Periodicity in Genome Architecture from Bacteria to Humans

Atsushi Fukushima

The Graduate School of Biological Sciences,
NARA INSTITUTE of SCIENCE and TECHNOLOGY

2003

バイオサイエンス研究科　博士論文要旨

所属 (主指導教官)	遺伝子教育研究センター・生体情報学講座 (森 浩禎 教授)		
氏名	福島 敦史	提出	平成　15年　1月　7日
題目	Periodicity in Genome Architecture from Bacteria to Humans (バクテリアからヒトまでのゲノム周期構造)		

要旨

A deep understanding of the structural characteristics of living organisms requires exhaustive analyses of the DNA sequences that serve as the blueprint for organic architecture. One of the characteristics of genomic DNA is latent periodicity in base pair organization, the elucidation of which may play an important role in revealing basic genome architecture and provide a way to discriminate species on the basis of there sequence alone. For this thesis work, I performed analysis of the periodicities in the nucleotide sequences of several prokaryotes, *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Drosophila melanogaster*, *Anopheles gambiae*, and *Homo sapiens*, with a methodology based on power spectra. Power spectrum analysis is widely used in physics, especially in signal recognition theory, to measure the periodicities of sound and light signals. The application of this methodology to analysis of DNA sequences has lead to the finding of long-range correlations in genomic DNA. To characterize periodical regions in genomic DNA, I proposed that the periodic nucleotide distribution is represented by a parameter F_k computed as the frequency of periodic pairs separated by k nucleotides. For the *C. elegans* genome, a periodicity of 68 bp was found on chromosome I, a 59-bp periodicity was found on chromosome II, and a 94-bp periodicity was present on chromosome III. For *A. thaliana*, I detected three periodicities (248, 167, and 126 bp) on chromosome 3, three (174, 88, and 59 bp) on chromosome 4, and four (356, 174, 88, and 59 bp) on chromosome 5. These findings are consistent with open reading frames encoding Gly-rich sequences. Additionally, in the human genome, 84- and 167-bp periodicities were detected on chromosomes 21 and 22. A periodicity of 167 bp is identical to the length of DNA that comprises two complete helical turns in a nucleosome. For *D. melanogaster* and *A. gambiae*, the G and C spectra contained flat regions in the middle frequency, which indicates randomness of the base sequence composition. Such flat regions have not been observed in *Saccharomyces cerevisiae*, *C. elegans*, *A. thaliana*, or *H. sapiens*. The present study goes beyond the simple analysis of periodical correlations and attempts to clarify the biological and physiological implications of the genomic structure.

Contents

1. Introduction

1.1. Fractals in nature

Fractal geometry is the fundamental design of nature. Most of natural configurations, such as clouds, mountains, coastlines, and trees, can be explained by self-similarity. To explain these configurations, Mandelbrot (1982) proposed the fractal concept of self-similar structures; the components of a structure resemble the whole structure. The fractal concept also exhibits scale-invariant geometrical features and can be found by the power-law. Basically, numerous power-law correlations in nature show self-similarity and are typically called "long-range correlations." A power-law has also been observed in the realm of biology. It was recently reported that the complexities of metabolic networks, including enzymes and metabolites, protein-protein interactions, and genetic relations included in some paralogous families, are described by power-law (Jeong *et al.*, 2000; Jeong *et al.*, 2001; Luscombe *et al.*, 2002).

Fractal geometry construction is illustrated in Figure 1a. First, let us consider a copy machine that is equipped with an image reduction feature. If we take an image of the word "GENE," put it on the machine, and push a button, we obtain a copy of the image. However, it is reduced uniformly by a factor of 1/2 (50%). The copy is similar to the original. Such reduction is achieved by a lens system. A machine with three reduction lenses, each reducing an image by 50%, generates a triangle form of the word "GENE" (see Figure 1a). After some iteration, the word "GENE" forms self-similar triangles (see Figure 1a). This geometry is a famous fractal pattern known as the Sierpinski Gasket. Chromatin packing based on current ideas is illustrated in Figure 1b. The DNA helix is the first coil. It turns twice around a histone core unit, and histone cores are attached to each other by DNA linkers (see Figure 1b). A nucleosome consists of the histone core, the linker, and the DNA helix. Histone H1 attaches nucleosomes to each other, and the strand is 30 nanometers (nm) thick (see Figure 1b). The next order of chromatin packing is provided by folded loops in the 30-nm strand. This is 300-nm thick. The 300-nm strand, which contains the looped domains, forms a chromatid spiral. In the spiral, some regions are compacted more tightly than others. When a chromosome is extended, as during prophase stages of meiosis, it appears to comprise different sizes of condensed chromeres and less condensed regions between them. In this manner, the coiled-coil structure generates self-similarity.

Genome DNA sequences are important targets for investigations of fractal properties. Long-range correlations are found in DNA sequences (Li and Kaneko, 1992; Peng *et al.*, 1992; Voss, 1992). Long-range correlation means that the base composition tends to vary in a similar manner in regions of the DNA sequence separated by very long distances (Li *et al.*, 1994). The typical study of long-range correlations is applicable for numerical sequences x_j (e.g. if base 'A' occurs at position j along the DNA strand, $x_j = 1$, otherwise $x_j = 0$). The correlation between sequences x_j and x_{j+l} separated by l bases was examined, and significant correlations were obtained for 10 bp to 10000 bp. This correlation is observed primarily in intronic regions; the correlation in exonic regions is unclear (Li and Kaneko, 1992; Peng *et al.*, 1992; Voss, 1992). In contrast, Azbel

(1995) suggested that there are no long-range correlations in genomic DNA sequences. However, most studies have indicated that long-range correlations are present in both protein-coding and non-coding sequences. Long-range correlations in DNA sequence have been examined by many methods. The autocorrelation function (see Section 2.1) is the direct measure of correlations in DNA sequences (Herzel *et al.*, 1998). To clarify the length of long-range correlation of a given DNA segment, Vieira (1999) studied long-range correlations in 13 bacterial genomes by autocorrelation function and power spectrum (see Section 2.2). These methods detected a hidden periodicity. For lower frequency f (1/bp) (see Figure 2), the power spectrum presents a power-law behavior with exponent approximately equal to –1 (corresponds to $1/f$ noise). When a cutoff of the power-law in DNA exists in high frequencies (Figure 2, left), the spectrum is called 'partial power-law.' As a result, the power-law correlation is not always found across the entire DNA strand (Vieira, 1999). Buldyrev *et al.* (1995) observed the different properties of coding and noncoding sequences by power spectrum and reported that detrended fluctuation analysis (Peng *et al.*, 1994) discriminates the pseudo-correlation property from the real correlation property. Detrended fluctuation analysis is a powerful method adapted specifically to deal with non-stationary processes. The statistical property of this process is not invariant under a shift of the time origin. Li *et al.* (1998) investigated the correlation between x_j and x_{j+l} of 16 complete yeast chromosomes by power spectrum.

The biological and physiological significances of periodicity, including long-range correlation, in genomes are unclear. Power spectrum analysis would be helpful for examining long-range correlations and detecting periodicity in DNA sequences. In the present study, I examined periodicities and long-range correlations in DNA sequences by power spectrum analysis. The power spectrum corresponds to the square of the Fourier transformation (Section 2.1), and the power spectrum is more suitable than correlation function because of the high calculation efficiency. There is a relation between power spectrum and autocorrelation. The Fourier transformation of power spectrum corresponds to the autocorrelation function, and the inverse Fourier transformation of autocorrelation corresponds to power spectrum as put forth by the Wiener-Khinchin Theorem.

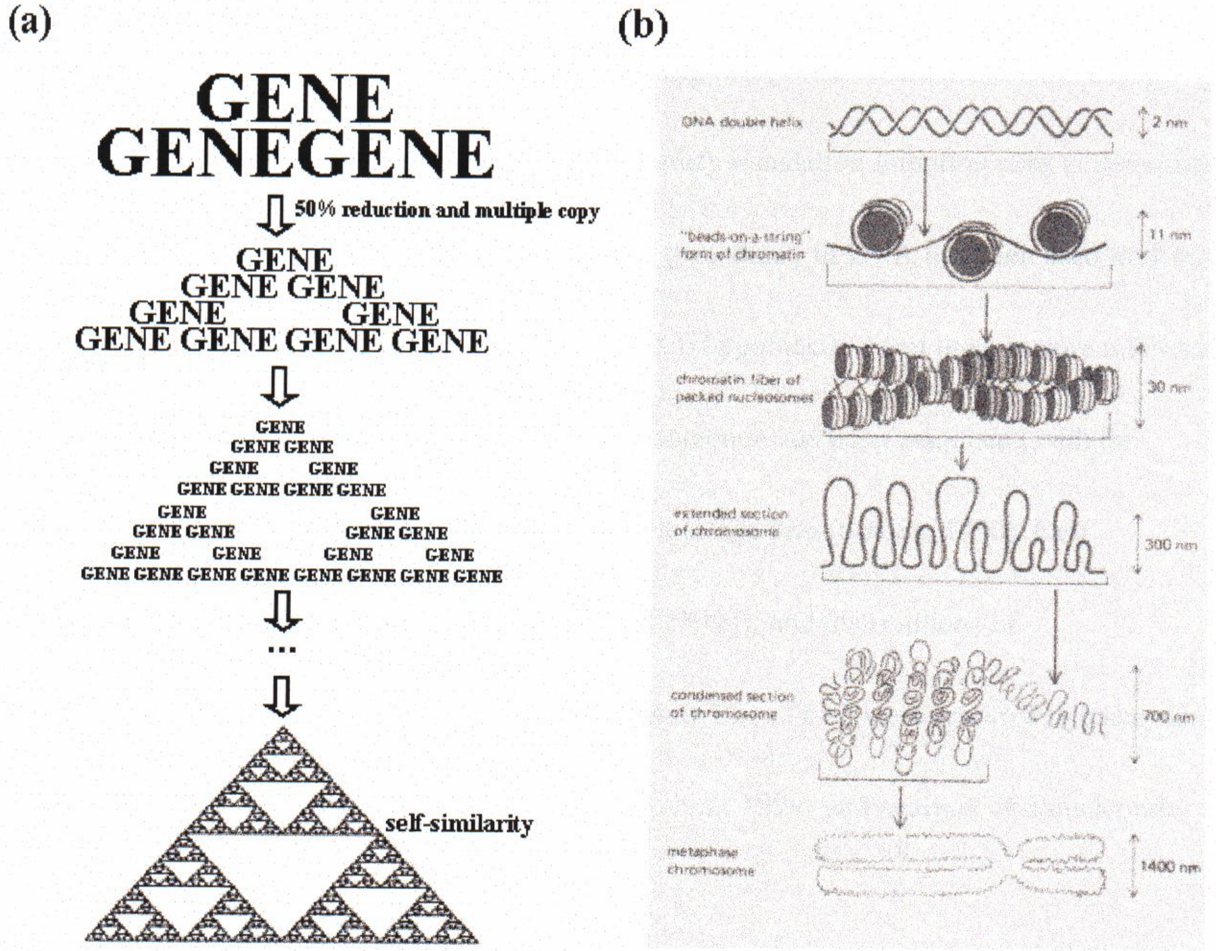
(a)
(b)
GENE
GENEGENE
50% reduction and multiple copy
GENE
GENE GENE
GENE GENE
GENE GENE GENE GENE
...
self-similarity
DNA double helix
2 nm
"beads-on-a-string" form of chromatin
11 nm
chromatin fiber of packed nucleosomes
30 nm
extended section of chromosome
300 nm
condensed section of chromosome
700 nm
metaphase chromosome
1400 nm

Figure 1. Relation between fractal property and chromatin packing. (a) Sierpinski gasket by word "GENE." (b) diagram of different orders of chromatin packing (Fig. 8-24 from Alberts *et al.*, Molecular Biology of the Cell, 1983).

1.2. Periodicity in DNA

Since the early 1970s, biologists have tried to distinguish protein-coding and non-coding regions in DNA sequences. They have also tried to identify translation initiation sites to determine protein-coding regions in DNA sequences. Three-base periodicity in DNA has been used to detect protein-coding regions (Fickett, 1982). Shepherd (1981a) hypothesized that most mature mRNAs have a strong rhythms with 3-bp periodicity. This phenomenon has been associated with the statistical properties of coding sequences, such as codon usage bias (Staden, 1990), base composition of protein-coding regions (Gutierrez *et al.*, 1994), and distributions of purines/pyrimidines (Shepherd, 1981a, b). Periodicities of 10 to 11 bp were reported in bacterial genomes on the basis of a correlation function (Herzel *et al.*, 1999) and analysis of dinucleotide pairs (Tomita *et al.*, 1999).

Genomic DNA contains repetitive sequences. Repeat sequences are organized physically into tandem arrays of simple sequences that are widely dispersed within a relatively long unit. Repetitive sequences yield various periodicities in genomes. The functions of repetitive sequences in disease and higher-order structures are reviewed in the following.

Minisatellite and satellite DNAs

Minisatellite DNA

Minisatellites are typically 0.5 kb to several kb regions composed of sequences, and are found in a wide variety of species from bacteria to humans. Minisatellite repeats are thought to contribute to genome functions associated with polymorphism, gene expression, recombination, and mutability.

Some minisatellites are found on open reading frames and may or may not be polymorphic in the human population (Bois and Jeffreys, 1999). Minisatellites located in the 5' untranslated region of genes may participate in regulation of transcription (Kennedy *et al.*, 1995). Others located within introns can interfere with splicing (Turri *et al.*, 1995). Minisatellites in imprinted loci are thought to play a role in imprint control (Chaillet *et al.*, 1995; Neumann *et al.*, 1995). Minisatellites have been proposed as intermediate in chromosome pairing initiation in some eukaryote genomes (Ashley, 1994, Sybenga, 1999), which might be related to their proposed recombinogenic properties (Boan *et al.*, 1998; Wahls and Moore, 1998). Minisatellites may constitute chromosome fragile sites (Sutherland *et al.*, 1998) and have been found near a number of common translocation breakpoints and in the switch recombination sites in the immunoglobulin heavy chain genes (Brusco *et al.*, 1999). The high degree of length polymorphism among minisatellites suggests that they are rapidly evolving sequences, and newly mutated alleles have been observed in several loci. Roughly 300 human minisatellites have been typed across families (Nakamura *et al.*, 1987; Armour *et al.*, 1990; Armarger *et al.*, 1998) and approximately ten forms have been classified as hypermutable. All hypermutable minisatellites characterized so far possess internal variants (Jeffreys *et al.*, 1991; Buard and Vergnaud, 1994).

Satellite DNA

Satellite DNAs are tandemly repeated sequences, organized in long, typically megabase

(Mb)-sized arrays and are located in regions of pericentrometric and/or telomeric heterochromatin (Charlesworth *et al.*, 1994). In kangaroo rats (*Dipodomys ordii*) and beetles (from the coleopteran family Tenebrionidae), satellite DNAs comprise the majority of the genomic DNA (Hatch and Mazrimas, 1974). Satellite DNAs are associated with complex organizational features, such as heterochromatic compartments for proper chromosomal behavior in mitosis and meiosis, necessary for the function of eukaryotic genomes (Csink and Henikoff, 1998). Satellite DNAs are major constituents of functional centromeres in humans (Schueler *et al.*, 2001). Centromeric satellite differ sequence even among closely related organisms, and these differences are reflected as changes in corresponding centromeric histones, such as CENP-A in mammals and Cid in *Drosophila* (Henikoff *et al.*, 2001). Satellite DNAs can increase in copy number by replication slippage, rolling circle replication, conversion-like mechanisms, or other presently unknown mechanisms in relatively short evolutionary time (Charlesworth, 1994).

Microsatellites composed of simple tandem DNA repeats

Genomes contain simple tandem repeats of mono-, di-, tri-, tetra-, and penta-nucleotide repeat units. Many simple tandem repeat sequences in human populations are polymorphic in copy number and have been utilized widely for study of the human genome. Many of single tandem repeats have imperfections in the repeat unit, and the degree of instability of such repeats is directly related to the length of the perfect repeat (Weber, 1990). Sequence-specific DNA-binding proteins have been identified for di- and tri-nucleotide repeats (Richards *et al.*, 1993), and one type of repeat can act as the preferred site for nucleosome assembly *in vitro* (Wang *et al.*, 1994). Simple trinucleotide tandem repeats can undergo dynamic mutation. Dynamic mutation is change of the genetic material that occurs over several generations (Richards, 1993). Three trinucleotide repeats (AAC, CCG, and AGC repeats) often undergo dynamic mutation that results in disease, and homopurine/homopyrimidine repeats form unusual structures, including triplexes and hairpins, by intrastrand interactions of purine stretches. NGG repeats form tetraplex structures. The structures genome functions, and diseases caused by repeats are summarized in Table 1.

(a) Homopurine/homopyrimidine sequences

Homopurine/homopyrimidine DNA sequences are widely distributed in eukaryotic genomes, and they block DNA replication by the formation of triple helical structures (Rao, 1996). Amplification of cellular DNA may be regulated by homopurine/homopyrimidine sequences (Rao *et al.*, 1988; Lapidot *et al.*, 1989; Baran *et al.*, 1991). Imperfect repeats of GAAG that are larger than 100 bases in length can form a triplex DNA structure through Watson-Crick and Hoogsteen base pairings. Such sequences are associated with recombination hot spots (Wells *et al.*, 1988). AG repeats are homopurine sequences and are associated with triplex DNA formation. The d(GA/TC)$_{22}$ sequence can form two conformers: a pyr/pur/pur triplex (referred to as *H-triplex) and pur/pur hairpin (*H-hairpin), depending on the zinc concentration (Beltran *et al.*, 1993). GAA repeats can also form intramolecular triplex DNA structures (Hanvey *et al.*, 1988).

(b) AC repeats

AC repeats on one DNA strand and the corresponding GT repeats on the complementary strand are the simplest and most common repeats. Instability of the AC repeats is proportional to the length of repeat (Weber, 1990). AC repeats can induce the Z-conformation of DNA (Hamada *et al.*, 1984).

(c) CCG repeats

CCG repeats comprise one group of fragile sites and may occur in excess of 1000 copies. Fragile X syndrome is a genetic disease associated with expansion of the CCG trinucleotide repeat in the 5'-untranslated region of the FMR1 gene (Verkerk *et al.*, 1991; Kremer *et al.*, 1991; Yu *et al.*, 1992). The following mechanism has been proposed for CCG repeat expansion. The change from pre to full mutation occurs during oogenesis, the mutation is transmitted via the ovum, and the expansion becomes unstable and breaks during the very early postzygotic cell divisions. In males, the length of the repeat is reduced or remains within the permutation range during spermatogenesis (Sutherland *et al.*, 1995). CCG repeats form unusual structures such as hair-pins in single-strand DNA (Mitas, 1997) and slipped-strand DNA structures (Pearson and Sinden, 1996; Pearson *et al.*, 1997).

(d) AGC repeats

AGC repeats are involved in a number of neurological disorders. They can expand to high copy numbers when in the 5'untranslated region of a gene (as in myotonic dystrophy), and in coding regions in which the copy numbers are typically less than 100 repeats. Several human neurological disorders such as myotonic dystrophy (Spara *et al.*, 1991; Brook *et al.*, 1992) and spinobulbar muscular atrophy or Kennedy disease (La Spada *et al.*, 1991) are due to mutation of AGC trinucleotide repeats. An excessive number of AGC repeats in codons is associated with

X-linked spinal and bulbar muscular atrophy (La Spada *et al.*, 1991), Huntington disease (Huntington Disease Collaborative Research Group, 1993), type 1 spinocerebellar ataxia (Orr *et al.*, 1993), dentatorubral-pallidolusian atrophy (Koide *et al.*, 1994), and Machado-Joseph disease (Kawaguchi *et al.*, 1994). AGC repeats also form hairpins in single-stranded DNA, slipped-strand DNA structures.

(e) CAA repeats

Hereditary nonpolyposis colon cancer is caused by 1-base deletions in a region containing three direct CAA repeats through polymerase errors that persist as a result of a deficiency in mismatch repair. A slipped mismatch-repair mechanism has been proposed to account for the observed mutations in this repeat sequence (Bissler *et al.*, 1994).

(f) NGG repeats

All three NGG-repeats (CGG, TGG, and AGG) form tetraplexes, and the stabilities of these tetraplexes decrease in the order $(AGG)_{20} > (TGG)_{20} > (CGG)_{20}$ (Usdin, 1998).

Long interspersed nuclear elements (LINEs) and short intersperse nuclear elements (SINEs)

In vertebrates, most transposable sequences are LINEs and SINEs (Lander *et al.*, 2001; Venter *et al.*, 2001). LINEs comprise approximately 20% of the human genome, and SINEs such as *Alu* and MIR elements in murine and human cells that are clearly capable of transposition in genome, suggesting that these elements are active transposons (Dombroski *et al.*, 1991; Naas *et al.*, 1998). Unlike LINEs, SINEs do not encode enzymes that explain their mobility, and their transposition may be due to the enzyme encoded by LINEs (Kajikawa and Okada, 2002).

Nucleosome structure

The DNA helix is characterized by a repeated structure consisting of 10 to 11 bp. The basic structural unit of chromatin is the nucleosome, which contains 147 bp of DNA wrapped in a left-handed super helix 1.7-2.5 times around a core histone octamer (Horn and Peterson, 2002). Metazoan chromatin contains additional linker histones that bind to nucleosomes and protect an additional 20 bp of DNA from nuclease digestion at the core particle boundary. Thousands of nucleosomes are organized in a continuous fashion on the DNA helix and are separated by 10 to 60 bp of linker DNA. The smallest functional unit of chromatin might be the "nucleosomal array," which consists of 12 tandem repeats of a 208-bp nucleosome-positioning sequence (Hansen, 2002; Horn and Peterson, 2002).

Table 1. Simple tandem repeats (Nucleotide sequences are represented by the direction from 5'- to 3'-termini. Complement sequences are represented after '/').

Core sequences (double strand)	Genome functions including diseases
homopurine/homopyrimidine (=A/T- G/C- AG/CT-, AAG/CTT-, AGG/CCT-repeats and so on)	Triplex DNA structure
AC/GT	Instability in genome is proportional to their perfect repeat length, Z-conformation of DNA
AAC/GTT (= ACA/TGT, CAA/TTG)	Hereditary nonpolyposis colon cancer disease
AGC/GCT (= GCA/TGC, CAG/CTG)	unusual helical structure neurological disorders
CCG/CGG (= CGC/GCG, GCC/GGC)	unusual helical structure, tetraplexes fragile sites, fragile X syndrome
GAA/TTC (= AGA/TCT, AAG/CTT)	triplex DNA structure
AGG/CCT (= GAG/CTC, GGA/TCC)	triplex DNA structure, tetraplexes
TGG/CCA (= GGT/ACC, GTG/CAC)	tetraplexes

1.3. Genome architecture

The concepts of "base periodicity" and "repetitive sequences" are different; however, repetitive sequences influence base periodicity. Base periodicity can be generated by amplification of core units. Repetitive sequences including minisatellites, satellites, LINEs, and SINEs may be derived from core units. Paralogous protein coding sequences and multiplication of tRNAs and rRNAs can also act as core units. The distributions of core units appear to be regulated in genomes. Minisatellites with total lengths greater than 100 bp have a distribution strongly biased toward the long arm of human chromosome 22 in the region of the telomere (Vergnaud and Denoeud, 2000). Minisatellites similar to the chi sequence of lambda phage (GCTGTGG) are located within the terminal 10% of chromosome 22 and are present at a much higher frequency in terminal R bands of human chromosomes (Amarger *et al.*, 1998). Chromosome ends appear to be relatively poor in recombinatory units during human male meiosis, and the very high male recombination rates observed toward chromosome ends. Specific mechanisms must be activated during male meiosis, and subtelomeric minisatellites appear to be involved in chromosome pairing either directly or via interactions with pairing proteins (Ashley, 1994; Sybenga, 1999). Region-specific low-copy repeats (LCRs) (Lupski, 1998) are distinguished from highly repetitive sequences in the human genome. In contrast to other repeats, LCRs often appear to be located preferentially near the centromeres and telomeres of human chromosomes (Eichler *et al.*, 1999).

1.4. Objective of the present study

Computational approaches for identifying protein-coding sequences are available, whereas identifying regulatory elements or hidden signals in genomic DNA, which is composed primarily of noncoding sequences, is very difficult. Because power spectrum analysis is very sensitive method for detecting hidden periodicities in a genome, it can be used to study repetitive sequences associated with biological functions and to identify hidden signals in noncoding sequences. Recently, the complete sequences of the genomes of many organisms have been determined, and thus, it will be possible to examine species-specific periodicities and self-similarity in the base organization of individual genomes. The purpose of present study was to investigate the hidden periodicities (from short to long periodicities) in a wide variety of prokaryotic and eukaryotic genomes for which complete sequence is available. I used a power spectrum (described in Section 2.1) and to characterize their periodicities at nucleotide sequence level based on a parameter ($F_k(N_1,N_2)$; described in Section 2.3). Unique repetitive sequences can be identified by power spectrum analysis. This method also reveals characteristic elements that are not detected by homology searches because of relatively low levels of sequence homology.

In Section 3.1, I characterize short periodicities in view of periodicity reflected in codon usage and DNA helical repeat structures in prokaryote and eukaryote genomes. An 11-bp periodicity is found in eubacteria, whereas a 10-bp periodicity is observed in archaebacteria and eukarya. I discuss this difference in periodicities on the basis of histone and histone-like protein structures.

Next, I analyze the longer hidden periodicities, which are species-specific, in *C. elegans*, *A. thaliana*, *D. melanogaster*, and *H. sapiens* (in Section 3.2). In *C. elegans*, 68-, 59-, and 94-bp periodicities were detected. I also discuss the centric function of worm chromosomes. In *A. thaliana*, ten periodicities were identified. I found that two periodicities (126 and 174 bp) are related to ORFs that consist of Gly-rich amino acid sequences. In *D. melanogaster*, a common 5-bp periodicity in adenine and thymine spectrum is present. I also describe the common periodicities associated with nucleosome structure in the human genome. In *H. sapiens*, 167- and 84-bp periodicities were detected along the entire lengths of chromosomes 21 and 22. Interestingly, 167-bp is the length of DNA that forms two complete helical turns in nucleosome organization. Moreover, these periodicities were associated with NGG-repeat forming self-assembly DNA.

The relation of fractal property observed in long-range correlation to gene organization of genomes is described in Section 3.3. The slope of the logarithm of power (log $S(f)$) to the logarithm of frequency (log(f)) is associated with a fractal property in genomic sequences (Table 2). A flat power spectrum corresponds to a random sequence. In other words, if the slope is 0, the sequence is representative of those produced by random processes. When the slope is close to –1, the nucleotide sequence has a characteristic fractal correlation, and a slope between 0 and –1 indicates a long-range correlation, meaning that the self similarity is reflected in the base organization. Extreme flat spectra in middle frequencies were observed in genomes of two insects: *D. melanogaster* and *A. gambiae*. The significance of long-range correlation in genomic sequences is unclear. Because DNA sequences show the long-range correlations, it is worth examining the evolutionary origin of genomes. In Section 3.4, I present a model for describing long-range correlation with respect to the flat spectra of insect genomes. This model is based on a mathematical system consisting of three changes: exchange of a nucleotide, expansion of the DNA sequence, and insertion of a core sequence. These changes correspond to mutation, duplication or slippage, and foreign gene insertion event, respectively. The flat spectra of insect genomes can be explained by this model. This thesis provides evidence that periodicity, including long-range correlation, is common to all genomes, suggesting the basic strategy for organization of a genome in nature.

Table 2. Interpretation of the exponent β ($S(f) \propto f^{\beta}$).

Exponent β	interpretation
-2	Brownian motion or random walk
-1	1/f noise (= Fractals)
$-1 < \beta < 0$	long-range correlation
0	white noise (random sequence)

2. Methods

2.1. Long-range correlation

The term 'power-law' can be explained as follows. Let us consider one-dimensional movement of an object. The object is located at position *x(t)* at time *t*. The correlation between *x(t)* and *x(t+τ)*, which are separated by time lag *τ*, is represented by the autocorrelation function

$$C_x(\tau) = \langle x(t)x(t+\tau) \rangle \tag{1}$$

where the brackets (< >) denote the average over the positions along the time. When the object is located in a positive region at both *t* and *t+τ* or in a negative region at both *t* and *t+τ*, $C_x(\tau)$ has a positive value. When the positions *x(t)* and *x(t+τ)* are inversely correlated, $C_x(\tau)$ is negative. Statistical independence between positions separated by time lag *τ* implies that $\langle x(t)x(t+\tau) \rangle = \langle x(t) \rangle^2$. In the case that $C_x(\tau)$ is larger than $\langle x(t) \rangle^2$, *τ* is called the correlation length. The relation between $C_x(\tau)$ and *τ* expressed in Eq. (2) is referred to as the "power law"

$$C_x(\tau) \sim \tau^{\alpha}. \tag{2}$$

In general, most physical processes can be characterized by exponential decay ($C_x(\tau) \sim e^{-\tau}$) of the correlation function $C_x(\tau)$. This can be explained as follows. Let us consider the magnetism of iron. At the atomic level, the magnetization consists of a spin with two states. One state is 'up' (or magnetic), whereas the other is 'down' (or non-magnetic). Where the spins are aligned uniformly, iron is transformed into magnetism. The degree of the order of the spin array is characterized by the correlation function. Exponential decay means that components of a spin pair, s_i and s_j, are separated by a distance, l, that is not correlated. The transition between the magnetic and non-magnetic states occurs at a critical temperature. However, the correlation function is expressed by Eq. (2) at the critical point. This is called the 'critical phenomena' (Stanley, 1971) because the power-law ($C_x(\tau) \sim \tau^{\alpha}$ with a < 0) decays more slowly than does the exponential function ($C_x(\tau) \sim e^{-\tau}$). Such power-law correlation is typically called a long-range correlation. Moreover, many systems evolve spontaneously to the critical state, refer to as self-organized criticality (Bak *et al.*, 1988; Bak and Chen, 1991). Long-range correlations are now considered as signatures of self-organized criticality. Thus, the long-range correlation is very important for understanding the behavior of a system, for example, nucleotide organization in a genome.

2.2. Power spectrum analysis

The power spectrum is the transformation of a sequence of variables in the "frequency space." It has the advantage that any hidden or latent periodic patterns existing in the original data become evident after transformation; hidden periodic signals are visible as peaks in the spectrum. Because the power spectrum can be applied only to numerical sequences, each DNA sequence was transformed into a binary sequence x_j. First, the base of the jth position was represented by b_j, where $j = 0, 1, 2, \ldots, N-1$ and b = A, T, G, or C. In the case that base A is targeted for transformation, if b_j is equal to A, then x_j is set to 1, otherwise, x_j is set to 0. Thus, the DNA sequence is transformed by binary sequence (Voss, 1992; Vieira, 1999).

Consequently, four sets of binary sequence are generated from a DNA sequence. The power spectrum of a binary sequence x_u of length N is by definition

$$S(f_j) = \left| \frac{1}{N} \sum_{u=0}^{N-1} x_u \exp(-2\pi i u f_j) \right|^2 \qquad (1)$$

where $i^2 = -1$ and frequency with j-base length $f_j = j/N$ ($j = 0, \ldots, N-1$). In other words, the power spectrum is the square of the Fourier transformation of x_u. Fourier transformation is explained as follows. In general, a process can be described in the time domain by the values of some quantity y as a function of time t (e.g. $y(t)$). However, a process can also be described in the frequency domain, where the process is specified as amplitude Y as a function of frequency f, that is $Y(f)$. It is useful to think of $y(t)$ and $Y(f)$ as two different representations of the same function. This relation between $y(t)$ and $Y(f)$ is the Fourier transformation. The two representations are described as Eq. (2a) and Eq. (2b).

$$Y(f) = \int_{-\infty}^{\infty} y(t) \exp(-2\pi i f t) dt \qquad (2a)$$

$$y(t) = \int_{-\infty}^{\infty} Y(f) \exp(2\pi i f t) df \qquad (2b)$$

Eq. (1) contains the discrete transformation of Eq. (2a). When the DNA sequence is regarded as a time series $y(t)$, Fourier transformation can be used to study the periodicity in genomes.

From Eq. (1), it can be seen that $S(f_0) = \langle x_u \rangle^2$, where the brackets (< >) denote the average along the sequence. As a result, this quantity carries no information regarding the relative positions of the nucleotides. Therefore, this quantity was neglected throughout the calculations; that is, only frequencies with $j > 0$ were used in the present study. Because the power spectrum for subsequences of real numbers is symmetric with respect to the axis $f = 0.5$, the plot of the power spectrum is only for frequencies in the interval of 0 to 0.5.

The average power spectrum was computed by

$$\bar{S}(f_j) = \frac{1}{n} \sum_{s=j-n/2}^{j+1+n/2} S(f_s). \qquad (3)$$

The power spectrum can be averaged by calculating it for the entire sequence of N points and plotting it by averaging over n neighboring points. In the present study, I used a fast Fourier transformation algorithm (Cooley and Tukey, 1965) that accelerates calculation of the power spectrum. The length of DNA sequence analyzed should be a power of 2, that is, 2^m nucleotides, where m is an integer.

Typically, the spectrum landscape can be classified into three types: flat, sharp peaks, and slope. Figure 2a, b, and c provide examples of the three types of spectra. A flat spectrum is obtained in frequency range from approximately 10^{-3} to 0.5. Two peaks are observed at the frequencies with ≈ 0.091 (= 1/11) and 0.333 (= 1/3), which correspond to 11-base and 3-base periodicities in DNA sequence. A slope spectrum is observed when the frequency range is smaller than 10^{-3} (1/bp), which is associated with fractal property. Computation of the slope of the logarithm of power (log $S(f)$) to the logarithm of frequency ($\log(f)$) provides us with information regarding nucleotide organization in DNA sequences. Again, a flat spectrum, $S(f) \propto$ constant, represents random sequences, and this spectrum is known as 'white noise,' which represents true random phenomena. Integration of white noise produces a Brownian motion or random walk with $S(f) \propto f^{-2}$. The spectrum underlying a periodic pattern in the DNA sequence has sharp peaks that correspond to periodicities. For examples, the sine function of trigonometry has a sharp peak at $1/2\pi$. The slope spectrum $S(f) \propto f^{\beta}$ is known as the "power-law" distribution and is related to fractal properties. In particular, $S(f) \propto f^{\beta}$ with $\beta \approx -1$ is called '$1/f$ noise.' Many naturally occurring fluctuations, for example, from electronic voltages, time standards, and meteorological, biological, traffic, economic, and musical quantities, have nontrivial correlations (long-range correlations). Interpretation of the exponent β is summarized in Table 2. In the case of the *Escherichia coli* genome shown in Figure 2, the exponent $\beta = -0.73$ (for adenine) is obtained in the frequency range smaller than 10^{-3} (1/bp). There is a long-range correlation in this frequency range.

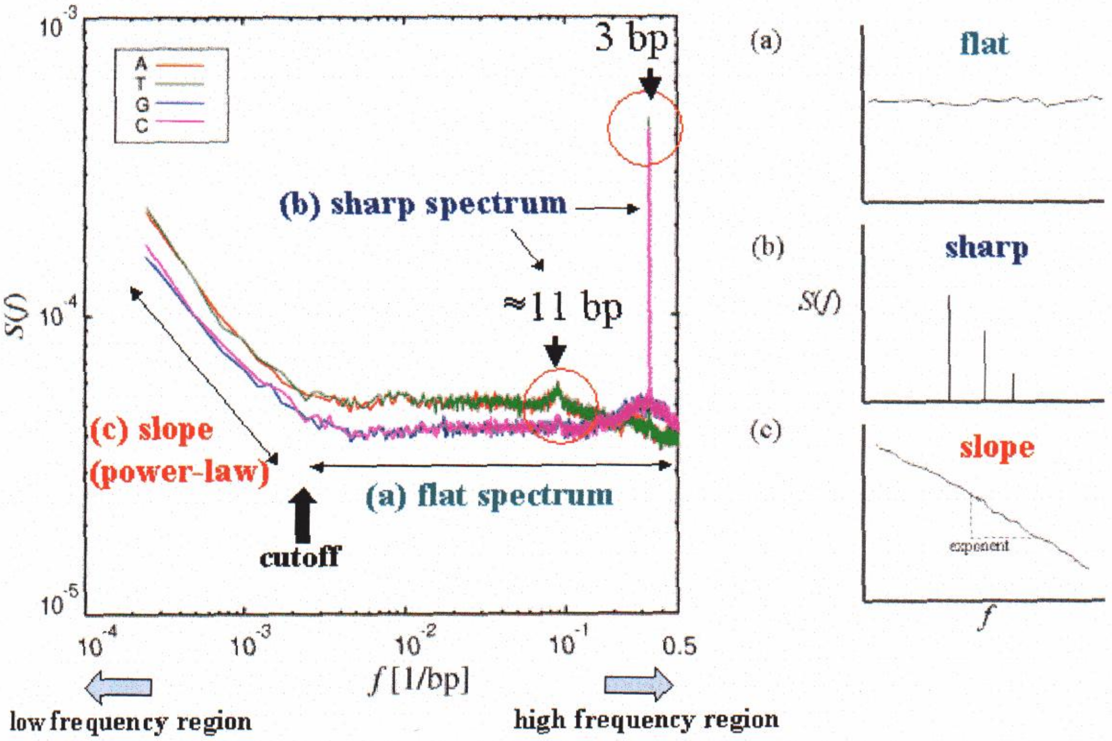

Figure 2. Typical spectra of *Escherichia coli* genome and classification of spectrum patterns. (a) flat spectrum, (b) sharp spectrum, and (c) slope spectrum. A flat spectrum, $S(f) \propto$ const, represents a random sequence and is known as 'white noise.' The spectra with sharp peaks correspond to periodicities. A slope spectrum, $S(f) \propto f^{\beta} (\beta < 0)$, is known as the power-law distribution and is related to fractal properties of the sequence. The power spectrum in genomes is typically called the partial power-law.

2.3. Periodic nucleotide distribution parameter

The role of power spectrum in the analysis of genomic DNA sequences is to identify underlying periodicities in DNA. Basically, the power spectrum is calculated by applying Fourier transformation. To identify regions that contribute to a certain periodicity, the periodic nucleotide distribution parameter $F_k(N_1,N_2)$ (Fukushima *et al.*, 2002a) is calculated by

$$F_k(N_1,N_2)=\frac{f_k(N_1,N_2)}{f(N_1)\cdot f(N_2)}. \tag{4}$$

Here, $f_k(N_1, N_2)$ denotes the number of the nucleotide pair N_1 and N_2 at distance k bp in window L, and $f(N_s)$ denotes the number of the single nucleotide N_s ($s = 1, 2$) in window L. When the occurrences of nucleotides N_1 and N_2 at distance k bp are statistically independent, $f_k(N_1, N_2) = f(N_1) \cdot f(N_2)$. If $F_k(N_1, N_2)$ is significantly larger than 1, the nucleotide pair is highly abundant in some regions of the genome. Bias in the nucleotide composition of the genome is decreased in $F_k(N_1, N_2)$, that is $F_k(N_1, N_2) = 1$. In the present analysis, N_1 and N_2 were set as identical nucleotides (A, T, G, and C); that is, they were examined by $F_k(N_1, N_1)$ for individual DNA sequences.

The parameter $F_k(N_1,N_2)$ in Eq. (4) can be extended from individual nucleotide occurrences to sequence occurrences. In the uth nucleotide sequence, the number of the four nucleotides (A, T, G, and C) in the sequence NS_u are denoted by $b_A(NS_u)$, $b_T(NS_u)$, $b_G(NS_u)$, and $b_C(NS_u)$; $f_k(NS_1, NS_2)$ represents the number of the pair of nucleotide sequences NS_1 and NS_2 separated by k bp in window L. $f(NS_u)$ represents the number of the nucleotide sequence NS_u estimated statistically from single nucleotide numbers ($u = 1, 2$) in window L. The periodic nucleotide distribution $F_k(NS_1,NS_2)$ is therefore calculated by

$$F_k(NS_1, NS_2) = \frac{f_k(NS_1, NS_2)}{f(NS_1) \cdot f(NS_2)} \tag{5}$$

where, $f(NS_u) = f(A)^{b_A(NS_u)} \cdot f(T)^{b_T(NS_u)} \cdot f(G)^{b_G(NS_u)} \cdot f(C)^{b_C(NS_u)}$.

2.4. DNA sequences analyzed

I analyzed available genomic DNA sequences retrieved from GenBank web site (http://www.ncbi.nlm.nih.gov/Genbank/). The genomes analyzed are listed in Table 3. The longer contigs of human, fly, and mosquito sequences were retrieved from genome draft sequences: ftp://ncbi.nlm.gov/genomes/.

Table 3. Genome sequences used.

Eubacteria	Archaebacteria	eukarya
Escherichia coli	*Archaeoglobus fulgidus*	*Homo sapiens*
Bacillus subtilis	*Methanococcus jannaschii*	*Drosophila melanogaster*
Neisseria meningitidis	*Methanobacterium thermoautotrophicum*	*Anopheles gambiae*
Helicobacter pylori	*Pyrococcus horikoshii*	*Arabidopsis thaliana*
Rickettsia prowazekii	*Aeropyrum pernix*	*Caenorhabditis elegans*
Chlamydophila pneumoniae	*Halobacterium* sp.	*Saccharomyces cerevisiae*
Mycobacterium tuberculosis	*Thermoplasma volcanium*	
Borrelia burgdorferi	*Thermoplasma acidophilum*	
Synechocystis sp.		
Deinococcus radiodurans		
Aquifex aeolicus		
Thermotoga maritima		

3. Results and Discussion

3.1. Short periodicities in prokaryotic and eukaryotic genomes

The power spectra for high frequency regions of genomes that have been sequenced completely are shown in Figure 3. All prokaryotic and eukaryotic genomes have a 3-bp periodicity (corresponding to frequency $f = 1/3$), that corresponds to the periodicity associated with codon usage. This periodicity has been reported by Shepherd (1981a, b), Fickett (1982), Staden (1990), Tsonis *et al.* (1991), and Gutierrez *et al.* (1994). The 10- and 11-bp periodicities are explained by the DNA helical repeat structure (10.55 ± 0.01 bp) as reported by Trifonov and Sussman (1980) and Tomita *et al.* (1999).

The power of the 10-11 bp periodicities is much smaller than that of the 3 bp (Figure 3). The 11-bp periodicity was observed in six eubacteria (*E. coli*, *Helicobacter pylori*, *Chlamydophila pneumoniae*, *Synechocystis* sp., *Deinococcus radiodurans*, and *Thermotoga maritime*), two archaea (*Aeropyrum pernix* and *Halobacterium* sp.), and three eukarya (*Saccharomyces cerevisiae*, *C. elegans*, and *A. thaliana*). The 10-bp periodicity was observed in one eubacteria (*Aquifex aeolicus*) and six archaea (*Archaeoglobus fulgidus*, *Methanococcus jannaschii*, *Methanobacterium thermoautotrophicum*, *Pyrococcus horikoshii*, *Thermoplasma acidophilum*, and *Thermoplasma volcanium*). The 10-bp periodicity was prevalent in the hyperthermophilic bacteria *A. aeolicus* and archaebacteria (see Figure 3b), and the 11-bp periodicity was also prevalent in eubacteria (see Figure 3a). These results are consistent with those of other reported spectral analyses (Trifonov, 1998; Herzel *et al.*, 1999). If the sequence periodicities reflect the characteristic superhelical densities of genomic DNA, the differences in periodicities between hyperthermophilic bacteria, archaebacteria, and eubacteria can be explained as follows. Archaeal histones are structurally similar to eukaryotic core histones, that is, eukaryotic and archaeal DNAs are packed as nucleosomes in negatively-constrained supercoils (Sandman, 2000). Some periodicities are known to cause curvature of the DNA. The genomic sequences in eukarya and archaea are organized and stabilized by interactions between histones and nucleotides, and the 10-bp periodicity contributes to this nucleosome organization. The 11-bp periodicity was consistent with the occurrence of negative supercoiling in bacterial DNAs (Trifonov and Sussman, 1980; Vologodsky, 1992). The 10-bp periodicity was observed in all eukaryotes examined; and the periodicity observed in *C. elegans* was more prevalent than that in *S. cerevisiae* or *A. thaliana* (see Figure 3c). Consequently, the 10-bp periodicity is prevalent in eukarya and archaea, and the 11-bp periodicity is prevalent in eubacteria. This difference is explained mainly by the high-order structure associated with packing of DNA in the cells.

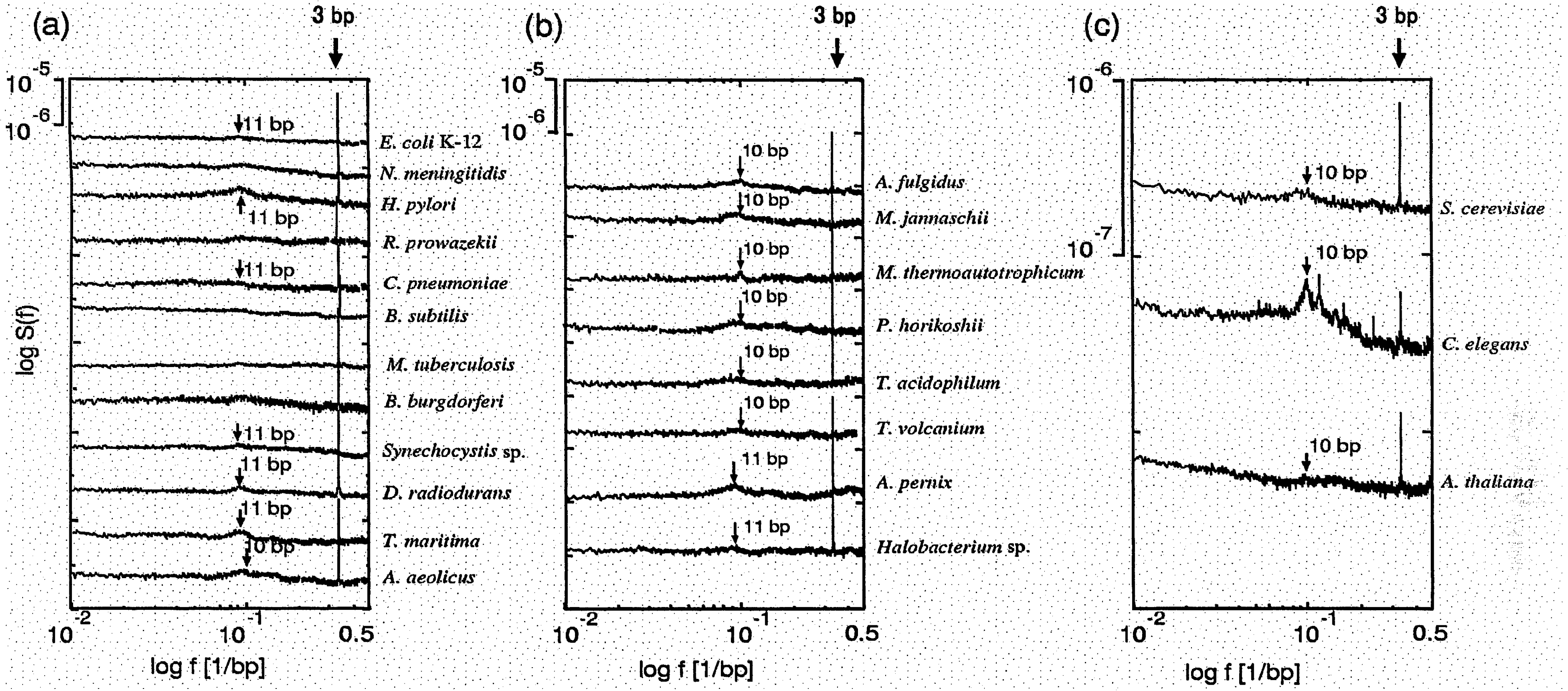

Figure 3. Power spectra (in log-log scale) of genomes for (a) eubacteria, (b) archaea, and (c) eukarya in high frequencies. All bacterial genomes have a 3-bp periodicity (corresponding to frequency $f = 1/3$), and most genomes have a 10-11 bp periodicity ($f \approx 1/10\text{–}1/11$).

3.2. Periodicities in eukaryote genomes

The sizes of the *C. elegans*, *A. thaliana*, *D. melanogaster*, *A. gambiae*, and *H. sapiens* genomes are 97, 130, 180, 278, and 3,000 Mb, respectively (*C. elegans* Sequencing Consortium, 1998; Lin *et al.*, 1999; Theologis *et al.*, 2000; European Union Chromosome 3 Arabidopsis Sequencing Consortium, 2000; The European Union *Arabidopsis* Genome Sequencing Consortium & The Cold Spring Harbor, Washington University in St. Louis and PE Biosystems *Arabidopsis* Sequencing Consortium, 1999; The Kazusa DNA Research Institute, The Cold Spring Harbor and Washington University in St. Louis Sequencing Consortium & The European Union *Arabidopsis* Genome Sequencing Consortium, 2000; Adams *et al.*, 2000; Lander *et al.*, 2001; Venter *et al.*, 2001; Holt *et al.*, 2002). The nematode *C. elegans* was the first multicellular organisms for which sequenced complete genomic sequence was available; it contains five autosomes and the sex chromosome X. Its genome has 36% G+C content. The genome of the flowering plant *A. thaliana* has five chromosomes that are remarkably uniform in 36% G+C content and a small repeat sequence. Both *D. melanogaster* and *A. gambiae* have five major chromosomes (X, 2L, 2R, 3L, and 3R), but *D. melanogaster* contains a small chromosome 4. The genomic G+C content is 41% for *D. melanogaster* and 35% for *A. gambiae*. The human genome, like the genomes of most warm-blooded vertebrates in general has long-range G+C% mosaic structures or isochores that are related to chromosome bands (Bernardi *et al.*, 1985; Ikemura, 1985; Ikemura and Aota, 1988; Bernardi, 1989).

The total number of tandem repeats is not proportional to chromosome length in three genomes

(*C. elegans* chromosome 1 (12.75 Mb), *A. thaliana* chromosome 4 (17.8 Mb), and human chromosome 22 (34.6 Mb)) (Dunham *et al.*, 1999). The number of positive minisatellites is similar in the three species, although the chromosomal sizes are different. A strong telomeric bias is observed in *C. elegans* chromosome, similar to that of human chromosome 22. In contrast, the distribution of minisatellites in *A. thaliana* is strikingly different from the distribution in human and in *C. elegans*. Tandem repeats are located mainly around centromeres. *C. elegans* chromosome 1 has telomeric bias for both short oligomers (6- and 12-bp units), resulting from the presence of many (TTAGGC)n telomere-like tandem arrays, and longer oligomers (> 18 bp). Human chromosome 22 also has telomeric bias for repeat units longer than 17 bp. In yeast, 16 bp is the threshold above which mismatch repair mechanisms are unable to correct DNA loops (Sia *et al.*, 1997). In contrast with *C. elegans*, the telomeric bias for human chromosome 22 appears only for arrays longer than 120-140 bp. This threshold is comparable to the triplet repeat instability observed above 40-50 repeats. No telomeric bias is observed for *A. thaliana* chromosome 4 (Vergnaud and Denoeud, 2000). In this section, periodicities in genomes from power spectrum analyses (Section 2.2) are characterized by nucleotide sequence level on the basis of the periodic nucleotide distribution parameter described in Section 2.3.

(a) *C. elegans*

The power spectra for all *C. elegans* chromosomes are shown in Figure 4. Each chromosome is divided into 2^{22}-bp subsequences (approximately 4.2 Mb) along the DNA strand registered in GenBank with a moving step-size of 2.1 Mb (spectra from top to bottom in Figure 4). These include a 68-bp periodicity in chromosome I, a 59-bp periodicity in chromosome II, and a 94-bp periodicity in chromosome III (marked in Figure 4). Although there are many peaks in regions with frequencies f larger than 2×10^{-2} (i.e., periodicity shorter than 50 bp), I focused on several distinct periodicities found in regions with frequencies less than 2×10^{-2} (i.e., periodicity longer than 50 bp) that had not been characterized previously (Figure 4). Short repeats, such as tandem repeats, in eukaryotic genomes have been described previously (Katti *et al.*, 2001).

To relate these periodicities to nucleotide sequences, the genomic distribution of nucleotide pair N_1 and N_1 separated by k bp was examined with parameter $F_k(N_1, N_1)$ (see Section 2.3). Figure 5 shows the distributions of the 68-, 59-, and 94-bp periodicities in a 10-kb window along each chromosome. $F_k(N_1, N_1)$ is higher than 1.5 in 11 regions that are designated by ID numbers CE1 to CE11 (Figure 5). The consensus sequences comprising the individual periodicities are listed in Table 4. The 68-bp periodicities were found for four regions of chromosome I (CE1, 1.34–1.35 Mb; CE2, 8.53–8.55 Mb; CE3, 12.32–12.33 Mb; CE4, 14.86–14.87 Mb). Interestingly, the consensus sequences were not necessarily similar between individual regions even if the pitches of the periodicities were identical. The consensus sequence of CE2, which is a cluster composed of 219 copies of the 68-bp periodicity on chromosome I, is similar to that of CE3 but is very different from those of CE1 and CE4 (Table 4). It should be noted that CE1 contained as a 12-bp core element sequence, CeRep45 (TTGGTTGAGGCT) that was characterized previously by Sanford *et al.* (2001).

Chromosome-specific periodic segments of 11-16 bp have been reported by Sanford *et al.*

(2001). These sequences were found primarily near telomeres and were predicted to be responsible for meiotic pairing. The three periodicities found in the present study (CE2, CE7, and CE9) are larger than those reported previously (Katti *et al.*, 2001), and they are distributed along the chromosomes. Several are located near the centers of chromosomes. *C. elegans* has holocentric instead of monocentric chromosomes. Diffuse kinetochores are formed along the entire length of each chromosome, and clear centromeric sequences are lacking in *C. elegans* (Comings and Okada, 1972). Though it is unclear if the periodic sequences observed in this study are related to centromere function, the strategy proposed in this study may have the power to detect the hidden periodic sequences, if present, that are related to centromere function.

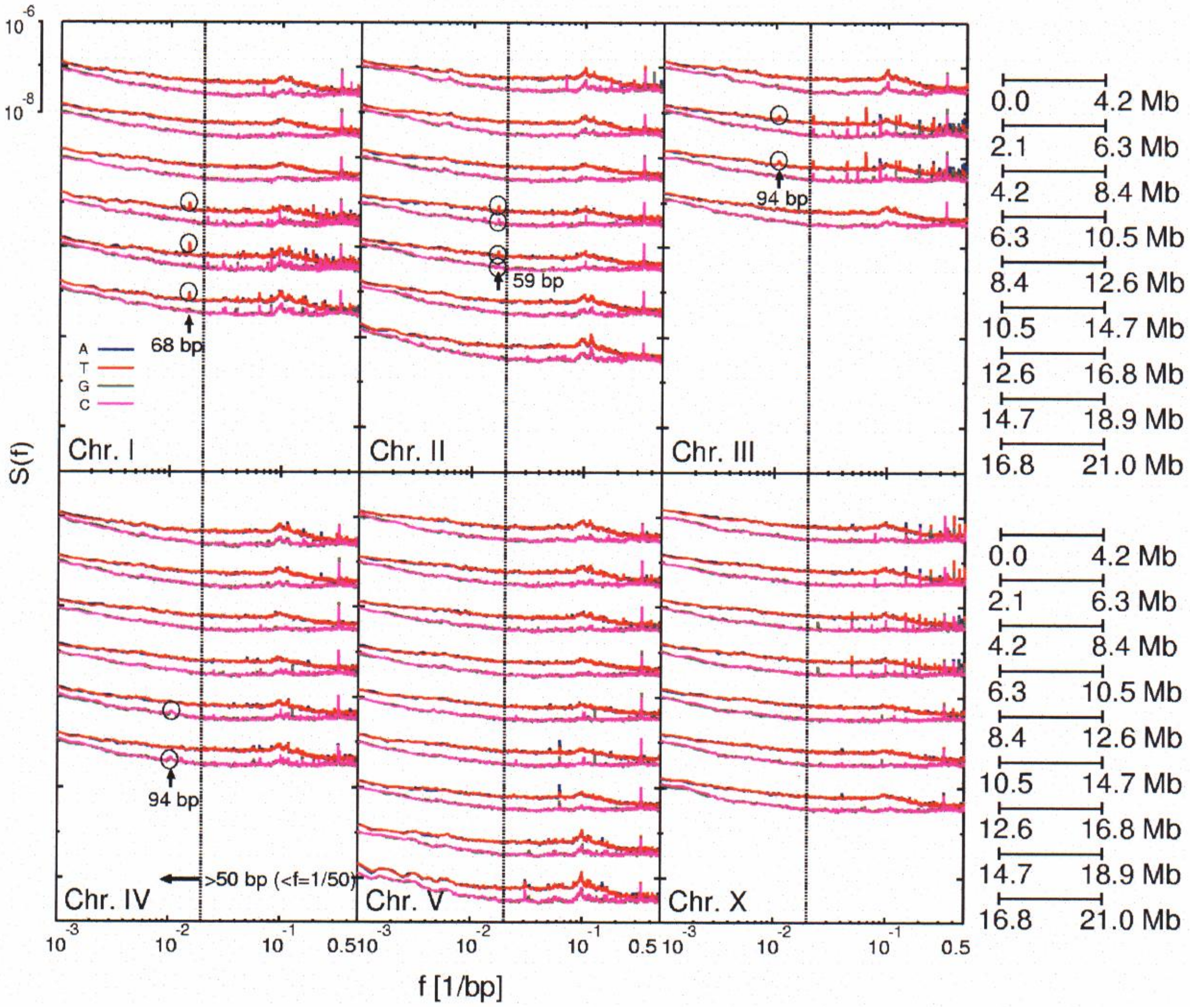

Figure 4. Power spectra (in log-log scale) of *C. elegans* chromosomes. Each DNA sequence is divided into subsequences of length 2^{22} bp (approximately 4.2 Mb) with a step-size of 2.1 Mb. Circle indicates location of periodicity length in the genomic sequences.

Table 4. Periodic elements in *C. elegans*.

ID	Chr	Region (Mb)	Period (bp)	The number of consensus core sequences	Consensus core sequence
CE1	I	1.34-1.35	68	6	TTGCTGATCTCGGTAAATATGCCAAATTTCCCGTTTGCCGACATCGGCAAATTTGCGGAATTCGCCGT
CE2	I	8.53-8.54	68	219	TTTGTGTTTTCTTTCTGAAATTCTAAGAATTTTGGTAAAAGAAAACCATTGTCAACTGAATAGGTTGA
CE3	I	12.32-12.33	68	29	TTTGTGTTTTCTTTCTGAAATTCTAAGAATTTTGTTAAAAGAAAACCATTGTCAACTGAATAGGTTGA
CE4	I	14.86-14.87	68	17	TTAATTTTGGTTGAGGCTAACACACTACAAACTACAACATTTTCTAGCCTCAACCAATTAAAAAAAAA
CE5	II	0.68-0.69	59	76	GGTGAGACCCATCGCGGTGAGACCCATCGTGACGAGACCTTTCGTGGTGAGACCCATCGT
CE6	II	1.02-1.03	59	194	TTCGTGGTGAGACCC
CE7	II	10.19-10.21	59	297	TTTGAAAACCAGTGCACAATTGAAACTCCATATTCTCAATAATTCTCAGTTTAAAAAAA
CE8	III	5.38-5.39	94		(none)
CE9	III	6.22-6.27	94	329	TTTTCCCATTGATTTGTCTACAAAGGGCATCGAAAAGCACCCAATATTTAGAGAACAGAAGATTTTGAGAATTACTGCCTCCAGAAATTGATGA
CE10	III	10.54-10.55	94	151	TTTGCGGTTTGC
CE11	IV	3.16-3.17	94	121	TTCATCTAATGGTCTAACTTTGGAAA

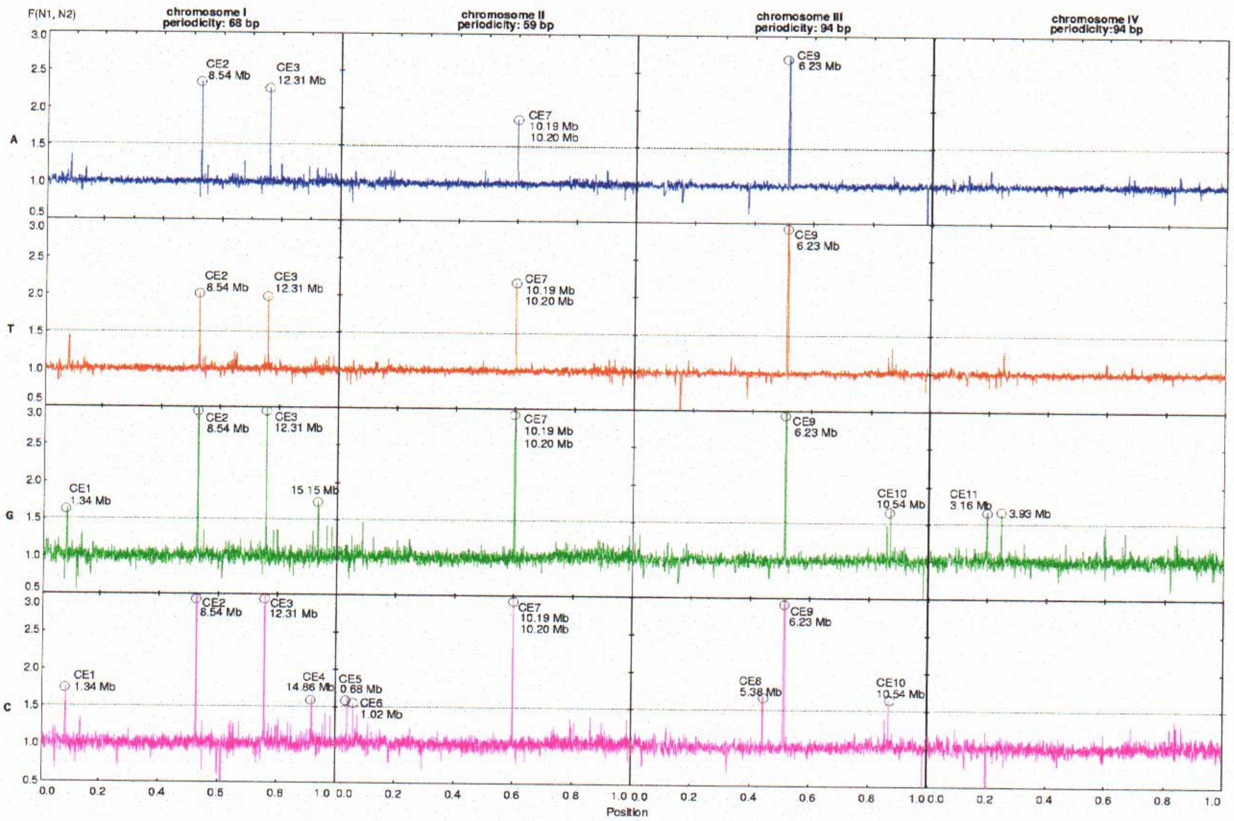

Figure 5. Periodic nucleotide distributions based on parameter $F_k(N_1, N_1)$ values with a 10-kb window in *C. elegans* chromosomes. Eleven regions with $F_k(N_1, N_1)$ values higher than 1.5 are designated by ID numbers CE1 to CE11 (corresponding to Table 4). The total length of each chromosome is normalized to 1.0.

(b) *A. thaliana*

The power spectra for the *A. thaliana* genome are shown in Figure 6. Three peaks are detected in the middle of *A. thaliana* chromosome 3, and many sharp peaks are detected in chromosomes 4 and 5. Several regions with frequencies less than 2×10^2 (i.e., periodicity larger than 50 bp) were found: Three periodicities (248-, 167-, and 126-bp) were present on chromosome 3; three periodicities (174-, 88-, and 59-bp) were present on chromosome 4; and four periodicities (356-, 174-, 88-, and 59-bp) were present on chromosome 5 (marked in Figure 6).

To describe these periodicities at the nucleotide sequence level, the genomic distribution of periodic nucleotide sequences was examined with parameters $F_k(N_1, N_1)$ and $F_k(NS_1, NS_1)$ (in Section 2.3). The distribution of $F_k(N_1, N_1)$ is shown in Figure 1. Five areas with region-specific distribution of a 126-bp periodicity was observed on chromosome 3 (Figure 7a, see G and C). These regions were designated AT1–AT5. The region-specific distributions of the 59-, 88-, and 174-bp periodicities were examined on chromosome 4, and only the 174-bp periodicity was detected. These areas were designated AT6 and AT7 (Figure 7b). In chromosome 5, I obtained two region-specific distributions (AT8 and AT9) for the 174-bp periodicity, and two (AT10 and AT11) for the 356-bp periodicity (Figure 7c). The distribution of the 356-bp periodicity designated by AT10 extends from 11.38 to 11.60 Mb and that designated by AT11 extends from 12.73 to 12.75 Mb. Core sequences comprising trinucleotide structure with $F_k(N_1, N_1)$ or $F_k(NS_1, NS_1)$ larger than 2 are listed in Table 5. NGG-type sequences were detected on several chromosomes (AT1, AT3, and AT5 for chromosome 3; AT6 and AT7 for chromosome 4; and AT8 and AT9 for chromosome 5). NGG-type sequences were found in AT1. It is well known that NGG-repeats (CGG, TGG, and AGG) form tetraplexes (summarized in Table 1). Perfect tandem NGG-repeats were not detected in those regions.

Short repeats consisting of AAG/CTT were detected in the AT4 and AT11 regions. These may be related to triplex formation. The AT10 region also contains short AGC-type repeats that are related to the unusual DNA structure. ORFs in the AT1 region have Gly at high frequency (Table 6). The Gly codons correspond to GGN. Thus, core sequences comprising periodic structures reflect the amino acid composition in an ORF. The periodic sequences in Table 6 reflect amino acid composition obtained from the present analysis. The common sequence, **SPPPPYVYSSPPPPYYS,** was also found in six regions AT2, AT3, AT5, AT6, AT8, and AT9 (Table 6).

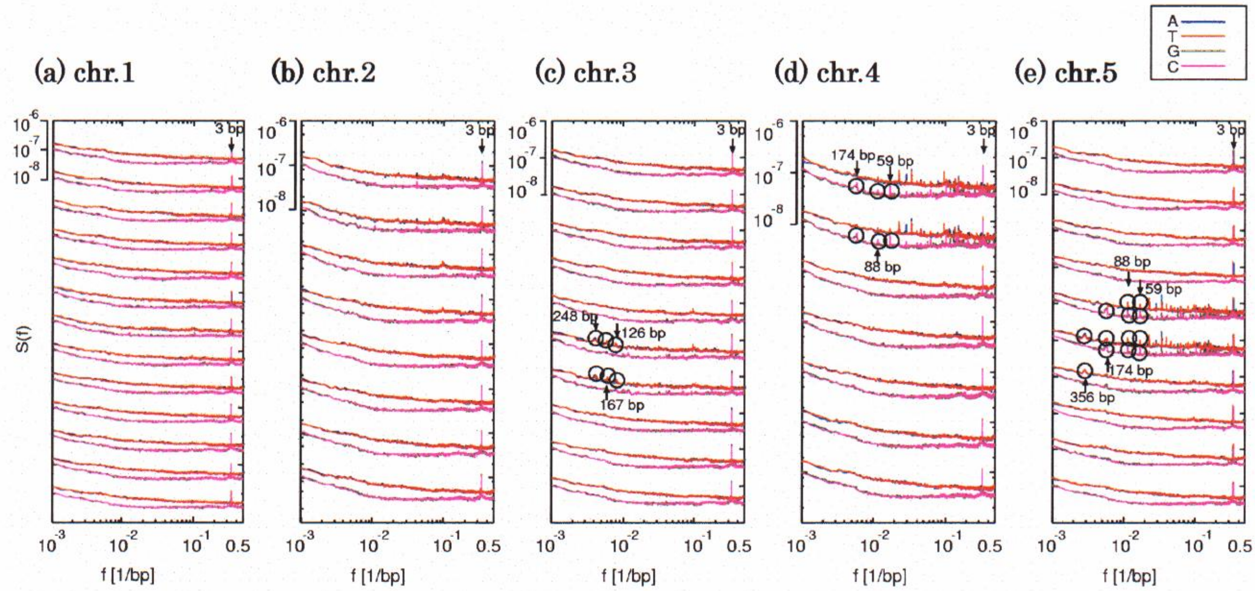

Figure 6. Power spectra of *A. thaliana* complete genome for short periodicities in log-log scale (a) Chromosome 1 (GenBank accession #: NC_003070), (b) chromosome 2 (NC_003071), (c) chromosome 3 (NC_003072), (d) chromosome 4 (NC_003073), and (e) chromosome 5 (NC_003074). The spectra of A resemble those of T, and the spectra of G curves are similar to those of C.

Table 5. Core sequences consisting of trinucleotide with $F_k(NS_1, NS_1)$ larger than 2.

chr.3	AT1	F_{126}	AT2	F_{126}	AT3	F_{126}	AT4	F_{126}	AT5	F_{126}
	GGC	3893.9	GTA	174.3	TGG	455.6	AAG	78.9	TAC	318.7
	TGG	726.6			GTA	223.9	AGA	57.0	CCA	533.2
	GGT	363.3			GAA	74.5	TAA	21.6	TTT	58.4
	TTT	29.4			TTT	29.4	AAA	9.4	TGG	1527.3
									GTA	239.9

Chr.4	AT6	F_{174}	AT7	F_{174}
	TGG	672.2	CCA	290.4
	TTT	53.3	CAC	193.6
	AAA	54.8	AAA	32.5

Chr.5	AT8	F_{174}	AT9	F_{174}	AT10	F_{356}	AT11	F_{356}
	TGG	736.7	TGG	909.5	AGC	278.6	TCT	131.0
	GTA	109.9	GTA	124.2	TGA	79.3	TAT	9.0
	AAA	47.9	AAA	56.2	ACT	78.5	ATA	6.7
			TTT	32.3	TAC	61.7		
			ATT	31.6	TGT	57.8		
					AGA	32.9		
					TAG	27.9		
					ATA	5.3		

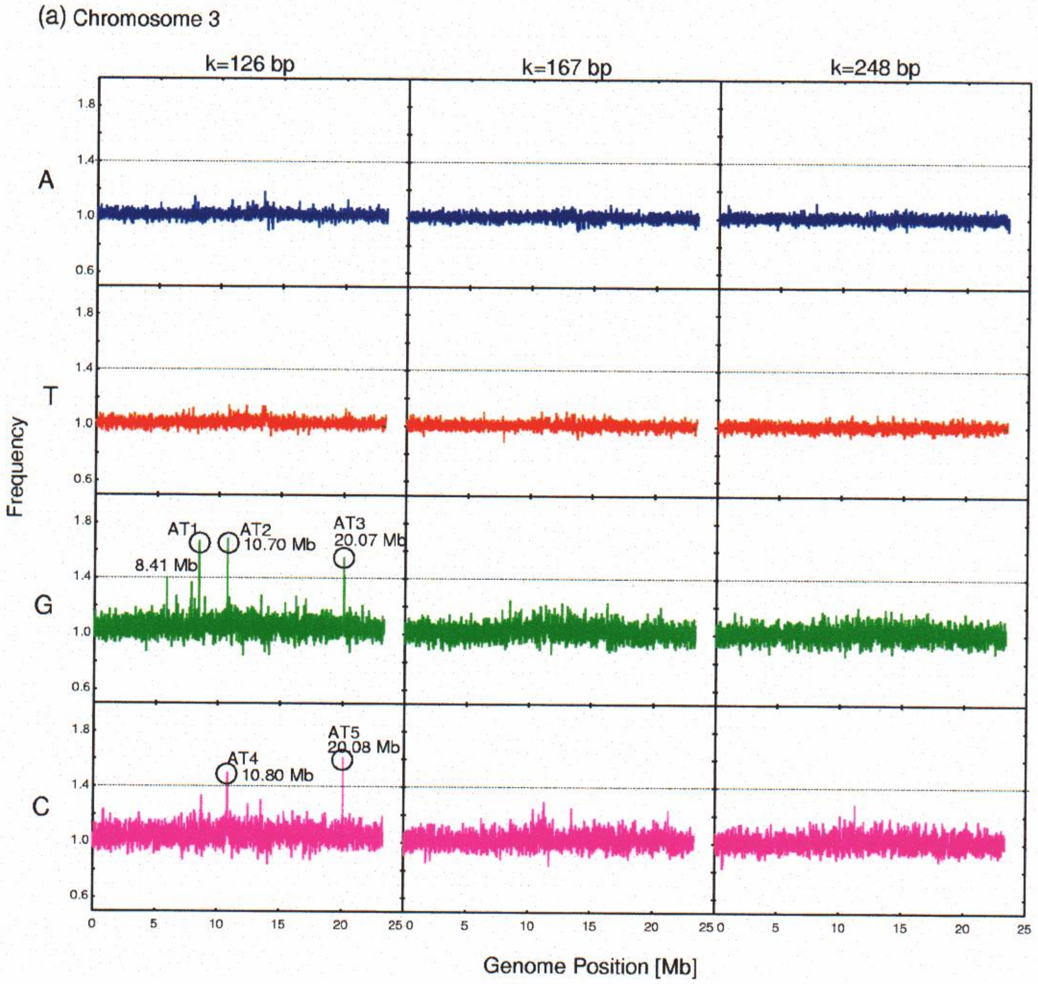

Figure 7a. Periodic nucleotide distributions based on $F_k(N_1, N_1)$ values with 10-kb window on *A. thaliana* chromosome 3. Five regions with $F_k(N_1, N_1)$ values higher than 1.4 are designated by ID numbers AT1 to AT5.

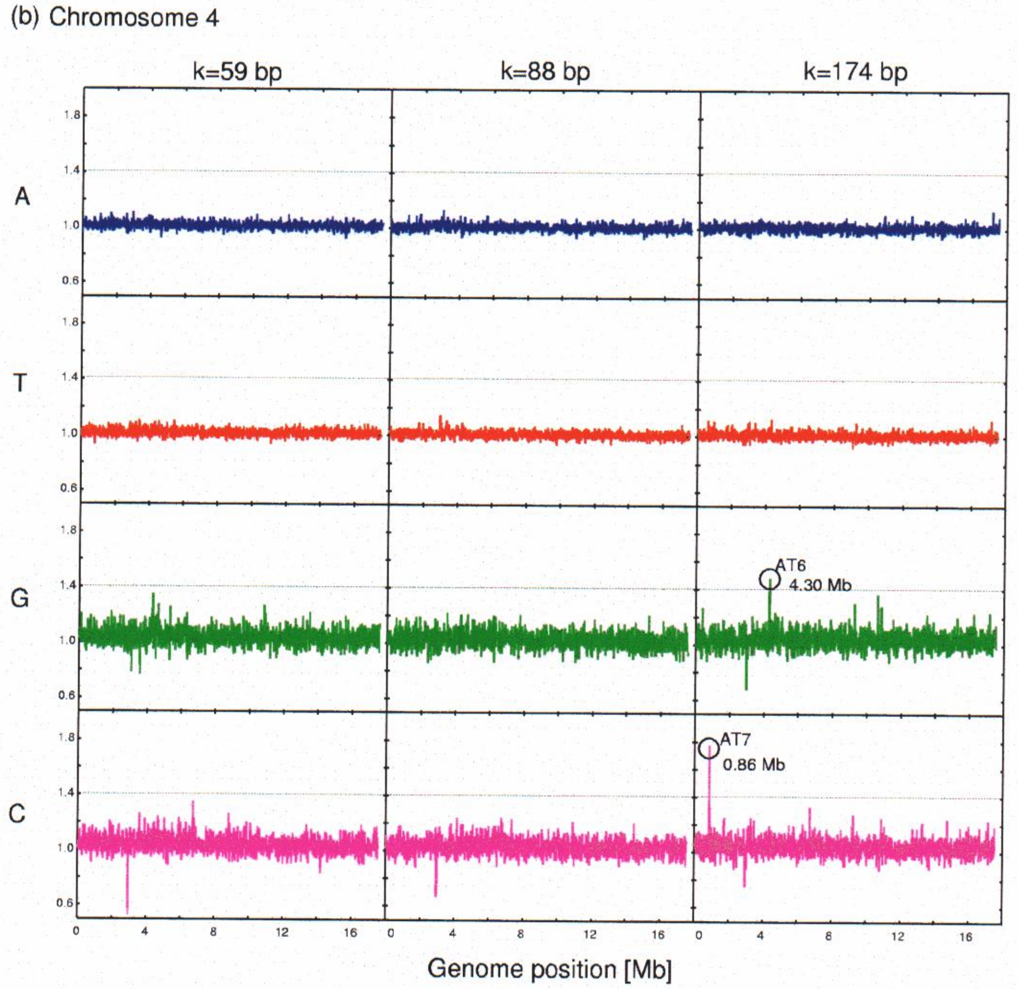

Figure 7b. Periodic nucleotide distributions based on $F_k(N_1, N_1)$ values with 10-kb window on *A. thaliana* chromosome 4. Two regions with $F_k(N_1, N_1)$ values higher than 1.4 are designated by ID numbers AT6 and AT7.

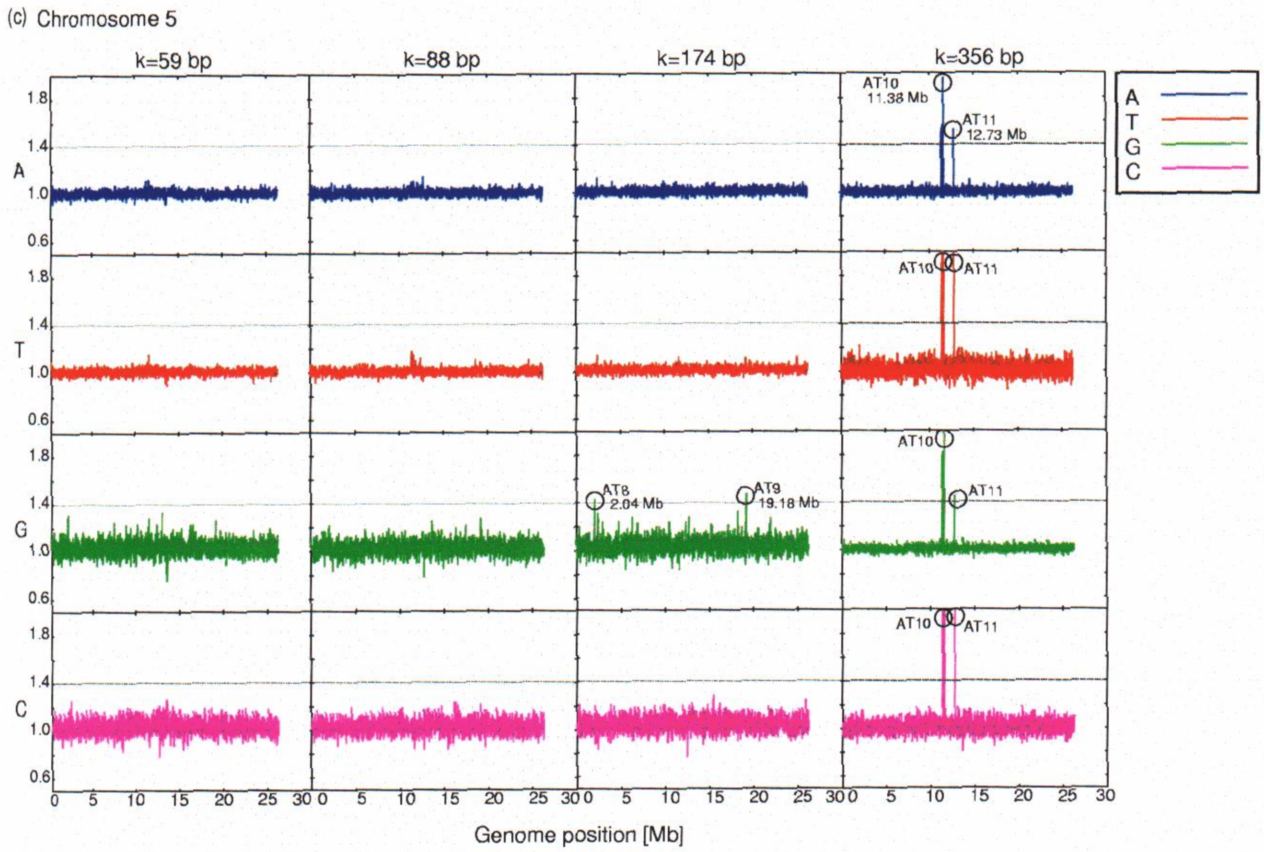

Figure 7c. Periodic nucleotide distributions based on $F_k(N_1, N_1)$ values with 10-kb window on *A. thaliana* chromosome 5. Four regions with $F_k(N_1, N_1)$ values higher than 1.4 are designated by ID numbers AT8 to AT11.

Table 6. Relation between periodicities and amino acid sequences in ORFs.

ID	chr.	aa–sequence characteristics	product protein [gene name]
AT1	3	Gly–rich*	hypothetical [At3g23450]
AT2	3	SPPPPYVYSSPPPPYYS-repetitive	unknown [At3g28550]
AT3	3	SPPPPYVYSSPPPPYYS-repetitive	extension precursor–like [At3g54580]
AT4	3	Gly, Ser, Ala–rich**	histone-H4-like [At3g28780]
AT5	3	SPPPPYVYSSPPPPYYS-repetitive	extension precursor –like [At3g54590]
AT6	4	SPPPPYVYSSPPPPYYS-repetitive	extension-like [At4g08410]
AT7	4	Not found	hypothetical protein [At4g01980]
AT8	5	SPPPPYVYSSPPPPYYS-repetitive	putative [At5g06640]
AT9	5	SPPPPYVYSSPPPPYYS-repetitive	putative [At5g49080]

[Gly–rich sequence (*)]

MGRLVSGATLLALLCFHVFVVNVVARDVSSGRDEDEKTLVGGGKGGGFGGGFGGGAGGGV
GGGAGGGFGGGAGGGFGGGGGGGGGGGGGGGGGFGGGGGFGGGHGGGVGGGVGGGHGGGV
GGGFGKGGGIGGGIGKGGGVGGGIGKGGGIGGGIGKGGGVGGGIGKGGGIGGGIGKGGGI
GGGIGKGGGIGGGIGKGGGIGGGIGKGGGVGGGFGKGGGVGGGIGKGGGVGGGFGKGGGV
GGGIGKGGGIGGGIGKGGGIGGGIGKGGGIGGGIGKGGGIGGGIGKGGGIGGGIGKGGGI
GGGIGKGGGIGGGIGKGGGIGGGGGFGKGGGIGGGIGKGGGIGGGGGFGKGGGIGGGIGK
GGGIGGGFGKGGGIGGGIGGGGGFGGGGGFGKGGGIGGGIGKGGGFGGGGGFGKGGGIGG
GGGFGKGGGFGGGGFGGGGGGGGGGGGGIGHH

[Gly, Ser, Ala–rich sequence (**)]

MGPSAHLISALGVIIMATMVAAYEPETYASPPPLYSSPLPEVEYKTPPLPYVDSSPPPTY
TPAPEVEYKSPPPPYVYSSPPPPTYSPSPKVDYKSPPPPYVYSSPPPPYYSPSPKVDYKS
PPPPYVYNSPPPPYYSPSPKVDYKSPPPPYVYSSPPPPYYSPSPKVEYKSPPPPYVYSSP
PPPYYSPSPKVDYKSPPPPYVYSSPPPPYYSPSPKVEYKSPPPPYVYSSPPPPYYSPSPK
VDYKSPPPPYVYSSPPPPYYSPSPKVDYKSPPPPYVYSSPPPPYYSPSPKVDYKSPPPPY
VYSSPPPPYYSPSPKVDYKSPPPPYVYSSPPPPYYSPSPKVDYKSPPPPYVYSSPPPPTY
SPSPKVDYKSPPPPYVYSSPPPPYYSPSPKVEYKSPPPPYVYSSPPPPTYSPSPKVYYKS
PPPPYVYSSPPPPYYSPSPKVYYKSPPPPYVYSSPPPPYYSPSPKVYYKSPPPPYVYSSP
PPPYYSPSPKVYYKSPPPPYVYSSPPPPYYSPSPKVYYKSPPPPYVYSSPPPPYYSPSPK
VHYKSPPPPYVYSSPPPPYYSPSPKVHYKSPPPPYVYNSPPPPYYSPSPKVYYKSPPPPY
VYSSPPPPYYSPSPKVYYKSPPPPYVYSSPPPPYYSPSPKVYYKSPPPPYYSPSPKVYYK
SPPHPHVCVCPPPPPCYSPSPKVVYKSPPPPYVYNSPPPPYYSPSPKVYYKSPPPPSYYS
PSPKVEYKSPPPPSYSPSPKTEY

(c) *D. melanogaster*

The power spectra of all chromosomes of *D. melanogaster* have short periodicities. All chromosomes contained a 3-bp periodicity (frequency $f = 1/3$) and a 10-bp periodicity ($f = 1/10$) (Figure 8). These results are consistent with the eukaryotic periodicities associated with chromatin structure (see also Figure 3c). I also found that all *D. melanogaster* chromosomes have ≈ 5-bp periodicity, $f \approx 1/5$, indicated by arrows in Figure 8. Interestingly, the broad weak peaks centered at 5-bp periodicities are present across entire genomes and are especially prevalent for A and T (Figure 8).

The distributions of 3-, 4-, and 5-bp periodicities based on $F_k(N_1, N_1)$ values with 10-kb window for the *D. melanogaster* X chromosome are shown in Figure 9. For nucleotide G or C, the $F_k(N_1, N_1)$ average is the highest for the 3-bp periodicity. The average of these three periodicities was higher than 1.0 (Figure 9). I have already indicated that this 3-bp periodicity is associated with codon structure (Section 3.1). In contrast, the average of A and T is the lowest in 3-bp periodicity (Figure 9 in A and T), suggesting that 4- and/or 5-bp periodicities consist of A/T-rich sequences.

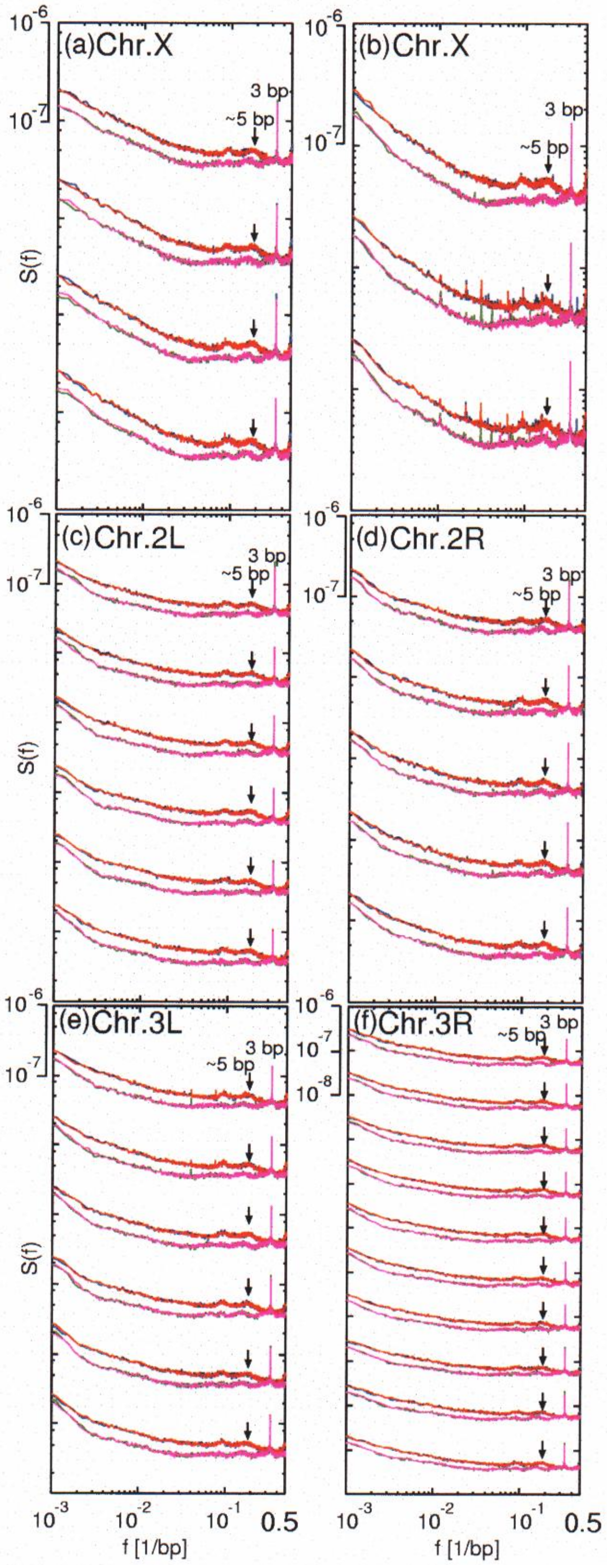

Figure 8. Power spectra of *D. melanogaster* genomic DNA at short periodicity in log-log plot (a) Chromosome X (AE002566), (b) chromosome X (AE002593), (c) chromosome 2L (AE002690), (d) chromosome 2R (AE002787), (e) chromosome 3L (AE002602), and (f) chromosome 3R (AE002708).

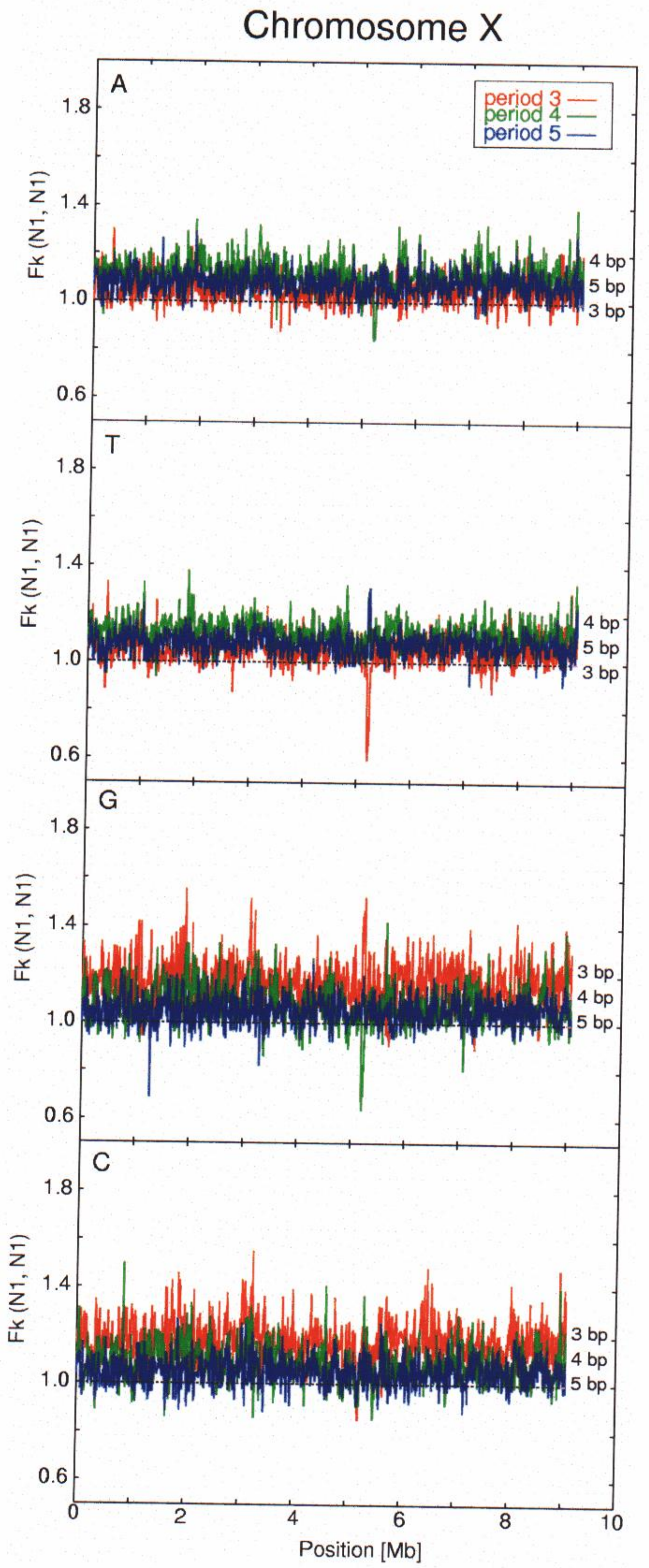

Figure 9. Periodic nucleotide distributions based on $F_k(N_1, N_1)$, values with 10-kb window in *D. melanogaster* X chromosome. Period sizes k set to 3, 4, and 5 bp.

(d) *H. sapiens* chromosomes 21 and 22

It is of interest how the human genome differs from those of other species. From a periodical point of view, examination of the *H. sapiens* genome is very important. Power spectra of human chromosomes 21 (Hattori *et al.*, 2000) and 22 (Dunham *et al.*, 1999) are shown in Figure 10. Two broad peaks centered at the 167- and 84-bp periodicities are found across the entire lengths of both chromosomes (Figure 10). Interestingly, the 167-bp periodicity is identical to the length of DNA that is known to form two complete helical turns in one nucleosome with H1 histone (Sinden, 1994). It is possible that the respective sequences form contiguous arrays of a specific compact form of nucleosome. The distributions of the 84- and 167-bp periodicities are shown in Figure 11 as the periodic nucleotide density for a 10-kb window on human chromosomes 21 and 22. These periodicities are present across the entire chromosomes because the baselines of these distributions along the chromosomes are shifted to a level clearly higher than 1.0 (Figure 11). The core elements corresponding to evident peaks (marked in Figure 11) contained a high frequency of TGG (Table 7). In the case of 42 copies of a 167-bp periodic element clustered in the 3.49-3.50 Mb region of chromosome 22 (ID number H3), each element was composed of TGG-containing sequences such as GGCTGG, CTGGCT, and GCTGGC when represented by hexanucleotides (Table 7). On chromosome 22, the high frequencies of TGG were found near the centromere (region HS2, 317 copies of 84-bp element, 0.39 to 0.40 Mb) (see Table 7 for the periodic elements). On chromosome 21, a cluster of the same 84-bp elements was detected in the region near the telomere (HS1) (Table

7). Short TGG-repeats are known to form a specific subset of folded DNA structures (see Table 1) and to be associated with the self-assembly phenomenon (Chen, 1997; Usdin, 1998). Elements including short TGG-repeats observed in the present study may form specific higher-order structures, and the organization of these elements may be important for constructing nucleosomes.

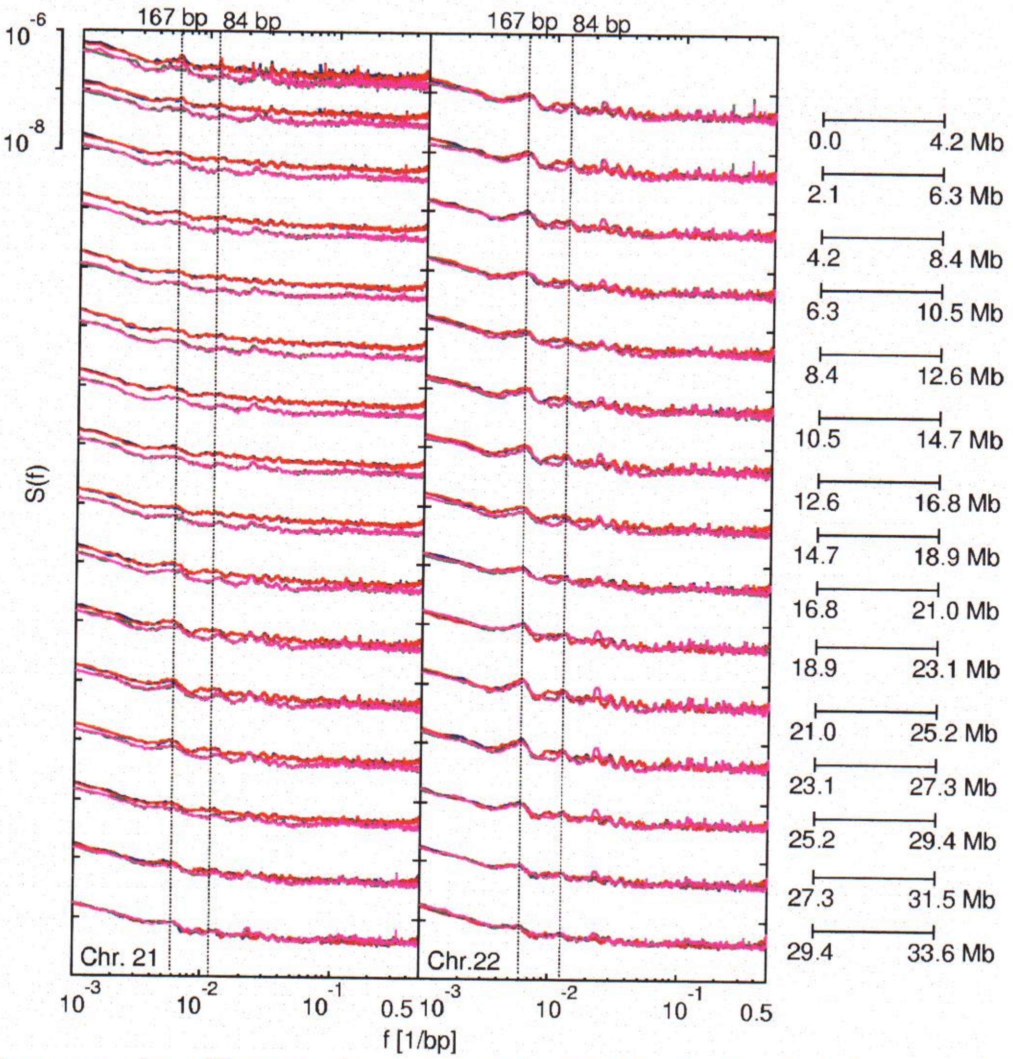

Figure 10. Power spectra of human chromosomes 21 and 22. For the high frequency, the spectral behaviors differ from those of *C. elegans*. In the middle frequency region, broad peaks centered at 84- and 167-bp periodicities are present for all subsequences of both chromosomes.

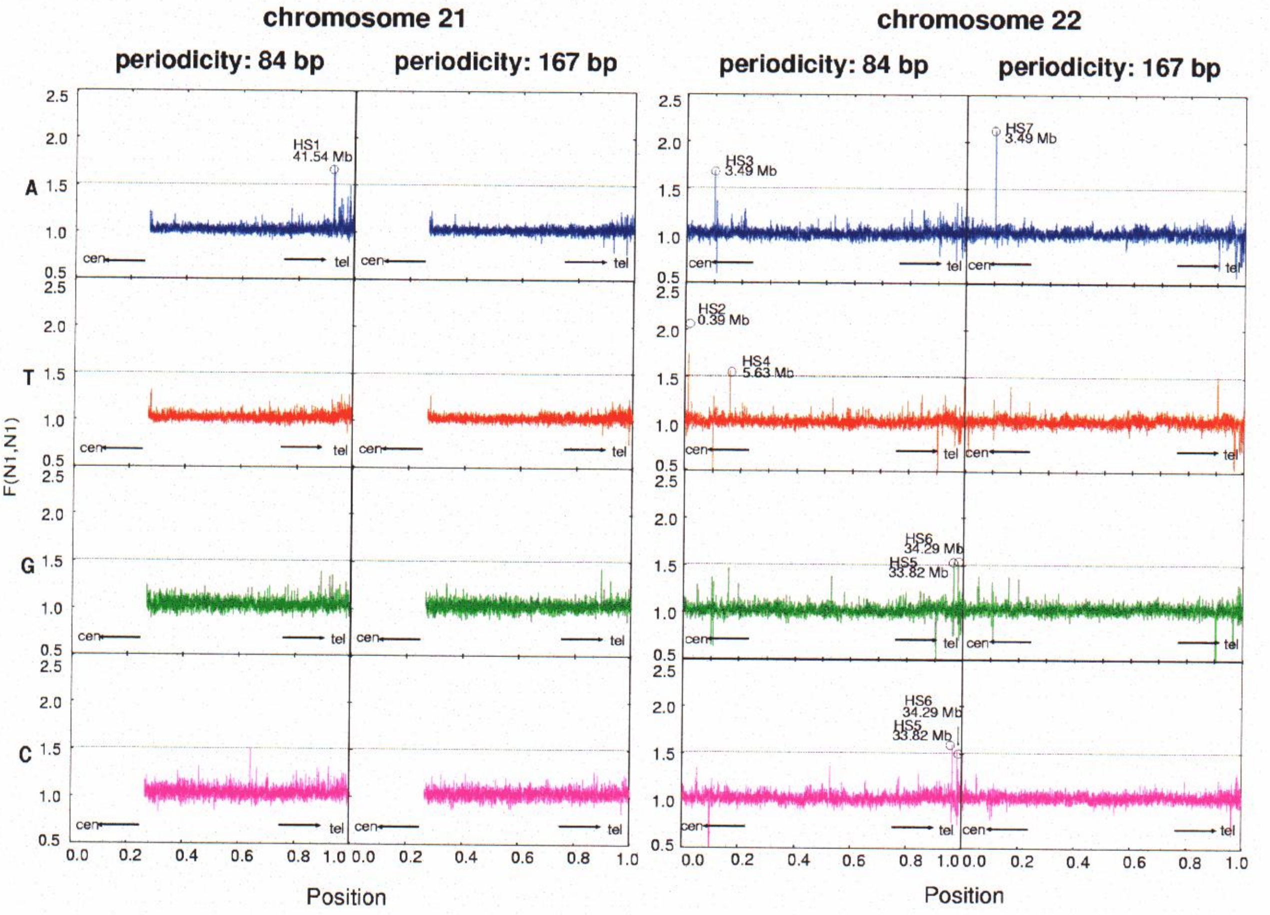

Figure 11. Periodic nucleotide distributions based on $F_k(N_1, N_1)$ values with 10-kb window on human chromosomes 21 and 22. Seven regions with $F_k(N_1, N_1)$ values higher than 1.5 are designated by ID numbers HS1 to HS7 (corresponding to Table 7). The total length of each chromosome is normalized to 1.0.

Table 7. Periodic elements in *H. sapiens*.

ID	Chr	Region (Mb)	Period (bp)	Consensus core represented with hexanucleotide composition	The number of consensus core(> 20 pairs)
HS1	21	41.54-41.55	84	GTGGTG	167
				TGGTGG	167
				GGTGGT	166
				TAGTGG	97
				TGGTGA	96
				GTGATG	94
				ATGGTG	92
				TGATGG	92
				GATGGT	91
HS2	22	0.39-0.40	84	TGGTGG	317
				GGTGGT	309
				GTGGTG	281
				TGATGG	88
				GATGGT	84
				ATGGTG	70
				GTGATG	74
HS3	22	3.49-3.50	84	TGGCTG	42
				GGCTGG	38
				CTGGCT	27
				GCTGGC	20
HS4	22	5.63-5.64	84	ATTTCA	34
				TTTCAT	31
				TCATTT	30
				CATTTC	27
HS5	22	33.82-33.83	84	AATGTG	27
HS6	22	34.29-34.30	84	AATGTG	22
HS7	22	3.49-3.50	167	TGGCTG	42
				GGCTGG	38
				CTGGCT	27
				GCTGGC	20

3.3. Spectrum landscape of low frequency in genomes

Relation between fractal and gene organization

Long-range correlation related with self-similarity is the focus of this section. Because large or even complete genomic sequences are available for many species, calculation of the spectra for several genomes is possible. It has been reported that the slope of the logarithm of power (log $S(f)$) to the logarithm of frequency (log(f)) is associated with fractal properties (Li, 1991; Li and Kaneko, 1992; Voss, 1992). The slope is generally referred to as the exponent. As described in Section 2.1 and summarized in Table 2, a flat power spectrum corresponding to the slope of 0 can be associated with random sequences. In the case of a slope close to –1 ($S(f) \propto f^{-1}$), the nucleotide sequence has the signature of fractal correlation, and a slope between 0 and –1 indicates long-range correlation (Li, 1991; Li, 1992).

The human genome has heterogeneous properties characterized by two distinct slopes that have been designated α for the region with larger than 10^5 bp periodicity (frequency < 10^{-5}) and β for the region with 10^4 to 10^5 bp periodicity (Figure 12a). Whereas GC composition is known to be homogeneous within genomes of most prokaryotes and unicellular eukaryotes, the genomes of higher vertebrates have mosaic GC% structure, referred to as "isochores" (Bernardi *et al.*, 1985; Ikemura, 1985; Ikemura and Aota, 1988; Bernardi, 1989), that appear to be related to replication timing (Holmquist, 1989; Bernardi, 2000; Watanabe *et al.*, 2002). This complexity should be reflected in the heterogeneous nature of the slopes observed in human chromosomes. The relations between GC% and the slopes (α and β of each human chromosome) are shown in Figure 12b. All chromosomes had similar α slopes, which were close to –1 regardless of the GC% (correlation coefficient = 0.18). It should be noted that the β slope was observed in the range of 10 to 100 kb and was clearly correlated with the GC% (correlation coefficient = –0.84) for each chromosome. The range from 10 to 100 kb is roughly the size of many genes. Furthermore, GC% in the human genome is known to be related to gene density. For example, human chromosomes 19 and 22 have high GC% (49% and 48%, respectively) and high gene density (23 and 17 genes/Mb, respectively). Conversely, chromosomes 4 and 13 have low GC% (both 38%) and low gene density (6 and 5 genes/Mb, respectively) (Lander *et al.*, 2001; Venter *et al.*, 2001). Chromosomes with a high GC% and high gene density tend to have a β slope closer to –1 (Figure 12b). Gene density and GC% may be important factors producing fractal structures in chromosomes. In addition, the highly variegated landscape of GC-poor and GC-rich isochores typical of these chromosomes (Pavlíček *et al.*, 2001; Oliver *et al.*, 2001) and gene organization (and presumably exon and intron organization) may also contribute to fractal structure. The cause of fractal property in genomes can be explained with the following model (Figure 13a). First, a genetic unit is duplicated, and a structure consisting of repeats of the unit is organized in the genome. The repetitive structure is again duplicated and organized in the genome. The size of the unit now roughly corresponds to the size of a gene. Thus, gene duplication may be one factor generating fractal property. The construction of the Koch curve, which proceeds in stages, is shown in Figure 13b. The initial object is a straight line (Figure 13b, on the top). This line is partitioned into three equal parts. Then an equilateral triangle is replaced by the middle third. In each stage the number of line segments reduces by a factor of 4. After some iteration, the fractal geometry with self-similarity is obtained (Figure 13b, on the bottom). The duplication process in the genome is very similar to the mathematical process of generating fractal properties. Further interpretation of fractals in genomes with a proposed mathematical model is discussed in Section 3.4.

Power spectra in low frequencies have been examined for five chromosomes of *A. thaliana* (Figure 14). The slopes of spectra for the four kinds of nucleotides have similar behaviors in this range, so the representative relation between gene number and the slope (exponent) for adenine is shown in Figure 15. The exponent shows a strong correlation with the number of genes on the respective chromosomes. Chromosomes with high gene numbers tend to have a slope closer to –1. These findings also suggest that there is a relation between fractal property and gene organization. These results are consistent with those for *H. sapiens* (Figure 12b).

Flat spectra in two insects (*D. melanogaster* and *A. gambiae*)

The power spectra of *D. melanogaster* and *A. gambiae* are shown in Figure 16 and 17. The spectra of A resemble those of T, and the spectra of G are similar to those of C. Interestingly, the G and C spectral curves had a flat region in the middle frequency range from $f = 2.0 \times 10^{-4}$ to 10^{-3} (corresponding to a period size of 1 kb–5 kb) in fly, whereas the mosquito genome had two flat regions. This type of correlation is called "partial power-law" (see Figure 2). In the high frequencies, the power spectrum is roughly flat, whereas the spectrum in low frequencies, the spectrum is expressed as the power-law decay (f^{β}) with exponent approximately equal to β (β < 0). The "1/*f*" noise (β = –1) of the given frequency range reveals the existence of a fractal structure (Mandelbrot, 1982; Li, 1991, 1992; Voss, 1992). A recent observation with DNA sequences revealed that the behavior of the power spectrum as a function of the frequency is visible three different regions on the logarithmic scale: a flat region, a power-law region, and another flat region from low to high frequency (Vieira, 1999). A flat power spectrum indicates randomness, that is, lack of correlation (Mandelbrot, 1982; Li, 1991, 1992; Voss, 1992). Flat power spectra in the middle frequencies have not been observed in eukaryotes such as *S. cerevisiae*, *C. elegans*, *A. thaliana*, and *H. sapiens* or prokaryotes (Figures 14, 16, and 17). Taking these findings into consideration, our result for the *D. melanogaster* genome is very important for understanding genome architecture (Fukushima *et al.*, 2002b). The exponents in eukaryotic and prokaryotic genomes are shown in Figure 18. In general, genomes with high gene density tend to have exponents close to –1. The exponents of prokaryotic genomes, except that of *Chlamydophila pneumoniae* (β = –0.44) are distributed around β = –1. Long-range correlations are observed in the genomes of many species. Thus, the base organization of genomes contains fractal properties. Exponents for the all genomic sequences analyzed are listed in the Appendix.

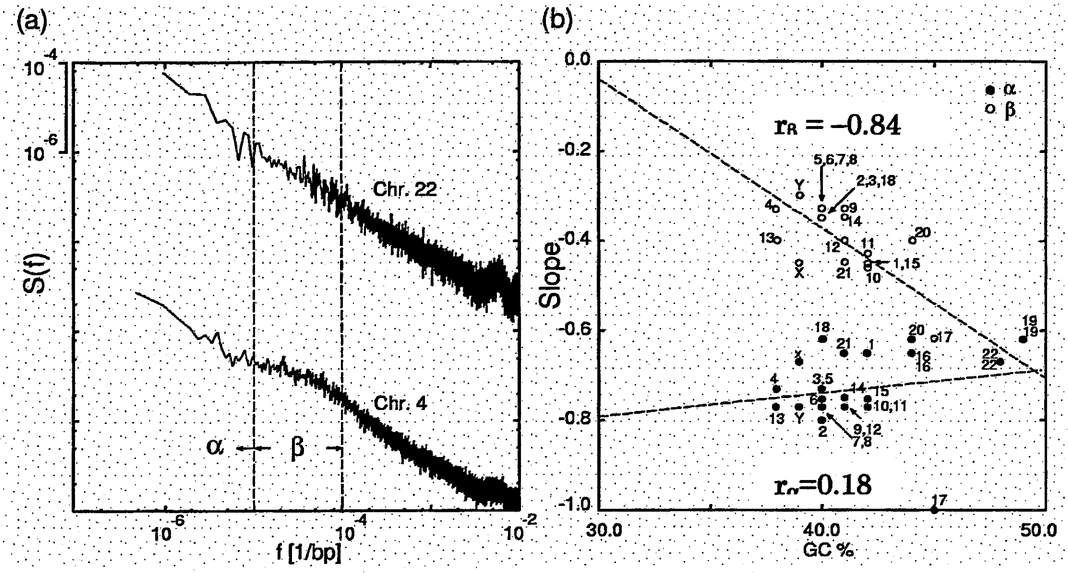

Figure 12. (a) Representative power spectra of human chromosomes and description of slopes α and β. Slope α is defined for the region with larger than 10^5 bp periodicity (frequency < 10^{-5}) and slope β for the region of 10^4 to 10^5 bp periodicity. Power spectra are drawn for human chromosomes 4 and 22. (b) The relation between genome GC% and slopes (α and β for all human chromosomes; human draft genomic sequences compiled from GenBank for individual chromosomes) were analyzed. All chromosomes had similar α slopes close to –1 that are independent of GC% (correlation coefficient = 0.18). In contrast, the β slope was highly correlated with the GC% of the respective chromosome (correlation coefficient = –0.84).

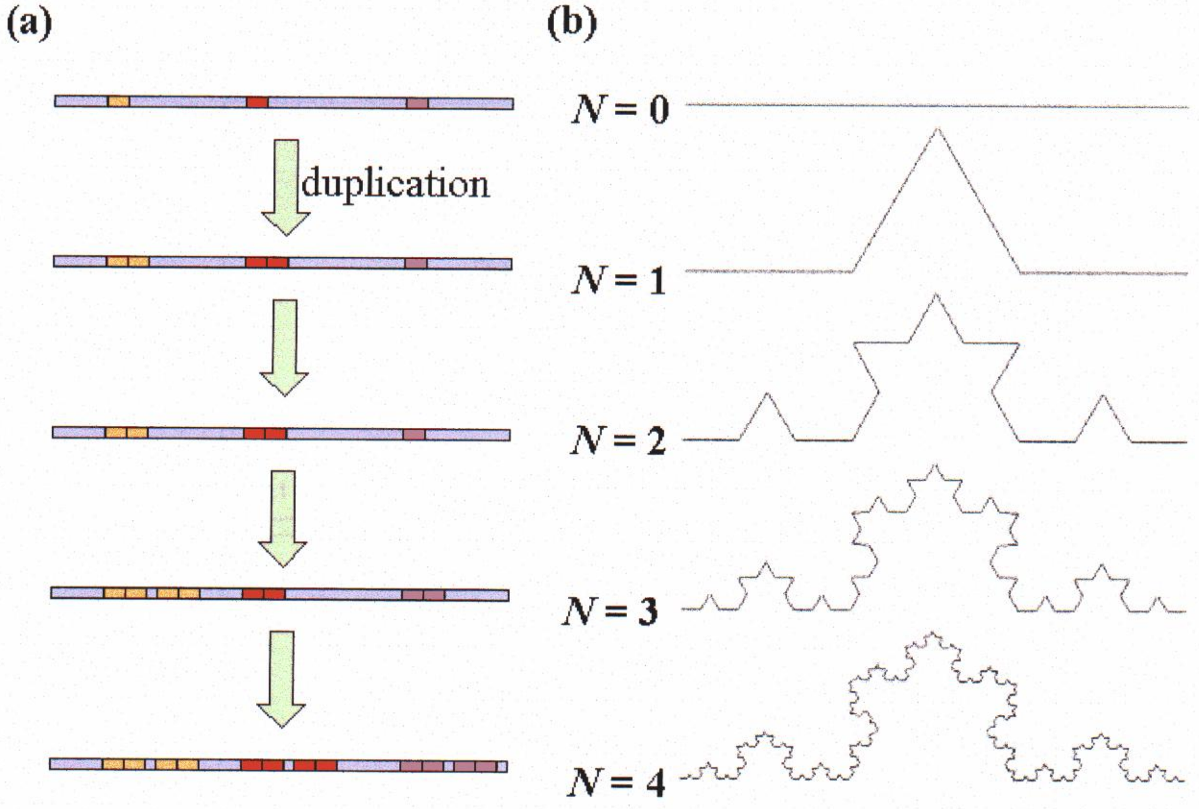

Figure 13. Relation between gene duplication and fractal property. (a) gene duplication and (b) Koch curve construction.

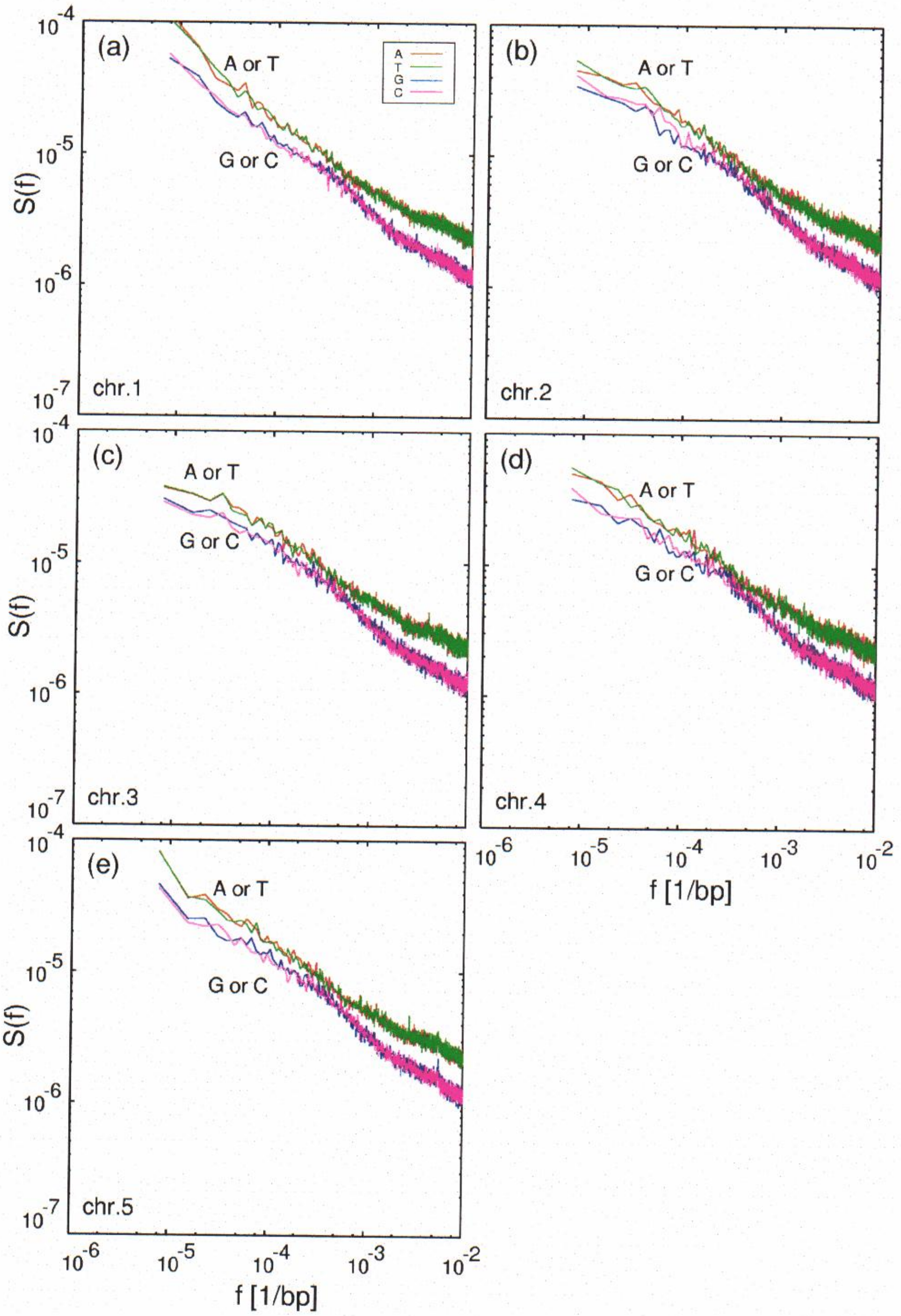

Figure 14. Power spectra of the *A. thaliana* genome in low frequency in log-log scale. (a) Chromosome 1, (b) chromosome 2, (c) chromosome 3, (d) chromosome 4, and (e) chromosome 5. For all chromosomes the slopes of spectra were very similar in this range.

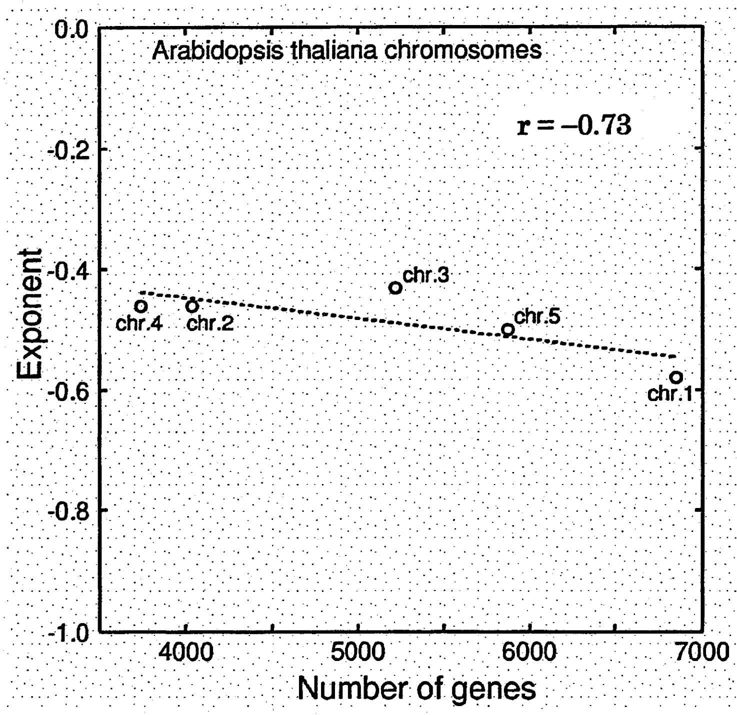

Figure 15. Relation between gene number and exponent [the slope of the logarithm of power (log *S*(*f*)) to the logarithm of frequency (log(*f*)] for A in *A. thaliana* chromosomes. The exponent is correlated with the gene number of the respective chromosome (correlation coefficient = –0.73). Slopes of other the nucleotide spectra are similar.

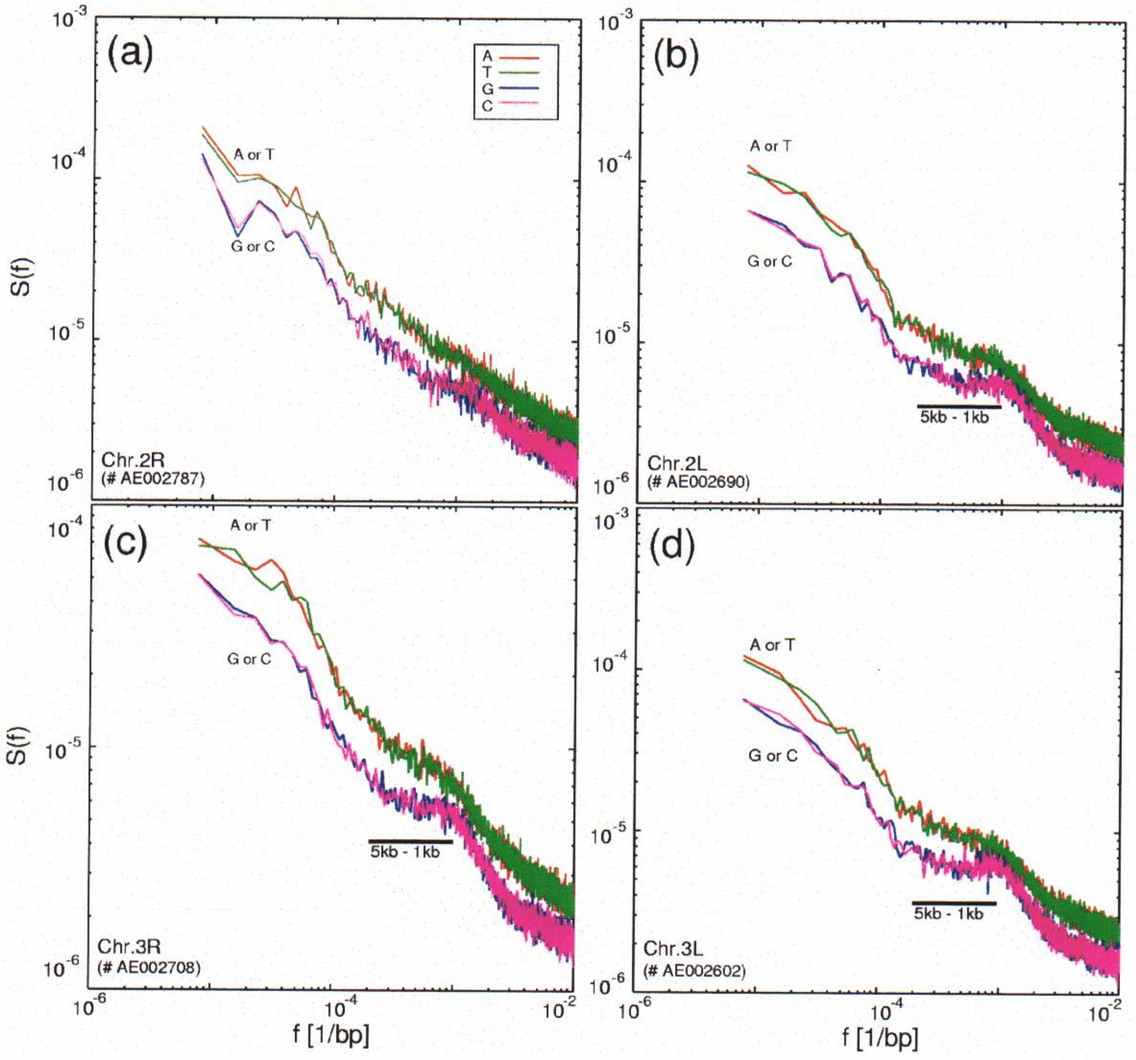

Figure 16. Power spectra of *D. melanogaster* genomic DNA sequences in low frequencies in log-log plot. (a) Chromosome 2R (accession #: AE002787), (b) chromosome 2L (AE002690), (c) chromosome 3R (AE002708), and (d) chromosome 3L (AE002602). Spectra of A resemble those of T, and the G curves are similar to those of C. Interestingly, the G and C spectral curves contained flat regions at range from *f* = 10^{-4} to 10^{-3} (corresponding to a period of 1 kb–5 kb).

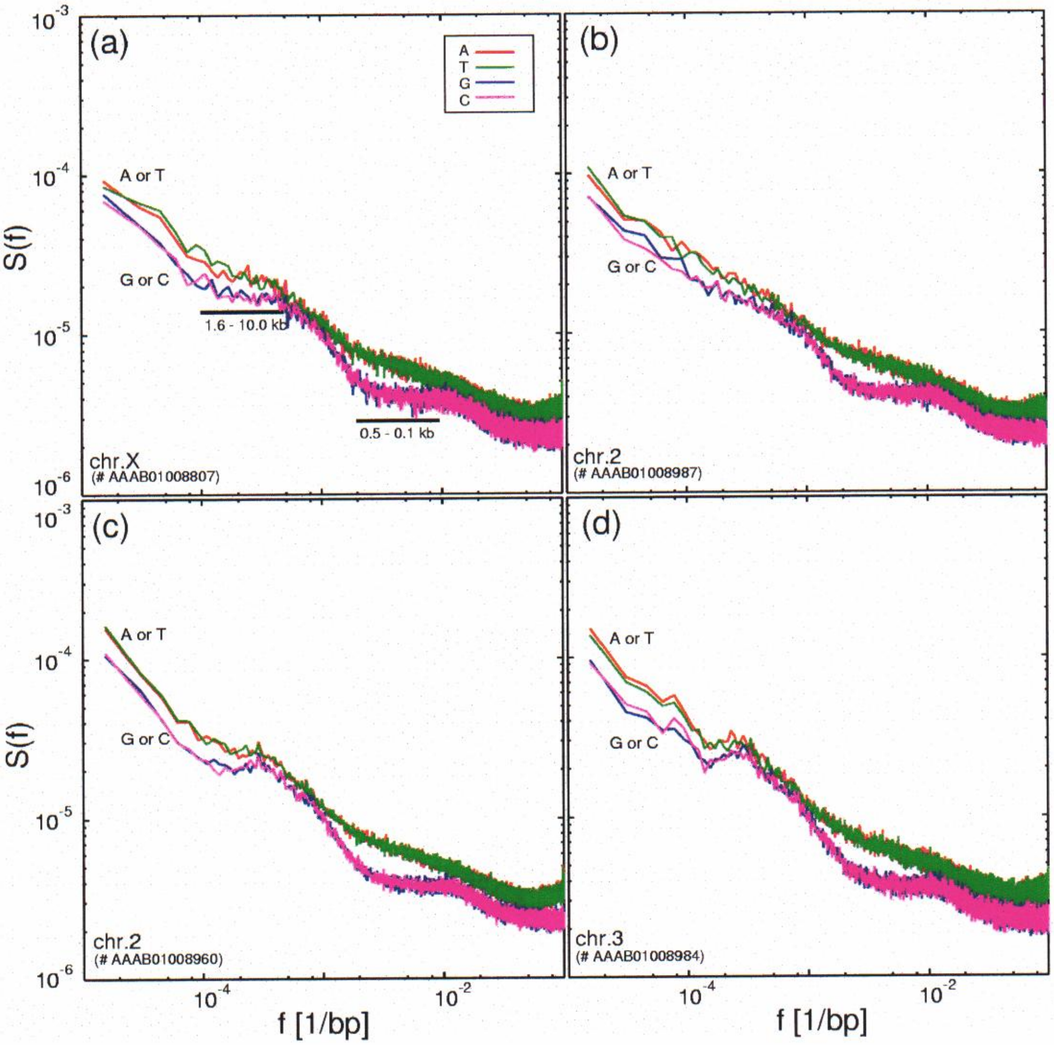

Figure 17. Power spectra of *A. gambiae* genomic DNA sequences in low frequencies in log-log scale. (a) Chromosome X (accession #: AAAB01008807), (b) chromosome 2 (AAAB01008987), (c) chromosome 2 (AAAB01008960), and (d) chromosome 3 (AAAB01008984). The spectra of A resemble those of T, and the G curves are similar to those of C. Interestingly, the G and C spectral curves contained two flat regions at ranges corresponding to periods of 0.1 kb to 0.5 kb and of 1.6 kb to 10.0 kb.

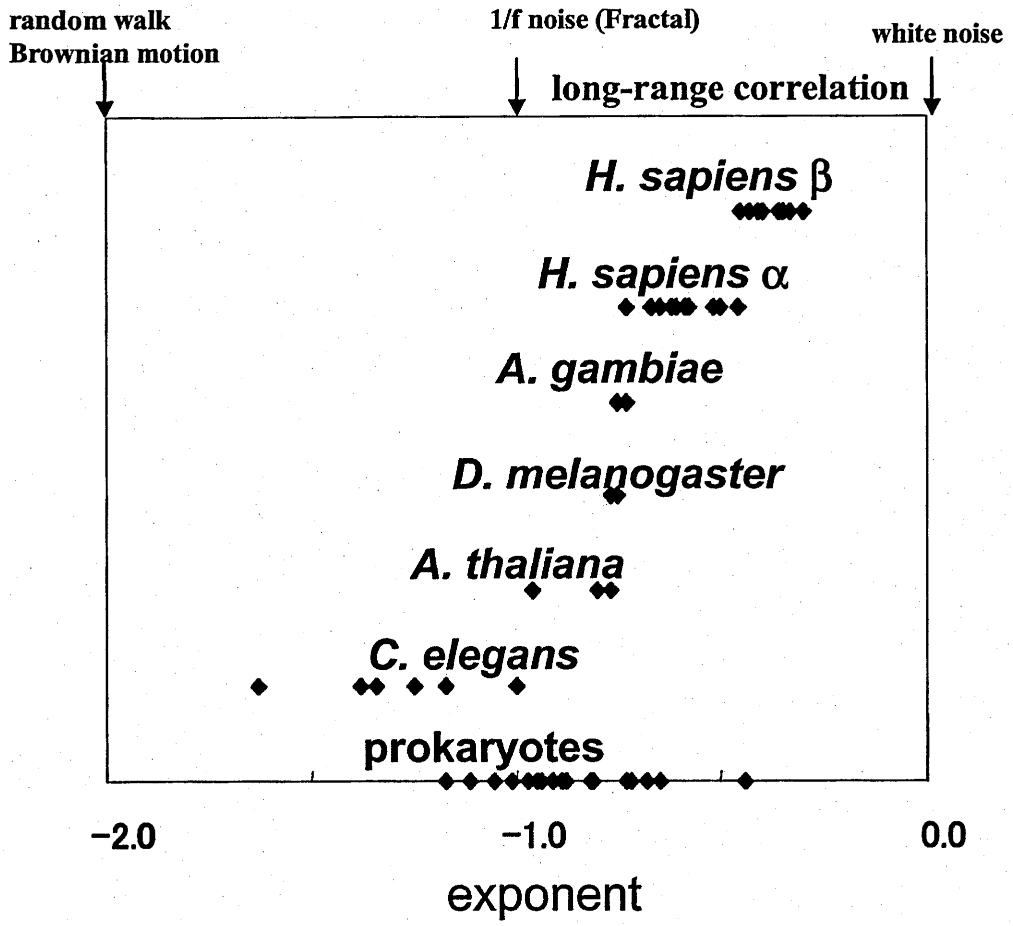

Figure 18. Comparison of the exponents in the wide range of species.

3.4. Possible model of genome evolution

Considering the implications of long-range correlations, system behavior is characterized with the exponent of the spectra. The fractal property of eukaryotic genomes which is represented by the exponent, is associated with gene organization (in Section 3.3). In particular, human chromosomes have heterogeneous power spectra characterized by two distinct slopes (Figure 12a). Moreover, in insect genomes (*D. melanogaster* and *A. gambiae*), the G and C spectral curves have flat regions in the middle frequency, which indicates random base sequence composition (Figures 16 and 17). Consequently, the fractal property represented by the exponent of the spectrum is specific to the species (or chromosomes) (Figure 18 and Appendix). I have attempted to develop a model to describe these properties.

Previously, Li (1991) proposed a model for $S(f) \propto f^{\beta}$ with $\beta = -1$ ($1/f$ spectra) called an 'Expansion-modification systems.' This model provides a convenient method for my purpose. 'Expansion-modification systems' is based on expanding a cellular automaton and has two types of manipulations, mainly modification or *mutation* and expansion or *duplication* (Li, 1991, 1992). Here, a cellular automaton is a mathematical and computational construct that consists of an array. A cell is defined as an element of this array. A cell has different states (dead = 0 and living = 1). The cellular automaton permits transition from one state to another at one step in a time frame. The simplest expansion-modification system in the two-symbol system in which the expansion process rewrites one symbol to two identical symbols, and the modification process switches one symbol to another symbol. The process of modification of 1 to 0 and of 0 to 1 occurs in probability p. The process of expansion of 1 to 11 and of 0 to 00 occurs in probability $1-p$. After a limited period of time passes, the statistical property of the binary sequence depends on the probability p. If the probability of modification process is close to 1, then $p \approx 1$, 'Expansion-modification systems' generates random sequences (Li, 1991, 1992). This means that most of the processes are random modifications in the calculation. In contrast, if the probability of the expansion process $(1-p)$ is close to 1, then $p \approx 0$, and the sequences generated display a $1/f$ spectrum, i.e. $S(f) \sim f^{-1}$. Thus, long-range correlations are generated by 'Expansion-modification systems' (Li, 1991, 1992).

Insect genomes may have evolved in a different manner from other species. Insect genomes show strange behaviors for long-range correlations. The cause of such species-specific fractal properties can be explained by a new model that consists of three processes: mutation, duplication, and sequence insertion. Transposable elements are a major source of genetic variation, including creation of novel genes, alteration of gene expression, and major genomic rearrangements. Indeed, the important role of host factors in the regulation of transposable elements has been revealed by recent studies of several systems in *D. melanogaster* (Lozovskaya *et al.*, 1995). Genome rearrangements, including sequence insertion or transposition, are important in genome evolution.

The extreme flat spectra observed in *D. melanogaster* and *A. gambiae* are not generated in the 'Expansion-modification system.' I have tried to generate flat spectra in middle frequencies by extending this system. In the model proposed in the present thesis, three factors, mutation (which occurs at probability p), duplication (probability q), and sequence insertion (probability q), are taken into consideration. Here, $p + q + r = 1$. This model can be considered an extension of that of 'Expansion-modification systems.' To examine the flat region of the spectrum in the middle frequencies, I designed a computational model.

The power spectra (in log-log scales) of the sequences generated by the computational experiment are shown in Figure 19. For the expansion-modification systems (probability $p = 0.1$), a spectrum with a slope is obtained (Figure 19a). In contrast, a sequence calculated with tentative parameters ($p = 0.100$ for mutation, $q = 0.882$ for duplication, and $r = 0.018$ for sequence insertion (of 100 bp) shows flat spectra similar to those of insects (Figure 19b). The results when two .operations (mutation and sequence insertion) are shown in Figure 19c. In this case, a flat spectrum was generated. Consequently, long-range correlations in genomes can be described by mutation and duplication, and sequence insertion, indicating that genome structure is determined mainly by these processes in particular insect genomes.

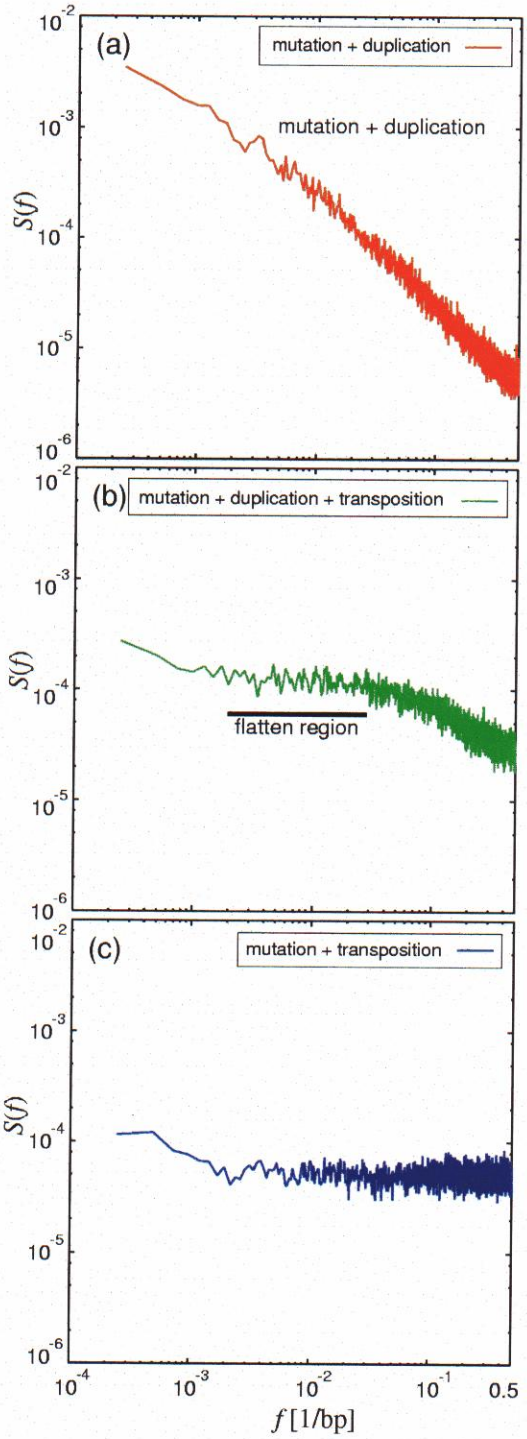

Figure 19. Power spectra for sequences generated by (a) mutation and duplication, (b) mutation, duplication, and sequence insertion, and (c) mutation and sequence insertion. The sequence length is longer than 2^{17} = 131072. Probability parameters are p = 0.100 for mutation, q = 0.882 for duplication, and r = 0.018 for sequence insertion (length 100 bp). The three manipulations yield a remarkably flat region in the spectrum.

4. Conclusion

Fractal concepts have provided a new approach for analyzing and interpreting species-specific periodicities in genomic sequences. In the present study, I found that such periodicities have biological and physiological significance. I also found that the gene organization has fractal properties. My results can be summarized as follows:

(1) Periodicities of 10 to 11 bp were observed in most prokaryotes and eukaryotes analyzed. The 11-bp periodicity was dominant in eubacteria (Figure 3a). The 10-bp periodicity was dominant in hyperthermophilic bacteria, archaebacteria, and all eukarya tested (Figures 3b and c). Because the 10-11 bp periodicities reflect the characteristic superhelical densities of genomic DNAs, the differences in periodicities between hyperthermophilic bacteria, archaebacteria, and eubacteria can be explained as a function of the archaeal histones, which are structurally similar to eukaryotic core histones.

(2) In *C. elegans*, a 68-bp periodicity on chromosome I, a 59-bp periodicity on chromosome II, and a 94-bp periodicity on chromosome III were detected (Figure 4). The sequences with 68-bp periodicities (CE4) contained a 12 bp core element sequence, CeRep45 (TTGGTTGAGGCT), that was previously characterized by Sanford *et al.* (2001) (Table 4). Though it is unclear if the periodic sequences observed are related to centromeric function in *C. elegans*, the strategy proposed in this thesis has the potential to detect hidden periodic sequences.

(3) In *A. thaliana*, three periodicities (248, 167, and 126 bp) on chromosome 3, three periodicities (174, 88, and 59 bp) on chromosome 4, and four periodicities (356, 174, 88, and 59 bp) on chromosome 5 were detected (Figure 6). Two periodicities (126 and 174 bp) are related to ORFs that encode Gly-rich amino acid sequences (Table 6). These periodicities include histone proteins, which consist of Gly-, Ser-, and Ala-rich amino acid sequences.

(4) In *H. sapiens,* 167 and 84 bp periodicities were detected along the entire length of chromosomes 21 and 22 (Figure 10). The 167-bp periodicity is identical to the length of DNA that forms two complete helical turns in nucleosome organization (Sinden, 1994). These periodicities were associated with NGG-repeat forming self-assembly DNA (Table 7).

(5) The fractal property of genomes represented by the long-range correlation in eukaryotic genomes is associated with gene organization. The human genome has heterogeneous properties in power spectra characterized by two distinct slopes (Figure 12).

(6) In the *D. melanogaster* genome, the G and C spectral curves contain a flat region in the middle frequency range from $f = 2.0 \times 10^{-4}$ to 10^{-3} (corresponding to a period of 1 kb–5 kb), which is associated with random base sequence composition (Figure 16). For *A. gambiae,* the G and C spectral curves have two flat regions at ranges corresponding to base periodicities of 0.1 kb to 0.5 kb and of 1.6 kb to 10.0 kb (Figure 17).

(7) On the basis of the observation that long-range correlations are present in DNA sequences, I developed a model to explain the evolutionary origin of genome. I used this model to explain

that the presence of spectra with flat regions in the middle frequency in insect genomes is due to mutation, duplication, and sequence insertion (Figure 19b).

Herein, I present a wide variety of periodicities for genomes of prokaryotes and eukaryotes. The periodicities were associated with biological and physiological properties such as nucleosomes, centromeres, and self-assembly of DNA. Long-range correlation was detected in all genomes analyzed. I have described the relation between the fractal property and gene organization and also proposed a model with a new interpretation of the long-range correlation in DNA sequence based on power spectrum analysis of genomes of higher eukaryote. This model is helpful for understanding genome evolution. Power spectrum is a useful tool for detecting hidden periodicities in a genome. The parameter F_k proposed in this thesis is applicable for detecting core elements based on periodicity. With progress of the genome projects, this analysis will play a vital role for understanding genome architecture.

Acknowledgements

I would like to express my sincere gratitude to Professor Hirotada Mori for his continuous guidance and support during all the stages of my work. Gratitude expressed to Associate Professor Shigehiko Kanaya for his supervision during my doctoral course. I also acknowledge Professor Toshimichi Ikemura at National Institute of Genetics for his kind advice and helping me with my research.

I thank Professor Yoshihiro Kudo, Dr. Makoto Kinouchi at Yamagata University and Dr. Taku Oshima for helpful discussions and valuable comments. I also acknowledge Dr. Wentian Li at North Shore LIJ Research Institute, Dr. Md. Arifuzzaman and Dr. Carlos A. Del Carpio at Toyohashi University of Technology for helpful comments. Throughout the stages of my research, I am grateful to all members of the Mori laboratory and the Kudo laboratory for their support and their kind help. I appreciate the support and encouragement of all my friends. Last but not least, I would like to thank my parents for their spiritual support to finish this program; and I dedicate them the present work.

References

Adams, M.D., Celniker, S.E., Holt, R.A. *et al.* (2000). The genome sequence of *Drosophila melanogaster*. Science 287, 2185–2195.

Alberts, B., Bray, D., Lewis, J. *et al.* (1983). Molecular biology of the cell. Garland, New York.

Amarger, V., Gauguier, D., Yerle, M., Apiou, F., Pinton, P., Giraudeau, F., Monfouilloux, S., Lathrop, M., Dutrillaux, B., Buard, J., and Vergnaud, G. (1998). Analysis of distribution in the human, pig, and rat genomes points toward a general subtelomeric origin of minisatellite structures. Genomics 52, 62-71.

Armour, J.A., Povey, S., Jeremiah, S., and Jeffreys, A.J. (1990). Systematic cloning of human minisatellites from ordered array charomid libraries. Genomics 8, 501-512.

Ashley, T. (1994). Mammalian meiotic recombination: a reexamination. Hum. Genet. 94, 587-593.

Bak, P.C., Tang C. and Wiesenfeld, K. (1988). Self-organized criticality, Phys. Rev. Ser. A38, 364-374.

Bak, P.C. and Chen, K. (1991). Self-organized criticality, Sci. Amer. 264, 26-33.

Baran, N., Lapidot, A., and Manor, H. (1991). Formation of DNA triplexes accounts for arrests of DNA synthesis at d(TC)n and d(GA)n tracts. Proc. Natl. Acad. Sci. USA 88, 507-511.

Beltran, R., Martinez-Balbas, A., Bernues, J. Bowater, R., and Azorin, F. (1993). Characterization of the zinc-induced structural transition to *H-DNA at a d(GA.CT)22 sequence. J. Mol. Biol. 230, 966-978.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., and Rodier, F. (1985). The mosaic genome of warm-blooded vertebrates. Science 228, 953-958.

Bernardi, G. (1989). The isochore organization of the human genome. Annu. Rev. Genet. 23, 637-661.

Bernardi, G. (2000). Isochores and the evolutionary genomics of vertebrates: organization of the human genome. Gene 241, 3–17.

Bissler, J.J., Cicardi, M., Donaldson, V.H., Gatenby, P.A., Rosen, F.S., Sheffer, A.L., and Davis, A.E. (1994). A cluster of mutations within a short triplet repeat in the C1 inhibitor gene. Proc. Natl. Acad. Sci. USA 91, 9622-9625.

Boan, F., Rodriguez, J.M., and Gomez-Marquez, J. (1998). A non-hypervariable human minisatellite strongly stimulates *in vitro* intramolecular homologous recombination. J. Mol. Biol. 278, 499-505.

Bois, P., and Jeffreys, A.J. (1999). Minisatellite instability and germline mutation. Cell Mol. Life Sci. 55, 1636-1648.

Brook, J.D., McCurrach, M.E., Harley, H.G., Buckler, A.J., Church, D. *et al.* (1992). Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member. Cell 68, 799-808.

Brusco, A., Saviozzi, S., Cinque, F., Bottaro, A., and DeMarchi, M. (1999). A recurrent breakpoint in the most common deletion of the Ig heavy chain locus (del A1-GP-G2-G4-E). J. Immunol. 163, 4392-4398.

Buldyrev, S.V., Goldberger, A.L., Havlin, S., Mantegna, R.N., Matsa, M.E., Peng, C.-K., Simons, M., and Stanley, H.E. (1995). Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. Phys. Rev. E 51, 5084-5091.

C. elegans Sequencing Consortium. (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282, 2012-2018.

Chaillet, J.R., Bader, D.S., and Leder, P. (1995). Regulation of genomic imprinting by gametic and embryonic processes. Genes Dev. 9, 1177-1187.

Charlesworth, B, Sniegowski, P., and Stephan, W. (1994). The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 371, 215-220.

Chen, F.M. (1997). Supramolecular self-assembly of d(TGG)4, synergistic effects of K+ and Mg2+. Biophys. J. 73, 348-56.

Comings, D.E., and Okada, T.A. (1972). Holocentric chromosomes in Oncopeltus: kinetochore plates are present in mitosis but absent in meiosis. Chromosoma 37, 177-92.

Cooley, J.W., and Tukey, J.W. (1965). An algorithm for machine computation of complex Fourier series. Math. Computation 19, 297-301.

Csink , A.K., and Henikoff, S. (1998). Something from nothing: the evolution and utility of satellite repeats. Trends Genet. 14, 200-204.

Dombroski, B.A., Mathias, S.L., Nanthakumar, E., Scott, A.F., and Kazazian, H.H., Jr. (1991). Isolation of an active human transposable element. Science 254, 1805-1808.

Dunham, I., Hunt, A.R., Collins, J.E., Bruskiewich, R. *et al.* (1999). The DNA sequence of human chromosome 22. Nature 402, 489-495.

European Union Chromosome 3 *Arabidopsis* Sequencing Consortium, The Institute for Genomic Research & Kazusa DNA Research Institute. (2000). Sequencing and analysis of chromosome 3 of the plant *Arabidopsis thaliana.* Nature 408, 820-822.

European Union *Arabidopsis* Genome Sequencing Consortium & The Cold Spring Harbor, Washington University in St Louis and PE Biosystems *Arabidopsis*

Sequencing Consortium. (1999). Sequencing and analysis of chromosome 4 of the plant *Arabidopsis thaliana.* Nature 402, 769-777.

Fickett, J. W. (1982). Recognition of protein coding regions in DNA sequences. Nucleic Acids Res. 10, 5303-5318.

Fukushima, A., Ikemura, T., Kinouchi, M., Oshima, T., Kudo, Y., Mori, H., and Kanaya, S. (2002a). Periodicity in prokaryotic and eukaryotic genomes identified by power spectrum analysis, Gene 300, 203-211.

Fukushima, A., Ikemura, T., Oshima, T., Mori, H., and Kanaya, S. (2002b). Detection of periodicity in eukaryotic genomes on the basis of power spectrum analysis. Genome Informatics Series No. 13, 21-29.

Gutierrez, G, Oliver, J. L., and Marin, A. (1994). On the origin of the periodicity of three in protein coding DNA sequences. J. Theor. Biol. 167, 413-414.

Hamada, H., Petrino, M.G., Kakunaga, T., Seidmann, M., and Stollar, B.D. (1984). Characterization of genomic poly(dT-dG).poly(dC-dA) sequences: structure, organization, and conformation.. Mol. Cell. Biol. 4, 2610-2621.

Hatch, F.T., and Mazrimas, J.A. (1974). Fractionation and characterization of satellite DNAs of the kangaroo rat (*Dipodomys ordii*). Nucleic Acids Res. 1, 559-575.

Hansen, J.C. (2002).Conformational dynamics of the chromatin fiber in solution: determinants, mechanisms, and functions. Annu. Rev. Biophys. Biomol. Struct. 31, 361-92.

Hanvey, J.C, Shimizu, M., and Wells, R.D. (1988). Intramolecular DNA triplexes in supercoiled plasmids. Proc. Natl. Acad. Sci. USA 85, 6292-6296.

Hattori, M., Fujiyama, A., Taylor, T.D., *et al.* (2000). The DNA sequence of human chromosome 21. Nature 405, 311-319.

Herzel, H., Trifonov, E.N., Weiss, O., and Gloße, I. (1998). Interpreting correlations in biosequences. Physica A 249, 449-459.

Herzel, H., Weiss, O., and Trifonov, E.N. (1999). 10-11 bp periodicities in complete genomes reflect protein structure and DNA folding. Bioinformatics (CABIOS) 15, 187-193.

Holmquist, G.P. (1989). Evolution of chromosome bands: molecular ecology of noncoding DNA. J. Mol. Evol. 28, 469-486.

Holt, R.A., Subramanian, G.M., Halpern, A. *et al.* (2002). The genome sequence of the malaria mosquito *Anopheles gambiae.* Science 298, 129–149.

Horn, P.J., and Peterson, C.L. (2002). Chromatin higher order folding--wrapping up transcription. Science 297, 1824-1827.

Huntington Disease Collaborative Rersearch Group. (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell 72, 971-983.

Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. 2, 13-34.

Ikemura, T., and Aota, S. (1988). Global variation in G+C content along vertebrate genome DNA: possible correlation with chromosome band structures. J. Mol. Biol. 203, 1-13.

Jeffreys, A.J., MacLeod, A., Tamaki, K., Neil, D.L., and Monckton, D.G. (1991). Minisatellite repeat coding as a digital approach to DNA typing. Nature, 354, 204-209.

Jeong, H., Mason, S.P., Barabási, A.L., and Oltvai, Z.N. (2001). Lethality and centrality in protein networks. Nature 411, 41-42.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., and Barabási, A.L. (2000). The large-scale organization of metabolic networks. Nature 407, 651-654.

Kajikawa, M., and Okada, N. (2002). LINEs mobilize SINEs in the eel through a shared 3' sequence. Cell 111, 433-444.

Katti, M.V., Ranjeker, P.K., and Gupta, V.S. (2001). Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol. Biol. Evol. 18, 1161-1167.

Kawaguchi, Y., Okamoto, T., Taniwaki, M., Aizawa, M., Inoue, M., *et al.* (1994). CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. Nature Genet. 8, 221-228.

Kazusa DNA Research Institute, The Cold Spring Harbor and Washington University in St Louis Sequencing Consortium & The European Union Arabidopsis Genome Sequencing Consortium. (2000). Sequencing and analysis of chromosome 5 of the plant *Arabidopsis thaliana.* Nature 408, 823-826.

Kennedy, G.C., German, and M.S. Rutter, W.J. (1995). The minisatellite in the diabetes susceptibility locus IDDM2 regulates insulin transcription. Nature Genet. 9, 293-298.

Koide, R., Ikeuchi, T., Onodera, O., Tanaka, H., Igarashi, S., Endo, K. *et al.* (1994). Unstable expansion of CAG repeat in hereditary dentatorubral-pallidoluysian atrophy (DRPLA). Nature Genet. 6, 9-13.

Kremer, E.J., Pritchard, M., Lynch, M., Yu, S., Holman, K., Baker, E., Warren, S.T., Schlessinger, D., Sutherland, and G.R., Richards, R.I. (1991). Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CCG)n. Science 252, 1711-1714.

Lander, E.S., Linton, L.M., Birren, B. *et al.*, 2001. Initial sequencing and analysis of the human genome. Nature 409, 860-921.

Lapidot, A., Baran, N., and Manor, H. (1989). (dT-dC)n and (dG-dA)n tracts arrest single stranded DNA replication *in vitro*. Nucleic Acids Res. 17, 883-900.

La Spada, A.R., Wilson, E.M., Lubahn, D.B., Harding, A.E., and Fischbeck, K.H. (1991). Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. Nature 352, 77-9.

Li, W. (1991). Expansion-modification systems: a model for spatial 1/f spectra. Phys. Rev. A 43, 5240-5260.

Li, W., and Kaneko, K. (1992). Long-range correlation and partial $1/f^{\alpha}$ spectrum in a noncoding DNA sequence. Europhys. Lett. 17, 655-660.

Li, W., Marr, T.G., and Kaneko, K. (1994). Understanding long-range correlations in DNA sequences. Physica D 75, 392-416.

Li, W. (1992). Generating nontrivial long-range correlations and 1/f spectra by replication and mutation. Int. J. Bifurcation and Chaos. 2, 137-154.

Li, W., Stolovitzky, G., Bernaola-Galvan, P., and Oliver, J.L. (1998). Compositonal heterogeneity within, and uniformity between, DNA sequences of yeast chromosomes. Gen. Res. 8, 916-918.

Lin, X., Kaul, S., Rounsley, S. *et al.* (1999). Sequencing and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. Nature 402, 761-768.

Lozovskaya, E.R., Hartl, D.L., and Petrov, D.A. (1995). Genomic regulation of transposable elements in *Drosophila*. Curr. Opin. Genet. Dev. 5, 768-73.

Luscombe, N., Qian, J., Zhang, Z., Johnson, T., and Gerstein, M. (2002). The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. Genome Biol. 3, research0040.1-0040.7.

Mandelbrot, B.B. (1982). The Fractal Geometry of Nature. Freeman, W.H. and Co., New York.

Mitas, M. (1997). Trinucleotide repeats associated with human disease. Nucl. Acids Res. 25, 2245-2254.

Naas, T.P., DeBerardinis, R.J., Moran, J.V., Ostertag, E.M., Kingsmore, S.F., Seldin, M.F., Hayashizaki, Y., Martin, S.L., and Kazazian, H.H. (1998). An actively retrotransposing, novel subfamily of mouse L1 elements. EMBO J., 17, 590-597.

Nakamura, Y., Leppert, M., O'Connell, P., Wolff, R., Holm, T., Culver, M., Martin, C., Fujimoto, E., Hoff, M., Kumlin, E., *et al.* (1987). Variable number of tandem repeat (VNTR) markers for human gene mapping. Science 235, 1616-1622.

Neumann, B. Kubicka, P., and Barlow, D.P. (1995) Characteristics of imprinted genes. Nature Genet. 9, 12-13.

Oliver, J., Bernaola-Galva´n, P., Carpena, P., and Roma´n-Rolda´n, R. (2001). Isochore chromosome map of eukaryotic genomes. Gene 276, 47-56.

Orr, H.T., Chung, M.Y., Banfi, S., Kwiatkowski, T.J., Jr., Servadio, A., Beaudet, A.L., McCall, A.E., Duvick, L.A., Ranum, L.P., and Zoghbi, H.Y. (1993). Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. Nature Genet. 4, 221-226.

Pavlíček, A., Jabbari, K., Pačes, J., Pačes, V., Hejnar, J., and Bernardi, G. (2001). Similar integration but different stability of Alus and LINEs in the human genome. Gene 276, 39-45.

Pearson, C.E., and Sinden, R.R. (1996). Alternative structures in duplex DNA formed within the trinucleotide repeats of the myotonic dystrophy and fragile X loci. Biochem. 35, 5041-5053.

Pearson, C.E., Ewel, A., Acharya, S., Fishel, R.A., and Sinden, R.R. (1997). Human

MSH2 binds to trinucleotide repeat DNA structures associated with neurodegenerative diseases. Hum. Mol. Genet. 6, 1117-1123.

Peng, C.-K., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Sciortino, F., Simons, M., and Stanley, H.E. (1992). Long-range correlations in nucleotide sequences. Nature 356, 168-170.

Peng, C.-K., Buldyrev, S.V., Havlin, S., Simons, M., Stanley, H.E., and Goldberger, A.L. (1994). Mosaic organization of DNA nucleotides. Phys. Rev. E 49, 1685-1689.

Rao, B.S., Manor, H., and Martin R.G. (1988). Pausing in simian virus 40 DNA replication by a sequence containing (dG-dA)27.(dT-dC)27. Nucleic Acids Res. 16, 8077-8094.

Rao, B.S. (1996). Regulation of DNA replication by homopurine/homopyrimidine sequences. Mol. Cell. Biochem. 156, 163-168.

Richards, R.I., Holman, K., Yu, S., and Sutherland, G.R. (1993). Fragile X syndrome unstable element, p(CCG)n, and other simple tandem repeat sequences are binding sites for specific nuclear proteins. Hum. Mol. Genet. 2, 1429-1435.

Sandman, K., and Reeve, J.N. (2000). Structure and functional relationships of archaeal and eukaryal histones and nucleosomes. Arch. Microbiol. 173, 165-169.

Sanford, C., and Perry, M.D. (2001). Asymmetrically distributed oligonucleotide repeats in the *Caenorhabditis elegans* genome sequence that map to regions important for meiotic chromosome segregation. Nucleic Acids Res. 29, 2920-2926.

Shepherd, J. C. (1981a). Periodic correlations in DNA sequences and evidence suggesting their evolutionary origin in a comma-less genetic code. J. Mol. Evol. 17, 94-102.

Shepherd, J. C. W. (1981b). Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. Proc. Natl. Acad. Sci. USA. 78, 1596-1600.

Schueler, M.G., Higgins, A.W., Rudd, M.K., Gustashaw, K., and Willard, H.F. (2001). Genomic and genetic definition of a functional human centromere. Science 294, 109-15.

Sinden, R.R. (1994). DNA structure and function. Academic Press, Inc., San Diego.

Staden, R. (1990). Finding protein coding regions in genomic sequences. Methods Enzymol. 183, 163-80.

Stanley, H.E. (1971). Introduction to phase transitions and critical phenomena. Oxford University Press, London.

Sutherland, G.R., and Richards, R.I. (1995). Simple tandem DNA repeats and human genetic disease. Proc. Natl. Acad. Sci. USA 92, 3636-3641.

Sutherland G.R., Baker, E., and Richards, R.I. (1998). Fragile sites still breaking. Trends Genet. 14, 501-506.

Sybenga, J. (1999). What makes homologous chromosomes find each other in meiosis? A review and a hypothesis. Chromosoma 108, 209-219.

Theologis, A., Ecker, J.R., Palm, C.J. *et al.* (2000). Sequencing and analysis of chromosome 1 of the plant *Arabidopsis thaliana*. Nature 408, 816-819.

Tomita, M., Wada, M., and Kawashima, Y. (1999). ApA Dinucleotide Periodicity in prokaryote, eukaryote, and organelle genomes. J. Mol. Evol. 49, 182-192.

Trifonov, E.N. (1998). 3-, 10.5-, 200- and 400-base periodicities in genome sequences. Physica A 249, 511-516.

Trifonov, E.V., and Sussman, J.L. (1980). The pitch of chromatin DNA is reflected in its nucleotide sequence. Proc. Natl. Acad. Sci. USA 77, 3816-3820.

Tsonis, A. A., Elsner, J. B., and Tsonis, P. A. (1991). Periodicity in DNA coding sequences: implications in gene evolution. J. Theor. Biol. 151, 323-331.

Turri, M.G., Cuin, K.A., and Porter, A.C. (1995). Characterisation of a novel minisatellite that provides multiple splice donor sites in an interferon-induced transcript. Nucleic Acids Res. 23, 1854-1861.

Usdin, K. (1998). NGG-triplet repeats form similar intrastrand structures: implications for the triplet expantion diseases. Nucleic Acids Res. 26, 4078-4085.

Venter, J.C., Adams, M.D., Myers, E.W. *et al.* (2001). The sequence of the human genome. Science 291,1304-51.

Vergnaud, G., and Denoeud, F. (2000). Minisatellites: Mutability and Genome Architecture. Genome Res. 10, 899-907.

Verkerk, A.J., Pieretti, M., Sutcliffe, J.S., Fu, Y.H., Kuhl, D.P., Pizzuti, A., Reiner, O., Richards, S., Victoria, M.F., Zhang, F.P., *et al.* (1991). Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell 65, 905-914.

Vieira, M. (1999). Statistics of DNA sequences: a low-frequency analysis. Phys. Rev. E 60, 5932-5937.

Vologodsky, A. (1992). Topology and Physics of Circular DNA. CRC Press, Boca Raton.

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

Wahls, W.P., and Moore, P.D. (1998). Recombination hotspot activity of hypervariable minisatellite DNA requires minisatellite DNA binding proteins. Somat. Cell. Mol. Genet. 24, 41-51.

Wang, Y.H., Amirhaeri, S., Kang, S., Wells, R.D., and Griffith, J. D. (1994). Preferential nucleosome assembly at DNA triplet repeats from the myotonic dystrophy gene. Science 265, 669-671.

Watanabe, Y., Fujiyama, A., Ichiba, Y., Hattori, M., Yada, T., Sakaki, Y., and Ikemura, T. (2002). Chromosome-wide assessment of replication timing for human chromosomes 11q and 21q: disease-related genes in timing-switch regions. Hum. Mol. Genet. 11, 3-21.

Weber, J.L. (1990). Informativeness of human (dC-dA)n.(dG-dT)n polymorphisms. Genomics 7, 524-30.

Wells, R.D., Collier, D.A., Hanvey, J.C., Shimizu, M., and Wohlrab, F. (1988). The chemistry and biology of unusual DNA structures adopted by oligopurine.oligopyrimidine sequences. Fed. Am. Soc. Expt. Biol. J. 2, 2939-2949.

Yu, S., Mulley, J., Loesch, D., Turner, G, Donnelly, A., *et al.* (1992). Fragile-X syndrome: unique genetics of the heritable unstable element. Am. J. Hum. Genet. 50, 968-980.

Appendix

Spectral exponents for genomes

(a) prokaryotes

species	exponent β[#]
Escherichia coli	−0.73
Bacillus subtilis	−0.68
Neisseria meningitidis	−0.88
Helicobacter pylori	−0.73
Rickettsia prowazekii	−0.97
Chlamydophila pneumoniae	−0.44
Mycobacterium tuberculosis	−0.65
Borrelia burgdorferi	−0.95
Synechocystis sp.	−0.68
Deinococcus radiodurans	−0.77 (chr. 1)
	−0.81 (chr. 2)
Aquifex aeolicus	−1.11
Thermotoga maritima	−0.91
Thermoplasma acidophilum	−0.91
Archaeoglobus fulgidus	−0.82
Methanococcus jannaschii	−1.05
Methanobacterium thermoautotrophicum	−0.89
Pyrococcus horikoshii	−0.94
Aeropyrum pernix	−0.72
Halobacterium sp.	−1.01

[#] The exponent β of spectrum is fitted in *f* ranged from 2.4×10^{-5} to 10^{-4}.

(a) *C. elegans*

species	chr.	base	exponent β[#]
C. elegans	I	A	−1.38
		T	−1.41
		G	−1.23
		C	−1.25
	II	A	−0.96
		T	−0.97
		G	−0.73
		C	−0.73
	III	A	−1.63
		T	−1.60
		G	−1.37
		C	−1.38
	IV	A	−1.34
		T	−1.38
		G	−1.24
		C	−1.23
	V	A	−1.25
		T	−1.24
		G	−0.95
		C	−0.96
	X	A	−1.17
		T	−1.14
		G	−0.86
		C	−0.84

[#] The exponent β of spectrum is fitted in *f* ranged from 1.5×10^{-5} to 10^{-1}.

(b) ***A. thaliana***

species	chr.	base	exponent $\beta^{\#}$
A. thaliana	1	A	−0.58
		T	−0.57
		G	−0.55
		C	−0.55
	2	A	−0.46
		T	−0.47
		G	−0.50
		C	−0.52
	3	A	−0.43
		T	−0.43
		G	−0.49
		C	−0.48
	4	A	−0.46
		T	−0.47
		G	−0.49
		C	−0.50
	5	A	−0.50
		T	−0.50
		G	−0.52
		C	−0.51

The exponent β of spectrum is fitted in f ranged from 1.5×10^{-5} to 10^{-2}.

(c) *D. melanogaster*

species	chr.	base	exponent β[#]
D. melanogaster	2L	A	−0.61
		T	−0.60
		G	−0.56
		C	−0.56
	2R	A	−0.58
		T	−0.58
		G	−0.55
		C	−0.54
	3L	A	−0.59
		T	−0.58
		G	−0.54
		C	−0.54
	3R	A	−0.52
		T	−0.52
		G	−0.51
		C	−0.51
	X*	A	−0.61
		T	−0.62
		G	−0.60
		C	−0.62
	X**	A	−0.58
		T	−0.57
		G	−0.57
		C	−0.56

[#] The exponent β of spectrum is fitted in *f* ranged from 7.6×10^{-6} to 10^{-2}.

* ACCESSION number AE002593

** ACCESSION number AE002566

(d) ***A. gambiae***

species	chr.	base	exponent β[#]
A. gambiae	X	A	−0.36
		T	−0.36
		G	−0.36
		C	−0.35
	2*	A	−0.36
		T	−0.36
		G	−0.35
		C	−0.34
	2**	A	−0.40
		T	−0.41
		G	−0.41
		C	−0.41
	3***	A	−0.41
		T	−0.40
		G	−0.41
		C	−0.41

The exponents β of spectrum is fitted in *f* ranged from 1.5 × 10^{-5} to 10^{-2}.

* ACCESSION number AAAB01008987

** ACCESSION number AAAB01008960

*** ACCESSION number AAAB01008984

(e) *H. sapiens*

species	chr.	exponent α[#]	exponent β[##]
H. sapiens	1	−0.65	−0.45
	2	−0.80	−0.35
	3	−0.73	−0.35
	4	−0.73	−0.33
	5	−0.73	−0.33
	6	−0.75	−0.33
	7	−0.77	−0.33
	8	−0.77	−0.33
	9	−0.77	−0.33
	10	−0.77	−0.46
	11	−0.77	−0.43
	12	−0.77	−0.40
	13	−0.77	−0.40
	14	−0.75	−0.35
	15	−0.75	−0.45
	16	−0.65	−0.65
	17	−1.00	−0.62
	18	−0.62	−0.35
	19	−0.62	−0.62
	20	−0.62	−0.40
	21	−0.65	−0.45
	22	−0.67	−0.67
	X	−0.67	−0.45
	Y	−0.77	−0.30

[#] The exponent α of spectrum is fitted in f ranged smaller than 10^{-5}.

[##] The exponent β of spectrum is fitted in f ranged from 10^{-5} to 10^{-4}.

[*] all data shows two exponents for adenine.

[**] correlation coefficient 0.18 for exponent α, while −0.84 for exponent β.