Chapter

Visible Evolution from Primitive Organisms to *Homo sapiens*

Kenji Sorimachi

Abstract

The ratios of amino acids to the total amino acids deduced from the complete genome and those of nucleotides to the total nucleotides in the genome are useful indexes to characterize various large genomes among different species from bacteria to *Homo sapiens*. These indexes are not only independent of species but also of genome size. Using these indexes, the following results were obtained: (1) primitive life forms appeared to have similar amino acid compositions to present day organisms; (2) cellular amino acid compositions that are similar among various species and between whole cells and complete genomes; (3) genome structure that is homogeneously constructed from putative small units encoding proteins of similar amino acid compositions, followed by synchronous mutations over the genome; (4) all organisms can be classified into two groups, "GC-rich" and "AT-rich," based on their nucleotide contents, or "terrestrial" and "aquatic vertebrates" based on natural selection by cluster analyses using amino acid contents as the traits; and (5) evolution based on nucleotide content alterations can be expressed by definitive equations. Thus, the ratios of amino acids or nucleotides to their total contents are useful indexes for characterizing genomes, regardless of species differences and genome sizes. The two normalized nucleotide contents are universally expressed regression line.

Keywords: genome, mitochondria, codons, Chargaff's parity rules, cluster analysis, normalization, phylogenetic trees, evolution

1. Introduction

The origin of life has long been interested to human since old times. Indeed, Aristotle proposed "spontaneous generation" more than 2000 years ago, although this idea was disproved by Louis Pasteur in experiments using "swan neck flasks." Our great interest in the origin of life might be expressed by the following philosophical words: *Where do we come from? What are we? Where are we going?* These words were written by French artist Paul Gauguin on his canvas in Tahiti in 1897.

The development of nucleotide sequencing technology [1, 2] has contributed to progress in molecular biology, including the analysis of a complete bacterial genome first carried out in 1995 [3], and, subsequently, the draft human genome, which was reported in 2001 [4, 5]. At present (June 19, 2019), 498 eukaryote, 5159 bacterial, and 296 archaeal complete genomes were determined. However, the origin of life is still unclear. Assuming that the replacement rates of nucleotides or amino acids in genes are constant [6], phylogenetic trees were drawn [6–11]. However, we know that their exact replacement rates differ between genes and between species. Studies based on nucleotide or amino acid sequences are applicable to genes whose nucleotide or amino acid numbers are much smaller than those of complete genomes, but not to genomes consisting of huge numbers of nucleotides and many genes. Of course, simple comparison of sequence differences between genes in the same species and the same genes in different species is useful.

2. Normalization

Intraspecies nucleotide contents were first analyzed in 1950 by Chargaff, who reported that G = C, A = T, and [(G + A) = (C + T)] [12], which was named as Chargaff's first parity rule. This rule is understandable based on the double-stranded DNA structure [13]. Additionally, this rule is applicable to single-stranded DNA obtained from a single species nucleus, termed Chargaff's second parity rule [14]. As the rules are based on normalized values to 1 (G + C + A + T = 1), nucleotide contents are expressed by their ratios. However, the second parity rule is more difficult to understand because we could not image how G and C or T and A pairs are formed in the single DNA strand. Recently, this puzzle has been solved mathematically, using the similarity of the forward and reverse strands and homogeneity of the DNA strand over the genome structure [15]. Although Chargaff's parity rules represent original intra-species phenomena, the rules can be expanded to inter-species phenomena using data from a large number of complete genomes [16]: the second parity rule is applicable only to a single DNA strand from a double-stranded DNA molecule.

Sueoka [17] was the first to analyze the cellular amino acid composition in bacteria, and our laboratory has independently analyzed the cellular amino acid compositions of bacteria, archaea, and eukaryotes [18]. Graphical representation or a diagrammatic approach to the study of complicated biological systems can provide an intuitive picture and provide useful insights [19, 20]. Using certain graphical presentations, huge data sets from genomes can be easily recognized as simple patterns representing complicated organisms. Indeed, using a radar chart to express cellular amino acid compositions, their patterns, a "star-shape," are similar among various organisms, and their differences seem to reflect biological evolution [18]. In addition, the amino acid compositions deduced from complete genomes resemble those obtained from amino acid analyses of cell lysates [21]. These results suggest that the ratios of amino acids to the total amino acids and those of nucleotides to the total nucleotide content are useful indices to characterize whole genome structures [21].

3. Patternalization of amino acid compositions

In general, there are 20 amino acids that can form proteins, and the amino acid sequences are strictly controlled by 64 codons consisting of three nucleotides, a triplet. Thus, differences in amino acid sequences of the same kind of proteins reflect biological evolution among species, although differences among different kinds of proteins seem not to be significant. Furthermore, sequence comparisons of protein mixtures are theoretically too complex to consider given currently available tools. Conversely, the amino acid composition predicted from protein(s) can characterize protein(s) from a different point of view, not only among the same organisms, but also among different organisms. In fact, the cellular amino acid compositions of various bacteria have been analyzed [17]. Based on the 20 amino acids that comprise proteins, there were 20 traits that could be evaluated, which, at first glance, seemed too many to provide meaningful information for cells.

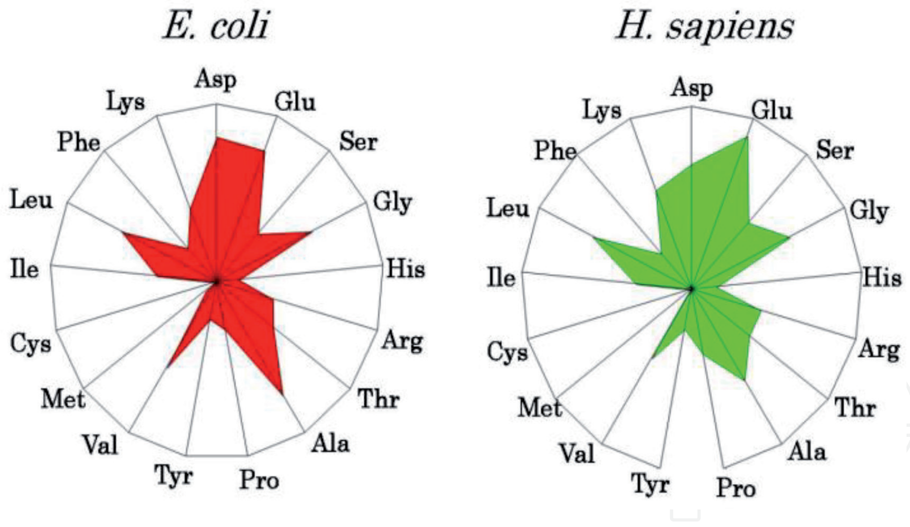

Figure 1.
Radar charts of cellular amino acid compositions of Escherichia coli *and* Homo sapiens. *Amino acid compositions are expressed as the percentage of total amino acids. Gln and Asn are combined with Glu and Asp, respectively, because the former two are converted into the latter two during hydrolysis [18].*

However, using a radar chart to present the amino acid compositions, the data could be patternalized, and the amino acid composition was observed to represent certain cellular characteristics, as shown in **Figure 1**. The patterns of bacteria (*Escherichia coli*) and of humans (*Homo sapiens*) resemble each other, although there is a great evolutionary distance between these two organisms. Microorganisms' fossils were found in 550–2800-million-year-old rocks [22–24], and it is thought that bacteria are evolutionarily close to primitive life forms. Therefore, it seemed that the primitive life forms might have similar amino acid compositions [21]. This "star-shape" cellular amino acid composition pattern must have been conserved from primitive organisms to those current organisms.

4. Chronological precedence of protein formation over codon formation

To understand the establishment of primitive organisms, the chronological precedence of protein and codon formation is a very important subject in biological evolution. Unfortunately, this theory has not yet been proven, because primitive organisms were formed under so many unknown factors an extremely long time ago. However, a simulation analysis based on a random choice of amino acids or nucleotides was carried out, which assumed that their polymerization depended on their free monomer concentrations, according to the chemical reaction rule that governs natural phenomena. Amino acid polymerizations produced a protein which reflected original free amino acid concentrations without codons, while nucleotide polymerizations did not produce functional proteins, even after considering the codon table, as shown in **Figure 2** [25]. Therefore, it seems difficult to predict "the RNA world" which presumes that RNA polymers formed primitive life forms [26]. Additionally, the possibility of the accumulation of RNA, which has a UV absorbance at around 250 nm, might be very low under the strong UV irradiation present on the primitive Earth. These results suggest that protein formation might chronologically precede codon formation at the end of prebiotic evolution, although we have no explanation of how the nucleotide sequence information necessary for proteins might have been transmitted to the nucleotide polymerization that established the codons. The

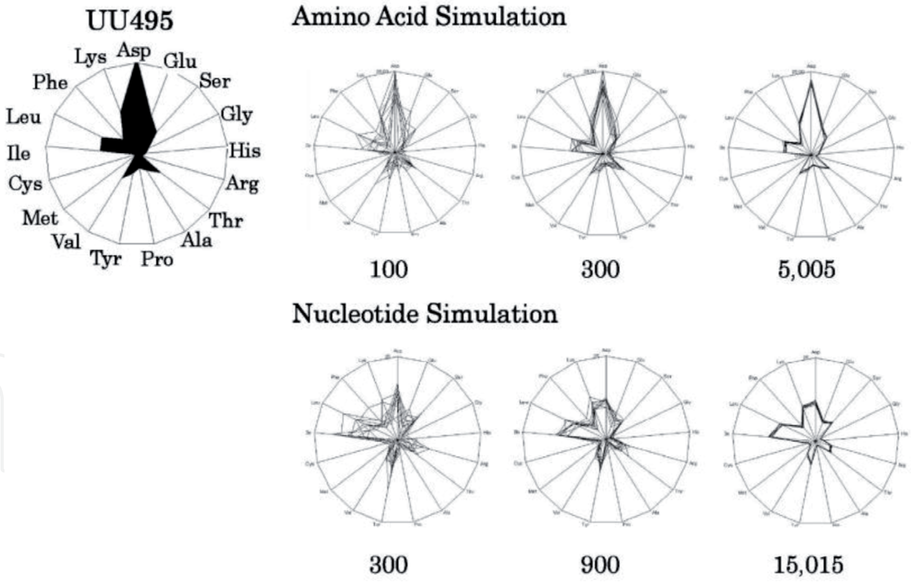

Figure 2.
Computational amino acid compositions of an Ureaplasma urealyticum *gene. Upper panel: random choice of amino acids was carried out in the original gene (5005 amino acid pool). Lower panel: random choice of nucleotides was carried out in the original gene (15,018 nucleotides). In the simulation using nucleotides, the stop codon and Trp were discarded from the calculation of amino acid compositions, and a triplet formed was immediately counted as an amino acid. This figure was adapted from Sorimachi and Okayasu [25].*

“amino acid world” [21] seems a better fit for primitive life forms rather than the “RNA world.” There are several hypotheses for codon formation [27–29], but the process of codon formation has not yet been determined.

According to our simulation analyses [25], proteins that were components of primitive life forms might reflect the free amino acid concentrations on the primitive Earth. As shown in **Figure 1**, the cellular basic amino acid composition, the “star-shape,” is characterized by comparatively high concentrations of hydrophobic amino acids, such as valine, leucine, and isoleucine. The glycine and alanine contents were also comparatively high. The former might contribute to self-aggregation of proteins via hydrophobicity to form primitive life forms under low protein concentrations, and the latter might reflect their easy formation on the primitive Earth. In fact, simple amino acids such as glycine and alanine have been identified in meteorites [30, 31] and can be formed by electrical discharge in an atmosphere presumed to reflect primitive Earth [32]. Conversely, the phenylalanine, tryptophan, and tyrosine content, which can absorb ultraviolet light, were quite low. Strong ultraviolet irradiation might induce photodegradation of these amino acids. The differences in amino acid contents in cellular amino acid compositions seem to reflect the presumed free amino acid concentrations on the primitive Earth and eventually resulted in the formation of the “star-shaped” cellular amino acid compositions (**Figure 1**).

5. Amino acid compositions deduced from complete genomes

Initially, amino acid compositions were deduced from complete genomes by assuming that each gene is equally expressed in a whole cell [21]. This resulted in the amino acid composition deduced from the complete genome resembling the cellular amino acid composition obtained from the amino acid analyses of cell lysates [21], as shown in **Figure 3**. This coincidence is difficult to understand because of the different origins of both values, until the genome structure has been clarified, as shown in the next section.

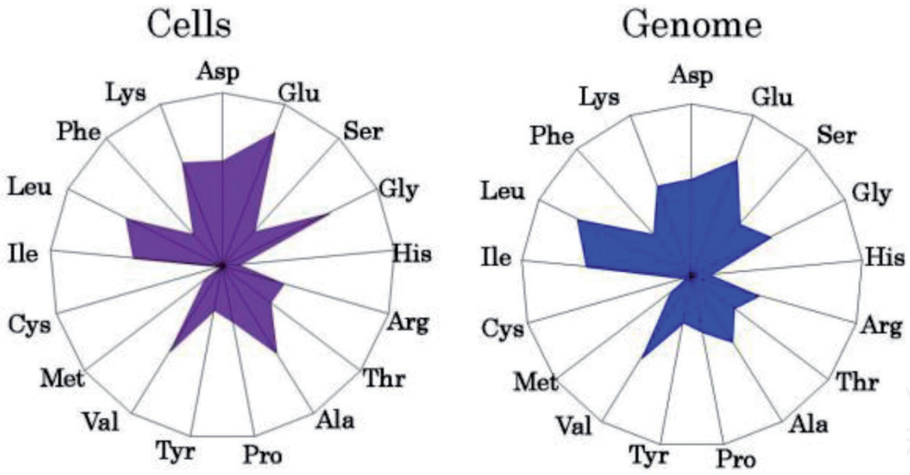

Figure 3.
Radar charts of cellular and genomic amino acid compositions. Values are expressed as the percentages of total amino acids. Pyrococcus horikoshii *was examined. The cellular amino acid composition was obtained from three independent analyses. In genomic calculations, Gln and Asn were also incorporated into Glu and Asp, respectively, to compare with data based on amino acid analysis.*

6. Homogeneity of genome structure

Each gene has its characteristic amino acid or nucleotide sequence, and its amino acid or nucleotide composition differs not only in inter-species but also in intraspecies. Conversely, gene assemblies encoding 3000–7000 amino acid

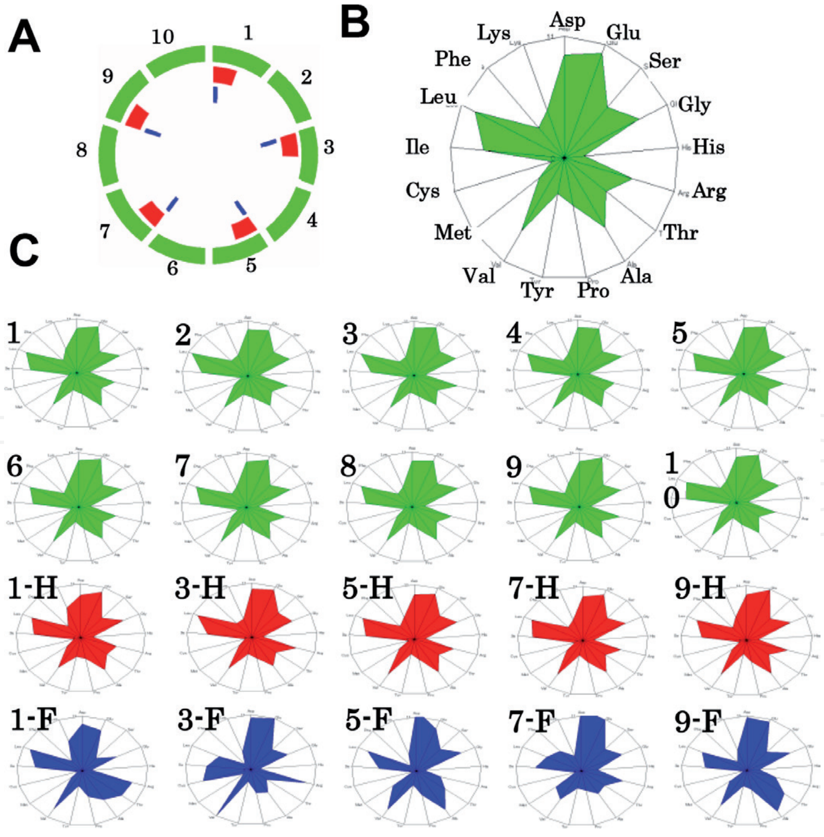

Figure 4.
Radar charts of amino acid compositions calculated from various units of the complete genome of Methanobacterium thermoautotrophicum*. (A) The complete genome structure of* M. thermoautotrophicum *(B) radar charts of amino acid compositions calculated from the complete genome, and (C) from various units. The complete genome, comprising 1869 protein genes, was divided into 10 or 20 units. Ten units (1–10); based on 186 and 195 genes, half size units (1-H–9-H); based on 93 genes, single genes (1-F–9-F); based on the first single gene of each unit. Glutamine and asparagine were calculated as glutamic acid and aspartic acid, respectively, and tryptophan (<1%) was omitted in the radar charts [18]. This figure was adapted from Sorimachi [36].*

residues show very similar amino acid compositions [33] and nucleotide compositions [34] in intraspecies examinations. Consistent results were obtained from whole chromosomes consisting of putative small units of 3000–7000 amino acid residues [33]. Additionally, it has been shown mathematically that 3000–7000 amino acid residues represent the amino acid composition of a certain amino acid pool [35]. Thus, genome structure, which is constructed homogeneously from putative similar small units, can be represented by a “pearl-necklace,” as shown in **Figure 4**. The fact that the structure of a genome is homogeneously constructed with putative similar small units indicates that micro-alterations of nucleotide sequences are canceled out within the small unit and that the small unit represents the whole genome characteristics. Macro-alterations represented by the small unit, and based on species differences, occur synchronously over the genome [33]. This conclusion has never been obtained from the analysis of nucleotide or amino acid sequences of actual genes. Based on these results, the ratios of amino acids to the total amino acids or those of nucleotides to the total nucleotides form useful indices for characterizing a genome whose nucleotide numbers differ among species.

7. Nucleotide compositions

As described above, the intraspecies rule of nucleotide composition was reported by Chargaff in 1950, as the first parity rule [12], and a similar parity rule regarding the single DNA strand was reported by the same group in 1968, as the second parity rule [14]. Using the normalized values to 1 (G + C + T + A = 1), the following relationships are obtained: G = C, T = A, and [(G + A) = (C + T)]. Recently, Mitchell and Bridge [16] reported that Chargaff’s second parity rule is applicable to a single DNA strand comprising a double-stranded DNA, based on many complete genome data among various species. Conversely, we showed that chloroplast and plant mitochondrial DNA and nuclear DNA obey Chargaff’s second parity rule as an inter-species rule [37], and that the second parity rule was applicable to the nucleotide relationships not only in the coding region, but also in non-coding regions compared with those of the complete single DNA strand [37, 38]. When invertebrate mitochondrial DNA is classified into two groups, high C/G and low C/G ratios, nucleotide content relationships may be expressed by linear formulae [37]. However, organellar DNA deviated from Chargaff’s second parity rule and nucleotide relationships were heteroskedastic [16, 39, 40]. The fact that all regression lines based on different kingdoms closed at the same single point suggests that all species descended from a single origin [41]. This is the first demonstration based on scientific evidence that all species were descended from a single origin of life. This concept has been presumed since Darwin’s theory “Origin of Species” was published in 1859. Charles Darwin discussed evolution over the course of generations via the presence of “Natural Selection” in “On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life”; however, he discussed neither “a single origin” nor “a common ancestor” of species. The two regression lines of nucleotide relationships based on coding and non-coding regions closed to form a wedge-shape, because both fragments exist on the same DNA strand [37]. Similarly, the two regression lines based on chloroplast and plant mitochondrial DNA also closed to form a wedge-shape [37]. Thus, both organellar DNA independently descended from the same origin in biological evolution. Quite recently, it has been shown that vertebrates are descended from a certain invertebrate [42]. However, although the phylogenetic trees [7–11] have an apparent single origin, these "facts" are merely mathematical calculation results.

8. Diagonal genome universe

Chargaff's parity rules were originally based on intraspecies phenomena [12, 14], and the rules are applicable to inter-species evolutionary phenomena for nuclear, chloroplast, and plant mitochondria as mentioned above. The rules are represented by the following equations: G = C, T = A, [(G + A) = (C + T)]. As all values are normalized to 1, Chargaff's parity rule can also be represented as: 2G + 2A = 1, A = 0.5 – G, T = 0.5 – G, C = G, G = (G). The lines G and C overlap and the lines A and T overlap, and the former is line symmetrical to the latter against the line y = 0.25, as shown in **Figure 5**. These equations mean that four nucleotide contents can be expressed by just one nucleotide content using regression lines (**Figure 5**), and the two duplicate nucleotide contents (G or C and T or A) are symmetrical. Thus, the four nucleotide contents (two duplicate points) move strictly on the diagonal of 0.5 of a square in nuclear, chloroplast, and mitochondrial DNA, which obey Chargaff's second parity rule. Therefore, biological evolution caused by nucleotide alterations is expressed on the diagonal of a 0.5 square: the "diagonal genome universe" [36], although biological evolution shows a wide spectrum of phenotypic expressions over a 3.5-billion-year period.

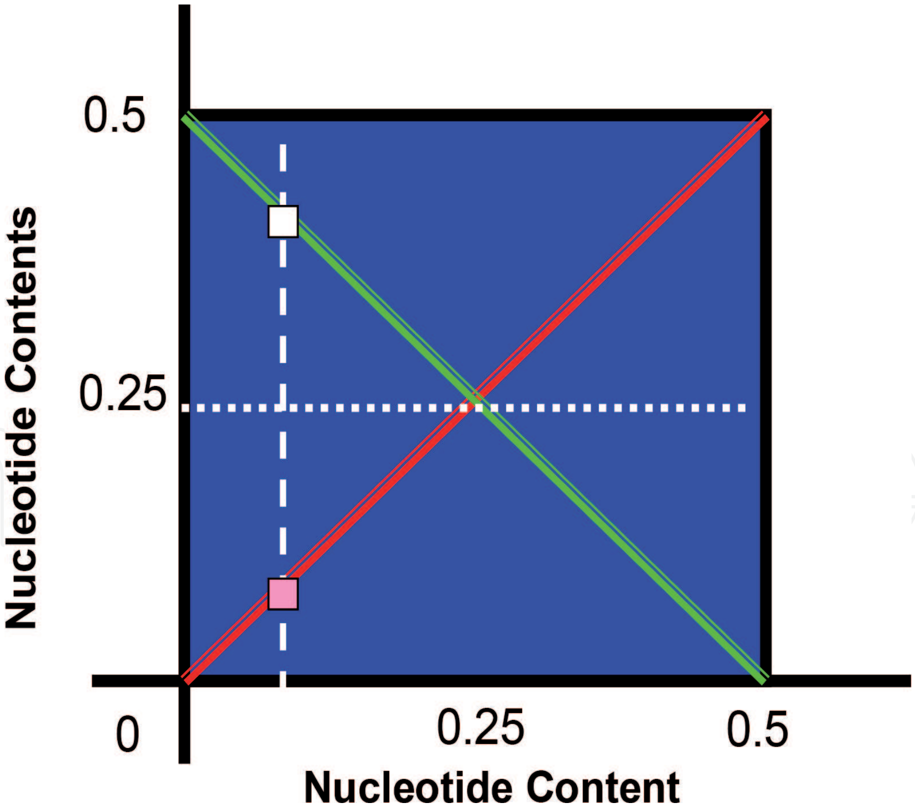

Figure 5.
The "Diagonal Genome Universe." Plotting four nucleotide contents normalized to 1 against certain nucleotide content (i.e., G or C content), G and C contents are expressed by (G = G) and (G = C), respectively, and T and A contents are expressed by (T = 0.5 − G) and (A = 0.5 − G), respectively. For example, if G = 0.1 (white dashed line), C = 0.1, T = 0.4, and A = 0.4. White open square, A or T; pink closed square, C or G. The white dotted line represents the line of symmetry (y = 0.25). Similarly, plotting nucleotide contents against T or A content, (T = T), (T = A), (C = 0.5 – T or A), and (G = 0.5 − T or A) are obtained. This figure was adapted from Sorimachi [36].

9. Codon evolution

The 20 amino acids are encoded by genes using nucleotide triplets; therefore, these sequences are determined according to triplet sequences. Additionally, amino acid sequences differ not only inter-gene but also intraspecies. These facts indicate that a comparison of codon evolution based on the complete genome, which comprises large numbers of different genes, would not be significant. Indeed, no clear evaluation has been obtained, despite the attempted explanations of many scientists [27–29]. However, as described in the previous section, it has been clarified that a whole genome is constructed from putative small units that encode proteins of similar amino acid composition. This suggests that the total codon usage deduced from the complete genome is stable and represents the whole genome characteristic. According to this concept, correlationships of nucleotide contents in a complete genome can be expressed by the linear formula, y = ax + b; where "y" and "x" are nucleotide contents, and "a" and "b" are constant values. In addition, as each codon usage is expressed by a linear formula among various organisms, the determination of any one nucleotide content in certain organism can essentially estimate other three nucleotide contents and, therefore, the 64 codon usages (**Figure 6**). The estimated codon usage patterns and amino acid compositions are almost the same between the original experimental results and estimated results. The codon usage patterns clearly indicate that codon usages changed synchronously among the 64 codons during biological evolution.

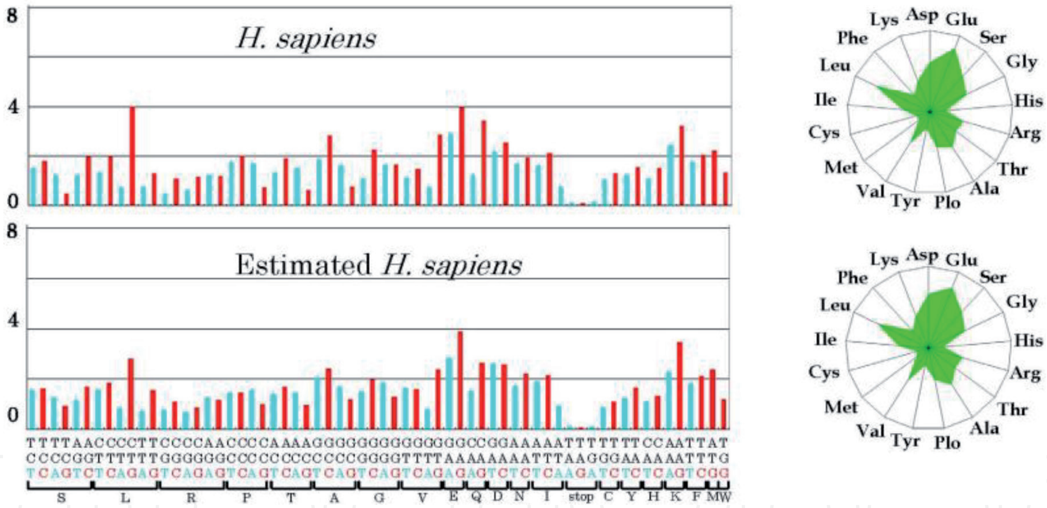

Figure 6.
Codon usage patterns and amino acid compositions of Homo sapience. *Codon usage (bar) and amino acid composition (radar chart) are expressed as a percent of total codons and amino acids, respectively. Upper and lower panels represent genomic and estimated data, respectively. This figure was reproduced from Sorimachi and Okayasu [38].*

10. Natural selection in biological evolution based on amino acid contents

The above mentioned theories have been described in previous review articles [36, 43]; therefore, in this section, unique applications based on the amino acid compositions or nucleotide contents in the construction of phylogenetic trees to study evolution are presented using recent data.

The theory of natural selection was promoted by Charles Darwin and Alfred Wallace 150 years ago. This theory was derived from specific differences or similarities in the phenotypes of organisms that lived on geologically isolated islands.

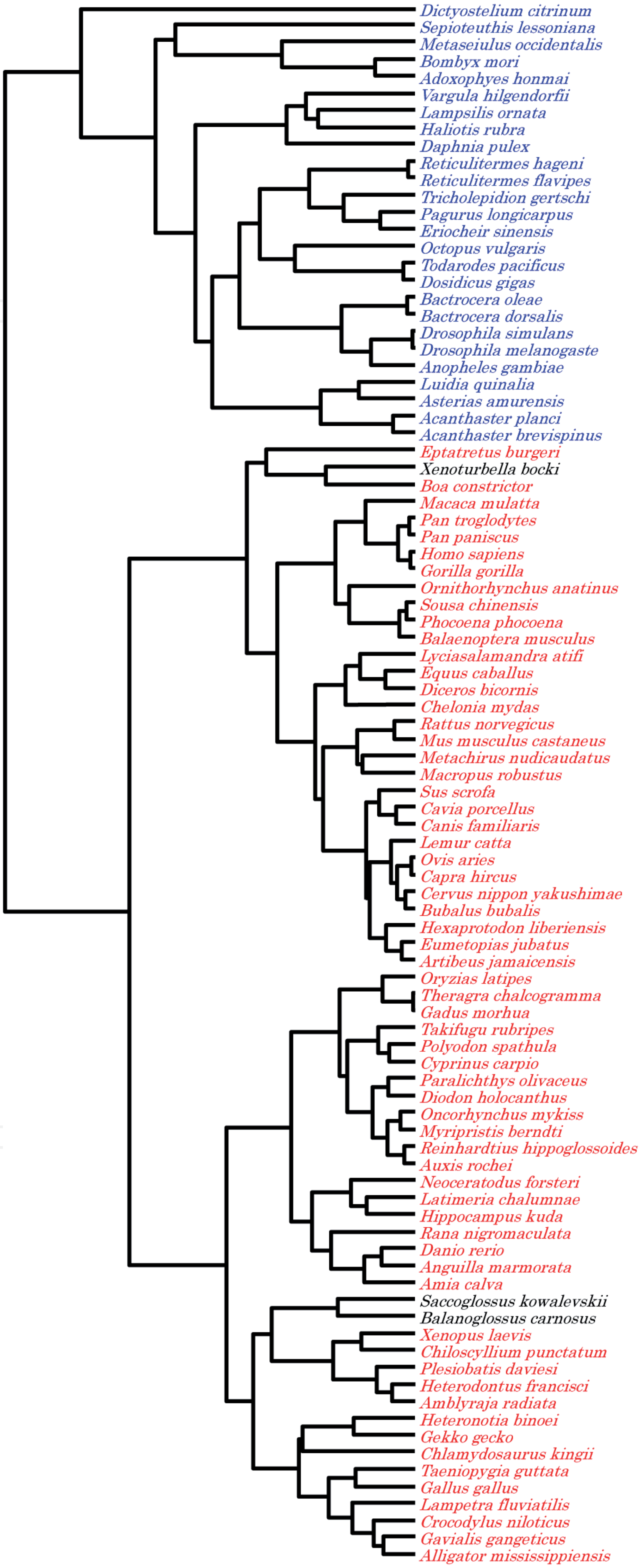

Figure 7.
Phylogenetic tree generated using Ward's cluster analysis method [48] from the predicted amino acid composition of the complete mitochondrial genomes of 26 invertebrates (blue), 3 hemichordates (black), and 63 vertebrates (red). This figure first appeared in Ref. [49] and is reproduced with permission.

The theory of biological evolution has been further developed by paleontology [44], using phenotypic changes in fossils, and by molecular biology [6], using genotypic modifications (nucleotides or amino acids) of genes in living organisms.

Generally, the nucleotide or amino acid sequences of a particular gene or genes have been the focus of biological evolution studies, and many phylogenetic trees have been constructed using nucleotide or amino acid sequences [7–11, 27, 29, 45]. Conversely, the amino acid compositions or nucleotide contents have been rarely used for whole genome research. However, these indices have been used to classify bacteria, archaea, and eukaryotes [46] and recently vertebrate evolution [47]. In those studies, all organisms could be classified into two types, "GC-rich" and "AT-rich," and the vertebrates examined were further classified into two groups: terrestrial and aquatic vertebrates, based on natural selection. A similar result was obtained from an analysis based on 16S rRNA sequences [45, 47].

When the normalized amino acid compositions of vertebrate and invertebrate complete mitochondrial genomes were used, the groups were separated cleanly into two large clusters, vertebrates and invertebrates (**Figure** 7). In invertebrates, starfish (Echinodermata) formed a small cluster, and squids and octopus (Mollusca) were grouped into the same cluster. Vertebrates were further classified into three major clusters, mammals, fish, and a mixture of reptiles and amphibians. For example, primates (human, chimpanzee, and gorilla) formed a small cluster. Thus,

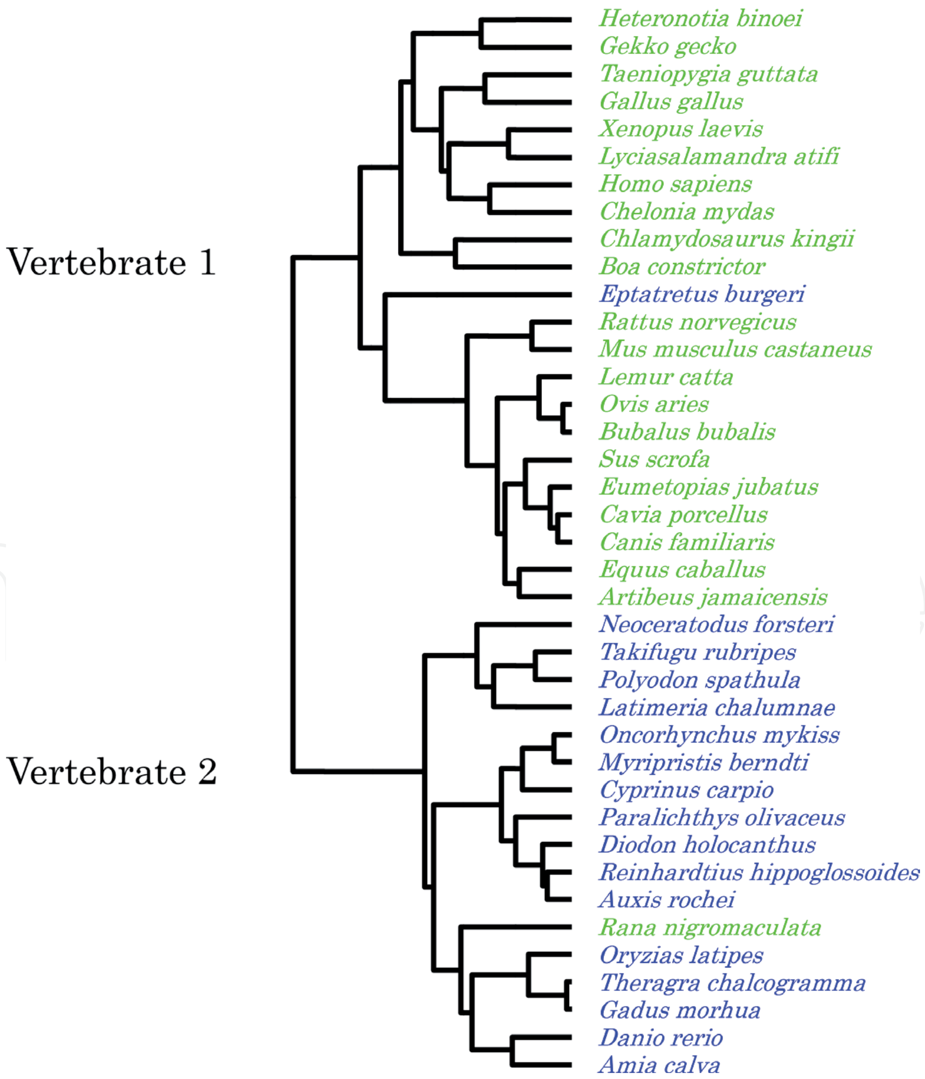

Figure 8.
Phylogenetic tree of complete vertebrate mitochondrial genomes based on cluster analysis [51] using amino acid compositions as the trait. Green and blue characters represent terrestrial and aquatic vertebrates, respectively. This figure was adapted from Sorimachi et al. [47].

close species fell into the same cluster and did not split into different clusters. These results indicate that the normalized values of amino acid and nucleotide contents calculated from complete genomes could be used to characterize organisms and to construct phylogenetic trees. Our results based on complete mitochondrial genomes revealed that hemichordates (*Balanoglossus carnosus* and *Saccoglossus kowalevskii*) and *Xenoturbella bocki*, which were classified into the low G/C content invertebrates group, were closer to vertebrates than to invertebrates [49]. Protists (*Monosiga brevicollis*) and cephalochordate (*Branchiostoma belcheri*) were classified into the low G/C and high G/C content invertebrate groups, respectively [49].

In a previous study to classify vertebrates [49, 50], as organisms were chosen at random without any preposition, it was difficult to evaluate whether the classification results were reasonable in the phylogenetic trees. Using the amino acid composition as the trait, the vertebrates examined were separated into two major clusters (**Figure 8**), terrestrial and aquatic vertebrates. The exceptions were the hagfish (*Eptatretus burgeri*), which fell into the terrestrial vertebrate cluster, and the black spotted frog (*Rana nigromaculata*), which clustered with the aquatic vertebrates [47]. The clustering of the

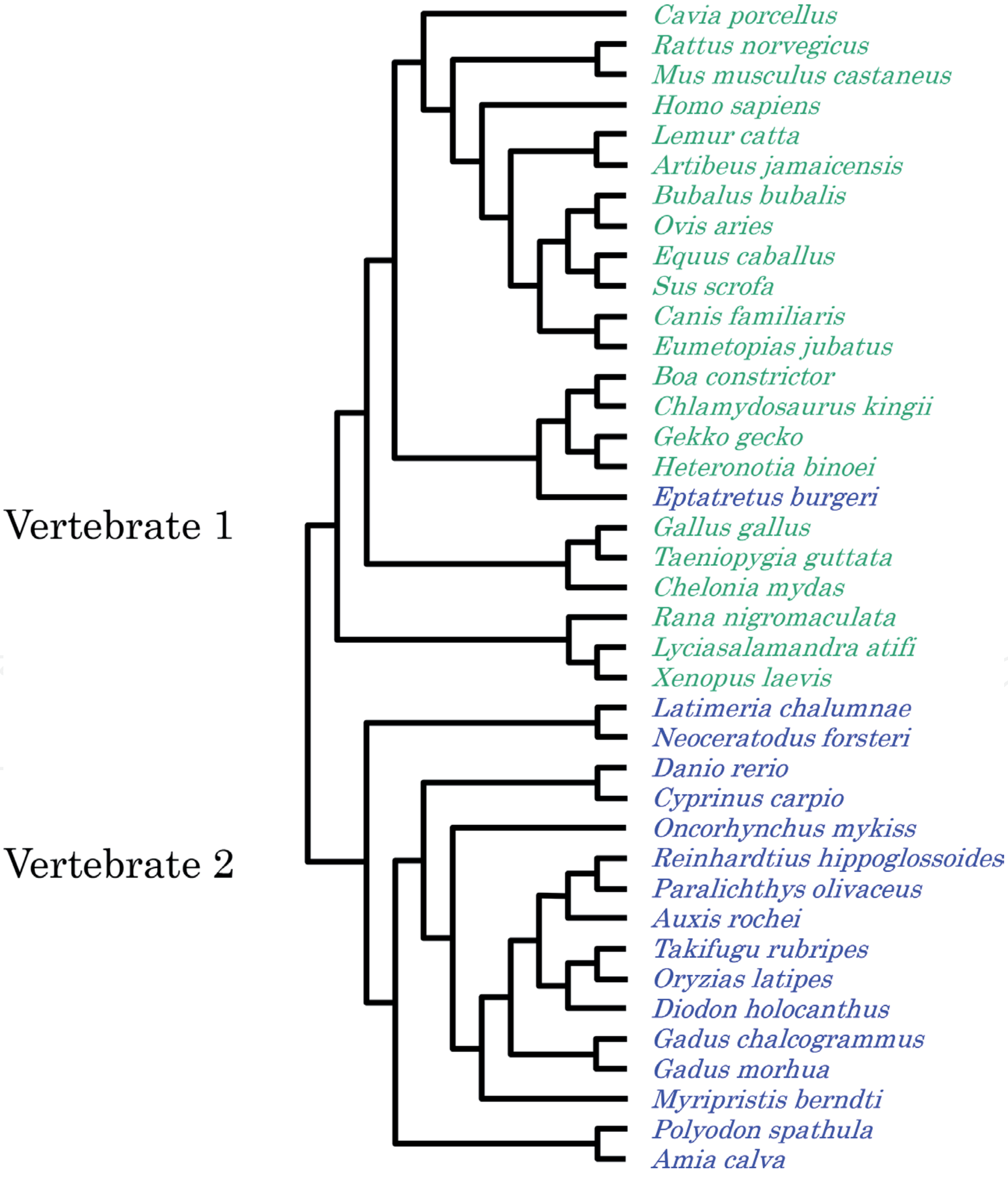

Figure 9.
Phylogenetic tree of 16S rRNA. The phylogenetic tree was constructed by the neighbor-joining method [48] using nucleotide sequences. Green and blue characters represent terrestrial and aquatic vertebrates, respectively. This figure was adapted from Sorimachi et al. [47].

hagfish (*E. burgeri*) with the terrestrial vertebrates may reflect the controversy over the classification of this fish [52]. If the hagfish truly belongs to the terrestrial group, it suggests that hagfish still possesses some primitive mitochondrial characteristics that were present before its evolution. The frog (*R. nigromaculata*) was consistently grouped with the aquatic vertebrates which may reflect the conservation of tadpole characteristics after metamorphosis. The coelacanth (*Latimeria chalumnae*), the Queensland lungfish (*Neoceratodus forsteri*), which is a living fossil and one of the oldest living vertebrate genera, and the American paddlefish (*Polyodon spathula*), which is the oldest living animal species in North America, all belonged to an additional small cluster. Using the G, C, A, and T content of the coding regions, non-coding regions, and complete mitochondrial genomes as the traits in cluster analyses, similar results were obtained, but with some additional exceptions [50].

Single genes have been used to construct phylogenetic trees [7–11], and 16S rRNA has been frequently examined [27, 29]. The phylogenetic tree based on 16S rRNA sequences of various vertebrates is shown in **Figure 9**. The tree is consistent with that based on nucleotide contents. The hagfish (*E. burgeri*) fell into the terrestrial vertebrates, while the black spotted frog (*R. nigromaculata*) belonged to the terrestrial vertebrates. These results indicate that vertebrate evolution is controlled by natural selection under both an internal bias resulting nucleotide replacement rules and by an external bias caused by environmental biospheric conditions. In addition, based on amino acid composition or nucleotide content of complete mitochondrial genomes, Hemichordates (*Balanoglossus carnosus* and *Saccoglossus kowalevskii*) and Xenoturbella were classified into vertebrates not into invertebrates [49].

11. Organelle evolution

In Chargaff's first parity rule [12], G = C and A = T in a double DNA strand, while in the second parity rule [14], G ≈ C and A ≈ T in a complete single DNA strand. Based on Chargaff's second parity rule, nucleotide content differences such as (G – C) and (A – T) reflect biological evolution. In addition, the other nucleotide content differences, (G – A, G – T, C – A, and C – T), also reflect biological evolution [34, 53].

Six nucleotide content differences among the complete mitochondria of the four species (*M. brevicollis, P. pallidum, D. discoideum*, and *R. Americana*) were examined (**Figure 10**, left panel). The GC and AT skew are expressed by the ratios of (G – C)/(G + C) and (A – T)/(A + T), respectively [54]. The skew seems to be due to differences in replication processes between the leading and lagging strands [55]. In the replication of the lagging strand, the deamination of cytosine increases the probability of mutations, and the inversion of nucleotide content differences reflects biological divergence. Similarly, these phenomena are observed in mitochondria, consisting of heavy (H) and light (L) chains [56–58]. When the GC skew was plotted against G content, animal mitochondria were classified into two groups: high and low C/G [59].

To allow simple comparison of inter- and intraspecies genome structures, genomes were divided into three fragments throughout subsequent analyses, from which three separate patterns emerged. There is no inversion of nucleotide content differences that was observed in the mtDNA of *M. brevicollis* (G: 0.081, C: 0.059), the mycetozoan *Polysphondylium pallidum* (G: 0.143, C: 0.085), or *Dictyostelium discoideum* (G: 0.171, C: 0.104) (**Figure 10**), whereas differences in (G – C) and (T – A) values for *M. brevicollis* mtDNA were the lowest among these species. Choanoflagellates are most closely related to animals based on genome sequencing [60]. The fact that the nucleotide content difference patterns of the three fragments were almost identical for these three species indicates that their nucleotide distributions were homogeneous, and that the nucleotide content was symmetrical.

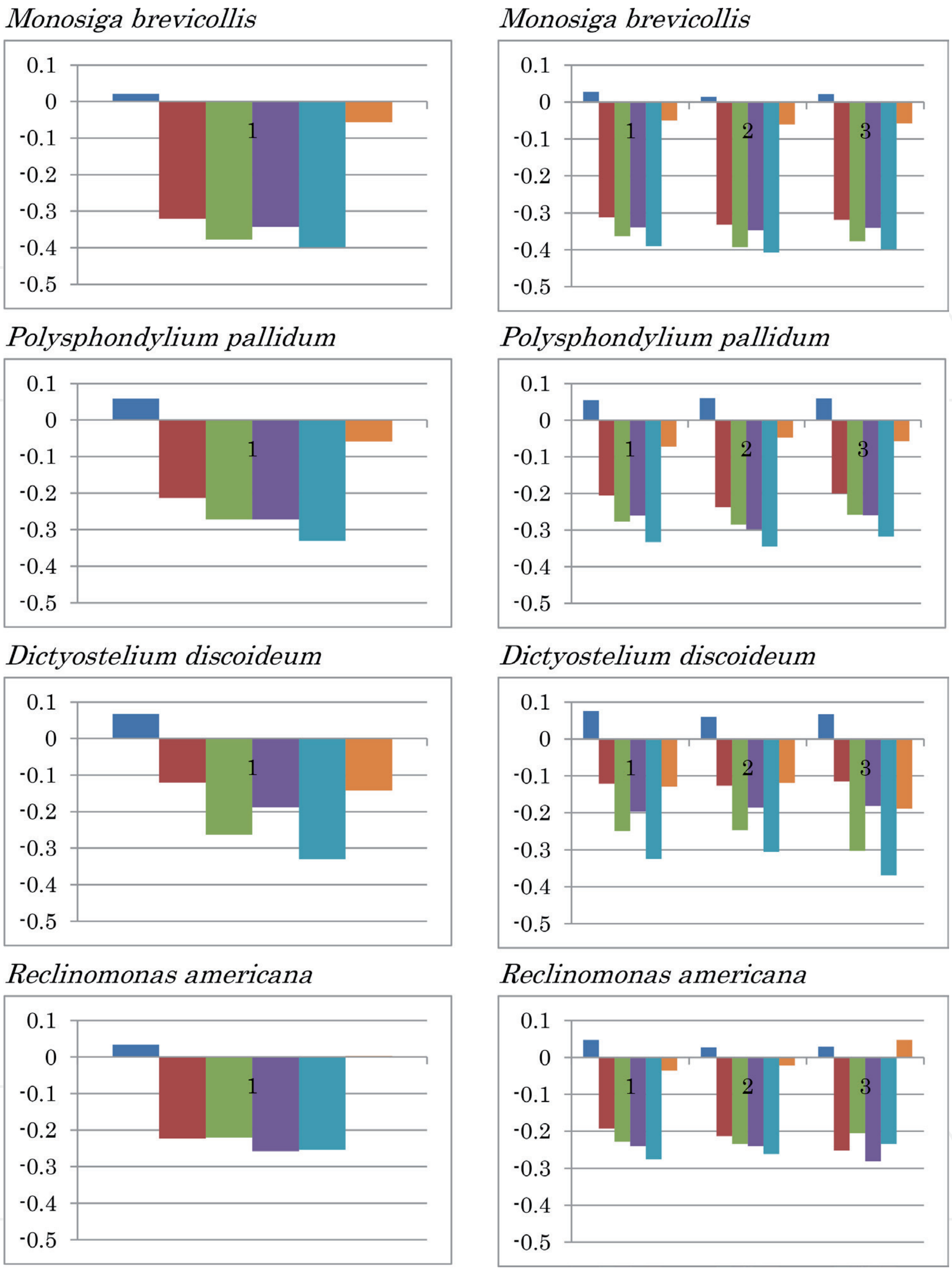

Figure 10.
Nucleotide content differences in complete mitochondrial genomes (left side) and the three fragments of each mitochondrial genome (right side). Left to right: (G – C), (G – T), (G – A), (C – T), (C – A), and (T – A).

Based on these results, these mitochondria are likely to be primitive. Consistent results were obtained from Ward's clustering analysis using amino acid compositions predicted from complete mitochondrial genomes as traits [59]. Thus, the *M. brevicollis* mitochondrion is the most primitive among the three. Although the *Reclinomonas americana* mtDNA (G: 0.148, C: 0.114) has previously been proposed as a mitochondrial ancestor [61], AT inversion was observed in the third fragment. In addition, differences in (G – C) and (T – A) values in *R. americana* mtDNA were smaller than those in the mtDNA of the previous three organisms. The unsymmetrical nucleotide content causes significant differences in nucleotide content patterns as a result of nucleotide content inversion. Judging from these results, the *R. americana* mitochondrion is probably more evolved than the former three mitochondria. In addition, AT inversion occurred in the following more highly evolved organisms: Mollusca species, squid (*Todarodes pacificus*), octopus (*Octopus vulgaris*), Echinodermata species, sea urchin (*Paracentrotus lividus*), water flea (*Daphnia pulex*), hermit crab (*Pagurus longicarpus*), and Humboldt squid (*Dosidicus gigas*) [53, 62]. In addition, large positive (G – A) values in the three fragments were observed in *Paragonimus westermani*, while large positive (G – C) and (A – T) values in the three fragments were observed for the mtDNA of representatives of the following phyla: Cnidaria (*Pavona clavus*), Platyhelminthes (*Schistosoma mansoni*), Porifera (*Geodia neptuni*), Arthropoda (*Tigriopus californicus*), and Chordata (*Branchiostoma belcheri*) [53]. Furthermore, the following invertebrate

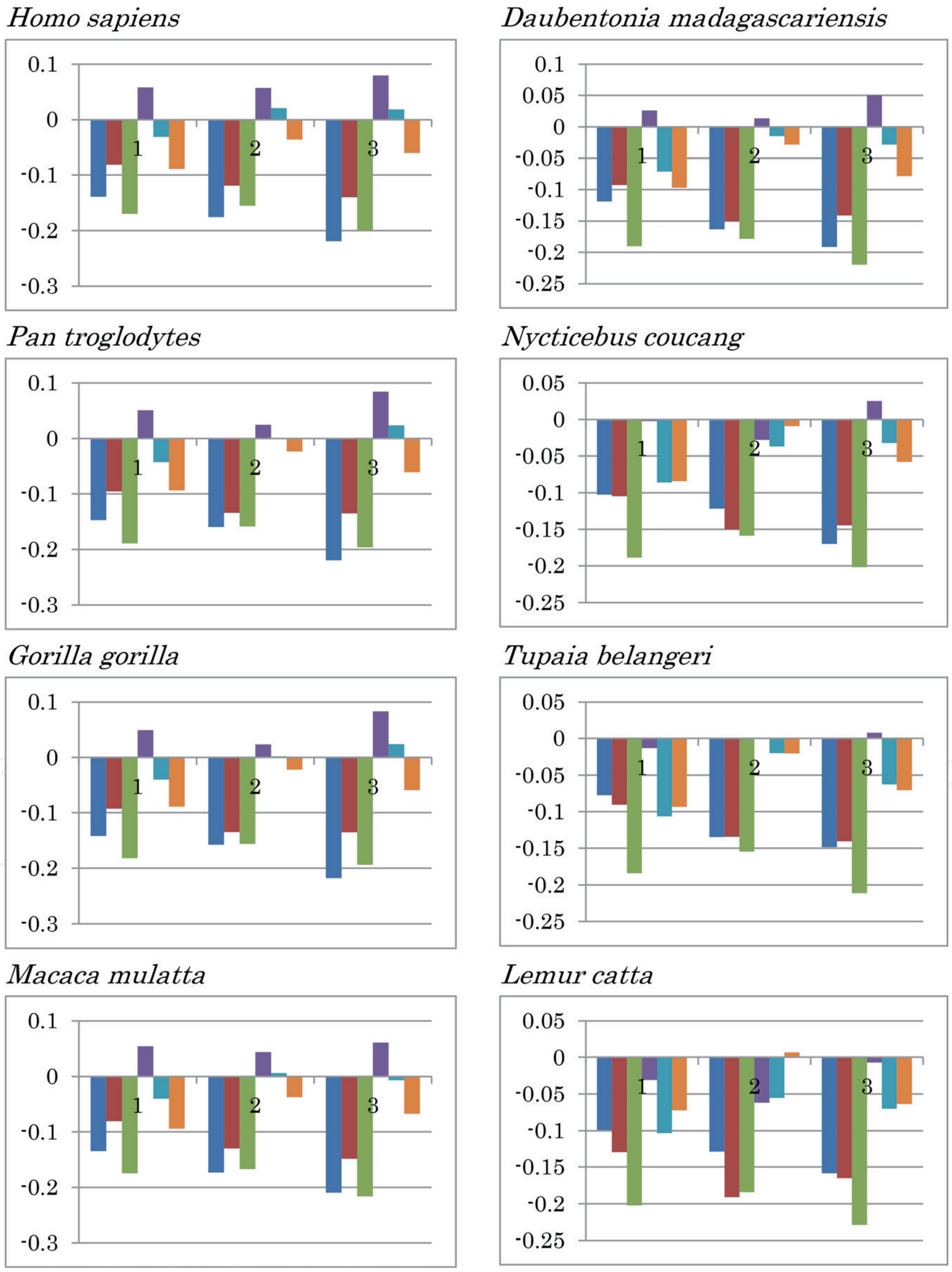

Figure 11.
Nucleotide differences in the three fragments of each primate mitochondrial genome. Left to right: (G – C), (G – T), (G – A), (C – T), (C – A), and (T – A).

mitochondria were also examined: *Acanthaster planci*, *Haliotis rubra*, *Lampsilis ornate,* and the mtDNA of hemichordates, *Saccoglossus kowalevskii, Balanoglossus carnosus*, and *Xenoturbella bocki* was examined [53].

In the mtDNA of primate species *H. sapiens*, *P. troglodytes*, *G. gorilla*, *Macaca mulatta*, *Daubentonia madagascariensis*, *Nycticebus coucang*, and *Tupaia belangeri*, nucleotide content difference patterns were quite similar in the first four species, and large positive increases in (C – T) differences in the three fragments clearly indicated evolutionary divergence (**Figure 11**). The positive (C – T) differences in all three fragments were characteristic of these four primate mitochondria, while positive increases in (C – T) values were only observed in the third fragment of *N. coucang* and *T. belangeri* mtDNA. In contrast, nucleotide content difference patterns of the prosimian *Lemur catta* completely differed from those of the primates, although TA inversion was observed in the second fragment. The primate mtDNA nucleotide content patterns were also completely different from that of hemichordate *B. carnosus*, although their C contents were the highest among all organisms examined [59]. This finding indicates that mitochondrial structures respect epigenomic evolutionary functions.

12. Definitive universal equations

In the normalization of nucleotide contents (G + C + A + T = 1), as (G = C) and (A = T) based on Chargaff's parity rules, (2G + 2A = 1) is obtained. This equation is altered to (A = 0.5 – G) and then (A – G = 0.5 – 2G). Finally, G – A = 2G – 0.5. The relationship between (G – A) and (G) is linear when both (G) and (A) are expressed by linear functions. In animal mitochondria, only the correlations between the two purines (A versus G) or the two pyrimidines (C versus T) are linear, while the correlations between purines and pyrimidines (A or G versus T or C) are weak or not correlated at all [62]. For example, when plotting (G – C), (G – T), (G – A), (C – T), (C – A), and (T – C) against G content, only (G – A) versus G content was linear in vertebrate mitochondria [59]. In invertebrate mitochondria, plotting nucleotide content differences against G content was weakly linear.

Plotting (X – Y)/(X + Y) against (X – Y), the following linear relationship was obtained in mitochondria, chloroplasts, and chromosomes (**Figure 12**): (X – Y)/(X + Y) = a (X – Y) + b, where X and Y are nucleotide contents, and (a) and (b) are constants. As (b) was almost null and (a) was ~2.0, (X – Y)/(X + Y) ≈ 2.0 (X – Y). In these genome analyses, which are independent of Chargaff's parity rules, the values of (a) for (G, C), (G, A), (G, T), (C, T), (C, A), and (A, T) were 2.5858, 1.85558, 1.9908, 1.9771, 1.9968, and 1.5689, respectively, in our previous results [53, 54]. Based on these results, (G + C), (G + A), (G + T), (C + A), (C + T), and (A + T) were 0.39, 0.54, 0.50, 0.51, 0.50, and 0.64, respectively. In virus genome analyses [53, 54], the constant values for (a) were 1.9–2.1, and the values for (X + Y) were 0.47–0.53. In contrast, in the normalization of nucleotide contents (G + C + A + T = 1), as (G = C) and (A = T) based on Chargaff's parity rules, (2G + 2A = 1) is obtained. This equation is altered to (G + A = 0.5). This value is consistent with the value obtained above from genome analyses. Similarly, (G + T = 0.5), (C + A = 0.5), and (C + T = 0.5), although (G + C) and (A + T) cannot be determined. Therefore, the four nucleotide contents are expressed by the following regression lines, plotted against G content: A = 0.5 – G, T = 0.5 – G, C = G, and G = G. Lines G and C overlap, as do lines A and T, and the former line is symmetrical to the latter against line (y = 0.25). The intercepts of lines G and C are close to the origin, while those of lines A and T are close to 0.5 at the vertical and horizontal axes. All organisms from bacteria to *H. sapiens* are located on the diagonal lines of a 0.5 square, termed the “Diagonal Genome Universe,” using the normalized values that obey Chargaff’s first parity rule [12]. These relationships lead to (G or C) + (A or T) = 0.5. The present results indicate that a linear regression line equation, (X – Y)/(X + Y) = a (X – Y) + b, universally represents all normalized values, including the values deviating from Chargaff’s parity rules. This newly discovered equation clearly reflects not only Chargaff’s first parity rules, based on hydrogen bonding between two nucleotides, but also natural rule.

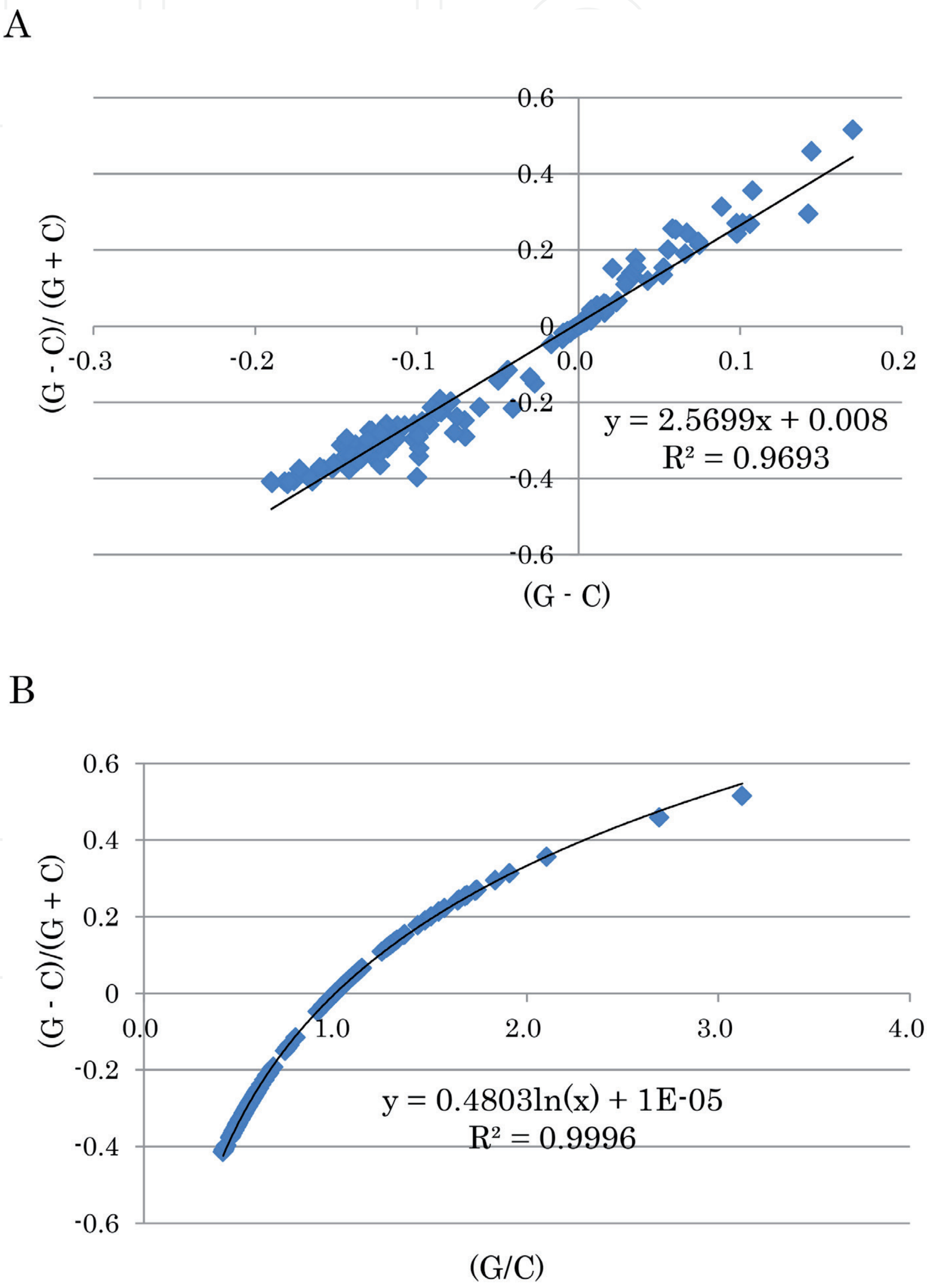

Figure 12.
Universal rules. The following genome samples were examined: mitochondria of vertebrates (65), invertebrates (54), and non-animals (42), chloroplasts (28), prokaryote chromosomes (21), and eukaryote chromosomes (15). Left side: relationship between (X – Y) and (X – Y)/(X + Y) and right side: relationship between (X/Y) and (X – Y)/(X + Y).

A linear regression line was not obtained when using randomly chosen value (**Figure 12A**). Furthermore, plotting (X – Y)/(X + Y) against (X/Y), the following logarithmic function was obtained for all tested genomes as well as when using randomly chosen values (**Figure 12B**): (X – Y)/(X + Y) = a ln (X/Y) + b. As (b) was almost null and (a) was ~0.5, (X – Y)/(X + Y) ≈ 0.5 ln (X/Y). The ratio between two values, (X/Y), can be expressed by a logarithmic function, ~0.5 ln (X/Y) ≈ (X – Y)/(X + Y). Plotting the GC skew vs. G content, animal mitochondria were classified into two groups: high and low C/G [59]. This fact indicates that the ratio C/G and the GC skew are evolutionarily related to each other. Any change can be expressed universally by a definitive logarithmic function, (X – Y)/(X + Y) = a ln (X/Y) + b. The present results indicate that cellular organelle evolution is strictly controlled under these characteristic rules, although non-animal mitochondria, chloroplasts, and chromosomes are controlled under Chargaff's parity rules [12, 14]. The present study clearly shows that biological evolution, which seems to be based on complicated processes, is governed by simple universal equations.

13. Conclusions

The ratios of amino acids to the total amino acids or of nucleotides to total nucleotides predicted from complete genomes consisting of huge number of nucleotides can characterize a whole organism. In addition, as these values are independent of species and genome size, these indexes are very useful for genome research, as well as single gene research. The validity of these indexes is clearly based on the homogeneity of genomic structures. In addition, patternalization of values after simple calculations based on large data sets can provide an intuitive picture and provide useful insights, revealing the homogeneity of genomic structures followed by synchronous alterations over the genome. In addition, any change between two values, X and Y, including biological evolution can be expressed definitively by a linear regression line equation, (X – Y)/(X + Y) = a (X – Y) + b, where X and Y are nucleotide contents, and (a) and (b) are constants, and by a logarithmic function, (X – Y)/(X + Y) = a′ ln (X/Y) + b′, where (a′) and (b′) are constants. As the present review is based on the endeavors and data of numerous scientists from all over the world, the author would like to express finally his following feeling as one of scientists. (Human being is an organism of huge numbers of organisms on the Earth, and we are not ranked as a special species above all organisms as a result of long evolution.) However, we have made the present modern civilization based on fossil energy usage which seems to induce climate changes. Thus, we must be responsible to establish sustainable development not only for Human being but also for other organisms. The Earth is for all organisms, not only for Human being.

Acknowledgements

The author greatly acknowledges President Hiroyuki Okada of Shinko Sangyo, Co. Ltd., Takasaki, Gunma, Japan for his financial support and Dr. Teiji Okayasu who was one of collaborators in Dokkyo Medical University for his excellent computer analyses.

Author details

Kenji Sorimachi[1,2]

1 Educational Support Center, Dokkyo Medical University, Tochigi, Japan

2 Research Laboratories, Gunma Agriculture and Forest Development Com., Ltd., Takasaki, Gunma, Japan

*Address all correspondence to: kenjis@jcom.home.ne.jp

References

[1] Sanger F, Coulson AR. A rapid method for determining sequences in DNA by printed synthesis with DNA polymerase. Journal of Molecular Biology. 1975;**94**:441-446

[2] Maxam AM, Gilbert W. A new method for sequencing DNA. Proceedings of the National Academy of Sciences of the United States of America. 1977;**74**:560-564

[3] Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science. 1995;**269**:496-512

[4] Lander ES, Linton ML, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;**409**:860-921

[5] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 2001;**291**:1304-1351

[6] Zuckerkandl E, Pauling LB. Molecular disease, evolution, and genetic heterogeneity. In: Kasha M, Pullman B, editors. Horizons in Biochemistry. New York: Academic Press; 1962. pp. 189-225

[7] Dayhoff MO, Park CM, McLaughlin PJ. Building a phylogenetic trees: Cytochrome C. In: Dayhoff MO, editor. Atlas of Protein Sequence and Structure. Vol. 5. Washington, D.C.: National Biomedical Foundation; 1977. pp. 7-16

[8] Sogin ML, Elwood HJ, Gunderson JH. Evolutionary diversity of eukaryotic small subunit rRNA genes. Proceedings of the National Academy of Sciences of the United States of America. 1986;**83**:1383-1387

[9] Doolittle WF, Brown JR. Tempo, mode, the progenote, and the universal root. Proceedings of the National Academy of Sciences of the United States of America. 1994;**91**:6721-6728

[10] Maizels N, Weiner AM. Phylogeny from function: Evidence from the molecular fossil record that tRNA originated in replication, not translation. Proceedings of the National Academy of Sciences of the United States of America. 1994;**91**:6729-6734

[11] DePouplana L, Turner RJ, Steer BA, Schimmel P. Genetic code origins: tRNAs older than their synthetases? Proceedings of the National Academy of Sciences of the United States of America. 1998;**95**:11295-11300

[12] Chargaff E. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. Experientia. 1950;**VI**:201-209

[13] Watson JD, Crick FHC. Genetical implications of the structure of deoxyribonucleic acid. Nature. 1953;**171**:964-967

[14] Rundner R, Karkas JD, Chargaff E. Separation of *B. subtilis* DNA into complementary strands. 3. Direct analysis. Proceedings of the National Academy of Sciences of the United States of America. 1968;**60**:921-922

[15] Sorimachi K. A proposed solution to the historic puzzle of Chargaff's second parity rule. Open Genomics Journal. 2009;**2**:12-14

[16] Mitchell D, Bridge R. A test of Chargaff's second rule. Biochemical and Biophysical Research Communications. 2006;**340**:90-94

[17] Sueoka N. Correlation between base composition of deoxyribonucleic acid

and amino acid composition in proteins. Proceedings of the National Academy of Sciences of the United States of America. 1961;**47**:1141-1149

[18] Sorimachi K. Evolutionary changes reflected by the cellular amino acid composition. Amino Acids. 1999;**17**:207-226

[19] Chou K-C. Review: Applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady state systems. Biophysical Chemistry. 1990;**35**:1-24

[20] Qi XQ, Wen J, Qi ZH. New 3D graphical representation of DNA sequence based on dual nucleotides. Journal of Theoretical Biology. 2007;**249**:681-690

[21] Sorimachi K, Itoh T, Kawarabayasi Y, Okayasu T, Akimoto K, Niwa A. Conservation of the basic pattern of cellular amino acid composition during biological evolution and the putative amino acid composition of primitive life forms. Amino Acids. 2001;**21**:393-399

[22] Schopf JW, Barghoorn ES, Maser MD, Gordon RO. Electron microscopy of fossil bacteria two billion years old. Science. 1965;**149**:1365-1367

[23] MacGregor IM, Truswell JF, Eriksson KA. Filamentous alga from the 2,300 m.y. old Transvaal dolomite. Nature. 1974;**247**:538-539

[24] Nagy LA, Zumberge JE. Fossil microorganisms from the approximately 2800 to 2500 million-year-old Bulawayan stromatolite: Application of ultramicrochemical analyses. Proceedings of the National Academy of Sciences of the United States of America. 1976;**73**:2973-2976

[25] Sorimachi K, Okayasu T. Mathematical proof of the chronological precedence of protein formation over codon formation. Current Topics in Peptide & Protein Research. 2007;**8**:25-34

[26] Gilbert W. The RNA world. Nature. 1986;**319**:618

[27] Crick FHC. The origin of genetic code. Journal of Molecular Biology. 1968;**38**:367-379

[28] Wong JT-F. A co-evolutionary theory of the genetic code. Proceedings of the National Academy of Sciences of the United States of America. 1975;**72**:1909-1912

[29] Woese CR. Order in the genetic code. Proceedings of the National Academy of Sciences of the United States of America. 1965;**54**:71-75

[30] Kvenvolden K, Lawless J, Pering K, Peterson E, Flores J, Ponnamperuma C, et al. Evidence for extraterrestrial amino-acids and hydrocarbons in the Murchison meteorite. Nature. 1970;**228**:923-926

[31] Wolman Y, Haverland W, Miller SL. Nonprotein amino acids from spark discharges and their comparison with the Muchison meteorite amino acids. Proceedings of the National Academy of Sciences of the United States of America. 1972;**69**:809-811

[32] Miller SL. A production of amino acids under possible primitive earth conditions. Science. 1953;**117**:528-529

[33] Sorimachi K, Okayasu T. Gene assembly consisting of small units with similar amino acid composition in the *Saccharomyces cerevisiae* genome. Mycoscience. 2003;**44**:415-417

[34] Sorimachi K, Okayasu T. An evolutionary theories based on genomic structures in *Saccharomyces cerevisiae* and *Enchephalitozoon cuniculi*. Mycoscience. 2004;**45**:345-350

[35] Sorimachi K, Okayasu T, Ebara Y, Nakagawa T. Mathematical proof of genomic amino acid composition homogeneity based on putative small units. Dokkyo Journal of Medical Sciences. 2005;**32**:99-100

[36] Sorimachi K. Evolution based on genome structure: The "diagonal genome universe". Natural Science. 2010;**2**:1104-1112

[37] Sorimachi K, Okayasu T. Universal rules governing genome evolution expressed by linear formulas. Open Genomics Journal. 2008;**1**:33-43

[38] Sorimachi K, Okayasu T. Codon evolution is governed by linear formulas. Amino Acids. 2008;**34**:661-668

[39] Bell SJ, Forsdyke DR. Deviations from Chargaff's second parity rule with direction of transcription. Journal of Theoretical Biology. 1999;**197**:63-76

[40] Nikolaou C, Almirantis Y. Deviations from Chargaff's second parity rule in organelle DNA insights into the evolution of organelle genomes. Gene. 2006;**381**:34-41

[41] Sorimachi K. Genomic data provides simple evidence for a single origin of life. Natural Science. 2010;**2**:519-525

[42] Sorimachi K, Okayasu T, Ohhira S, Fukasawa I, Masawa N. Evidence for the independent divergence of vertebrate and high C/G ratio invertebrate mitochondria from the same origin. Natural Science. 2012;**4**:479-483

[43] Sorimachi K. Evolution from primitive life to *Homo sapiens* based on visible genome structures: The amino acid world. Natural Science. 2009;**1**:107-119

[44] Cobbett A, Wilkinson M, Wills M. Fossils impact as hard as living taxa in parsimony analyses of morphology. Systems Biology. 2007;**17**(2007):753-766

[45] Weisburg WG, Barns SM, Pelletier DA, Lane DJ. 16S ribosomal DNA amplification for phylogenetic study. Journal of Bacteriology. 1991;**173**:697-703

[46] Okayasu T, Sorimachi K. Organisms can essentially be classified according to two codon patterns. Amino Acids. 2009;**36**:261-271

[47] Sorimachi K, Okayasu T, Ohhira S, Masawa N, Fukasawa I. Natural selection in vertebrate evolution under genomic and biosphere biases based on amino acid content: Primitive vertebrate hagfish (*Eptatretsu burgeri*). Natural Science. 2013;**5**:221-227

[48] Saitou N, Nei M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molecular Biology and Evolution. 1987;**4**:406-425

[49] Sorimachi K, Okayasu T, Ebara Y, Furuta E, Ohhira S. Phylogenetic position of *Xenoturubella Bocki* and Hemichordates *Balanoglossus carnosus* and *Saccoglossus kowalevskii* based on amino acid composition or nucleotide content of complete mitochondrial genomes. International Journal of Biology. 2014;**6**:82-94

[50] Sorimachi K, Okayasu T. Claasification of non-animals and invertevrates based on amino acid composition of complete mitochondrial genomes. International Journal of Biology. 2014:1-6

[51] Ward JH. Hierarchic grouping to optimize an objective function. Journal of the American Statistical Association. 1963;**58**:236-244

[52] Janvier P. Micro RNAs revive old views about jawless vertebrate divergence and evolution. Proceedings

of the National Academy of Sciences of the United States of America. 2010;**107**:19137-19138

[53] Sorimachi K. The most primitive extant ancestor of organisms and discovery of definitive evolutionary equations based on complete genome structures. Natural Science. 2018;**10**(9):338-369

[54] Lobry JR. Asymmetric substitution patterns in the two DNA strands of bacteria. Molecular Biology and Evolution. 1996;**13**:660-665

[55] Tillier ER, Collins RA. The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. Journal of Molecular Evolution. 2000;**50**:249-257

[56] Anderson S et al. Sequence and organization of the human mitochondrial genome. Nature. 1981;**290**:457-465

[57] Fonceca MM, Harris DJ, Posada D. The inversion of the control region in three mitogenomes pro vides further evidence for an asymmetric model of vertebrate mtDNA replication. PLoS One. 2014;**9**:e106654

[58] Seligmann H. Coding constraints modulate chemically spontaneous mutational replication gradients in mitochondrial genomes. Current Genomics;**13**:37-54

[59] Sorimachi K. Origine of life in the ocean: Direct derivation of mitochondria from primitive organisms based on complete genomes. Current Chemical Biology. 2015;**9**:23-35

[60] King N et al. The genome of the choanoflagellates Monosigarevicollis and the origin of metazoans. Nature. 2008;**451**:783-788

[61] Andersson SG et al. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature. 1998;**396**:133-140

[62] Sorimachi K. Codon evolution in double-stranded organelle DNA: Strong regulation of homonucleotides and their analog alternations. Natural Science. 2010;**2**:846-854