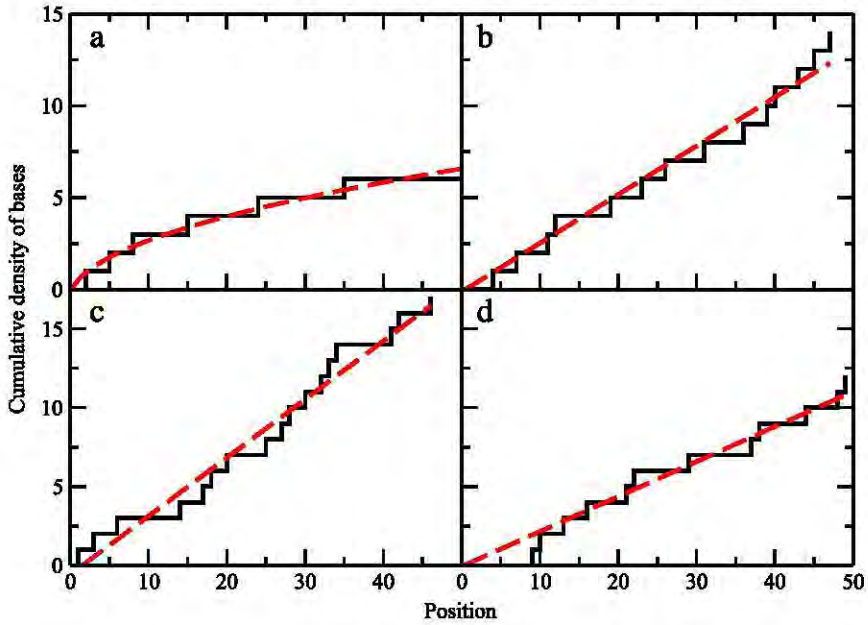



**Fig. 15.** (Color online) Cumulative density of bases for the sequence of the example (a) adenine, (b) cytosine, (c) thymine and (d) guanine. The dashed (red) line corresponds to (a) $\langle N^A(x)\rangle = \sqrt{x} - 1/2$, (b) $\langle N^T(x)\rangle = 0.26x - 0.13$, (c) $\langle N^G(x)\rangle = 0.37x - 0.59$ and (d) $\langle N^C(x)\rangle = 0.22x - 0.08$.

Note that $\langle d_i\rangle \approx 1$. Finally a histogram of the distances can be done to get the "detrended distance distribution" or "nearest neighbor spacing histogram".

The same steps used earlier, but now for guanine, thymine and cytosine can be done easily. Since the positions are given at random, the averaged cumulative densities of bases are obtained numerically fitting straight lines by least squares. In this case we get

$$\langle N^G(x)\rangle = 0.37x - 0.59, \tag{A.2}$$

$$\langle N^T(x)\rangle = 0.26x - 0.13, \tag{A.3}$$

$$\langle N^C(x)\rangle = 0.22x - 0.08, \tag{A.4}$$

for guanine, thymine and cytosine, respectively. The detrended positions and detrended distances are

$$\{y_i^G\} = \{-0.2, 0.5, 1.6, 4.6, 5.7, 6.1, 6.8, 8.7, 9.4, 9.8, 10.5, 11.2, 11.6, 12, 14.6, 15, 16.4\},$$

$$\{y_i^T\} = \{0.9, 1.7, 2.7, 3, 4.8, 5.9, 6.6, 7.9, 9.2, 10, 10.3, 11.1, 11.6\},$$

$$\{y_i^C\} = \{1.9, 2.1, 2.8, 3.4, 4.5, 4.8, 6.3, 8.1, 8.3, 9.6, 10.5, 10.7\},$$

and

$$\{d_i^G\} = \{0.3, 1.1, 3, 1.1, 0.4, 0.7, 1.8, 0.7, 0.4, 0.7, 0.7, 0.4, 0.4, 2.6, 0.4, 1.5\},$$

$$\{d_i^T\} = \{0.8, 1, 0.3, 1.8, 1, 0.8, 1.3, 1.3, 0.8, 0.3, 0.8, 0.5, 0.5\},$$

$$\{d_i^C\} = \{0.2, 0.7, 0.7, 1.1, 0.2, 1.5, 1.8, 0.2, 1.3, 0.9, 0.2\},$$

for guanine, thymine and cytosine, respectively.

## References

[1] H.S. Shapiro, E. Chargaff, Studies on the nucleotide arrangement in deoxyribonucleic acids II. Differential analysis of pyrimidine nucleotide distribution as a method of characterization, Biochim. Biophys. Acta 26 (1957) 608–623.
[2] C.K. Peng, et al., Long-range correlations in nucleotide sequences, Nature 356 (1992) 168–170.
[3] J.R. Lobry, Asymmetric substitution patterns in the two DNA strands of bacteria, Mol. Biol. Evol. 13 (1996) 650–665.
[4] R. Berezney, et al., Heterogeneity of eukaryotic replicons, replicon clusters, and replication foci, Chromosoma 108 (2000) 471–478.
[5] A. Arneodo, et al., Characterizing long-range correlations in DNA sequences from wavelet analysis, Phys. Rev. Lett. 74 (1995) 3293–3296.
[6] V. Afreixo, et al., Genome analysis with inter-nucleotide distances, Bioinformatics 25 (2009) 3064–3070.
[7] T.A. Brody, et al., Random-matrix physics: spectrum and strength fluctuations, Rev. Modern Phys. 53 (1981) 385–479.
[8] M.L. Metha, Random Matrices, 2nd ed., Academic, New York, 1991.
[9] T. Guhr, et al., Random matrix theories in quantum physics: common concepts, Phys. Rep. 299 (1998) 189–425.
[10] P. Šeba, Random matrix analysis of human EEG data, Phys. Rev. Lett. 91 (2003) 198104.
[11] M. Müller, et al., Detection and characterization of changes of the correlation structure in multivariate time series, Phys. Rev. E 71 (2005) 046116.
[12] F. Luo, et al., Application of random matrix theory to microarray data for discovering functional gene modules, Phys. Rev. E 73 (2006) 031924.

[13] M.F. Higarerda, et al., Nearest-neighbor spacing distribution of bases in some intron-less and intron-containing DNA sequences, Physica A 372 (2006) 368–373.
[14] A.K. Mohanty, A.V.S.S. Narayana-Rao, Factorial moments analyses show a characteristic length scale in DNA sequences, Phys. Rev. Lett. 84 (2000) 1832–1835.
[15] W. Li, W. Kaneko, Long-range correlation and partial $1/f^{\alpha}$ spectrum in a noncoding DNA sequence, Europhysics 17 (1992) 655–660.
[16] R.F. Voss, Evolution of long-range fractal correlations and 1/f noise in DNA base sequences, Phys. Rev. Lett. 68 (1992) 3805–3808.
[17] C.K. Peng, et al., Mosaic organization of DNA nucleotides, Phys. Rev. E 49 (1994) 1685–1689.
[18] Y. Nakamura, et al., Codon usage tabulated from the international DNA sequence databases: status for the year 2000, Nucleic Acids Res. 28 (2000) 292.
[19] J. Vegelius, S. Janson, F. Johansson, Measures of similarity between distributions, Qual. Quant. 20 (1986) 437–441.
[20] J. Sánchez, M.V. José, Bilateral inverse symmetry in whole bacterial chromosomes, Biochem. Biophys. Res. Commun. 299 (2002) 126–134.
[21] M.V. José, et al., Statistical properties of DNA sequences revisited: the role of inverse bilateral symmetry in bacterial chromosomes, Physica A 351 (2005) 477–498.
[22] J.C. Biro, Does codon bias have an evolutionary origin? Theor. Biol. Med. Modelling 5 (2008) 16.

## 4.2. Resultados adicionales

### 4.2.1. Universal Statistical Properties in DNA: Beyond Random Sequences

Basado en le modelo de máxima entropía se construyó un modelo de las fluctuaciones de distancia entre pares de bases. Se comparo este modelo con los datos experimentales observando que se ajustan a una distribución de Poisson.

# Universal Statistical Properties in DNA: Beyond Random Sequences

M. F. Higareda[1] and R. A. Méndez-Sánchez[2]

[1]*Instituto de Investigaciones Biomédicas,*
*Universidad Nacional Autónoma de México, A. P. 70228,*
*Ciudad Universitaria, C.P. 04510 México D. F. México*

[2]*Instituto de Ciencias Físicas, Universidad Nacional Autónoma de México,*
*A.P 48-3, 62210, Cuernavaca, Morelos, México*

## Abstract

Based on the maximum entropy approach, a model for the fluctuations of distances between pairs of basis in DNA is obtained. As a starting point, a Poissonian distribution with a short-range repulsion is used. Within this model, the nearest-neighbor distribution and the number variance or spectral rigidity are obtained. To test the validity of this model we analyze some sequences of what can be called the *genomic ensemble*. Agreement between the model and the statistical properties of some sequences of the genomic ensemble is obtained in several decades for the nearest-neighbor spacing distribution although the statistical properties of other sequences are not correctly described. Since the fluctuating properties obtained are universal, the genomic densities of basis can be used as a classification scheme for the genomic sequences. Some biological consequences are discussed.

PACS numbers: 87.15.Cc, 87.10.+e, 87.14.Gg

In recent times sequencing of several genomes opened new questions in theoretical biology. Large data of genomic sequences are now available and ready to be analyzed. The statistical properties of those sequences are of broad interest several biological properties can be inferred from them [1]. Several authors studied different statistical measures like correlations in position [2–7] and distance [8]; mutual information function [9–12], wavelets [2, 14], Fourier spectra [3], and frequency distributions on $n$-tuples [16] among others. Unfortunately, models predicting the properties found in the statistical analysis are scarce [17].

In this paper we propose a model, based on the maximum entropy approach, for the distribution of the distances between basis. This model assumes that the secular part of the density of basis vary smoothly [18, 19]. The model will be compared with numerical analysis of genomic sequences. Thus, the fluctuations of the distances between similar basis in some genomic sequences will be analyzed using the tools of spectral statistics [20–22]. The theoretical nearest-neighbor spacing distribution obtained with the model will be compared with a few eukaryotic and prokaryotic genomes.

To obtain the nearest neighbor spacing distribution $p(s)$, between similar basis, its is neccesary to maximize the informatic entropy defined as

$$\mathcal{S} = -\int_0^\infty \mathrm{d}s\ p(s)\ln p(s), \tag{1}$$

where $s$ is the distance between basis. The maximization should be done subject to 3 constrictions: (a) the normalization of the distribution, the second (b) fixes the average distance $\langle s\rangle$ between basis. The third condition (c) will be taken as a repulsion at a distances shorter than $a$. This condition should reflect a minimal distance between basis given by $1/N$ where $N$ is the number of basis in the genome. Since the last condition is difficult to write as an integral ranging from 0 to $\infty$, the procedure that we follow is to put an artificial hole at short distances with a shifted Heaviside function $\Theta(s-a)$. This allows the maximization of the distribution is then maximized subject to constrictions (a) and (b) only.

The normalization of the distribution then reads

$$\int_0^\infty \mathrm{d}s\Theta(s-a)p(x) = 1, \tag{2}$$

and the average $\langle s\rangle$ is fixed with

$$\int_0^\infty \mathrm{d}s\ \Theta(s-a)sp(s) = \langle s\rangle. \tag{3}$$

Then, the distribution that maximizes the entropy, subject to the constrictions given in Eqs. (2) and (3), is given by

$$p(s) = e^{-1-\alpha\Theta(s-a)-\beta s\Theta(s-a)}, \tag{4}$$

where $\alpha$ and $\beta$ are Lagrange multipliers than can be evaluated obtained using Eqs. (2) and (3). The normalization condition yields

$$\frac{1}{\beta}e^{-1-\alpha-\beta a} = 1, \tag{5}$$

while the first moment gives

$$\frac{1}{\beta}e^{-1-\alpha-\beta a}\left[a + 1/\beta\right] = \langle x\rangle. \tag{6}$$

The solution of this system of equations is

$$\beta = \frac{1}{\langle s\rangle - a} \tag{7}$$

$$e^{-1-\alpha} = \frac{1}{\langle s\rangle - a}e^{\beta a}. \tag{8}$$

Then the nearest neighbor spacing distribution between basis is then a shifted exponential

$$p(s) = \frac{\Theta(s-a)}{\langle s\rangle - a}\exp\left(-\frac{s-a}{\langle s\rangle - a}\right). \tag{9}$$

This distribution yields the Poissonian result when $a \to 0$.

In the same way, the spacing distributions $p(n;s)$ of basis, with $n$ basis in between, can be calculated. To obtain the $n$-neighbor spacing distribution a normalization constriction is needed and the average should be fixed by

$$\int_0^\infty \mathrm{d}s\ \Theta(s-(n+1)a)xp(n;s) = (n+1)\langle s\rangle. \tag{10}$$

Notice that $p(s) = p(0;s)$. For this distribution the Lagrange multipliers are given by

$$\beta = \frac{1}{(n+1)(\langle s\rangle - a)} \tag{11}$$

and

$$e^{-1-\alpha} = \frac{1}{(n+1)(\langle s\rangle - a)}\exp\left(\frac{(n+1)a}{(n+1)(\langle s\rangle - a)}\right). \tag{12}$$

The resulting distributions are

$$p(n;s) = \frac{\Theta(s-(n+1)a)}{(n+1)\left(\langle s\rangle - a\right)}\exp\left(-\frac{s-(n+1)a}{(n+1)(\langle s\rangle - a)}\right). \tag{13}$$

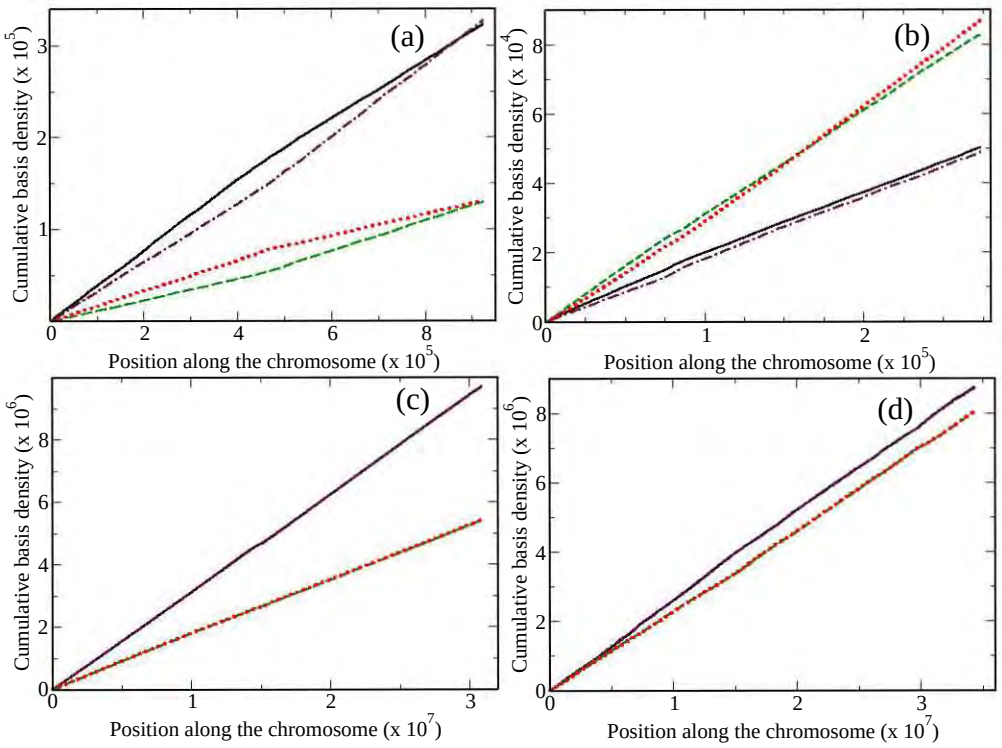



FIG. 1: (color online) Cumulative basis density of some genomic sequences (a) *Borrelia burgdorferi*, (b) *Leishmania major*, (c) *Arabidopsis thaliana*, chromosome 1, and (d) *Homo sapiens*, chromosome 22. The solid (black) line corresponds to adenine (A), the dash-dotted (maroon) line to thymine (T), the dotted (red) line to cytosine (C) and the dashed line (green) to guanine (G).

When $n = 0$ the last equation agrees with Eq. (9).

The number variance can be calculated using [22–24]

$$\Sigma^2(L) = L - 2\int_0^L ds(L-s)(1 - \sum_{n=0}^{\infty} p(n;s)). \tag{14}$$

The final result is

$$\Sigma^2(L) = L - L^2 - 2\left(\frac{(a-1)^2(a-e)e^{\frac{1+a(-2+L)}{a-1}}}{a^2} + \frac{(a-1)(a-e)e^{-1+a}(1+a^2-a(1+L))}{a^2}\right)$$

that agrees with Poisson's result for $a \to 0$. The last equation is valid only for distances in which long-range correlations are not present. Also, for short distances the result is quite similar to the Poisson result while at long distances, as expected, it goes down. This function is plotted in Fig. **??** for some values of the repulsion-hole parameter $a$.

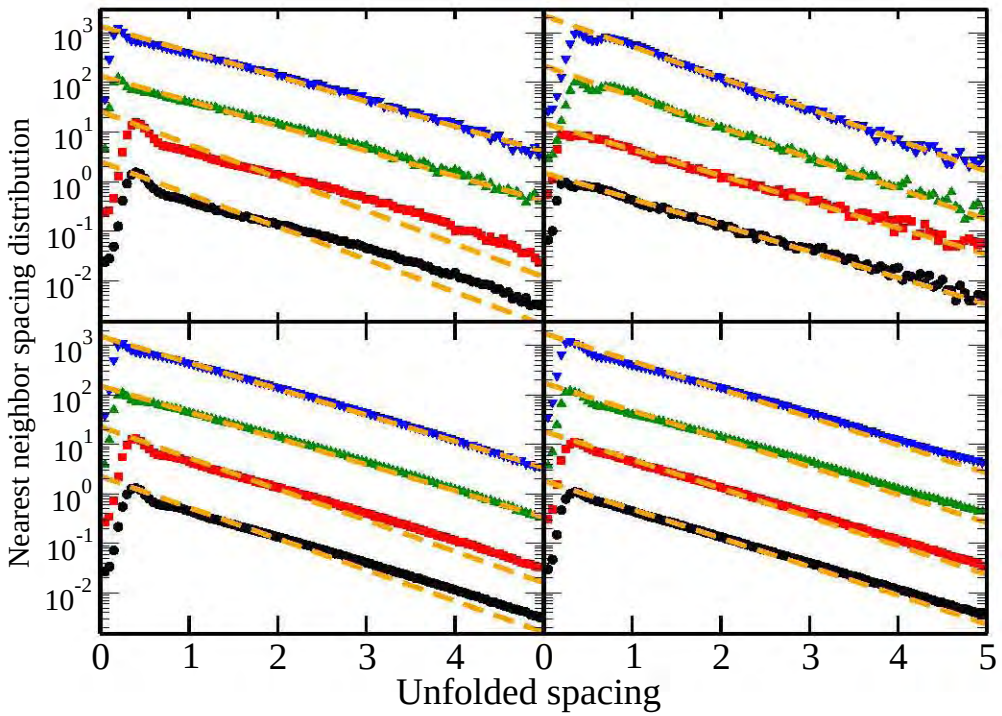



FIG. 2: (color online) Nearest neighbor spacing distribution for some genomic sequences for the same cases of Fig. 1. The solid (black), dash-dotted (maroon), the dotted (red) and the dashed lines correspond to the theoretical prediction given in Eq. (9) multiplied by 1, 10, 100 and 1000, respectively. The circles (black) correspond to adenine (A), the squares (red) to cytosine, the triangles up (green) to guanine and the triangles down (blue) to thymine.

In what follows, we will test the model with some genomic sequences. To this end, we define $\{x_i^\gamma\}_{i=1}^{N_\gamma}$ as the set of ordered $N_\gamma$ positions of the $\gamma$-basis. Here $\gamma$ stands for adenine (A), thymine (T), cytosine (C) or guanine (G) in DNA [27]. The cumulative density of basis (CDoB), or staircase function, is defined as

$$\mathrm{N}^\gamma(x) = \sum_{i=1}^{N_\gamma} \Theta(x - x_i^\gamma), \tag{15}$$

where $\Theta(x)$, as before, is the Heaviside function. Its derivative yields the density of basis (DoB).

The CDoB function is plotted in Fig. 1 for some genomes. In Fig. 1 (a) we see that for a bacteria, *borrelia burgdorferi*, mainly two slopes appear in the CDoB. This means that two average densities are present. The point in which the change of slope occurs corresponds to the origin of replication. Also, the densities of conjugated basis (A↔T and C↔G for

DNA) are very similar if one takes the reflexion symmetry around the origin of replication [8]. This symmetry appears in all sequenced bacterial chromosomes up to 2003 [25] and probably comes from the morphological (circular) symmetry of almost all bacterial genomes. The local densities (slopes) on each half of the bacterial genomes should give a measure of the stability of the DNA chain and could be used as a classification scheme. In Fig. 1 (b) the CDoB of a protozoan parasite, *Leishmania major*, is plotted. In (c) and (d) the CDoBs of chromosome 1 of *Arabidopsis thaliana* and chromosome 22 of *Homo sapiens* are plotted, respectively. Opposite to bacterial genomes, the genomes of *Arabidopsis thaliana* and *Homo sapiens* present long non-coding regions. In the insets of Fig. 1 some zooms of the CDoB are plotted. As can be seen, the CDoB shows kinks or cuadratic behavior. In spectral statistics, all quantities can be written as a sum of a secular and a fluctuating term. For the CDoB, the secular term is $\langle \mathrm{N}^\gamma \rangle$ while the fluctuating part is $\mathrm{N}^\gamma_{fluct}$,

$$\mathrm{N}^\gamma(x) = \langle \mathrm{N}^\gamma \rangle + \mathrm{N}^\gamma_{fluct}. \tag{16}$$

Thus, to compare sequences with different secular terms, a transformation yielding a CDoB with a constant(=1) DoB should be introduced. The transformation that does this job, i. e. that "unfolds the sequence", is

$$y_i^\gamma = \langle \mathrm{N}^\gamma(x_i^\gamma) \rangle. \tag{17}$$

This transformation yields a new sequence $\{y_i^\gamma\}_{i=1}^{N_\gamma}$ with an average level spacing $\langle s_i^\gamma \rangle = 1$, where $s_i^\gamma = y_i^\gamma - y_{i-1}^\gamma$. After the application of this transformation, it is possible to study the fluctuating properties of the different genomic sequences and compare one with another of a very different organism. Since there is not a known theoretical prediction for $\langle \mathrm{N}^\gamma \rangle$, a fitted piecewise-polynomial will be used for it. Also, since the sequences are very large, they were analyzed by windows. We checked that the results obtained in this way are stationary, i.e., that they do not depend on their position along the sequence, nor on the degree of the polynomial used in the fitting, nor on the length of the window. Nevertheless, when the mean level basis density is constant, the result given in Eq. (9) with $\langle s \rangle$ taken form the genomic sequence, is valid.

Fig. 2 shows the nearest neighbor spacing distribution $p(s)$ for the same cases of Fig. 1. As seen in this figure, at short distances ($\sim 12$), all distributions agree with the Poisson distribution. At larger distances ($\sim 15$) some of the distributions show small deviations from the model. Since the deviations are in the tail of the distribution, only few distances have

strong fluctuations. Thus, one can say that short range statistics are almost uncorrelated sequences notwithsanding that they come from very different organisms. As can be seen in the same figure, the nearest neighbor spacing distribution of chromosome 22 of the *Homo sapiens* has a non-exponential tail.

To conclude, we introduced a new ensemble, "the genomic ensemble" with strong evidence of the following conjectures: 1) The fluctuation properties of the DNA sequences are stationary and universal once the secular properties are subtracted, 2) Due to conjecture 1, it is possible to classify genomic sequences by number and size of the chromosome(s) and by the local densities of different basis: adenyne, thymine, guanine, and citosine. 3) The fluctuating properties are correctly described by Poissonian models for very short distances. For a first estimate, a model for the fluctuations can be obtained using the maximum entropy approach introducing repulsion at short distances. A second model should introduce some correlations starting from the Poissonian ensemble [26].

M.F.H. was supported during the Ph. D. program (Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México) by a scholarship from the Consejo Nacional de Ciencia y Tecnología (México)). This work was supported by PAPIIT DGAPA-UNAM under project IN111308. We are indebted to L. Mendoza and O. Geiger for useful comments and discussions as well as to H. Hernández-Saldaña for suggesting us the methodology of Ref. [24]. We thank X. A. Méndez-Báez for a carefull reading of the manuscript.

[1] B. Audit *et al* Phys. Rev. Lett. **99**,248102 (2007).

[2] C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H.E. Stanley, Nature 356 (1992) 168170.

[3] R. Voss, Phys. Rev. Lett. **68**, 3805 (1992).

[4] W. Li, K. Kaneko, Europhys. Lett. **17**, 655 (1992).

[5] S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, E. Matsa, C.-K. Peng, M. Simons and H. E. Stanley, Phys. Rev. E **51**, 5084 (1995).

[6] P. Bernáola-Galván, P. Carpena, R. Román-Roldán, and J. L. Oliver, Phys. Rev. E **53**, 5181 (1996).

[7] A. K. Mohanty and A. V. Narayana Rao, Phys. Rev. Lett. **84**, 1832 (2000).

[8] M. V. Jose, T. Govezensky and J. R. Bobadilla, Physica A **351** 477-498 (2005)

[9] R. Mansilla, N. Del Castillo, T. Govezensky, P. Miramontes, M. José and G. Cocho

[10] W. Li, Journal of Statistical Physics, **60** 823-837 (1990)

[11] I. Grosse *et al*, Phys. Rev. E, **61** 5624-5629 (2000)

[12] A. Kaskov *et al*, Phys. Rev. E, **69** 066138 (2004)

[13] J. R. Lobry, Mol. Biol. Evol. **13** 650-665 (1996)

[14] S. Nicolay et al, Phys. Rev. Lett. **93**, 108101 (2004).

[15] R. F Voss, Phys. Rev. Lett. **68**, 3805-3808 (1992)

[16] López, M. José and Sánchez, Biochem. Biophys. Res. Comm. **325** 467-478 (2004)

[17] A. E. Allahverdyan, Zh. S. Gevorkian, Chin-Kun Hu, and Ming-Chya Wu, Physical Review E **69**, 061908 (2004).

[18] Higarerda, M. F. *et al.*, "Nearest-neighbor spacing distribution of bases in some intron-less and intron-containing DNA sequences", *Physica A* **372** (2006) 368-373.

[19] Higarerda, M. F., Geiger O., Mendoza L., and Méndez-Sánchez R. A., Physica A 391 (2012) 12551269.

[20] T.A. Brody *et al*, Rev. Mod. Phys. **53**, 385 (1981).

[21] M. L. Metha, (1991) *Random Matrices*, 2nd. ed., Academic, New York.

[22] T. Guhr, A. Müller-Groeling and H. A. Weidenmüller, Phys. Rep. 299 (1998) 189-425.

[23] U. Gerland, Ph.D. thesis, 1998, University of Heidelberg, Germany.

[24] H. Hernández-Saldaña, J. Flores, and T. H. Seligman, Phys. Rev. E **60**, 449.

[25] J. Sanchez and M. V. José, Biochem. Biophys. Res. Comm. **299** 126 (2002).

[26] Seligman, T. H. in *Lecture Notes in Physics*, H.-J. Krappe and R. Lipperheide (Eds), **236**, (1985) 78-91.

[27] In viral genomes the same kind of analysis can be done changing the thymine by the uracil (U).

# Capítulo 5

# Discusión

En esta tesis se analizaron las propiedades estadísticas de diferentes genomas con una metodología llamada estadística espectral, explicada en la sección 4.2 que separa a las secuencias en la propiedad secular y en la fluctuante.

Primero se analizó la distribución de primeros vecinos de regiones codificantes y no codificantes de la secuencia de la mosca y del humano, viendo los resultados en la sección 4.1.1. Cuando comparamos la densidad acumulada de bases (CDoB) de la secuencia de la mosca y del humano de regiones codificantes y no codificantes se observó que son distintos para cada organismo, es decir depende de la concentración de cada uno de los nucleótidos como podemos ver en la figura 1 de la referencia [16]. En cambio cuando vemos la distribución de las fluctuaciones todas se acercan a una distribución exponencial. Esto dice que podemos encontrar con mayor probabilidad dos nucleótidos cercanos y con poca probabilidad dos nucleótidos a distancias largas, tanto en regiones codificantes como no codificantes en la mosca y en el humano. Este comportamiento se observa en la figura 2 de la Ref [16]. Es importante tomar el promedio como una función y no como una constante porque cuando lo tomamos como un promedio constante obtenemos un resultado que probablemente tomen características estadísticas diferentes a cada pedazo de la secuencia como lo podemos ver en la figura 3 de la Ref. [16] También en la Ref. [16], mostramos que una distribución exponencial concuerda razonablemente con los resultados observados en lagunas secuencias genómicas del humano y de la mosca. Este resultado teórico es también consistente con la varianza de número de la Ref [13].

En la sección 4.1.2 se analizó, con estadística espectral, una gran variedad de secuencias bacterianas de diferente géneros, tamaños, contenido de GC,

habitats, etc. La CDoBs de estas bacterias son diferentes porque dependen de la densidad de cada nucleótido y esto depende de la especie de bacteria, pero en general estas se agrupan en dos: las que tienen una distribución lineal es decir que tienen una sola pendiente y que podemos encontrar en la tabla II de la Ref. [17], y las que tiene un cambio de pendiente que se encuentran en la tabla I de la Ref. [17]. Aquí es muy importante resaltar un procedimiento llamado *desdoblamiento*. este procedimiento se aplica a cualquier cantidad matemática que esté compuesta de una parte suave y una parte fluctuante. El desdoblamiento es, en cierto límite, equivalente al análisis de fluctuaciones sin tendencia y separa la parte suave dejando sólo las fluctuaciones. Cuando comparamos los histogramas de las distancias entre bases iguales (A, T, G o C), podemos ver que los resultados son muy diferentes en las bacterias que presentan un CDoB lineal por partes. Como se observa en la figura 3 de la Ref. [17]. En el histograma de este tipo de bacterias se observó que la primera mitad de la secuencia tiene una distribución diferente, a la segunda mitad de la bacteria y a la distribución del genoma completo. Mientras que la bacterias que sólo tiene una pendiente, aún cuando se tomó la secuencia por partes, se ve en la figura 4 de la Ref. [17], el histograma tiene la misma forma de la primera, segunda y secuencia completa. Aquí se pude ver claramente que *Clostridium* y *Burkholderia* se comportan muy diferente. *Clostridium* es una bacteria donde su secuencia la tenemos que tomar por partes y saber exactamente que parte estamos tomando para hacer análisis estadístico, mientras que *Burkholderia* podemos trabajar con el genoma completo e incluso por partes sin importar que parte estamos tomando.

Con esta misma metodología se analizaron cuatro generos de bacterias que son *Bacillus, Burkholderia, Clostridium* y *Corynebacterium*. En la Fig. 2 de la Ref. [17] vemos que para la bacteria, *Borrelia burgdorferi*, la CDoB tiene dos pendientes lo que implica dos densidades de base. El punto en el cual cambia la pendiente corresponde, aproximadamente al origen de replicación. La densidad de bases (A$\leftrightarrow$T y C$\leftrightarrow$G for DNA) es muy similar si uno toma la simetría de reflexión alrededor del origen de replicación [122]. Esta simetría aparece en todas las secuencias de cromosomas bacterianos del trabajo de Sánchez y Jose (2003) [121]. Las densidades locales (pendientes) de cada mitad de los genomas bacterianos equivale a dar una medida de la estabilidad de la cadena de DNA. En las figuras 7 y 8 de la Ref [17] podemos observar que la CDoB de todas las bacterias, independientemente del género, son diferentes mientras que cuando vemos la distribución de las

fluctuaciones observamos que las bacterias del mismo género no importando la especie tienen la misma distribución. Sin embargo, cuando cambiamos de género, cambiamos la distribución. Las figuras 9 y 10 de la Ref. [17] muestran estos resultados. Cuando comparamos la distribución de *Burkholderia cepacia* y *Clostridium botulinum* en la figura 11 de la Ref. citeHigareda11 se observa una clara diferencia tanto a distancias pequeñas como en la pendiente de la distribución. La explicación biológica no es completamente clara pero tenemos fuertes evidencias que es debido al uso de codon de las bacterias. Cuando comparamos la distribución de las bacterias *Burkholderia cepacia* y *Clostridium botulinum* pero a *Clostridium botulinum* le cambiamos el uso de codon por el de *Burkholderia cepacia* observamos que la distribución de *Botulinum* cambia a una muy similar al de *Burkholderia*, esto se observa en la figura 14 de la Ref. [17].

En la última sección 4.2 se hizo el modelo matemático de máxima entropía para poder predecir como se comporta las secuencias, se ajustó el modelo a los datos de distribución de fluctuaciones de las secuencias de varios organismos como: la bacteria *Borrelia burgdorferi*, el protozoario *Leishmania major*, la planta *Arabidopsis thaliana* y el cromosoma 22 de *Homo sapiens*. Podemos ver en la figura 2 de la Ref. [123] que casi todas las secuencias tienen un ajuste a una distribución de Poisson a cortas distancias mientras que algunos de ellos a largas distancias tienen algunas desviaciones al modelo.

Quizá el resultado más importante de esta tesis es que al aplicar la estadística espectral al análisis de secuencias genómicas arrojó que a) las secuencias están compuestas por propiedades suaves y b) por propiedades fluctuantes, las primeras son una propiedad biológica y las segundas son un resultado más estadístico pero que está relacionado con el uso de codón.

que este reflejando características tan diferentes unas a otras. Las distribución de las fluctuaciones podemos decir que en general casi todas tiene una distribución exponecial con algunos cambios sobre todo a cortas distancias y en algunos caso al final de la distribución.

# Capítulo 6

# Conclusiones y perspectivas

Los resultados presentados en esta tesis nos permiten concluir que:

1. Se estudiaron diversas secuencias genómicas usando la herramienta llamada estadística espectral que permite separar la parte fluctuante y la parte secular, demostrándose que la parte secular depende de cada organismo y que la parte fluctuante es más o menos genérica.

2. Se demostró que no hay diferencia en la distribución de fluctuaciones cuando comparamos a la secuencia de una proteina llama miosina del humano y de la mosca de la fruta en la regiones que tiene introns y las que no. Se encontró que esta distribución es parecida una distribución azarosa (exponencial).

3. Se tiene que tener mucho cuidado al estudiar propiedades estadísticas de las secuencias genómicas porque estas pueden variar a lo largo de la cadena. Esto se demostró al estudiar la densidad acumulada de bases en bacterias; algunas tiene varias pendientes en su densidad acumulada y en otras solo tienen una pendiente. Se demuestra que distribución de fluctuaciones agrupa a las bacterias al nivel de género.

4. Se dieron algunas de aplicaciones biologícas como: a) Cómo el cambio de densidad de bases afecta los análisis estadísticos de las bacterias. y b) cómo depende la distribución de distancias a primeros vecinos del

mismo genero para varios generos de bacterias. Damos clara evidencia que el uso de codón es quien rige las distribuciones de distancias de primeros vecinos.

5. Se construyó el modelo matemático para poder predecir y explicar como se comportas las genomas.

Las investigaciones de esta tesis doctoral aún se pueden profundizar es aspectos como:

- Conocer como es la estructura de las secuencias genómicas es importante para poder analizarlas a futuro.
- Esta metodología permite comparar secuencias muy diferentes entre sí, como el tamaño, funciones, contenido de GC, número de genes, secuencias repetidas, etc., lo cual nos permite ver sus diferencias y semejanzas.
- Este tipo de metodología nos ayuda a saber con que tipos de secuencias estamos trabajando al conocer sus propiedades estadísticas.
- Las propiedades estadísticas nos indica alguna característica biológica del organismo por eso es importante seguir haciendo este tipo de análisis incluyendo no sólo la distribución de un solo nucleótido, si no de dinucleótidos y tripletes.

# Bibliografía


[1] Berezney, R. *et al.*, "Heterogeneity of eukaryotic replicons, replicon clusters, and replication foci," *Chromosoma* **108** (2000) 471−478.

[2] Lobry, J. R. and Sueoka, N., "Asymmetric directional mutation pressures in bacteria," *Genome Biology* **3** (2002) 58.

[3] Arneodo, A. *et al.*, "Characterizing long-range correlations in DNA sequences from wavelet analysis," *Phys. Rev. Lett.* **74** (1995) 3293−3296.

[4] Hackenberg, M. *et al.*, "WordCluster: detecting clusters of DNA words and genomic elements," *Algorithms for Molecular Biology* **6** (2011) 2.

[5] Zhao, B., Duan, V., Yau, S. S.-T., "A novel clustering method via nucleotide-based Fourier power spectrum analysis," *Journal of Theoretical Biology* **279** (2011) 83−89.

[6] Bose, R. and Soni, C., "Alternate measure information useful for DNA sequences," *Phy. Rev. E.* **83** (2011) 051918.

[7] Brody, T. A. *et al.*, "Random-matrix physics: spectrum and strength fluctuations," *Rev. Mod. Phys.* **53** (1981) 385−479.

[8] M. L. Metha, Random Matrices, 2nd. ed., Academic, New York, 1991.

[9] Guhr, T. *et al.*, "Random matrix theories in quantum physics: Common concepts," *Phys. Rep.* **299** (1998) 189−425.

[10] Šeba, P., "Random Matrix Analysis of Human EEG Data," *Phys. Rev. Lett.* **91** (2003) 198104.

[11] Müller, M. *et al.*, "Detection and characterization of changes of the correlation structure in multivariate time series," *Phys. Rev. E* **71** (2005) 046116.

[12] Luo, F. *et al.*, "Application of random matrix theory to microarray data for discovering functional gene modules," *Phys. Rev. E* **73** (2006) 031924.

[13] Mohanty, A. K., and Narayana-Rao A. V. S. S., "Factorial Moments Analyses Show a Characteristic Length Scale in DNA Sequences," *Phys. Rev. Lett.* **84** (2000) 1832−1835.

[14] Afreixo, V., *et al.*, "Genome analysis with distance to the nearest dissimilar nucleotide," *Journal of Theoretical Biology.* **275** (2011) 5258.

[15] Afreixo, V. *et al.*, "Genome analysis with inter-nucleotide distances," *Bioinformatics* **25** (2009) 3064−3070.

[16] Higareda, M. F. *et al.*, "Nearest-neighbor spacing distribution of bases in some intron-less and intron-containing DNA sequences," *Physica A* **372** (2006) 368−373.

[17] Higareda, M. F., Geiger, O., Mendoza, L., and Méndez-Sánchez, R. A., "Evidence of codon usage in the nearest neighbor spacing distribution of bases in bacterial genomes," *Physica A* **391** (2012) 1255−1269.

[18] Watson, J.D. and Crick, F.H., *Nature* **171** (1953) 737−738.

[19] Strachan T., Read A. (2006) *Human Molecular Genetics*, Fourth Edition. Cambridge University Press.

[20] Summers, D., "Timing, self-control and a sense of direction are the secrets of multicopy plasmid stability," *Molecular Microbiology,* **29** (1996) 1137−1145.

[21] C.F. Amábile-Cuevas, M.E. Chicurel., "Bacterial plasmids and gene flux," *Cell,* **70** (1992), pp. 189−199.

[22] Fleischmann, R. *et al.*, "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd," *Science,* **269** (1995) 449−604.

[23] Fraser C. M., *et al.* "The Minimal Gene Complement of *Mycoplasma genitalium*," *Science* **20** (1995) 397−404.

[24] Malcolm F. White, Stephen D. Bell., "Holding it together: chromatin in the Archaea" *Trends in Genetics* **18** (2002) 621−626.

[25] Blackburn, E. H., "Switching and signaling at the telomere," *Cell.* **106** (2001) 661−673.

[26] Huffman, K. E., Levene, S. D., Tesmer, V. M., Shay, J. W. & Wright, W.E., "Telomere Shortening is proportional to the siza of the G rich telomeric 3," *J Biol Chem* **275** (2001) 19717−19722.

[27] Griffith, J. D., Comeau, L., Rosenfields, S., Stansel, R. M., Bianchi, A., Moss, H. & De Lange T., "Mammalian telomeres end in a large duplex loop," *Cell* **97** (1999) 503−514.

[28] Chong, L., Van-Steensel, B., Broccoli, D., Erdjument-Bromage, H., Hanish, J., Tempst, P. & De Lange, T., "A human telomeric protein," *Science* **270** (1995) 1663−1667.

[29] Smogorzewska, A., van Steensel, B., Bianchi, A., Oelmann, S., Schaefer, M. R., Schnapp, G., & De Lange, T., "Control of human telomere length by TRF1 and TRF2," *Mol Cell Biol* **20** (2000) 1659−1668.

[30] Bailey, S. M., Cornforth, M. N., Kurimasa, A., Chen, D. J. & Goodwin, E. H., "Strand-specific postreplicative processing of mammalian telomeres," *Science* **293** (2001) 2462−2465.

[31] Stansel, R. M., de Lange, T. & Griffith, J. D., "T-loop assembly in vitro involves binding of TRF2 near the 3' telomeric overhang," *EMBO J* **20** (2001) 5532−5540.

[32] Li, G. Z., Eller, M. S., Firoozabadi, R., Puri, N. & Gilchrest, B. A., "Evidence that exposure ofthe telomeme 3 overhang sequence induces senescence," *Proc Natl Acad Sci* **100** (2003) 527−53.

[33] Baumann, P., Podell, E. & Cech, T. R., "Human Pot1 (protection of telomeres) protein: cytolocalization, gene structure, and alternative splicing," *Mol Cell Biol* **22** (2002) 8079−8087.

[34] Kim, S. H., Kaminker, P. & Campisi, J., "TIN2, a new regulator of telomere length in human cells," *Nat Genet* **23** (1999) 405−412.

[35] Smith, S., Giriat, I., Schmitt, A. & De Lange, T., "Tankyrase, a poly(ADP-ribose) polymerase at humantelomeres," *Science* **282** (1998) 1484−1487.

[36] Kaminker, P. G., Kim, S. H., Taylor, R. D., Zebarjadian, Y., Funk, W. D., Morin, G. B., Yaswen, P. & Campisi, J., "TANK2, a new TRF1-associated poly(ADP-ribose) polymerase, causes rapid induction of cell death upon overexpression," *J Biol Chem* **276** (2001) 35891−35899.

[37] Li, B., Oestreich, S. & De Langem, T., "Identification of human Rap1: implications for telomere evolution," *Cell* **101** (2000) 471−483.

[38] McElligottt, R. & Wellinger, R. J., "The terminal DNA structure of mammalian chromosomes," *EMBO J* **16** (1997) 3705−3014.

[39] De Lange, T., "Telomeres," *Nature* **392** (1998) 753−754.

[40] Baur, J. A., Zou, Y., Shay, J. W. & Wright, W. E., "Telomere position effect in Human Cells," *Science* **292** (2001) 2075−2077.

[41] Vincent Daubin and Guy Perrire., "G+C3 Structuring Along the Genome: A Common Feature in Prokaryotes," *Mol Biol Evol* **20** (2003) 471−483.

[42] Antequera, F. & Bird, A. "Number of CpG islands and genes in human and mouse," *Proc. Natl. Acad. Sci.,* **90** (1993) 11995−11999.

[43] Aissani, B. & Bernardi, G. "CpG islands: features and distribution in the genomes of vertebrates," *Gene* **106** (1991) 173−183.

[44] Bachellier S., Gilson E., Hofnung M., Hill C.W., "Repeated sequences" in *Escherichia coli and Salmonella: Cellular and molecular biology*, eds Neidhardt F. C., Curtiss III R., Ingraham J. L., Lin E. C. C., Low K. B., Magasanik B., Reznikoff W. S., Riley M., Schaechter M., Umbarger H. E. (ASM Press, Washington, D.C), 2nd ed. pp 2012−2040.

[45] Myers, S., Freeman, C., Auton, A., Donnelly, P. & McVean, G., "A common sequence motif associated with recombination hot spots and genome instability in humans," *Nat. Genet.* **40** (2008) 1124−1129.

[46] McVean, G. A., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R. & Donnelly, P., "The fine-scale structure of recombination rate variation in the human genome," *Science* **304** (2004) 581−584.

[47] Charlesworth, B., Langley, C. H. & Stephan, W., "The evolution of restricted recombination and the accumulation of repeated DNA sequences," *Genetics.* **112** (1986) 947−962.

[48] Rizzon, C., Marais, G., Gouy, M. & Biemont, C., "Recombination rate and the distribution of transposable elements in the Drosophila melanogaster genome," *Genome Res.* **12** (2002) 400−407.

[49] Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P., "A fine-scale map of recombination rates and hotspots across the human genome," *Science* **310** (2005) 321−324.

[50] Jeffreys, A. J., Murray, J. & Neumann, R., "Highresolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot," *Mol. Cell.* **2** (1998) 267−273.

[51] Jeffreys, A. J. & Neumann, R., "Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot," *Nat. Genet.* **31** (2002) 267−271.

[52] Deininger, P. L. & Batzer, M. A., "Alu repeats and human disease," *Mol. Genet. Metab.* **67** (1999) 183−193.

[53] Gu, W., Zhang, F. & Lupski, J. R., "Mechanisms for human genomic rearrangements," *Pathogenetics* **1** (2008) 4.

[54] Sen, S. K. *et al.*, "Human genomic deletions mediated by recombination between Alu elements," *Am. J. Hum. Genet.* **79** (2006) 41−53.

[55] Ptak, S. E., Roeder, A. D., Stephens, M., Gilad, Y., Paabo, S. & Przeworski, M., "Absence of the TAP2 human recombination hotspot in chimpanzees," *PLoS Biol.* **2** (2004) e155.

[56] Winckler, W. *et al.*, "Comparison of fine-scale recombination rates in humans and chimpanzees," *Science* **308** (2005) 107−111.

[57] Boulton, A., Myers, R. S. & Redfield, R. J., "The hotspot conversion paradox and the evolution of meiotic recombination," *Proc. Natl Acad. Sci. USA* **94** (1997), 8058−8063.

[58] Ptak, S. E., Hinds, D. A., Koehler, K., Nickel, B., Patil, N.,Ballinger, D. G., Przeworski, M., Frazer, K. A. & Paabo, S., "Fine-scale recombination patterns differ between chimpanzees and humans," *Nat. Genet.* **37** (2005) 429−434.

[59] Cooper, G. M., Nickerson, D. A. & Eichler, E. E., "Mutational and selective effects on copy-number variants in the human genome," *Nat. Genet.* **39** (2007) S22−S29.

[60] Burn, J. & Goodship, J., "Developmental genetics of the heart," *Curr. Opin. Genet. Dev.* **6** (1996), 322−325.

[61] Emanuel, B. S., "Molecular mechanisms and diagnosis of chromosome 22q11.2 rearrangements," *Dev. Disabil. Res. Rev.* **14** (2008) 11−18.

[62] Kurahashi, H. & Emanuel, B. S., "Long AT-rich palindromes and the constitutional t(11;22) breakpoint," *Hum. Mol. Genet.* **10** (2001) 2605−2617.

[63] Kurahashi, H., Shaikh, T. H., Hu, P., Roe, B. A,, Emanuel, B. S. & Budarf, M. L., "Regions of genomic instability on 22q11 and 11q23 as the etiology for the recurrent constitutional t(11;22)," *Hum. Mol. Genet.* **9** (2000) 1665−1670.

[64] Gotter, A. L., Nimmakayalu, M. A., Jalali, G. R., Hacker, A. M., Vorstman, J., Conforto, Duffy, D., Medne, L. & Emanuel, B. S., "A palindrome-driven complex rearrangement of 22q11.2 and 8q24.1 elucidated using novel technologies," *Genome. Res.* **17** (2007) 470−481.

[65] Sheridan, M. B., Kato, T., Haldeman-Englert, C., Jalali, G. R., Milunsky, J. M., Zou, Y., Klaes, R., Gimelli, G., Gimelli, S., Gemmill, R. M., *et al.*, "A palindrome-mediated recurrent translocation with 3:1 meiotic nondisjunction: the t(8;22)(q24.13;q11.21)," *Am. J. Hum. Genet.* **87** (2010) 209−218.

[66] Kurahashi, H., Inagaki, H., Hosoba, E., Kato, T., Ohye, T., Kogo, H. & Emanuel, B. S., "Molecular cloning of a translocation breakpoint hotspot in 22q11," *Genome. Res.* **17** (2007) 461−469.

[67] Gotter, A. L., Shaikh, T. H., Budarf, M. L., Rhodes, C. H. & Emanuel, B. S., "A palindrome-mediated mechanism distinguishes translocations involving LCR-B of chromosome 22q11.2," *Hum. Mol. Genet.* **13** (2004) 103−115.

[68] Nimmakayalu, M A., Gotter, A. L, Shaikh, T. H. & Emanuel, B. S., "A novel sequence-based approach to localize translocation breakpoints identifies the molecular basis of a t(4;22)," *Hum. Mol. Genet.* **12** (2003) 2817−2825.

[69] Akgun, E., Zahn, J., Baumes, S., Brown, G., Liang, F., Romanienko, P. J., Lewis, S. & Jasin, M., "Palindrome resolution and recombination in the mammalian germ line," *Mol. Cell. Biol.* **17** (1997) 5559−5570.

[70] Collick, A., Drew, J., Penberth, J., Bois, P., Luckett, J., Scaerou, F., Jeffreys, A. & Reik, W., "Instability of long inverted repeats within mouse transgenes," *EMBO J.* **15** (1996) 1163−1171.

[71] Leach, D. R., "Long DNA palindromes, cruciform structures, genetic instability and secondary structure repair," *Bioessays* **16** (1994) 893−900.

[72] Inagaki, H., Ohye, T., Kogo, H., Yamada, K., Kowa, H., Shaikh, T. H., Emanuel, B. S. & Kurahashi, H., "Palindromic AT-rich repeat in the NF1 gene is hypervariable in humans and evolutionarily conserved in primates," *Hum. Mutat.* **26** (2005) 332−342.

[73] Gordenin, D. A., Lobachev, K. S., Degtyareva, N. P., Malkova, A. L., Perkins, E. & Resnick, M. A., "Inverted DNA repeats: a source of eukaryotic genomic instability," *Mol. Cell. Biol.* **13** (1993) 5315−5322.

[74] Lu, L., Jia, H., Droge, P. & Li, J., "The human genome-wide distribution of DNA palindromes," *Funct. Integr. Genomics.* **7** (2007) 221−227.

[75] Kato, T., Inagaki, H., Yamada, K., Kogo, H., Ohye, T., Kowa, H., Nagaoka, K., Taniguchi, M., Emanuel, B. S. & Kurahashi, H., "Genetic variation affects de novo translocation frequency," *Science* **311** (2006) 971.

[76] Kogo, H., Inagaki, H., Ohye, T., Kato, T., Emanuel, B. S. & Kurahashi, H., "Cruciform extrusion propensity of human translocation-mediating palindromic AT-rich repeats," *Nucleic. Acids. Res.* **35** (2007) 1198−1208.

[77] Trinh, T. Q. & Sinden, R. R., "Preferential DNA secondary structure mutagenesis in the lagging strand of replication in *E. coli*," *Nature* **352** (1991) 544−547.

[78] Mirkin, S. M., Smirnova, E. V., "Positioned to expand," *Nat. Genet.* **31** (2002) 5−6.

[79] Mirkin, S. M., "Expandable DNA repeats and human disease," *Nature* **447** (2007) 932−940.

[80] Zhang, H. & Freudenreich, C. H., "An AT-rich sequence in human common fragile site FRA16D causes fork stalling and chromosome breakage in *S. cerevisiae*," *Mol. Cell.* **27** (2007) 367−379.

[81] Braun, R. E., "Packaging paternal chromosomes with protamine," *Nat. Genet.* **28** (2001) 10−12.

[82] Ward, W. S., "Function of sperm chromatin structural elements in fertilization and development," *Mol. Hum. Reprod.* **16** (2010) 30−36.

[83] Ward, W. S., "Regulating DNA supercoiling: sperm points the way," *Biol. Reprod.* **84** (2011) 841−843.

[84] Russo, V.E.A., Martienssen, R.A., & Riggs, A.D., (1996) "Mechanism of gene regulation Epigenetic," Laboratory Press, Cold Spring Harbor, NY.

[85] Bestor, T. H. & Coxon A., "Cytosine methylation; the pros and cons of DNA methylation," *Curr. Biol.* **6** (1993) 384−386.

[86] Colot, V., & Rossignol, J. L., "Eukariotic DNA methylation as an evolutionary device," *Bioessays* **5** (1999) 402−411

[87] Jones, P. A., & Takar, D., "The role of DNA methylations in mammalian epigenetics," *Hum. Mol. Genet.* **23** (2001) 2619−2626.

[88] Jeltsch, A., "Beyond Watson and Crick: DNA methylation and molecular enzymology of DNA methyltransferasas," *Chembiochem* **3** (2002) 274−293.

[89] Goll, M. G. & Bestor, T., "Eukariotic cytosine methyltransferasas," *Annu. Rev. Biochem.* **74** (2005) 481−514.

[90] Bird Adrian., "DNA methylation patterns and epigenetic memory," *Genes and Development* **16** (2002) 6−21.

[91] Fischle, W., *et al.*, "Molecular basis for the discrimination of repressive methyl lysine marks in histone H3 by Polycomb and HP1 chromodomains," *Genes* **17** (2003) 1870−1881.

[92] Hermann, A., Growher, H. & Jelsch, "Biochemistry and biology of mammalian DNA methyltransferasas," *CMLS* **61** (2004) 2571−2587.

[93] Susuki, T., Fujii, M. & Ayasawa, D., "Demethylation of classical instability in senescent human fibroblast," *Exp. Gerontology* **37** (2002) 1005−1014.

[94] Roberts, R. J., Belfort, M., Bestor, T., Bhagwat, A. S., Bickel, T., *et al.* "A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes," *Nucleic Acids Res.* **31** (2003) 1805−1812.

[95] Maison, C. & Almovzni, G., "HP1 and dynamics of heterochromatin maintenance," *Nature Reviews* **5** (2004) 296−304.

[96] Jeanpierre, M., "Human Satellites 2 and 3," *Revenues Generals Reviews* **39** (1994) 163−171.

[97] Sueoka, N., "A statistical analysis of the deoxyribonucleic acid distribution in density gradient centrifugation," *PNAS* **45** (1959) 1480−1490.

[98] Sueoka, N., *et al.*, "Heterogeneity in deoxyribonucleic acid II: dependece of the density of deoxiribonucleic acids on guanine-citosine contents," *Nature* **183** (1959) 1429−1433.

[99] Rolfe, R. and Mendelson M., "The relative homogeneity of microbial DNA" *PNAS* **45** (1959) 1039−1043.

[100] Li, W. and Kaneko, W., "Long-Range Correlation and Partial $1/f^{\alpha}$ Spectrum in a Noncoding DNA Sequence," *Europhys. Lett.* **17** (1992) 655−660.

[101] Li, W., "Expansion-modification system: a model for spatial $1/f$ spectra," *Phy. Rev. A.* **43** (1991) 5240−5260.

[102] Mandelbrot, B.B., (1982) The fractal geometry of nature. W.H. Freeman. San Francisco, USA.

[103] T. Vicsek 1996, Fractal geometry. En: *Fractal geometry in biological system. An analitical aprroach.* P.M. Iannacone y M. Khokha, eds New York, USA, CRC Press.

[104] H.E Stanley, S.V Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, C.K Peng, y M. Simons (1996), Fractal geometry in biological systems. An analytical approach. New York USA. CRC Press.

[105] Stanley, H.E., *et al.* "Statistical mechanics in biology: how ubiquitous are long-range correlations?" *Physica A* **205** (1994) 214−253

[106] Peng, C. K. *et al.*, "Mosaic organization of DNA nucleotides," *Phys. Rev. E.* **49** (1994) 1685−1689.

[107] Peng, C. K. *et al.*, "Long-range correlations in nucleotide sequences," *Nature* **356** (1992) 168−170.

[108] Stanley, H.E. *et al.*, "Fractal Landscapes in Physics and Biology," [30th Annual Saha Memorial Lecture] *Physica A* **191** (1992) 1−32.

[109] Karlin, S. and Brendel V., "Patchiness and correlations in DNA sequences," *Science* **259** (1993) 677−680.

[110] Voss, R. F., "Evolution of long-range fractal correlations and 1/f noise in DNA base sequences," *Phys. Rev. Lett.* **68** (1992) 3805−3808.

[111] de Sousa-Vieira., M., "Statistics of DNA sequences: a low frequency analysis," *Phys. Rev. E* **60** (1999) 5932−5937.

[112] Tavare, S. and Giddings, B. W., Mathematical Methods for DNA sequences, editor: M.S. Waterman, Boca Raton: CRC Press (1989).

[113] Buldyrev, S. V., *et al.*, "Long-range Correlation Properties of Coding and Noncoding DNA Sequences: GenBank Analysis," *Physical Review E,* **51** (1995) 5084−5091.

[114] Borsnik, B., *et al.*, "Analysis of apparent $1/f$ spectrum in DNA sequences," *Europphys Lett.* **23** (1993) 389−394.

[115] A. L. Edwards, "The Correlation Coefficient" Ch. 4 in An Introduction to Linear Regression and Correlation. San Francisco, CA: W. H. Freeman, pp. 33−46, 1976.

[116] R. Mansilla, G. Cocho. "Multiscaling in expansion-modification systems: an explanation for long correlation in DNA" *Comp. Syst.* **12** (2000) 207−240.

[117] Podobnik, B. *et al.*, "Similarity and Dissimilarity in Correlations of Genomic DNA," *Physica A* **373** (2007) 046107.

[118] Martignetti, L. and Caselle, M. "Universal power law behaviors in genomic sequences and evolutionary models," *Physical Review E* **76** (2007) 021902.

[119] P. Carpena, P. Bernaola-Galván, A. Coronado M. Hackenberg, and J.L. Oliver, "Identifying characteristic scales in the human genome," *Physical review E* **75** (2007) 032903.

[120] Bulyrev, S.V. *et al.*, "Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis," *Physical Review E* **51** (1995) 5084−5091.

[121] Sánchez, J. and José, M. V., "Bilateral inverse symmetry in whole bacterial chromosomes," *Biochem. Biophys. Res. Comm.* **299** (2002) 126−134.

[122] José, M. V. *et al.*, "Statistical properties of DNA sequences revisited: the role of inverse bilateral symmetry in bacterial chromosomes," *Physica A* **351** (2005) 477−498.

[123] Higareda, M. F., and Méndez-Sánchez, R. A., "Universal Statistical Properties in DNA: Beyond Random Sequences," (Artículo por ser enviado a *Physica A*).