Evolution of the Periodicity and the Self-Similarity in DNA Sequence: A Fourier Transform Analysis

Naoki NAGAI, Kazuo KUWATA, Tomoya HAYASHI, Hiromi KUWATA*, and Seiichi ERA

Department of Physiology, Gifu University School of Medicine, Gifu, 500–8705 Japan; and

*College of Medical Sciences, Gifu University, Gifu, 501–1193 Japan

Abstract: Fourier transform analysis was applied to elucidate the periodical and self-similar properties in the DNA sequences mainly of β -globin genes in different species, and the evolutionary change in those properties was then investigated. Map patterns of a two-dimensional DNA walk showed that the stretches of exons are significantly shorter than those of introns, suggesting that the evolution of exons is driven by natural selection, whereas that of introns is generated by unknown internal rules. Using a monomer analysis, we obtained the power spectra of four different bases, A, G, C, and T, in DNA sequences. Periodicities in the short- (2 to 10 base pairs [bp]), medium- (10 to 50 bp) and long-range order (50 to 300 bp) of β -globin gene sequences could be observed, and power spectral densities of these periodicities were increased with evolution. These results suggest the exis-

tence of the internal rules in the occurrence of the synonymous and nonsynonymous substitutions in the sequences, the destabilization of the interaction between DNA and histone protein, and the stabilization of the chromatin structure, respectively. Moreover, $1/f^\alpha$ analysis of the power spectra (log-log plot) in the far long-range region (160 to 16,000 bp) suggested the increase in the self-similarity (the fractal structure) of DNA sequences with evolution. A general trend of the increase in a 3 bp periodicity with evolution might be functionally related to the CAG trinucleotide repeat diseases such as Huntington chorea, where a marked periodicity of 3 bp could be observed. Fourier transform analysis applied to a DNA sequence offers a great new avenue for extracting information on the evolution of a DNA sequence. [Japanese Journal of Physiology, 51, 159–168, 2001]

Key words: evolution, Fourier transform analysis, globin gene, periodicity, self-similarity.

Elucidation of the statistical properties of the DNA nucleotide sequence is of growing interest [1–11] as an approach to discovering the rules governing the arrangement of genetic information [1, 3, 8]. As denoted by the recent sequencing of human chromosomes 21 [12] and 22 [13], the sequencing of the entire human genome should be completed in the immediate future*. We may not obtain much insight into the whole DNA sequence, however, since the protein or RNA coding region (called the “exon”) is expected to be merely 2 to 3% of the total sequence [14]. There is certainly a need for a theoretical means of extracting information from the DNA sequence itself without experimental complication. In fact, we cannot rule out

the possibility that there might be additional genetic information in the noncoding region (called the “intron”), which accounts for 97 to 98% of the whole DNA sequence [14]. Although the rules hidden in DNA sequences have not yet been clarified, it may be worth characterizing them in terms of the power spectral density and the self-similarity (the fractal structure) of the DNA sequence [11].

A DNA chain is represented by a sequence of four letters, corresponding to four different bases: adenine (A), guanine (G), cytosine (C), and thymine (T). We first extracted the characteristic pattern of exons and introns in various β -globin DNA sequences with different species, using a novel genetic mapping method

Received on August 18, 2000; accepted on December 26, 2000

Correspondence should be addressed to: Kazuo Kuwata, Department of Physiology, Gifu University School of Medicine, 40 Tsukasamachi, Gifu, 500–8705 Japan. Tel: +81–58–267–2226, Fax: +81–58–267–2962, E-mail: kuwata@cc.gifu-u.ac.jp

that we termed the “GeneMap” program, and investigated how these map patterns changed with evolution. Then employing the power spectra obtained by Fourier transformation of these DNA sequences, we tried to identify the periodicity in the short- (2 to 10 base pair [bp]), medium- (10 to 50 bp), and long-range order (50 to 300 bp) of DNA sequences, and focused our analysis on the evolutionary changes in the periodicity of DNA sequences together with its self-similar property.

MATERIALS AND METHODS

Materials. To examine the evolutionary changes in the periodicity and the self-similarity of the genomic sequence, it is necessary to obtain a set of complete data on DNA sequences from a wide variety of species. Furthermore, the long complete DNA sequences should be used for the analysis of the periodicity in the long-range order. For these reasons, the database for globin genes available from GenBank (NCBI, U.S.A.) was mainly used. We selected data on β -globin genes belonging to 15 different species as listed in Table 1, and α -globin, albumin, preproinsulin, and huntingtin, that is, the protein associated with Huntington’s disease [15], genes as listed in Table 2 for a search of the periodicity of short-range order (2 to 10 bp). We also selected data on longer globin gene cluster (either β -, γ -, δ -, ϵ -, η -, or ρ -globin) belonging to 5 different species as listed in Table 3 for a search of the periodicity of both medium- (10

to 50 bp) and long-range order (50 to 300 bp) as well as the self-similarity.

Values defined as aligned scores given in Tables 1–3 represent the percent amino acid sequence homology of the organisms relative to humans (100%) and were calculated by the multiple sequence alignment by using a program called CLUSTAL W (ver. 1.8) [16]. We then assumed that this value represents the evolutionary ordering of species development in the present work.

DNA walk analysis. The technique of “DNA walk” analysis reported by Peng *et al.* [5] provides a novel graphical representation of genes. We modified this technique to study the graphical characterization of exons and introns in β -globin genes. For mapping the two-dimensional walk of a DNA sequence in the x - y plane, the four base letters, A, G, C, and T, were represented by vectors (1, 0), (−1, 0), (0, −1), and (0, 1) in the x - y plane, respectively. To map it, the vector of the present base is added to the previous coordinate of the walk, and this process is repeated until the end of the individual DNA sequence is encountered. The origin of the walk is taken at (0, 0) in the x - y plane. Thus the horizontal axis represents the change in base within purine bases (A and G), and the vertical shows the change within pyrimidine bases (T and C), as in Fig. 1. If two hydrogen bonds (A-T pair) in the DNA double strand are dominant, the walk tends to be in the right upper direction of the figure, whereas if three hydrogen bonds (G-C pair) are dominant, it declines to the left lower direction.

Table 1. Fifteen β -globin genes obtained from GenBank.^a

Common name	Organism	Code ^b	Aligned score ^c (%)	Total length ^d (bp)	% intron ^e (%)	mRNA ^f (bp)
African clawed frog	<i>Xenopus laevis</i>	XELHBBC	45	1,033	70.1	441
Common carp	<i>Cyprinus carpio</i>	AB004740	52	660	32.3	447
Atlantic salmon	<i>Salmo salar</i>	SSBGLOB4	53	803	44.7	444
Chicken	<i>Gallus gallus</i>	CHKHBBRE	69	1,346	67.0	444
Muscovy duck	<i>Cairina moschata</i>	CMBGA2B2	69	1,580	71.9	444
North American opossum	<i>Didelphis virginiana</i>	OPOHBBE	73	2,411	81.6	444
Norway rat	<i>Rattus norvegicus</i>	RNGLOBM	78	1,199	63.0	444
House mouse	<i>Mus musculus</i>	MUSHBBMAJ	80	1,213	63.4	423
Lemur macaco	<i>Eulemur macaco</i>	LEMHBB	82	1,442	69.2	444
Sheep	<i>Ovis aries</i>	OABCGLOB	82	1,404	69.4	429
Cow	<i>Bos taurus</i>	BTGL02	84	1,464	70.1	438
Pig	<i>Sus scrofa</i>	SSBETAG	85	1,202	63.1	444
Rabbit	<i>Oryctolagus cuniculus</i>	RABBGLOB	90	1,143	61.2	444
Gorilla	<i>Gorilla gorilla</i>	GGBGLOBIN	99	1,344	72.9	364
Human	<i>Homo sapiens</i>	HUMHBB	100	1,470	69.8	444

^a Fifteen β -globin genes are ordered from top to bottom according to the values of the aligned score (%). ^b GenBank names. ^c Percent amino acid sequence homology of the organisms relative to human (100%) derived from multiple sequence alignment, using a program called CLUSTAL W (ver. 1.8) [16]. ^d Total number of nucleotides. ^e Ratio for intron length vs. total length. ^f All exons.

Table 2. α -Globin, albumin, preproinsulin, and huntingtin genes obtained from GenBank.^a

Common name	Organism	Code ^b	Aligned score ^c (%)	mRNA ^d (bp)
α -Globin				
Zebrafish	<i>Danio rerio</i>	DRU50382	53	346
African clawed frog	<i>Xenopus laevis</i>	XLGLAA1	56	429
Chicken	<i>Gallus gallus</i>	GGGL01	57	429
Atlantic salmon	<i>Salmo salar</i>	SSABGLOB6	58	432
Domestic pigeon	<i>Columba livia</i>	AB001981	69	429
Norway rat	<i>Rattus norvegicus</i>	RNU29528	78	429
Rabbit	<i>Oryctolagus cuniculus</i>	OCATGL1	82	429
House mouse	<i>Mus musculus</i>	MMAGL1	85	429
Horse	<i>Equus caballus</i>	ECPZA2GL	87	429
Rhesus monkey	<i>Macaca mulatta</i>	MACHBA	96	429
Orangutan	<i>Pongo pygmaeus</i>	ORAHBA01	97	429
Human	<i>Homo sapiens</i>	HSAGL4	100	426
Albumin				
Atlantic salmon	<i>Salmo salar</i>	SSSERALB	27	1,827
African clawed frog	<i>Xenopus laevis</i>	XEL74KSA	39	1,818
Chicken	<i>Gallus gallus</i>	GGPPALB	47	1,848
Nowray rat	<i>Rattus norvegicus</i>	RATALBM	73	1,827
Rabbit	<i>Oryctolagus cuniculus</i>	OCU18344	75	1,827
Sheep	<i>Ovis aries</i>	OASERALB	75	1,824
Cow	<i>Bos taurus</i>	BTBSA	76	1,824
Horse	<i>Equus caballus</i>	ECSERALB	76	1,824
Cat	<i>Felis catus</i>	FDALBUMIN	82	1,827
Rhesus monkey	<i>Macaca mulatta</i>	MACSERALB	93	1,803
Human	<i>Homo sapiens</i>	HSALBU	100	1,830
Preproinsulin				
Zebrafish	<i>Danio rerio</i>	AF036326	45	327
Chum salmon	<i>Oncorhynchus keta</i>	SSINSU	48	327
Chicken	<i>Gallus gallus</i>	GGINSMRNA	64	323
Fat sand rat	<i>Psammomys obesus</i>	POPREIN	80	333
House mouse	<i>Mus musculus</i>	MMINSIIG	81	333
Cow	<i>Bos taurus</i>	BOVPPI	83	318
Pig	<i>Sus scrofa</i>	AF064555	84	386
African green monkey	<i>Chlorocebus aethiops</i>	CEPPINS	97	333
Chimpanzee	<i>Pan troglodytes</i>	PTPPINS	98	333
Crab-eating macaque	<i>Macaca fascicularis</i>	MACINS	98	333
Human	<i>Homo sapiens</i>	HSINSU	100	333
Huntingtin				
Human	<i>Homo sapiens</i>	MW_002111		13,672

^a Twelve α -globin, 11 albumin, and 11 preproinsulin genes are ordered from top to bottom according to the values of the aligned score (%). ^b GenBank names. ^c Percent amino acid sequence homology of the organisms relative to human (100%) derived from multiple sequence alignment, using a program called CLUSTAL W (ver. 1.8) [16]. ^d All exons.

Table 3. Five globin gene cluster obtained from GenBank.^a

Common name	Organism	Code ^b	Globin	Aligned score ^c (%)	Total length ^d (bp)	% intron ^e (%)
Chicken	<i>Gallus gallus</i>	CHKHBBRE	β , ϵ , ρ	69	30,539	91.0
House mouse	<i>Mus musculus</i>	MMBGCXD	β , ϵ	80	55,856	94.7
Thick-tailed bush baby	<i>Otolemur crassicaudatus</i>	OCU60902	β , ϵ , γ , δ , η	89	57,113	96.9
Rabbit	<i>Oryctolagus cuniculus</i>	RABBGLOB	β , ϵ , γ , δ	90	44,594	97.0
Human	<i>Homo sapiens</i>	HUMHBB	β , ϵ , γ , δ	100	73,308	97.0

^a Five globin genes are ordered from top to bottom according to the values of the aligned score (%). ^b GenBank names. ^c Percent amino acid sequence homology of the organisms relative to human (100%) derived from multiple sequence alignment, using a program called CLUSTAL W (ver. 1.8) [16]. ^d Total number of nucleotides. ^e Ratio for intron length vs. total length.

Spectral analysis of the periodicity in the DNA sequence. For the periodicity analysis, power spectra were obtained by the Fourier transformation of the DNA nucleotide sequences (Tables 1–3), and the power spectral densities were plotted against the periodicity (bp) in the spectra. For the fast Fourier transformation (FFT), we developed a useful program based on the Cooley-Turkey algorithm [17].

In the monomer analysis, we conducted FFT on four kinds of data sets consisting of 0 or 1. A value of 1 appears only when the base corresponds to A, G, C, or T. For example, (AGTCAGTC...) yields (10001000...), (01000100...), (00010001...), and (00100010...), which are used for the periodicity analysis for A, G, C, and T, respectively. The power spectral density of (AGGAGGAGGAGG...) of 512 stretches has a peak in A with 0.276 at a periodicity of 3.0 (half-linewidth, 0.03). Similarly, in the dimer analysis, 1 appears only when the present dimer corresponds to AA, AG, AT, AC, GA, GG, GT, GC, TA, TG, TT, TC, CA, CG, CT, or CC. Thus there are 16 data sets for FFT.

To examine the periodicity of the DNA sequence, we conducted a monomer analysis on the genes listed in Tables 1–3 and a dimer analysis on those listed in Table 3.

Spectral analysis of self-similarity (the fractal structure) in the DNA sequence. For the self-similarity analysis, power spectra were also obtained by Fourier transformation of the globin DNA nucleotide sequences listed in Table 3. Here, the power spectral density was plotted against the frequency (Hz) in logarithmic scales ($P(f)$). $P(f)$ is also a power law function: $P(f)=1/f^\alpha$, where f is the frequency and α is the scaling exponent, in other words, the slope of this plot at the low frequency region of the spectrum. Thus using a linear regression, we calculated the α value in the region between 1 and 100 Hz (i.e., 160 to 16,000 bp) of the spectrum to examine the self-similarity (the fractal structure) of the globin gene sequence.

RESULTS

Two-dimensional map patterns for the β -globin gene

Figure 1 displays the two-dimensional walks generated by using the “GeneMap” program. Introns are colored, and exons are black. With the evolutionary order of species development defined by the aligned score (%) given in Table 1, especially the walking profiles of carp (52%), chicken (69%), pig (85%), rabbit (90%), and human (100%), in this order, rotated in a

Fig. 1. Two-dimensional walk patterns generated by using the GeneMap program. Introns are colored in blue (carp), green (chicken), yellow (pig), pink (rabbit), and red (human); exons are black. A, G, C, and T in the DNA sequence were replaced with vectors (1, 0), (−1, 0), (0, −1), and (0, 1) in the x-y plane, respectively. These vectors are added to the previous coordinate of the walk, and the origin of the walk is taken at (0, 0) in the x-y plane. The horizontal axis shows the change within purine bases (A and G), and the vertical axis shows the change within pyrimidine bases (T and C).

clockwise manner, indicating a relative increase in the amounts of two bases, A and T, especially T, at the expense of G and C with evolution. Using all the data on β -globin genes listed in Table 1, we calculated the walking distance for each exon as well as for each intron, which is defined as the linear distance between both ends of each exon (or each intron) normalized to the single nucleotide. The mean value for the walking distance of all exons (0.118 ± 0.040) was significantly smaller than that of all introns (0.176 ± 0.062) ($p < 0.05$). It is worth noting that the walking distance of introns becomes longer to increase the content of bases A and T with evolution, and that of exons remains almost unchanged, as shown in Fig. 1.

Periodicity in the short-range order (2 to 10 bp) of the α - and β -globins, albumin, preproinsulin, and huntingtin gene sequences

Using the monomer analysis, we obtained the power spectra of four bases, A, G, C, and T, in the human β -globin exon sequence (512 bp) and plotted each power spectral density as a function of the periodicity (bp). It follows that prominent peaks at 2.99 bp were found only for A and also for G marked by * in Fig. 2a and b, respectively, and the values for their spectral densities were 0.051 and 0.055, respectively. To assess the statistical significance of the observed spectral density, we generated the 20 random se-

Fig. 2. Power spectral densities in the short periodicity region (2 to 5 bp). (a) A and (b) G in the human β -globin exon (512 bp) are shown. Characteristic peaks in A (2.99 bp) and G (2.99 bp) are marked by *. (c) Power spectral densities of A (●) at 2.99 bp ($r=0.68$) and G (○) at 2.99 bp periodicity ($r=0.76$) as a function of the aligned score (%) listed in Table 1.

Fig. 3. Power spectral density as a function of aligned score. (a) Power spectral densities of A (●) ($r=0.87$) and G (○) ($r=0.85$) in α -globin exon at a 2.99 bp periodicity as a function of the aligned score (%) listed in Table 2. (b) Power spectral densities of A (●) ($r=0.54$) and G (○) ($r=0.92$) in

albumin exon at a 2.99 bp periodicity as a function of the aligned score (%) listed in Table 2. (c) Power spectral densities of G (○) ($r=0.85$), C (△) ($r=0.87$), and T (▲) ($r=0.91$) in preproinsulin exon at a 2.99 bp periodicity as a function of the aligned score (%) listed in Table 2.

quences (the length of each sequence is 512 bp) and obtained their spectral densities at a periodicity of 3.0 bp, whose mean value was 0.017 ± 0.012 . Thus the observed values for the spectral densities of the β -globin genes at 2.99 bp were significantly large.

Power spectral densities at 2.99 bp for A (●) and G (○) increased with an increasing aligned score (%) from 15 different species listed in Table 1, as shown in Fig. 2c (correlation coefficients, $r=0.68$ for A and 0.76 for G). This indicates that both A and G preferentially locate every 3 bp in the β -globin exon sequence with evolution. However, the A and G contents of the sequence changed little with evolution (A, ranged from 19.1 to 28.8%; G, ranged from 21.2 to 31.6%). The rates of synonymous substitutions over [synonymous substitutions+nonsynonymous substitutions] per unit length of aligned score were almost constant (data not shown).

To confirm a 3 bp periodicity in the DNA sequences, power spectral densities at 3 bp for α -globin,

albumin, and preproinsulin exon genes listed in Table 2 were also investigated as a function of aligned score (%). As shown in Figs. 3a, b, and c, the periodicity at 3 bp of purine bases A and G, especially G, had a strong tendency to increase with an increasing aligned score (%). Furthermore, we analyzed the huntingtin gene listed in Table 2. As shown in Fig. 4, power spectral densities of bases A (a), G (b), and C (c) were markedly large at 2.99 bp, and T (d) had no significant spectral density. Especially, G had the largest spectral density, and its value was 0.113.

Of note, a 3 bp periodicity was not observed in β -globin intron sequence (data not shown).

Periodicity in the medium-range order (10 to 50 bp) of the β -globin gene sequence

Prominent peaks in human β -globin exon were also found at 21.3 bp periodicities for A and G, marked by * in Figs. 5a and b, respectively. Power spectral densities at 21.3 bp for A (●) and G (○) increased with in-

Fig. 4. Power spectral densities in the short periodicity region (2 to 5 bp). (a) A, (b) G, (c) C, and (d) T in huntingtin exon (512 bp) are shown. Prominent peaks in A, G, and C are clearly observed.

Fig. 5. Power spectral densities in the medium periodicity region (5 to 30 bp). (a) A and (b) G in the human β -globin exon (512 bp) are shown. Characteristic peaks in A (21.3 bp) and G (21.3 bp) are marked by *. (c) Power spectral densities of A (●) at 21.3 bp ($r=0.78$) and G (○) at 21.3 bp periodicities ($r=0.82$) as a function of the aligned score (%) listed in Table 1.

creasing values for the aligned score (%), as shown in Fig. 5c ($r=0.78$ for A and 0.82 for G).

We further conducted a dimer analysis on a longer globin DNA sequence, i.e., the human β -globin region on chromosome 11 listed in Table 3. The result showed that the sequence has a periodicity of ~ 10 bp. In 16 kinds of dimer pairs, it was found that only AA (Fig. 6a) and TT (Fig. 6b) have remarkable peaks at 11.2 and at 10.0 bp, marked by *, respectively. Power spectral densities at ~ 10 bp for AA (●) and TT (▲) increased with increasing values for the aligned score (%), as shown in Fig. 6c ($r=0.99$ for AA and 0.95 for TT). However, the AA and TT dimer contents of the sequence changed little with evolution (AA, ranged from 5.7 to 10.8%; TT, ranged from 4.5 to 10.8%).

Periodicity in the long-range order (50 to 300 bp) of the globin gene sequence

A long complete DNA sequence should be used for the analysis of periodicity in the long-range order. Thus by using the monomer analysis, we obtained power spectra of four different bases, A, G, C, and T, in the full length of globin genes, including introns from 5 different species listed in Table 3, and plotted each power spectral density as a function of the periodicity (bp). The power spectral density for the human β -globin region on chromosome 11 (65,536 bp) is shown in Fig. 7. Prominent peaks were observed only in bases A and T. Base A had two maxima typically at 106.4 (*) and 204.8 bp (**) (Fig. 7a), and base T one at 105.4 bp (*) (Fig. 7b). Figure 7c shows that

Fig. 6. Power spectral densities of the representative dimers, in the human β -globin region on chromosome 11 (65,536 bp). (a) AA dimer. (b) TT dimer. Characteristic peaks in AA (11.2 bp) and TT (10.0 bp) are marked by *. (c) Power spectral densities of AA dimer ($\bullet$) ($r=0.99$) and TT dimer ($\blacktriangle$) ($r=0.95$) at a periodicity of 10.5 bp in β -globin genes (16,384 bp) as a function of the aligned score (%) listed in Table 3.

Fig. 7. Power spectral densities of the human β -globin region on chromosome 11 (65,536 bp). (a) A. (b) T. Characteristic peaks in A at 106.4 bp and at 204.8 bp are marked by * and **, respectively. A peak of T (105.4 bp) is marked by *. (c) Power spectral densities of A around 100 bp ($\bullet$) ($r=0.01$) and 200 bp ($\circ$) ($r=0.86$), and T ($\blacktriangle$) around 100 bp ($r=0.86$) periodicities in β -globin genes (16,384 bp) as a function of the aligned score (%) listed in Table 3.

~ 100 bp periodicity for T ($\blacktriangle$) and ~ 200 bp periodicity for A ($\circ$) increased with evolution. The A and T contents of the sequence changed little with evolution (A, ranged from 23.4 to 31.5%; T, ranged from 23.2 to 32.5%).

Self-similarity (the fractal structure) in the globin gene sequence

An analysis of the far low frequency region of power spectral density is less accurate. Thus for the observation of self-similarity, we used $1/f^\alpha$ analysis at the far low frequency region in a log-log plot of the power spectrum. We determined the α value of the $1/f^\alpha$ plot in the region from 1 to 100 Hz (160 to 16,000 bp) of the spectrum. Figures 8a and b, show

representative power spectra for base A in the chicken (aligned score=69%) β -globin gene region (16,384 bp) with $\alpha=0.29$, and in the human (aligned score=100%) β -globin gene (16,384 bp) with $\alpha=0.75$, respectively. Almost all introns in the human β -globin gene region have positive α values in general. The α values for only A ($\bullet$) and T ($\blacktriangle$) increased with an increasing the aligned score (%) (Fig. 8c), indicating the increase of the far long-range periodicity between 160 and 16,000 bp (i.e., the self-similarity or the fractal structure) with evolution. The increase of self-similarity with evolution can be partially explained by the increase of the intron content (% intron) with evolution as given in Table 3.

Fig. 8. Power spectral densities at far low frequency region. (a) Power spectra of A in the chicken β -globin gene (16,384 bp) with a slope, $\alpha=0.29$. (b) Power spectra of A in the human β -globin gene on chromosome 11 (16,384 bp) with a slope $\alpha=0.75$. (c) Plot of α values (●, ▲, ○, and △ indicate A [$r=0.89$], T [$r=0.91$], G [$r=0.49$], and C [$r=0.19$], respectively) for globin genes as a function of the aligned score (%) listed in Table 3.

DISCUSSION

Periodicity of 3 bp. It is reported that exon sequence commonly has a periodicity at 3 bp [6, 10]. Here we found that a 3 bp periodicity in the DNA exon sequences was observed, and it increased with evolution. Purine bases A and G, especially G, have a surprisingly strong tendency to possess increasing 3 bp periodicity with increase of the aligned score (%) for α -globin, albumin, and preproinsulin (Fig. 3) as well as for β -globin (Fig. 2). Thus a 3 bp periodicity in exons may generally increase with evolution.

It is quite intriguing to compare our results with the codon usage reported by Nakamura *et al.* [18]. Using their codon usage table, we calculated the probabilities for the first, second, and third base usage in a human codon. For the first base usage, G had the highest probability, 0.317. For the second, the probability for A was 0.313, and for the third, that for C was 0.308. On average, ([G]-[A]-[C]) occurred most frequently in a human codon. In other words, G or A appears once in a single codon. It should be noted that the codon usage of (GAC) is not the highest. This is consistent with our observation that the power spectral density of A and G has a maximum at 3 bp periodicity. These probabilities essentially remain unchanged with evolution, however, except for *Xenopus laevis* (data not shown).

The increase in a 3 bp periodicity with evolution has an implication to the CAG repeat disease. Since Huntington chorea is well known as a representative CAG repeat disease, we analyzed its causative huntingtin protein gene [15]. As shown in Fig. 4b, G had

the largest spectral density at 2.99 bp, 0.113, which is about half the density of AGGAGG repeat (0.276) and about twice as large as that of β -globin (0.055). The predominant 3 bp periodicity of G is the common feature for other proteins such as α - and β -globins, albumin, and preproinsulin, and the spectral densities of these exons are increasing with evolution (Figs. 2c and 3). These results strongly suggest that there might be internal rules in the DNA sequence that increase the 3 bp periodicity of G. These internal rules, whose nature is unknown though, might be related to the increase of the CAG repeat during the progress of generation.

Periodicity of more than 3 bp. A periodicity of about 10.4 bp in the DNA sequence was first reported by Prunell *et al.* [2], employing the DNA fragments produced by DNase I digestion of chromatin. Widom [10] also concluded by using a dimer analysis that the short-range order in the eukaryotic genome has a periodicity of about 10.2 bp. Furthermore, according to a report on the X-ray crystallographic structure of the nucleosome core particle of chromatin, the value for the overall double-helix twist is 10.2 bp per turn in nucleosome and 10.6 bp in free DNA [19]. Our results obtained by a monomer analysis, however, showed that the periodicity is not ~ 10 bp, but 21.3 bp. The periodicity of 21.3 bp may correspond to twice the value of 10.6. Furthermore, power spectral densities at 21.3 bp for A (●) and G (○) increased with increasing aligned score (%) (Fig. 5c), possibly indicating that a 10.6 bp periodicity (free DNA) has been conserved and even strengthened with evolution. Thus nucleosome positioning may be less

relevant to the periodicity of the β -globin exon, and the interaction of the exon with the histone protein complex may become weaker and weaker and more exposed to the surface, thereby easily transcribed by RNA polymerase, according to the evolution.

Our results obtained by the dimer analysis (Fig. 6) indicate that ~ 10 bp periodicity has been conserved and even strengthened with evolution. They also indicate that AA and TT in DNA sequence including introns preferentially locate every 10 bp.

It is reported that the DNA sequence including introns has a periodicity of about 200 bp [14]. Periodicities of 100 and 200 bp correspond to those of the nucleosome phasing; i.e., a single turn of DNA around histone protein in the nucleosomal histone core is about 89 bp [14], and the single repeat of nucleosome, including the linker region, is about 200 bp [14]. These periodicities might be attributed to the stability of the chromatin structure [20]. The nucleosome phase might have also been stabilized with evolution as shown in Fig. 7c.

Self-similarity (the fractal structure) of DNA sequence. The intron sequence has remarkable self-similarity [5, 8]. The evolution of the self-similarity could be explained as follows. The lengths of present-day DNA sequences are believed to be much longer than those of ancient DNA when life started. One of the mechanisms for sequence elongation may be gene or oligonucleotide duplication [1, 4], during which copies of a segment of the sequence are made and then inserted back into the original sequence. According to a class of mathematical models called expansion-modification systems [4], the sequence lengths become longer and longer driven by only simple mechanisms, duplication and mutation. It was found that if the two operations are applied repeatedly, the elongated sequence has $1/f^\alpha$ power spectra, i.e., $\alpha=1.0$ [4].

In fact, Peng *et al.* [5] demonstrated the existence of a long-range correlation in the DNA sequence based on the purine/pyrimidine DNA walks. The long-range correlation dominates not in exons [8], but in introns [5]. Using 13 microbial complete genomes, de Sousa Vieira [11] reported that the self-similarity of DNA chains does not always prevail through the entire DNA molecule. However, self-similarity may be predominant in introns instead of exons, since natural selection may be imposed mainly on the exon rather than the intron. Especially, the functional role of the intron, which occupies more than 97% of the DNA [14], is essentially unknown. Mantegna *et al.* [9] reported that noncoding regions (introns) in eukaryotes have a smaller entropy and a larger redundancy simi-

lar to natural languages than the coding regions (exons) do.

The value for α seems to be approaching 1.0 with evolution as suggested in Fig. 8c. White noise yields an α value of 0, and Gaussian noise yields a value of 2.0. A value of 1.0 is characteristic of a self-similar sequence, e.g., fractal pattern [21]. The value for α seems to have been increased to increase the self-similarity. The information theory suggests that one can package the largest amount of information into characters of constant length when $\alpha=1.0$. Our results also suggest that introns may carry biological information—for the higher regulation of transcription [9], for example—and that the amount of information packaged in the DNA sequence might gradually increase with evolution.

Evolution of the periodical and the self-similar properties. According to the introns-early theory [22, 23], introns locate between the functional domains of proteins, but the introns-late theory [24, 25] states that introns form at random positions within the open reading frame. With the latter, boundaries between exons and introns are expected to be smooth, but with the former theory, they might be rough. Figure 1 shows that the last boundaries between exon and intron for pig, rabbit, and human β -globin genes are generally rough, supporting the introns-early theory [22, 23]. On the other hand, these boundaries are smooth for the chicken β -globin gene, suggesting the predominance of the latter mechanism [24, 25]. More mapping patterns must therefore be accumulated to discuss further this intriguing issue.

In conclusion, we have studied the periodical and the self-similar properties of DNA sequences by using Fourier transform analysis and characterized the evolutionary changes for β -globin genes from 15 different species. Two-dimensional DNA walk maps showed that the stretches of exon sequences were significantly shorter than those of intron sequences (Fig. 1), suggesting that the evolution of exons is mainly driven by natural selection (random forces), whereas that of introns is governed by unknown internal rules (nonrandom forces). Power spectral densities of bases A and G in exons have a characteristic peak at a periodicity of 3.0, which increases with an increasing aligned score (%) (Fig. 2). This tendency is independent of the base content or the length of exons. The general increase of the 3 bp periodicity observed also for other exons (Fig. 3) indicates the existence of some internal rules in DNA sequence and a possible connection to the cause of CAG repeat disease (Fig. 4). Power spectral density also has a peak at a periodicity of 21.3 bp, which corresponded to twice the

value of 10.6 bp, and it increases with an increasing aligned score (%) (Fig. 5). It is intriguing, the 10.6 bp periodicity corresponds to the length of the single turn of the free DNA double strand [19], suggesting that the interaction between DNA and histone protein has been destabilized with evolution. Power spectral densities of bases A and T in the DNA sequences including introns have two maxima at ~ 100 and at ~ 200 bp, which correspond to the periodicities of the nucleosome phasing [14]. Both peaks increased with the aligned score (%) shown in Fig. 7c, indicating that the chromatin structure has also been stabilized with evolution. A far longer periodicity (160 to 16,000 bp) in DNA sequence was examined by using the $1/f^\alpha$ analysis. The α values for bases A and T increased with an increasing aligned score (%), as shown in Fig. 8c, suggesting an increase in the amount of information in the nucleotide sequence with evolution.

The DNA double strand and the chromatin structure have been stabilized with evolution. The amount of information implemented in a genetic sequence has increased with evolution. These findings obtained through Fourier transformation of the DNA sequence suggest that a purely theoretical interpretation of the DNA nucleotide sequence is feasible. We should note again that this information is essentially independent of the knowledge about the amino acid sequence of proteins.

**Note Added in Proof:* The sequence of the entire human genome has been published: Lander ES *et al.*: Initial sequencing and analysis of the human genome. *Nature* 409: 860–921, 2001; Venter JC *et al.*: The sequence of the human genome. *Science* 291: 1304–1351, 2001.

The author (N.N.) is grateful to Dr. Y. Hattori, president of Hattori ENT Practiced Institute, Gifu, Japan, for continuous encouragement throughout the period of research in this laboratory.

REFERENCES

- Ohno S: Evolution by Gene Duplication, Springer-Verlag, New York, pp 1–146, 1970
- Prunell A, Kornberg RD, Lutter L, Klug A, Levitt M, and Crick FHC: Periodicity of deoxyribonuclease I digestion of chromatin. *Science* 204: 855–858, 1979
- Ohno S: Universal rule for coding sequence construction: TA/CG deficiency-TG/CT excess. *Proc Natl Acad Sci USA* 85: 9630–9634, 1988
- Li W: Expansion-modification systems: a model for spatial $1/f$ spectra. *Phys Rev A* 43: 5240–5260, 1991
- Peng C-K, Buldyrev SV, Goldberger AL, Havlin S, Sciortino F, Simons M, and Stanley HE: Long-range correlations in nucleotide sequences. *Nature* 356: 168–170, 1992
- Voss RF: Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences. *Phys Rev Lett* 68: 3805–3808, 1992
- Widom J: A relationship between the helical twist of DNA and the ordered positioning of nucleosomes in all eukaryotic cells. *Proc Natl Acad Sci USA* 89: 1095–1099, 1992
- Li W and Kaneko K: Long-range correlation and partial $1/f^\alpha$ spectrum in a noncoding DNA sequence. *Europhys Lett* 17: 655–660, 1992
- Mantegna RN, Buldyrev SV, Goldberger AL, Havlin S, Peng C-K, Simons M, and Stanley HE: Linguistic features of noncoding DNA sequences. *Phys Rev Lett* 73: 3169–3172, 1994
- Widom J: Short-range order in two eukaryotic genomes: relation to chromosome structure. *J Mol Biol* 259: 579–588, 1996
- de Sousa Vieira M: Statistics of DNA sequences: a low-frequency analysis. *Phys Rev E* 60: 5932–5937, 1999
- Hattori M, Fujiyama A, Taylor TD, Watanabe H, *et al.*: The DNA sequence of human chromosome 21. *Nature* 405: 311–319, 2000
- Dunham I, Shimizu N, Roe BA, Chissole S, *et al.*: The DNA sequence of human chromosome 22. *Nature* 402: 489–495, 1999
- Alberts B, Bray D, Lewis J, Raff M, Roberts K, and Watson JD: *Molecular Biology of the Cell*, 3rd Ed, Garland Publishing Inc, New York and London, 1294 pp, 1994
- The Huntington's Disease Collaborative Research Group: A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* 72: 971–983, 1993
- Thompson JD, Higgins DG, and Gibson TJ: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22: 4673–4680, 1994
- Press WH, Flannery BP, Teukolsky SA, and Vetterling WT: *Numerical Recipes*, Cambridge University Press, New York, pp 381–453, 1986
- Nakamura Y, Gojobori T, and Ikemura T: Codon usage tabulated from the international DNA sequence database: status for the year 2000. *Nucleic Acids Res* 28: 292, 2000
- Luger K, Mäder AW, Richmond RK, Sargent DF, and Richmond TJ: Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* 389: 251–260, 1997
- Lowary PT and Widom J: New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J Mol Biol* 276: 19–42, 1998
- Mandelbrot BB: *Fractals: Form, Chance, and Dimension*, W. H. Freeman & Co., San Francisco, 331 pp, 1977
- Gilbert W, de Souza SJ, and Long M: Origin of genes. *Proc Natl Acad Sci USA* 94: 7698–7703, 1997
- de Souza SJ, Long M, Klein RJ, Roy S, Lin S, and Gilbert W: Towards a resolution of the introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins. *Proc Natl Acad Sci USA* 95: 5094–5099, 1998
- Cavalier-Smith T: Intron phylogeny: a new hypothesis. *Trends Genet* 7: 145–148, 1991
- Palmer JD and Logsdon JM Jr: The recent origins of introns. *Curr Opin Genet Dev* 1: 470–477, 1991