

The Power-Law-Tail in the Distribution of the Nucleotides of Genomes Was Related to the Complexity of Organism: New Classification of Organisms

Masaharu Takeda^{1,2}

¹ Department of Materials and Biological Engineering, Tsuruoka National College of Technology, Tsuruoka, Yamagata, Japan

² Laboratory of Life Information Science, Yamagata, Japan

Correspondence: Masaharu Takeda, Department of Materials and Biological Engineering, Tsuruoka National College of Technology, Tsuruoka, Yamagata 997-8511, Japan. Tel: 81-23-525-9130; 81-23-641-4003. E-mail: mtakeda@tsuruoka-nct.ac.jp; takedamasaharu@gmail.com

Received: December 26, 2013 Accepted: February 16, 2014 Online Published: February 26, 2014

doi:10.5539/ijb.v6n2p77

URL: <http://dx.doi.org/10.5539/ijb.v6n2p77>

Abstract

We proposed a new index of the classification of organisms (cells) based on the appearance frequency of four nucleotides (bases) of various genomes. In double logarithmic plot of L (distance of a base to the next base, x-axis) vs F (frequencies of a base at L, y-axis), each value of four bases was expressed in $y = ae^{-bx}$ at $L = 1 \sim 15$, and $y = Ux + W$ (power-law-tail) at $L =$ more than 16 bases, respectively, in a single-strand of DNA. The a-, b- and U-values (slope) of four bases were resulted from the GC-content (%) and the size (nt) of the genome. Moreover, each value was identical as A to T, and as G to C, respectively, in one organism. The power-law-tail should be unique to the genomes of the same species, the eukaryotes, the prokaryotes. The eukaryotic genomes were essentially composed of great number of bases with plural long power-law-tail regions when compared with those of the prokaryotes. In the prokaryotes, the base-distribution was partitioned at $L = 20$, and the U-values (base-distribution in power-law-tail region) of the archaea were similar to the eukaryotes compared with those of the eubacteria. Thus, the power-law-tail of the genomic DNA should be come from the structural features of the cells, i.e., the size, the GC-content and other characteristics of the genomic DNA. These results indicated that the power-law-tail would be specific for the complexity of organisms in individual genome, and might be a new index for cells.

Keywords: appearance frequency of base sequence of genome, multiple fractality, power-law-tail, complexity of organism

1. Introduction

Recently, with the rapid progress of the genome-projects for many prokaryotic- and eukaryotic-cells (NCBI genome database, 2013; The Sanger Institute, 2013; KEGG Organisms, 2013), the whole-genome analyses of the biological phenomena for the individual cells were carried out to understand the organisms (Coleman et al., 2006; Eisen et al., 2006; Perfrey et al., 2006; Zmasek & Godzik, 2011).

The four bases in the genomic DNA were arranged sophisticatedly in all cells and distinguished the coding- and the non-coding region clearly on the genomic DNA, and analyzed the entire genome containing both the coding- and the non-coding sequences because the key sequences were distributed throughout the genome in cells (Takeda & Nakahara, 2009, 2013; Takeda, 2009, 2011, 2012; Nakahara & Takeda, 2010a, 2010b). In other words, all genes containing the non-coding base sequences of the genomic DNA could be maintained in the cells, and each gene in the genome could be expressed to be subjected to the living cells.

As described above, studies on the appearance frequency of the base sequence were very useful in clarifying the structural features of the genomic DNA, (1) the reverse-complement symmetry, (2) the bias of the distribution of four bases, (3) the multiple fractality of the distribution of the respective four bases co-existing even in a single-strand DNA of the genome (Takeda & Nakahara, 2009; Takeda, 2009, 2011), and were resulted from the above structural features of the genomic DNA to characterize or to identify the interactive regions between

chromosome-gene, protein-protein, protein-DNA and protein-tRNA from the structure (the base sequence) of the entire genome or chromosome (Nakahara & Takeda, 2010a, 2010b; Takeda, 2012; Takeda & Nakahara, 2013).

In early 1990s, many physicists interested in biology were reported to place four bases, A, T, G, C, with the self-similarity (fractality) in the genome base sequences (Peng et al., 1992; Voss, 1992; Bains, 1993; Weinberger & Stadler, 1993). But in those periods the data were limited for the complete genome and chromosome, therefore, these reports were based on the virus genomes, or a part of the prokaryotic- or a part of the small eukaryotic-chromosomes, the model genomes (chromosomes) based on the former speculation or the preliminary data for the genomes. At present, the base sequences of the genomic DNAs in many cells, including the eukaryotic- and the prokaryotic-cells, can be clearly observed and the base sequence of the genomic DNAs can be viewed the cells.

In 1992, the base sequence of chromosome III of *S. cerevisiae* were firstly determined in eukaryotic genome (Oliver et al., 1992), and since then the Genome Projects of *S. cerevisiae* (Mews et al., 1997), *H. sapiens* (Dunham et al., 1999) and many Genome Projects of the eukaryotic- and the prokaryotic-cells were accelerated and the base sequences were reported (NCBI genome database, 2013; The Sanger Institute, 2013; KEGG Organisms, 2013; Saccharomyces Genome Database, 2013). Based on these results, it was possible to carry out the detailed analyses of the base sequences of the genome, and reported. The genome of eukaryotic cells were possessed the large non-coding sequences in the base sequences of genome. The non-coding sequences were composed of many regulatory elements which might be necessary to express the genetic information precisely, rapidly and steady (Takeda & Nakahara, 2009; Takeda, 2009; Takeda, 2012; Takeda & Nakahara, 2013).

The distribution of four bases of the genomic DNA was generally arranged sophisticatedly with the symmetry, the bias and the multiple fractality with different fractal-dimension, i.e., an exponentially decreased-phase at short distances and a linearly decreased-phase (a power-law-tail, the fractality with the multi-fractal dimension expressed in $y = Ux + W$) at long distances in double logarithmic plot of base-distribution, respectively (Takeda & Nakahara, 2009; Takeda, 2009; Takeda, 2011). The exponentially decreased-phase with the uni-fractal dimensions (Barthelemy et al., 2000; Ordemann et al., 2000; Yu et al., 2004) of base(s) at short distances, expressed in $y = ae^{-bx}$, was observed in not only any genomes or chromosomes, but also even in their artificial genomes or chromosomes with the same genome-size (nt) and molar-contents of four bases as mentioned above. Thus, the uni-fractality expressed as $y = ae^{-bx}$ was the general phenomena in the arrangement of four bases not only in the genomic-DNA, but also in the artificial-DNA generated with the same molar-contents of the four bases and the base numbers as each single-strand of the genomic DNA. Furthermore, the artificial chromosomal sequences had only the reverse-complement symmetry, but not the base bias, and we could not find any Open Reading Frames (ORFs) in the artificial chromosomes (Takeda & Nakahara, 2009; Takeda, 2009; Takeda, 2011; Takeda, 2012).

By contrast, the linearly decreased-phase (a power-law-tail) observed at long distances, the power-law-tail was unique and specific in individual real (active) genome for the behavior of the base-distribution for each four bases. The power-law-tail could be observed only in the active genomes or chromosomes even in the virus and the organelle, but never in the distribution of base(s) of the artificial-genomes or -chromosomes, i.e., each gene in the living cells could be expressed when all the structural features of a single-strand of the genomic DNA were present (Takeda & Nakahara, 2009, 2013; Takeda, 2009, 2011, 2012).

The effect of the adjacent base sequences of the gene(s) was observed in living cells. In eukaryotic cells, most of them had the non-coding sequences (the regulatory elements) which were occupied 2 ~ 98% of the entire genomic DNA (Mattick, 2004; Taft et al., 2007), and they were deeply affected on the adjacent base sequences to express the gene(s) by the appropriate stage and organ according to the information of the generation-rule for each genome, the reverse-complement symmetry, the bias and the multiple fractality of the base(s)-distribution, i.e., most of the regulatory elements (base-sequences) for the gene-expression, the cellular complexity and the inheritance of living cells were placed sophisticatedly in the non-coding regions of genome. As mentioned above, to generate an active genome, the genomic base sequences should be arranged to be present in these three characters in the genome structure together at least.

Recently, there were many non-coding small RNAs on genome and participated in the regulation of the gene-expression (Lai et al., 2005; Martens et al., 2005; Taft et al., 2007; Molnar et al., 2010). In addition, the ratio of these non-coding RNAs in genome were regulated to the individual gene expression, and showed to be affected on the complexity of the living cells (Mattick, 2004; Lynch, 2007). Many other regulatory sequences, enhancers, promoters, poly (A)-binding signal, MAR (SAR), insulators, introns etc of individual genes were present on genome (Webb et al., 1992, 2002; Levine & Tjian, 2003). These regulatory elements (base sequences)

should be compact or adaptable. Therefore, the regulatory elements were either not evident or were little observed in prokaryotic genomes, and were apparently different regulation, i.e., the transcription, the other genomic- and the molecular-events in cells, the gene-organization in genome, and each gene-expression from eukaryotic cells (Kozak, 1983; Niehrs & Pollet, 1999).

On considering the various regulatory sequences of individual genes on chromosome, it was very suggestive that the genome-scale analysis of human Down Syndrome Critical Region (DSCR) consisted of approximately 1 Mb on chromosome 21 could not be caused sometimes the syndrome in mouse (Olson et al., 2004, 2007). In other words, the observation based on the gene-scale was not always in accord with that based on the genome-scale. Such researches suggested that the chromosomal (genomic) DNA might be also a molecule with a huge molecular weight and a higher-order structure of chromosome including both the coding- and the non-coding regions, and it would be necessary to analyze the entire base sequences of genomes to understand the gene expression in the living cells (Takeda & Nakahara, 2009, 2013; Takeda, 2009, 2011, 2012; Nakahara & Takeda, 2010a, 2010b).

In this report we focused on whether the base sequences of the whole-genome with the multiple fractality, especially the power-law-tail in more than 16 bases of L-value (Takeda & Nakahara, 2009; Takeda, 2009, 2011) at the tail-end region of the distribution-curve of base(s) in genomic DNA could be related to respective species, and to be prompted the progress of the chromosomes and to create the diversity, or the complexity of the living cells.

2. Materials and Methods

2.1 Distribution Curve of Base A, T, C, or G in Genome or Chromosome

In this paper, we had used the base-distribution as the frequency $F(L)$ instead of the probability $P(L)$. In the A-base, the x-axis was L and the y-axis was the base frequencies $F(L)$ in double logarithmic plot of the equation 1, $y = ae^{-bx}$. The same calculations for the other bases were performed as previously reported (Takeda & Nakahara, 2009; Takeda, 2009, 2011, 2012; Nakahara & Takeda, 2010a, 2010b).

Each base-distribution curve $F(L)$ was expressed the distribution of the distance L between a base and the next base, for the base “A”, the L -value was corresponded the base numbers from “A” to the next “A” in the genomic DNA, and $F(L)$ was the sum of the frequency with the same base-distance in the genomic DNA as follows;

Ex. Adenine bases: A

Base Sequence:

```

      1          10          20          30          40
5'-AGTCGCTGGAGTCGTCGCTCCGTGGGTCCATTCCACTCGA
      50          60          70          80          90
TTGCTTTGATGGTCTCTAAATGGACCGATTGCGTCCGCCTCCTTGGGACA-3'
```

In the above base sequence consisting of 90 nt, the ‘ L ’ and the ‘ $F(L)$ ’ for ‘A-base’ were calculated as follows.

L	Position A	F (L)
1:	(⁵⁸ A- ⁵⁹ A), (⁵⁹ A- ⁶⁰ A)	2
2:	(⁸⁸ A- ⁹⁰ A)	1
4:	(⁶⁰ A- ⁶⁴ A), (⁶⁴ A- ⁶⁸ A)	2
5:	(³⁰ A- ³⁵ A), (³⁵ A- ⁴⁰ A)	2
9:	(¹ A- ¹⁰ A), (⁴⁰ A- ⁴⁹ A), (⁴⁹ A- ⁵⁸ A)	3
20:	(¹⁰ A- ³⁰ A), (⁶⁸ A- ⁸⁸ A)	2

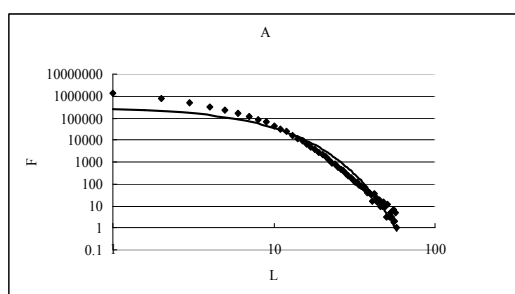
3. Results

3.1 Base-Distribution of the Various Chromosomes in Eukaryotic Cells

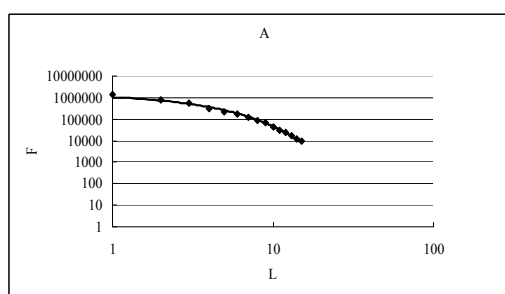
3.1.1 Adenine Base (A)

S. cerevisiae was maintained in three structural features in a single-strand of the genomic DNA of 16 chromosomes in nuclei as mentioned above, i.e., the GC-contents (%) for all of them were AT-rich, almost identical, approximately 38.0% irrespective their different chromosomal sizes. *S. cerevisiae* was basically an eukaryote consisting of a single-cell, and the 16 chromosomes of the cells were not large, a kind of small size of the genomic DNA (total base numbers were ca. 12,155,038 nt). In double logarithmic plot of L vs F, in the case of the L value was 1 through 15, the frequency F (L) of the base-distribution of the adenine base (A-base) was proportional to an exponential equation, $y = ae^{-bx}$ (eq.1, $x = \log L$, $y = \log F(L)$; a and b were constant). In the case of A-base in the *S. cerevisiae* genome, the a and b values were calculated from the equation 1 (eq.1) as $1.0 (E + 06)$ and 0.3446, respectively (Figure 1).

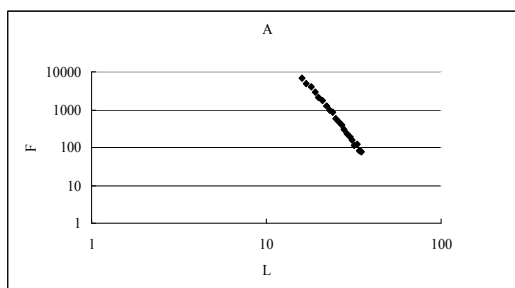
(1) *S. cerevisiae*, genome (12,155,038 nt)



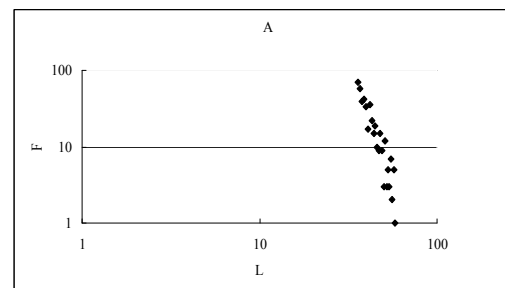
(a) $L = 1 \sim 58$



(b) $L = 1 \sim 15$, $y = (1.0E+06)e^{-0.3446x}$

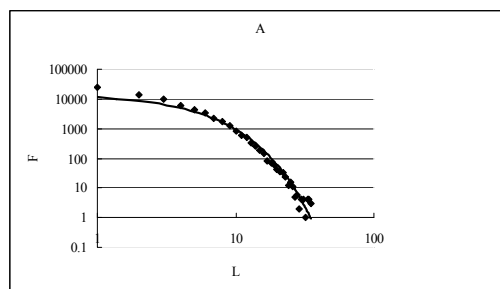


(c) $L = 16 \sim 35$, $y = -325.5x + W$

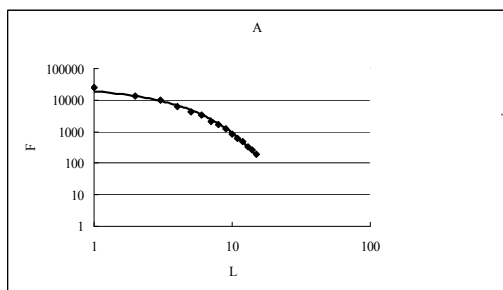
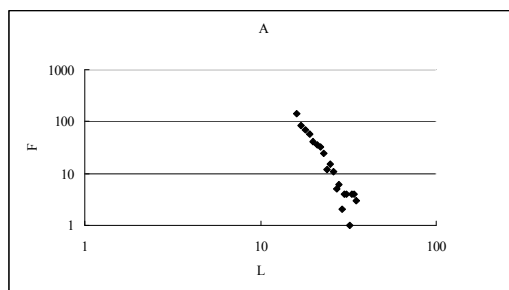
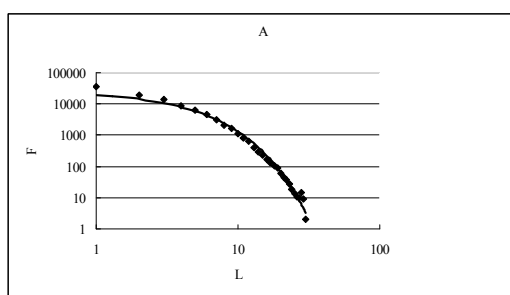
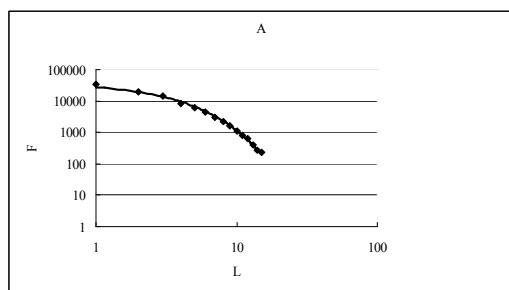
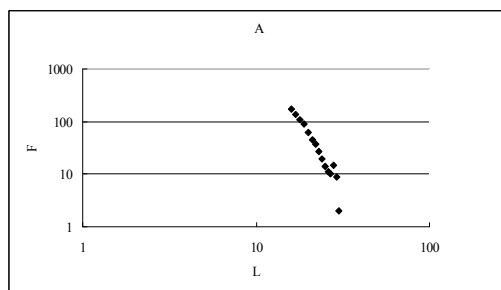
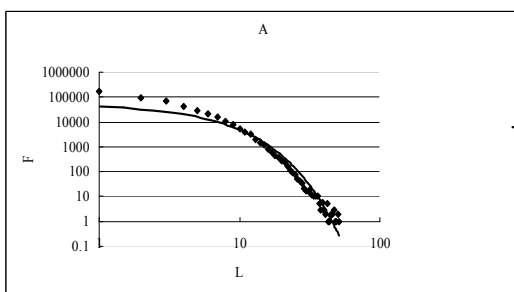
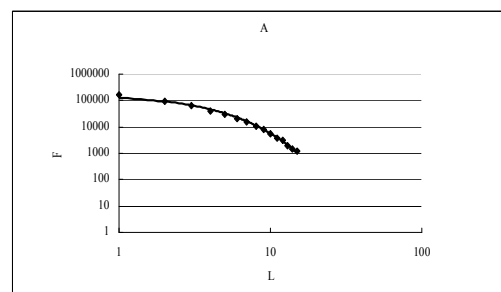


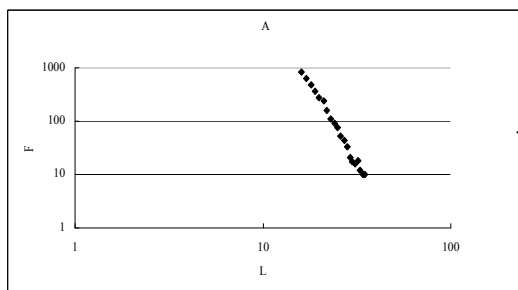
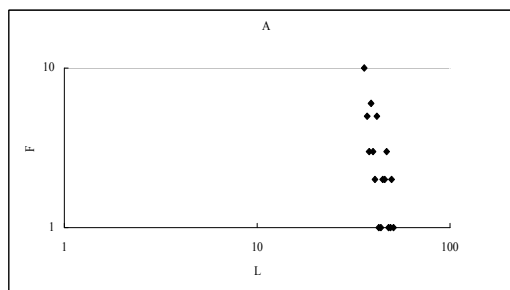
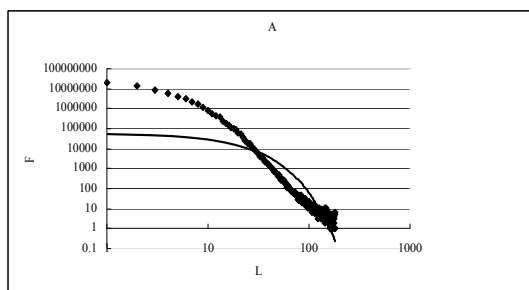
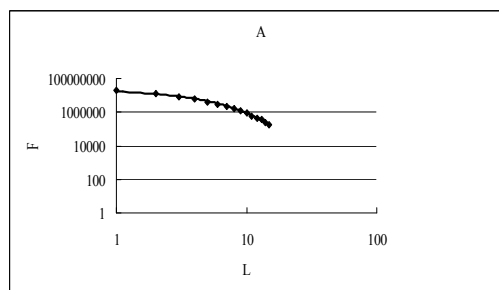
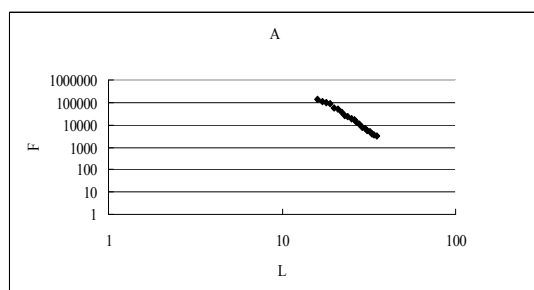
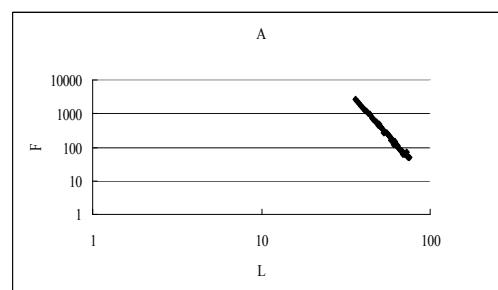
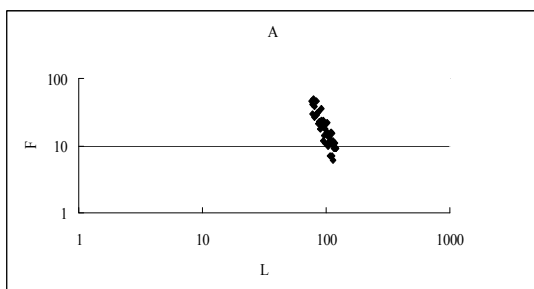
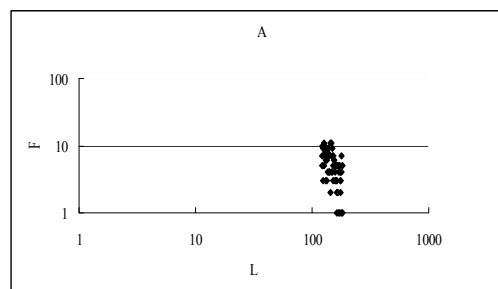
(d) $L = 36 \sim 58$, $y = -3.0x + W$

(2) *S. cerevisiae*, chromosome I (230,203 nt)



(a) $L = 1 \sim 35$

(b) $L = 1 \sim 15$, $y = 26086e^{-0.3365x}$ (c) $L = 16 \sim 35$, $y = -6.8x + W$ (3) *S. cerevisiae*, chromosome III (315,350 nt)(a) $L = 1 \sim 30$ (b) $L = 1 \sim 15$, $y = 38843 e^{-0.3516x}$ (c) $L = 16 \sim 30$, $y = -11.5x + W$ (4) *S. cerevisiae*, chromosome IV (1,531,929 nt)(a) $L = 1 \sim 51$ (b) $L = 1 \sim 15$, $y = 183748e^{-0.346x}$

(c) $L = 16 \sim 35$, $y = 40.3x + W$ (d) $L = 36 \sim 51$, $y = 0.5x + W$ (5) *H. sapiens*, chromosome I (221,348,412 nt)(a) $L = 1 \sim 182$ (b) $L = 1 \sim 15$, $y = (2.0E+07)e^{-0.3285x}$ (c) $L = 16 \sim 35$, $y = -7068.5x + W$ (d) $L = 36 \sim 75$, $y = -65.7x + W$ (e) $L = 76 \sim 120$, $y = -0.8x + W$ (f) $L = 121 \sim 182$, $y = -0.03x + W$

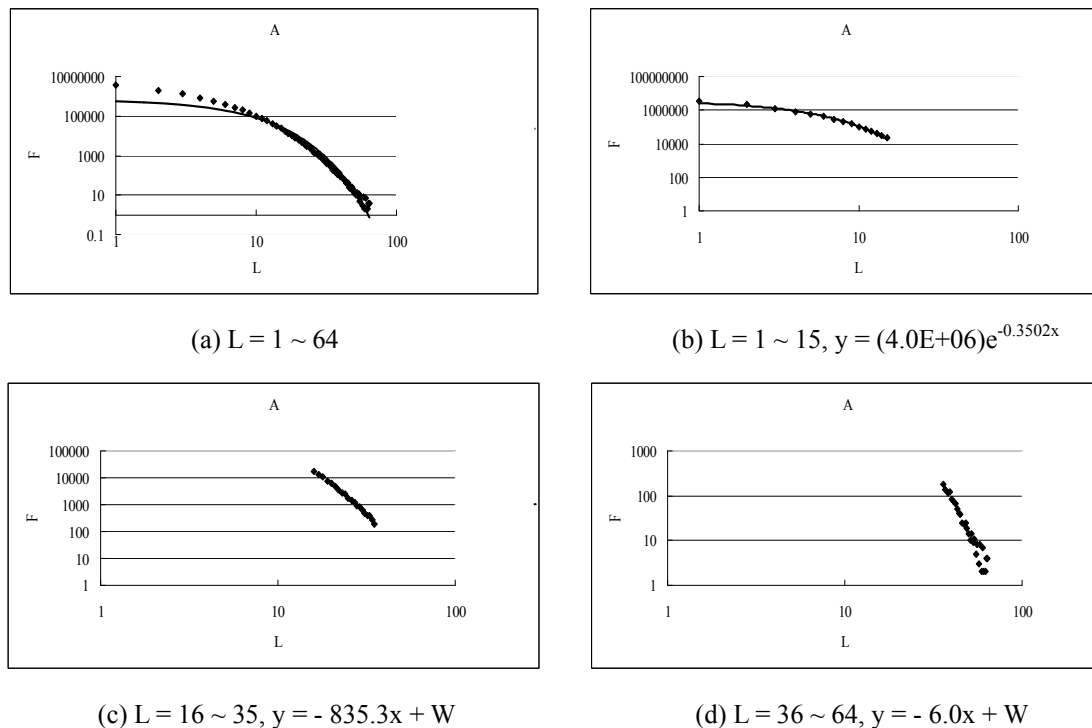
(6) *A. Thaliana*, chromosome I (30,263,279 nt)

Figure 1. Distribution curve of adenine nucleotide (A) in the base sequence

(1) *S. cerevisiae* genome, (2) *S. cerevisiae* chromosome I, (3) *S. cerevisiae* chromosome III, (4) *S. cerevisiae* chromosome IV, (5) *H. sapiens* chromosome I, (6) *A. thaliana* chromosome I.

In contrast, when the L value was more than 16, the distribution of base(s), $F(L)$ was different, i.e., gave a straight line, $y = Ux + W$ (Equation 2 = eq. 2; U was the slope and W was the intercept). Figure 1 showed the distribution of A-base in several eukaryotic chromosomes. In all chromosomes, when the L -value was used at maximum to be appeared the base-frequency in the genome, the distribution curve of the A-base was not proportional to the single-curved line expressed the uni-fractality (see the curved lines in (a)-panels of Figure 1-(1) ~ (6)). Therefore the L -value was partitioned (the detailed description about the L -partition as below). Figure 1-(1) showed the A-base in *S. cerevisiae* genome (12,155,038 nt, Figure 1-(1)) as described in Materials and Methods. Figure 1-(2) ~ (4) showed the A-base-distribution in several chromosomes of *S. cerevisiae*. Figure 1-(2) showed the A-base in chromosome I of *S. cerevisiae* (230,203 nt). In the case of the short-distance (Figure 1-(2)-(b), $L = 1 \sim 15$ bases) the frequencies of the A-base in chromosome I decreased exponentially, whereas in the case of the long-distance (Figure 1-(2)-(c), $L = \text{more than } 16$ bases), the frequencies of the A-base in chromosome I decreased linearly (power-law-tail). In addition, both regions were overlapped around $L = 11 \sim 15$ (Takeda & Nakahara 2009; Takeda 2009). The small chromosomes III (315,350 nt, Figure 1-(3)) of *S. cerevisiae* were also observed in both the exponentially- and the linearly-decreased boundary regions of the A-base such as chromosome I (Figure 1-(2)) in a single-strand of the DNA. The frequencies of the other three bases, T, G or C were basically similar to the A-base as A to T, and as G to C, respectively, in one organism. (Tables 1 and 2).

Table 1. The a, b and slope values of 4 nucleotides in eukaryotic genomes

	Size (nt)	GC (%)	Nucleotide	exponentially decreased		linearly decreased			
				L = 1~15		L = 16~35	L = 36~75		L = 76~120 L = 121~
				a	b	slope 1(-)	slope 2(-)	slope 3(-)	slope 4(-)
H. sapiens, chr.1 (Human)	221,229,757	41.7	A	2.00E+07	0.3285	7068.50	65.67	0.82	0.03
			T	2.00E+07	0.3293	6993.80	62.22	1.02	0.15
			C	1.00E+07	0.2419	14158.60	331.85	3.33	0.20
			G	1.00E+07	0.2418	14154.80	337.92	3.00	0.14
H. sapiens, chr. 2 (Human)	237,440,784	40.2	A	3.00E+07	0.3355	7081.15	57.80	0.69	0.07
			T	3.00E+07	0.3366	6999.65	56.25	0.82	0.08
			C	1.00E+07	0.2324	15765.55	369.75	4.02	0.21
			G	1.00E+07	0.2327	15646.50	364.42	3.75	0.26
H. sapiens, chr. 22 (Human)	33,476,901	47.8	A	3.00E+06	0.2987	1361.60	15.12	0.09	
			T	3.00E+06	0.2976	1381.80	15.77	0.16	
			C	2.00E+06	0.2762	1747.00	36.32	0.62	
			G	2.00E+06	0.2761	1756.45	34.07	0.29	
H. sapiens, chr. X (Human)	114,997,816	39.6	A	1.00E+07	0.3398	3315.50	24.47	0.53	
			T	1.00E+07	0.3399	3329.25	25.87	0.73	
			C	6.00E+06	0.2274	7916.90	169.90	1.71	0.06
			G	6.00E+06	0.2279	7967.90	175.10	1.51	0.17
R. norvegicus Chr.1 (Rat)	255,141,439	42.6	A	3.00E+07	0.3214	8646.80	86.67	5.04	0.39
			T	3.00E+07	0.3219	8584.55	87.85	4.42	0.38
			C	1.00E+07	0.2452	15370.10	314.62	7.91	0.40
			G	1.00E+07	0.2455	15384.55	315.35	7.78	0.25
M. musculus Chr.1 (Mouse)	184,262,511	41.0	A	2.00E+07	0.3303	5934.80	55.50	2.89	0.22
			T	2.00E+07	0.3299	5956.50	56.37	3.15	0.20
			C	1.00E+07	0.2375	11555.55	249.57	6.40	0.40
			G	1.00E+07	0.2373	11631.20	251.55	6.15	0.23
M. musculus Chr.2 (Mouse)	113,587,859	42.4	A	1.00E+07	0.3234	3891.85	36.12	1.42	0.42
			T	1.00E+07	0.3225	3928.70	36.70	1.49	0.48
			C	6.00E+06	0.2453	6800.65	147.27	3.44	0.25
			G	6.00E+06	0.2450	6821.15	147.17	2.91	0.20
B.taurus (Cow)	213,056,107	42.9	A	2.00E+07	0.3213	7640.30	66.32	1.64	0.07
			T	2.00E+07	0.3220	7642.70	91.85	2.66	0.15
			C	1.00E+07	0.2528	12418.05	255.35	4.64	0.15
			G	1.00E+07	0.2416	13639.60	357.37	6.82	0.50
X. tropicalis Raw (Frog)	52,589,696	45.9	A	5.00E+06	0.3012	2255.10	20.15	0.86	0.15
			T	4.00E+06	0.2938	2412.95	39.85	1.86	0.18
			C	3.00E+06	0.2597	3009.45	40.71	1.40	0.15
			G	4.00E+06	0.2664	2788.45	37.07	1.17	0.24
F. rubripes (Fish)	252,556,882	45.2	A	2.00E+07	0.3058	10521.30	109.05	1.93	0.13
			T	2.00E+07	0.3060	10438.05	95.15	2.33	0.16
			C	2.00E+07	0.2605	14103.25	182.69	2.46	0.15

			G	2.00E+07	0.2606	14114.10	182.38	3.13	0.20
Danio rerio	36,329,259	36.2	A	5.00E+06	0.3552	976.30	6.00	0.26	
(Fish)			T	4.00E+06	0.3531	998.15	6.90	0.09	
			C	2.00E+06	0.2159	2686.35	63.07	1.64	0.28
			G	2.00E+06	0.2154	2532.30	63.35	1.66	0.27
D. melanogaster	22,214,931	41.9	A	2.00E+06	0.3229	778.10	8.53		
Chr2L			T	2.00E+06	0.3225	775.50	8.95	0.07	
(Insect)			C	1.00E+06	0.2444	1324.50	25.55	0.28	
			G	1.00E+06	0.2447	1316.85	27.00	0.09	
D. melanogaster	20,298,206	43.4	A	2.00E+06	0.3164	738.00	8.77		
Chr2R			T	2.00E+06	0.3159	746.50	8.72		
(Insect)			C	1.00E+06	0.2545	1139.65	19.97	0.35	
			G	1.00E+06	0.2546	1144.25	19.20	0.53	
D. melanogaster	18,929,811	42.0	A	2.00E+06	0.3226	686.15	9.27		
Chr3L			T	2.00E+06	0.3233	670.00	8.60		
(Insect)			C	1.00E+06	0.2466	1121.55	22.80	0.33	
			G	1.00E+06	0.2458	1133.40	22.60	0.67	
D. melanogaster	27,900,790	42.9	A	3.00E+06	0.3194	1008.35	13.60	0.14	
Chr3R			T	3.00E+06	0.3182	1024.10	13.42	0.08	
(Insect)			C	2.00E+06	0.2517	1598.55	29.47	0.33	
			G	2.00E+06	0.2512	1603.35	31.07		
A. Gambiae Chr 2	109,306,628	44.9	A	1.00E+07	0.3075	4377.50	68.80	0.75	
(Insect)			T	1.00E+07	0.3076	4390.45	71.77	0.44	
			C	7.00E+06	0.2679	5773.85	110.52	0.95	0.04
			G	7.00E+06	0.2679	5821.80	108.15	0.84	0.13
A. Gambiae Chr 3	92,345,735	43.9	A	9.00E+06	0.3132	3593.70	54.52	0.51	0.06
(Insect)			T	9.00E+06	0.3127	3584.95	53.40	0.64	0.07
			C	6.00E+06	0.2614	5106.25	99.22	0.80	0.12
			G	6.00E+06	0.2614	5108.20	100.25	0.66	0.07
A. Gambiae Chr X	21,477,789	46.1	A	2.00E+06	0.3044	852.35	19.70	0.20	
(Insect)			T	2.00E+06	0.3050	849.70	20.15	0.39	
			C	2.00E+06	0.2765	1062.65	22.37	0.22	
			G	2.00E+06	0.2776	1054.50	22.72	0.62	
C. elegans Chr1	15,080,469	35.7	A	2.00E+06	0.3412	421.00	3.64		
(Worm)			T	2.00E+06	0.3424	421.90	4.04		
			C	6.24E+05	0.2131	1028.85	27.62	2.44	
			G	6.23E+05	0.2131	1048.30	28.47	0.67	
C. elegans Chr2	15,279,297	36.2	A	2.00E+06	0.3432	427.70	3.50		
(Worm)			T	2.00E+06	0.3428	422.05	4.24		

			C	6.47E+05	0.2147	1039.35	28.27	0.50
			G	6.47E+05	0.2150	1056.70	27.95	0.26
C. elegans Chr3	13,777,772	35.6	A	2.00E+06	0.3423	386.50	3.00	
(Worm)			T	2.00E+06	0.3391	392.35	3.30	
			C	5.62E+05	0.2117	962.45	26.37	0.24
			G	5.70E+05	0.2129	940.60	26.07	0.44
C. elegans Chr4	17,493,725	34.6	A	2.00E+06	0.3477	459.15	3.57	
(Worm)			T	2.00E+06	0.3488	452.05	2.50	
			C	6.76E+05	0.2066	1285.85	35.57	0.37
			G	6.67E+05	0.2042	1283.95	35.97	0.44
C. elegans Chr5	20,922,239	35.4	A	3.00E+06	0.3477	554.60	3.53	
(Worm)			T	3.00E+06	0.3480	546.55	3.69	
			C	8.41E+05	0.2081	1541.65	37.92	0.54
			G	8.35E+05	0.2073	1521.80	37.82	0.48
C. elegans ChrX	17,705,012	35.2	A	2.00E+06	0.3488	476.60	3.51	
(Worm)			T	2.00E+06	0.3482	480.85	3.73	
			C	6.96E+05	0.2061	1326.60	32.87	0.43
			G	6.95E+05	0.2059	1313.15	31.42	0.31
A. thaliana Chr1	30,263,279	35.8	A	4.00E+06	0.3502	835.35	6.00	
(Plant)			T	4.00E+06	0.3500	847.45	6.03	
			C	1.00E+06	0.2133	2102.95	54.90	1.04
			G	1.00E+06	0.2141	2094.95	55.97	0.76
A. thaliana Chr2	19,643,390	35.8	A	2.00E+06	0.3508	542.70	3.91	
(Plant)			T	2.00E+06	0.3507	537.40	4.58	
			C	8.18E+05	0.2138	1384.45	35.32	0.46
			G	8.10E+05	0.2130	1380.30	35.44	0.34
A. thaliana Chr3	23,463,327	36.3	A	3.00E+06	0.3490	642.85	5.32	
(Plant)			T	3.00E+06	0.3476	665.00	5.77	
			C	9.96E+05	0.2151	1648.55	37.06	0.88
			G	9.97E+05	0.2152	1636.60	36.98	0.63
A. thaliana Chr4	17,549,525	36.0	A	2.00E+06	0.3511	496.25	3.85	
(Plant)			T	2.00E+06	0.3486	493.20	4.00	
			C	7.35E+05	0.2142	1225.65	30.27	3.00
			G	7.32E+05	0.2137	1232.60	32.70	0.55
A. thaliana Chr5	26,689,407	35.9	A	3.00E+06	0.3500	748.20	5.62	
(Plant)			T	3.00E+06	0.3505	752.05	6.31	
			C	1.00E+06	0.2129	1870.95	51.22	1.14
			G	1.00E+06	0.2140	1877.05	48.00	0.71
S. cerevisiae Chr1	230,203	39.2	A	2.61E+04	0.3365	6.85		
(Fungi)			T	2.64E+04	0.3395	6.55		
			C	1.08E+04	0.2210	16.35	1.27	
			G	1.14E+04	0.2255	14.65	0.40	
S. cerevisiae Chr2	813,139	38.3	A	9.51E+04	0.3418	22.75		

(Fungi)				T	9.71E+04	0.3452	20.65	
				C	3.78E+04	0.2181	57.60	1.17
				G	3.60E+04	0.2127	54.10	2.16
S. cerevisiae Chr3	315,350	38.5		A	3.88E+04	0.3516	11.53	
(Fungi)				T	3.57E+04	0.3361	8.00	
				C	1.54E+04	0.2247	21.15	0.30
				G	1.38E+04	0.2126	23.35	0.87
S. cerevisiae Chr4	1,531,929	37.9		A	1.84E+05	0.3460	40.35	0.56
(Fungi)				T	1.81E+05	0.3432	41.55	0.60
				C	6.69E+04	0.2112	110.75	2.76
				G	6.82E+04	0.2133	108.10	2.34
S. cerevisiae Chr15	1,091,284	38.1		A	1.32E+05	0.3469	28.35	0.25
(Fungi)				T	1.26E+05	0.3396	29.00	0.07
				C	4.98E+04	0.2167	76.10	2.07
				G	4.84E+04	0.2126	74.75	1.75
S. pombe Chr1	5,566,797	36.0		A	7.00E+05	0.3532	137.45	1.07
(Fungi)				T	6.96E+05	0.3525	138.5	1.87
				C	2.29E+05	0.2075	402.95	8.31
				G	2.27E+05	0.207	410.4	7.5
S. pombe Chr2	4,467,299	36.0		A	5.60E+05	0.3529	109.2	0.82
(Fungi)				T	5.65E+05	0.3543	103.75	1.2
				C	1.83E+05	0.2074	324.4	5.97
				G	1.82E+05	0.2069	330.8	6.45
S. pombe Chr3	2,455,984	36.2		A	3.03E+05	0.3503	63.05	0.7
(Fungi)				T	3.04E+05	0.3509	60.3	0.36
				C	1.02E+05	0.2093	183.7	4.13
				G	1.03E+05	0.2107	177.7	3.57

Table 2. The a, b and slope values of 4 nucleotides in various genomes

	genome size(nt)*	GC (%)	a**	b	slope1(-)** *	slope2(-)** **	slope3(-)** ***	slope4(-)* *****	L-value *****
Adenine (A)									
Adenovirus	34,125	46.5	3169	0.3046	1.63E+00				26
P. aerophilum	2,222,430	51	151896	0.2603	5.30E+01	1.00E+00			52
E. coli	4,639,221	50.8	343678	0.2715	3.48E+00	2.00E+00			50
S. cerevisiae	12,155,038	38.1	1.00E+06	0.3446	3.26E+02	3.00E+00			58
P. falciparum	22,812,541	19.4	7.00E+06	0.4317	5.16E+02	1.39E+01			75
C. elegans	100,258,514	35.4	1.00E+07	0.3455	2.73E+03	1.46E+01			75
D. melanogaster	112,348,412	42.5	1.00E+07	0.3208	4.03E+03	5.45E+01	5.00E-01		95
A. Thaliana	117,226,760	36	1.00E+07	0.3502	3.27E+03	1.79E+01	2.00E-01		86
H. S. chr1	221,348,412	41.7	2.00E+07	0.3285	7.07E+03	6.57E+01	8.20E-01	3.00E-02	182
Thymine (T)									
Adenovirus			2720	0.283	1.39E+00				33

P. aerophilum	146698	0.2558	5.52E+01	1.65E+00			52
E. coli	341225	0.2706	2.09E+02	2.26E+00			50
S.cerevisiae	1.00E+06	0.3434	3.25E+02	1.00E+00			58
P. falciparum	7.00E+06	0.4315	5.11E+02	1.40E+01	3.30E-01		90
C. elegans	1.00E+07	0.3453	2.72E+03	1.59E+01	1.33E-01		90
D. melanogaster	1.00E+07	0.3205	4.33E+03	5.48E+01	3.00E-01		106
A. Thaliana	1.00E+07	0.3495	3.30E+03	1.99E+01	1.00E-01		95
H. S.chr1	2.00E+07	0.3293	6.99E+03	6.22E+01	1.02E+00	1.40E-01	186
Cytosine (C)							
Adenovirus	2500	0.2766	1.44E+00				33
P. aerophilum	155853	0.2656	5.12E+01	2.36E+00			49
E. coli	410983	0.2993	1.66E+02	1.47E+00			56
S.cerevisiae	549991	0.2159	8.45E+02	1.30E+00	2.10E-01		108
P. falciparum	485276	0.1379	2.32E+03	2.40E+02	2.08E+01	1.47E+00	225
C. elegans	4.00E+06	0.2096	7.18E+03	1.88E+02	1.40E+00		117
D. melanogaster	7.00E+06	0.2497	6.49E+03	1.29E+02	7.40E-01		114
A. Thaliana	5.00E+06	0.2138	8.23E+03	2.13E+02	3.13E+00	8.00E-01	135
H. S.chr1	1.00E+07	0.2419	1.42E+04	3.31E+02	3.35E+00	2.00E-01	205
Guanine (G)							
Adenovirus	2317	0.2662	1.57E+00				29
P. aerophilum	161008	0.2699	4.43E+01	1.08E+00			60
E. coli	409810	0.2986	1.60E+02	1.52E+00			56
S.cerevisiae	547570	0.2153	8.43E+02	1.37E+01	1.60E-01		100
P. falciparum	498353	0.1401	2.35E+03	2.38E+02	2.37E+01	2.86E+00	180
C. elegans	4.00E+06	0.2093	7.16E+03	1.88E+02	1.38E+00	1.60E-01	132
D. melanogaster	7.00E+06	0.2494	6.51E+03	1.32E+02	7.60E-01		109
A. Thaliana	5.00E+06	0.214	8.22E+03	2.18E+02	2.89E+00	4.10E-01	144
H. S.chr1	1.00E+07	0.2418	1.42E+04	3.38E+02	3.00E+00	1.60E-01	214

*: Base number;

** : Frequency (F) at L in genome (virus, L = 1-10; archaea & eubacteria, L = 1-20; eukaryotes, L = 1-15);

***: Slope at L = 16-35 (adenovirus, 11-26; archaea & eubacteria, L = 21-35; eukaryotes, L = 16-35);

****: Slope at L = 36-75;

*****: Slope at L = 75-120;

*****: Slope at L = 121 and more;

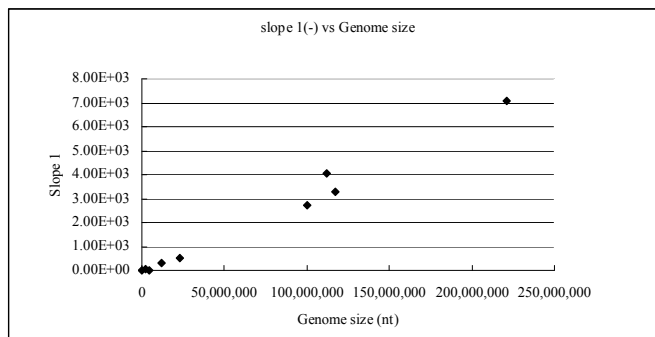
*****: Max. L-value estimated the base-frequency.

Another 13 chromosomes and mitochondrial DNA (mtDNA) of *S. cerevisiae* showed the same multiple fractality at the boundary of 11 ~ 15 bases as chromosome I (data not shown). In the largest chromosomes of *S. cerevisiae*, chromosome IV (1,531,929 nt, Figure 1-(4)) was needed to divide into the base-distribution.

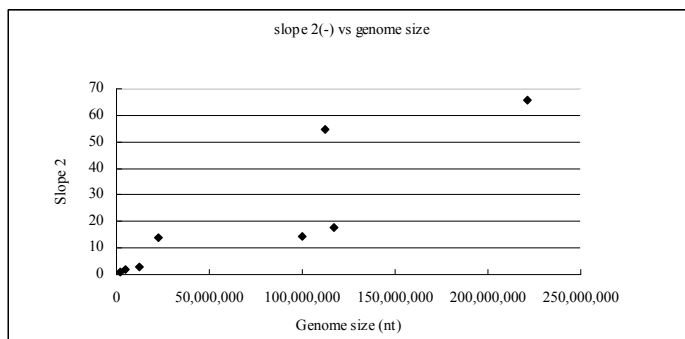
The presence of the plural power-law-tail regions was shown the frequencies of each four bases of the larger eukaryotic genomes and chromosomes (Figure 1-(5), 1-(6), Tables 1 and 2). They had the long power-law-tails, and partitioned at L = more than 16 bases. These long power-law-tail regions were observed in other eukaryotes composed of larger size genome than *S. cerevisiae* (Tables 1 and 2, Figure 1). In these genomes, the long power-law-tail regions were partitioned at L = 16 ~ 35 (designated as slope 1), 36 ~ 75 (slope 2), 76 ~ 120 (slope 3) and 121 ~ more than 121, usually max. L-value was used (slope 4). The L-value became larger, the

slope-value became smaller, for instance, slope 1 was usually larger than slope 2. In all genomes, the a- b- and U-values were shown identically as A to T, and as G to C in a single-strand of DNA, respectively, in one chromosome or genome (Tables 1 and 2, Figure 1).

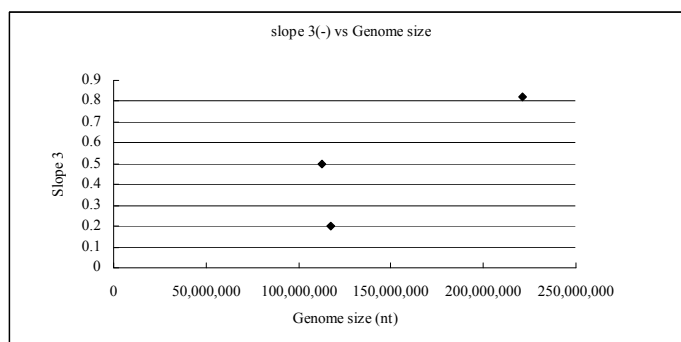
(1) Slope 1 vs Genome size



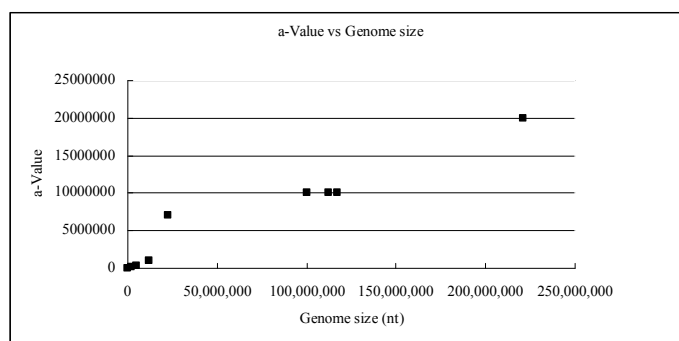
(2) Slope 2 vs Genome size



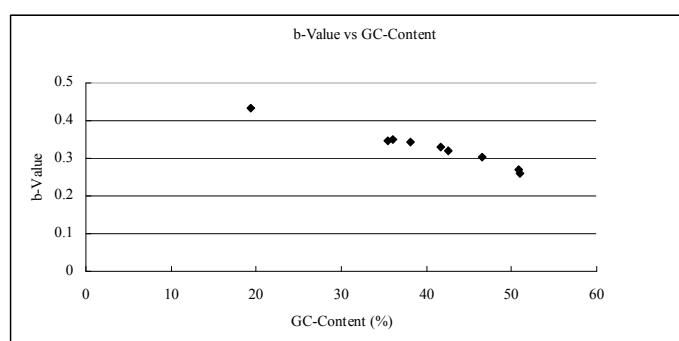
(3) Slope 3 vs Genome size



(4) a-Value vs Genome size



(5) b-Value vs GC-Content



(6) Slope 1 vs a-Value

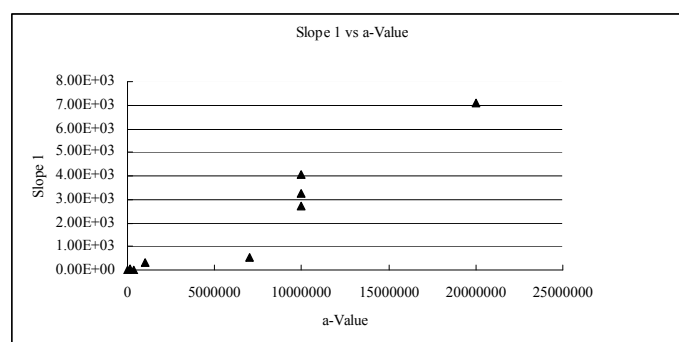


Figure 2. Relation between the a-, b- and the U-values of various species based on Table 2
U-values (slope) were used the slope 1, slope 2, and slope 3.

Figure 2 showed the relationship of each value in various genomic DNA based on the data in Table 2. The slope 1 (U-value) was proportional to the genome-size (nt) (Figure 2-(1)), and the b-value was proportional to the GC-content (%) of the genomic DNA (Figure 2-(5)). But the slope 2 and slope 3 were not proportional to the genome-size (nt) as the slope 1 (Figure 2-(2) and Figure 2-(3)). The a-value was also proportional to the genome-size (nt) as the slope 1 did (Figure 2-(4)). The results might indicate the difference of the base-frequencies in the non-coding, the power-law-tail regions of genome. These values might be variable on the account of the length and the base-sequences of the A-, T-, C- and G-bases in the regions, i.e., slope 2 might be GC-poor and AT-rich a non-coding region. The non-coding region contained the regulatory elements consisting of the specific base sequences and the repetitious base sequences.

Table 1 showed the a-, b- and the U-values in the large chromosome or genomes.

These large eukaryotic chromosome had the long power-law-tail region, and could be observed not less than two at the linearly-decreased region, although the frequency $F(L)$ of each base-distribution of A, T, G or C was proportional to an exponential equation, $y = ae^{-bx}$ (eq.1, $x = \log L$, $y = \log F(L)$; a and b were constant) at $1 \sim 15$ of the L-values. In other words, in the large chromosomes with the power-law-tail region, the region should be partitioned two or more to be appeared as above described (Tables 1 and 2, Figure 1).

Of the larger eukaryotic chromosomes, the power-law-tail region in plant-genome such as *A. thaliana* and *O. sativa*, might be two or three phases, although the plant-genomes were different from the genome-size (nt) of the animals (Tables 1 and 2, Figure 1).

Most eukaryotic cells had larger chromosomes or genomes than *S. cerevisiae*, and the each four bases distribution was shown in Tables 1 and 2. As shown in Table 2, the a-, b- and U-values were varied in both the eukaryotic- and the prokaryotic-cells. Interestingly, these values were almost identical with those of the complement bases, i.e., A- to the T-, the G- to the C, even in a single-strand of the genomic DNA of the individual cells. In addition, the b-value was proportionally related to the GC-content (%) of the genome or the chromosome, but the a-value was related to the chromosome-size (nt) of the individual cells (Figure 2). These features of the a- and b-values were observed in both the eukaryotic- and prokaryotic-cells (Table 2). In all cases, the a, b and U-values of A and T, G and C were almost identical in the genomic DNA (Table 2).

Generally, the genomes or the chromosomes of the eukaryotic cells were of two types. One possessed the long linearly-decreased regions (the power-law-tail), and the others possessed the longer linearly-decreased region (the large power-law-tail). The former eukaryotic cells were of plants, and the later eukaryotic cells were of primates, mammalians and vertebrates. The power-law-tail region could be partitioned two or more according to the base-distance, "L" (Tables 1 and 2). The longer the chromosome's L-values was, the fractality exhibited more phases.

Table 3 showed the a-, b- and the U-values of the archaea and the eubacteria genomes. These values were varied and diverse in the archaea in comparison with the eubacteria because most of archaea could be come to live under the severe conditions. In addition, the difference between the slope 1 and 2 might be derived from the dispersion of four bases in the non-coding regions containing the elements to be necessary to execute various extreme regulations in the cells. Thus, the archaea might be more similar to the eukaryotes in comparison with the eubacteria in the genome-level (Tables 1 - 3).

Table 3. The a-, b- and U (slope)-values of A-base in archaea and eubacteria genomes

(Archaea)			L = 1-20		L = 21-35	L = 36-	Lmax*
Species	genome size (nt)	GC (%)	a	b	slope 1(-)	slope 2(-)	
Afulgidus	2,178,400	48	176538	0.2832	38.8	1.22	44
Halo NRC1	2,014,239	67	63862	0.1855	156.5	4.37	79
MacetiC2	5,751,492	42	510846	0.2966	80.4	2.33	47
MdeltaH	1,751,377	49	150066	0.293	24.9	0.43	42
Mjannaschii	1,664,970	31	227298	0.3707	6.9		37
Mmazei	4,096,345	41	381333	0.3037	53.2	1.4	44
NequitansKin4M	490,885	31	56528	0.3407	4.2		31
Paerophlum	2,222,430	51	151896	0.2603	53	1	52
Phorikoshi	1,738,505	41	188211	0.3283	16.2		39
Ssolfataricus	2,992,245	35	411641	0.3715	13.8		37
Stokodaii	2,694,756	32	380434	0.3773	10.8	0.28	41
Tacidphilum	1,564,906	45	148272	0.3078	20		38
Tvolcanium	1,584,804	39	178119	0.3351	13.9		36

(Eubacteria)			a	b	slope 1(-)	slope 2(-)	Lmax*
Species	genome size (nt)	GC (%)					
A. tumefaciens (cir)	2,841,541	59.4	136974	0.22	115.73	3.54	61
A. tumefaciens (lin)	2,074,782	59.3	101140	0.2209	82.06	2	55
B. burgdorferi	910,681	28.6	121765	0.3651	5.33		38
B. garinii	904,246	28.3	123854	0.3699	4.4		36
B. halodurans	4,202,357	43.7	347701	0.2976	61.2	1.5	45
B. japonium	9,518,208	64.1	371126	0.1997	507.93	12.81	72
B. subtilis	4,214,811	43.5	350231	0.2869	71	1.21	49
C. acetobutylicum	3,940,880	30.9	590746	0.3864	16.8		36
C. muridarumNigg	1,072,931	40.3	109800	0.3161	10.67	0.25	39
C. trachomatisAHAR	1,044,459	41.3	104760	0.3135	10.8	0.28	42
C. trachomatisDUW	1,042,519	41.3	104704	0.3137	11		39
E. coli	4,639,221	50.8	336718	0.2685	91.53	2	50

*: Maximum L-value estimated A-base-frequency.

3.1.2 Thymine (T), Guanine (G) and Cytosine (C) Bases

The other three bases, the thymine base (T), the guanine base (G), and the cytosine base (C) of the eukaryotes, the eubacteria and the archaea were of a similar patterns as the A-base. In the T-base, the distribution-curve, a-, b- and U-values were respectively identical with the A-base, and also, the distribution-curve, a-, b- and U-values of the C-base were respectively identical with the G-base as previously mentioned (Tables 1 and 2). Table 2 showed the distribution of four bases in both the prokaryotic- and the eukaryotic cells reported (NCBI genome database, 2013; The Sanger Institute, 2013; KEGG Organisms, 2013; Saccharomyces Genome Database, 2013). Thus, the structural features (generation-rule) of the base contents of genomic DNA should have the same features in any organisms corresponding to those organisms, as shown in the previous paper (Takeda & Nakahara, 2009; Takeda, 2009, 2011).

3.2 Biological Meaning of *a*-, *b*-, and *U*-Values

Both *a*-, and *b*-values of each chromosomes of the cells were almost the same in an organism although in the case of *H. sapiens* chromosomes, the *a*- and *b*-values were slightly different from 24 chromosomes because these values might be related to the GC-content (%) of each chromosome (Table 1).

The *U*-value was the slope at the long distance of the base distribution in chromosome, and also related to the base-contents of each chromosome. But this power-law-tail region of the base(s) was unique and has not been found in any chromosome yet. In other words, the region was the power-law-tail distributed four bases in genome, and essential in individual chromosomes of the cells (Tables 1 and 2, Figure 1).

In the case of the distribution of the base(s) in genome, the linearly-decreased region of “L” (the power-law-tail) was present to the long distances in double logarithmic plot of *L* (long-foot region of the distance of a base to the next base) vs *F* (*L*) in all organisms. The power-law-tail was more readily observed in the genomes from the eukaryotic cells. The other three bases, the T, the G and the C were distributed with the similar multiple fractality (equation 1, equation 2 = power-law-tail) in each chromosome and mtDNA of *S. cerevisiae* (data not shown).

In every genome, the *a*-, *b*- and *U*-values of the A-base to those of the T-base were equal, and those of the G-base and the C-base were equal in the Equation 1 and the Equation 2 (Tables 1 and 2). The *b*-value was also correlated with the GC-content of each chromosome (Figure 2). In the AT-rich genomes such as fungi and large genomes such as the plants, the slope 3 was observed slightly in the A- and T-bases, but observed clearly in the C- and the G-bases although it was due to escape out of the slope 3 to consist of low base numbers. Thus, even in the smaller genomes and the plant genomes had two or three linearly-decreased regions (power-law-tails) (Tables 1 and 2, Figure 1).

In addition, the *a*-, *b*- and *U*-value might be maintained, parallel to the genome-size (nt) and the GC-content (%) in one chromosome. *H. sapiens* chromosome 1 composed of 247,249,719 nt were registered in NCBI as 39-divisions because of its largeness at July 30, 2009. Out of 39 divisions, we selected three contigs, NT_004610.18 (12,702,424 nt), NT_032977.8 (73,835,825 nt) and NT_004487.18 (56,413,061 nt) of *H. sapiens* chromosome 1 in this analysis.

The same facts that the *a*-, *b*- and *U*-values were equal in each contig were observed in other chromosomes of *H. sapiens*. Not only the large eukaryotic genomes such as mouse, rat, dog, chicken, cow, fish, insects, plants and so on, but also the small genomes such as fungi genomes were observed generally the slope 2 although they were not easy to distinguish the slope 2 because of the small base numbers (Table 1). Furthermore, the two slopes in the power-law-tails were specific to the individual eukaryotic genome, therefore, the power-law-tails were useful to classify cells based on the distribution of the four bases in genomes.

3.3 Power-Law-Tail in the Large Genome

In large genomes such as mouse and human, the power-law-tail could be observed different in the distribution of four bases in the genome base sequences. Thus, the large linearly-decreased region of “L” (power-law-tail) could be composed of two or more different slopes (Tables 1 and 2, Figure 2). These two regions were overlapped around at *L* = 35, and the slope 1 was usually larger than the slope 2. Slopes 1, 2 and 3 seemed to be distributed variable base-sequences. Therefore slopes 1, 2, 3 and 4 were reflected on the different base-distribution, respectively. Each slope-region might be performed respective different regulatory phenomena in the cells. The slope-values, especially the slope 1, might be reflected on the genome-size (nt), or the synergism of the genome-size (nt) and the GC-content (%) (Table 1, Figure 2). In AT-rich genomic DNA as *A. thaliana* chromosomes, the slopes 3 and 4 were not observed from the frequencies of the A- and T-base in the power-law-tail regardless the genome-size (nt) (Table 1).

In every case, the slopes 1 ~ 4 existed in the non-coding region, and the four bases in the regions were placed to form the power-law-tails, and the base-complement even in a single-strand of DNA (Tables 1 and 2, Figures 1 and 2).

4. Discussion

In large genomes, the power-law-tail could be partitioned into two or more with the different slopes around 35, 75, 120 bases-distances in double logarithmic plot of *L* (the distance of a base to the next base) vs *F* (*L*), and these might be related to the evolution of genomes, or the complexity of the cells. The larger genomes such as *H. sapiens* were big non-coding regions in the genome over 97% (Raphael et al., 2008; Loots & Ovcharenko, 2010), and the non-coding regions should be essential to express the gene(s) in precisely, rapidly and steady.

The chromosomes of fungi were displayed the power-law-tail clearly, but most of eubacterial genomes were

apparently obscure the existence of the power-law-tail region. The archaea genomes possessed longer power-law-tail regions than those of eubacteria (prokaryotes) and close to the eukaryotic genomes although they were prokaryotic cells. The eukaryotic genomes were essentially composed of a great number of bases, and revealed a long power-law-tail region. Of eukaryotic cells, in *A. thaliana* and *O. sativa* genomes as higher plants, the power-law-tails were observed not so large within 80 bases, and the genomes of the higher plants have shown two different slopes (ex. slopes 1 and 2) the linearity of the boundary around 35 bases (Tables 1 and 2). By contrast, not only in *H. sapiens* and *M. musculus* such as the mammals, but also in other eukaryotic cells, animals, fishes, insects, worms or protozoa genomes, the region with long power-law-tail was more than 100 bases, and it might be observed two or more phases around 75, and 120 bases with the linearity even in each chromosome itself (Tables 1 and 2).

As shown in Table 2, the a-value of *P. aerophilum* (archaea) was contrary in comparison with the genome-size (nt) because the a-, b- or U-values were variable on the account of the base-location in this region irrespective of the smaller genome-size. These results suggested that the archaea were also similar to the eukaryotes beside the eubacteria at genome-level.

Acknowledgements

The author wishes to thank to Dr. Peter Flaherty for his critical reading and advice of this manuscript.

References

- Bains, W. (1993). Local self-similarity of sequence in mammalian nuclear DNA is modulated by a 180 bp periodicity. *J. Theor. Biol.*, *161*, 137-43. <http://dx.doi.org/10.1006/jtbi.1993.1046>
- Barthelemy, M., Buldyrev, S. V., Havlin, S., & Stanley, H. E. (2000). Multifractal properties of random resistor network. *Phys. Rev. E.*, *61*, R3283-R3286. <http://dx.doi.org/10.1103/PhysRevE.61.R3283>
- Coleman, M. L., Sullivan, M. B., Martiny, A. C., Steglich, C., Barry, K., DeLong, E. F., & Chisholm, S. W. (2006). Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science*, *311*, 1768-1770. <http://dx.doi.org/10.1126/science.1122050>
- Dunham, I., Shimizu, N., & Roe, B. A. (1999). The DNA sequence of human chromosome 22. *Nature*, *402*, 489-495. <http://dx.doi.org/10.1038/990031>
- Eisen, J. A., Coyne, R. A., & Wu, M. (2006). Macronuclear genomes sequence of the Ciliate *Tetrahymena thermophila*, a model eukaryote. *PLoS Biol.*, *4*, e286. <http://dx.doi.org/10.1371/journal.pbio.0040286>
- KEGG Organisms. (2013). Retrieved from http://www.genome.jp/kegg/catalog/org_list.html
- Kozak, M. (1983). Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles. *Microbiol. Rev.*, *47*, 1-45.
- Lai, E. C., Roegiers, F., Qin, X., Jan, Y. N., & Rubin, G. M. (2005). The ubiquitin ligase *Drosophila* Mind bomb promotes Notch signaling by regulating the localization and activity of Serrate and Delta. *Development*, *132*, 2319-2332. <http://dx.doi.org/10.1242/dev.01825>
- Levine, M., & Tjian, R. (2003). Transcription regulation and animal diversity. *Nature*, *424*, 147-151. <http://dx.doi.org/10.1038/nature01763>
- Loots, G. G., & Ovcharenko, I. (2010). Human variation in short regions predisposed to deep evolutionary conservation. *Mol. Biol. Evol.*, *27*, 1279-1288. <http://dx.doi.org/10.1093/molbev/msq011>
- Lynch, M. (2007). The frailty of adaptive hypothesis for the origins of organismal complexity. *Proc. Natl. Acad. Sci. USA*, *104*, 8597-8604. <http://dx.doi.org/10.1073/pnas.0702207104>
- Martens, J. S., Wu, P.-Y.J., & Winston, F. (2005). Regulation of an intergenic transcript controls adjacent gene transcription in *Saccharomyces cerevisiae*. *Genes Dev.*, *19*, 2695-2704. <http://dx.doi.org/10.1101/gad.1367605>
- Mattick, J. S. (2004). RNA regulation: a new genetics? *Nat. Rev. Genet.*, *5*, 316-323. <http://dx.doi.org/10.1038/nrg1321>
- Mewes, H. W., Albermann, K., & Bahr, M. (1997). Overview of the yeast genome. *Nature*, *387*, 7-65. <http://dx.doi.org/10.1038/42755>
- Molnar, A., Melnyk, C. W., Bassett, A., Hardcastle, T. J., Dunn, R., & Baulcombe, D. C. (2010). Small silencing RNAs in plants are mobile and direct epigenetic modification in recipient cells. *Science*, *328*, 872-875. <http://dx.doi.org/10.1126/science.1187959>

- Nakahara, M., & Takeda, M. (2010a). Characterization of the sequence spectrum of DNA based on the appearance frequency of the nucleotide sequences of the genome - A new method for analysis of genome structure-. *J. Biomed. Sci. Eng.*, 3, 340-350. <http://dx.doi.org/10.4236/jbise.2010.34047>
- Nakahara, M., & Takeda, M. (2010b). Identification of the Interactive Region by the Homology of the Sequence Spectrum. *J. Biomed. Sci. Eng.*, 3, 868-883. <http://dx.doi.org/10.4236/jbise.2010.39117>
- NCBI genome data base. (2013). Retrieved from <http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome>
- Niehhs, C., & Pollet, N. (1999). Synexpression groups in eukaryotes. *Nature*, 402, 483-7. <http://dx.doi.org/10.1038/990025>
- Oliver, S. G., van der Aart, Q. J. M., & Agostoni-Carbone, M. L. (1992). The complete DNA sequence of yeast chromosome III. *Nature*, 357, 38-46. <http://dx.doi.org/10.1038/357038a0>
- Olson, L. E., Richtsmeiser, J. T., Leszl, J., & Reeves, R. H. (2004). A chromosome 21 critical region does not cause specific Down syndrome phenotype. *Science*, 306, 687-690. <http://dx.doi.org/10.1126/science.1098992>
- Olson, L. E., Roper, R. J., Sengstaken, C. L., Peterson, E. A., Aquino, V., Galdzicki, Z., ... Reeves, R. H. (2007). Trisomy for the Down syndrome "critical region" is necessary but not sufficient for brain phenotypes of trisomic mice. *Hum. Mol. Genet.*, 16, 774-782. <http://dx.doi.org/10.1093/hmg/ddm022>
- Ordemann, A., Porto, M., Roman, H. E., Havlin, S., & Bunde, A. (2000). Multifractal behavior of linear polymers in disordered media. *Phys. Rev. E*, 61, 6858-6865. <http://dx.doi.org/10.1103/PhysRevE.61.6858>
- Peng, C. K., Buldrev, S. V., Goldberger, A. L., Havlin, S., Sciortino, M., Simons, M., & Stanley, H. E. (1992). Long-range correlations in nucleotide sequences. *Nature*, 356, 168-170. <http://dx.doi.org/10.1038/356168a0>
- Perffrey, L. W., Barbero, E., Lasser, E., Dunthorn, M., Bhattacharya, D., Patterson, D. J., & Katz, L. A. (2006). Evaluating support for the current classification of eukaryotic diversity. *PLoS Genet*, 2, e220. <http://dx.doi.org/10.1371/journal.pgen.0020220>
- Raphael, B., Volik, S., & Yu, P. (2008). A sequence-based survey of the complex structural organization of tumor genomes. *Genome Biology*, 9, R59. <http://dx.doi.org/10.1186/gb-2008-9-3-r59>
- Saccharomyces Genome Database. (2013). Retrieved from <http://www.yeastgenome.org/>
- Taft, R. J., Pheasant, M., & Mattick, J. S. (2007). The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays*, 29, 288-299. <http://dx.doi.org/10.1002/bies.20544>
- Takeda, M. (2009). Several characteristics of a single-strand of genomic DNA. *Research Report of Tsuruoka National College of Technology* (No. 44 pp. 67-74). ISSN: 0286-1232. Retrieved from <http://repo.lib.yamagata-u.ac.jp/handle/123456789/9317>.
- Takeda, M. (2011). Structural features of the nucleotide sequences of virus and organelle genomes. *J. Biomed. Sci. Eng.*, 4, 719-733. <http://dx.doi.org/10.4236/jbise.2011.411089>
- Takeda, M. (2012). How is the biological information arranged in genome? *Am. J. Mol. Biol.*, 2, 171-186. <http://dx.doi.org/10.4236/ajmb.2012.23019>
- Takeda, M., & Nakahara, M. (2009). Structural Features of the Nucleotide Sequences of Genomes. *J. Comput. Aided. Chem.*, 10, 38-52. <http://dx.doi.org/10.2751/jcac.10.38>
- Takeda, M., & Nakahara, M. (2013). *Protean Genome*. Research Signpost, Trivandrum-695 023, Kerala, India. Retrieved from <http://www.signpostbooks.com>
- The Sanger Institute. (2013). Retrieved from <http://www.sanger.ac.uk>
- Voss, R. F. (1992). Evolution of long-range fractal correlation and 1/f noise in DNA base sequences. *Phys. Rev. Lett.*, 68, 3805-3809. <http://dx.doi.org/10.1103/PhysRevLett.68.3805>
- Webb, C. F., Das, C., Eneff, K., & Tucker, P. W. (1992). Identification of a matrix-associated region 5' of an immunoglobulin heavy chain variable region gene. *Mol. Cell Biol.*, 11, 5206-5211.
- Weinberger, E. D., & Stadler, P. F. (1993). Why some fitness landscapes are fractal. *J. Theor. Biol.*, 163, 255-275. <http://dx.doi.org/10.1006/jtbi.1993.1120>
- West, A. G., Gaszner, M., & Felsenfeld, G. (2002). Insulators: many functions, many mechanisms. *Genes Dev.*, 16, 271-288. <http://dx.doi.org/10.1101/gad.954702>
- Yu, Z. G., Anh, V., & Lau, K. S. (2004). Chaos game representation of protein sequences based on the detailed

HP model and their multifractal and correlation analysis. *J. Theor. Biol.*, 226, 341-348.
<http://dx.doi.org/10.1016/j.jtbi.2003.09.009>

Zmasek, C. M., & Godzik, A. (2011). Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires. *Genome Biology*, 12, R4.
<http://dx.doi.org/10.1186/gb-2011-12-1-r4>

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).