Send Orders for Reprints to reprints@benthamscience.ae

Current Genomics, **2015**, *16*, **99-106** **99**

Normalization of Complete Genome Characteristics: Application to Evolution from Primitive Organisms to *Homo sapiens*

Kenji Sorimachi[1,2,*], Teiji Okayasu[3] and Shuji Ohhira[4]

[1]*Educational Support Center, Dokkyo Medical University, Mibu, Tochigi 321-0293, Japan;* [2]*Life Science Research Center, Higashi-Kaizawa, Takasaki, Gunma 370-0041, Japan;* [3]*Center for Medical Informatics, Dokkyo Medical University, Mibu, Tochigi 321-0293, Japan;* [4]*Laboratory for International Environmental Health, Dokkyo Medical University, Tochigi 321-0293, Japan*

Abstract: Normalized nucleotide and amino acid contents of complete genome sequences can be visualized as radar charts. The shapes of these charts depict the characteristics of an organism's genome. The normalized values calculated from the genome sequence theoretically exclude experimental errors. Further, because normalization is independent of both target size and kind, this procedure is applicable not only to single genes but also to whole genomes, which consist of a huge number of different genes. In this review, we discuss the applications of the normalization of the nucleotide and predicted amino acid contents of complete genomes to the investigation of genome structure and to evolutionary research from primitive organisms to *Homo sapiens*. Some of the results could never have been obtained from the analysis of individual nucleotide or amino acid sequences but were revealed only after the normalization of nucleotide and amino acid contents was applied to genome research. The discovery that genome structure was homogeneous was obtained only after normalization methods were applied to the nucleotide or predicted amino acid contents of genome sequences. Normalization procedures are also applicable to evolutionary research. Thus, normalization of the contents of whole genomes is a useful procedure that can help to characterize organisms.

Keywords: Amino acid composition, Chargaff's parity rules, Cluster analysis, Evolution, Genome, Normalization, Nucleotide content, Phylogenetic trees.

INTRODUCTION

Molecular biology and parentology have contributed to our understanding of evolution using phylogenetic trees based on nucleotide or amino acid sequence changes and on morphological changes detected in fossils, respectively. In these studies, evolutionary divergences have been evaluated as the degree of similarity or difference in gene structures based on nucleotide sequences or on fossil shapes and sizes. Numerous phylogenetic trees have been drawn with single genes such as cytochrome c [1], tRNA [2-4], and rRNA [5], and fossils that indicate new species have been found in many different places. It has been almost two decades since the complete genome of *Haemophilus influenzae* was first analyzed in 1995 [6], and the first human genome draft was reported in 2001 [7, 8]. New data are continuously increasing and, to date, the complete genomes of 285 eukaryotes, 2898 bacteria, and 173 archaea have been sequenced. However, common data analysis methods that focus on the nucleotide or amino acid sequences of single gene(s) have been inadequate in evaluating whole-genome characteristics of many species, although, recently, whole genomes of *Homo sapiens*, *H. neanderthalensis*, the Denisova specimen, and chimpanzees (*Pan troglodytes*) were compared [9].

*Address correspondence to this author at the Life Science Research Center, Higashi-Kaizawa, Takasaki, Gunma 370-0041, Japan; Tel: 81-27-352-2955; E-mail: kenjis@jcom.home.ne.jp

Regarding the normalization of complete genomes consisting of a large number of different genes, the choice of target is particularly important: four nucleotide contents are simple, whereas 20 amino acid compositions are more complex because of many influencing factors. For example, the amino acid composition may differ according to transcriptional levels and analytical methods. To evaluate normalization methods for high-throughput RNA sequencing data analysis, Dillies *et al.* [10] compared the data obtained from real and simulated datasets of four species, *H. sapiens*, *Entamoeba histolytica*, *Aspergillus fumigatus*, and *Mus musculus*, and from seven different analyses. However, the ideal normalization method has not yet been obtained. Indeed, we encountered difficulties in our early studies regarding the gene expressions of whole genomes consisting of many different genes [11]. To our knowledge, that article is the first report that showed the amino acid composition predicted from the complete genome. Finally, we decided to assume that all genes are equally expressed in the entire genome, and that the data are coincidentally consistent between genomic and experimental analyses based on cell hydrolyzates [11], as shown in the later section of this review article. This assumption has been applied to our subsequent studies, to date. However, normalization, which is independent of target size and type, is a useful approach to characterize whole genomes. Particularly, in normalization based on nucleotide and amino acid sequences, experimental errors are theoretically excluded and the normalized values, which are accurate, can represent completely the characteristics of the target

1875-5488/15 $58.00+.00

organisms. In the normalization of small nucleotides or peptides, the normalized values are equal when the ratios of each of the four nucleotides to the total nucleotides are equal among samples, even though their sequences differ. However, for polynucleotide, polypeptide, or genome sequences that consist of large numbers of nucleotides, the normalized values are never equal among samples. In addition, by visualizing the values, complicated phenomena can be easily understood [12]. For example, investigations have been carried out based on molecular structure changes, DNA structure [13], evolution-driven protein structural changes [14], drug resistance mechanisms [15], cancer-associated single nucleotide polymorphisms [16] and protein–protein interactions involving point mutations [17]. In this review, we introduce the methods that have been used for the normalization of complete genome characteristics such as nucleotide and amino acid contents, and show some of the results related to genome structure and evolution.

Homogeneous Genome Structure

We first investigated whether it was reasonable to assume that the characteristics of a whole genome, which consists of a huge number of different genes with different nucleotide sequences, can be expressed simply by normalization of the nucleotide and amino acid contents. We found that when a mouse complementary DNA (cDNA) dataset [18] consisting of 10,465 genes was divided into two equal halves, the amino acid composition predicted from the first 5, 10, 50, 100, 500, 1,000, 5,232 and 5,233 genes, according to the order listed in the dataset, was similar in the two halves and within the same part (Fig. **1**). Indeed, the amino acid compositions based on more than 10 genes resembled each other and that of the whole dataset (Fig. **1**). This finding implies that genome structure is made up of putative small units that encode sequences that have similar amino acid compositions, even though each gene in these small units has a different nucleotide sequence, as shown in (Fig. **1**). To investigate this result further, the complete genome of *Treponema pallidum*, which consists of 1,031 genes [19], was divided into two equal halves (Fig. **2A**), and the amino acid composition of each half and of the complete genome was calculated. The amino acid compositions of each half and of the whole genome were very similar (Fig. **2B**). Next, we divided all the genes into nine groups of 103 genes each and one group of 104 genes, and then divided the 10 groups into 20 half-size groups consisting of 52 genes each (Fig. **2A**). The amino acid compositions for all the genes in each group were all similar to each other (Fig. **2C** and **2D**). Thus, we have shown that the genomes were homogeneously constructed with putative units with similar amino acid compositions encoded in the open reading frames, even though each gene encodes a clearly different amino acid sequence. We found that the amino acid composition of the 3,236 amino acid residues encoded by the first 10 genes in one of the units resembled the amino acid composition encoded by all the genes in the genome. The amino acid compositions calculated from the first gene encoding 151–677 amino acid residues in each of the 10 units differed from each other and from that of the complete genome (Fig. **2E**). However, the largest gene, which encoded 1,517 amino acid residues, also had an amino acid composition that was similar to that of the complete genome. In general, the size of the coding unit that showed similar amino acid composition was reported to be between 3,000 and 7,000 amino acid residues [20, 21]. This coding unit size is the same for bacteria, archaea, and eukaryotes. The fact that genome structure is homogeneous suggests that mutations occur synchronously over the whole genome. This idea led to the construction of phylogenetic trees based on different genes in the same organism. However, phylogenetic trees are not absolute because their form depends on the analytical methods and traits that are used to construct them.

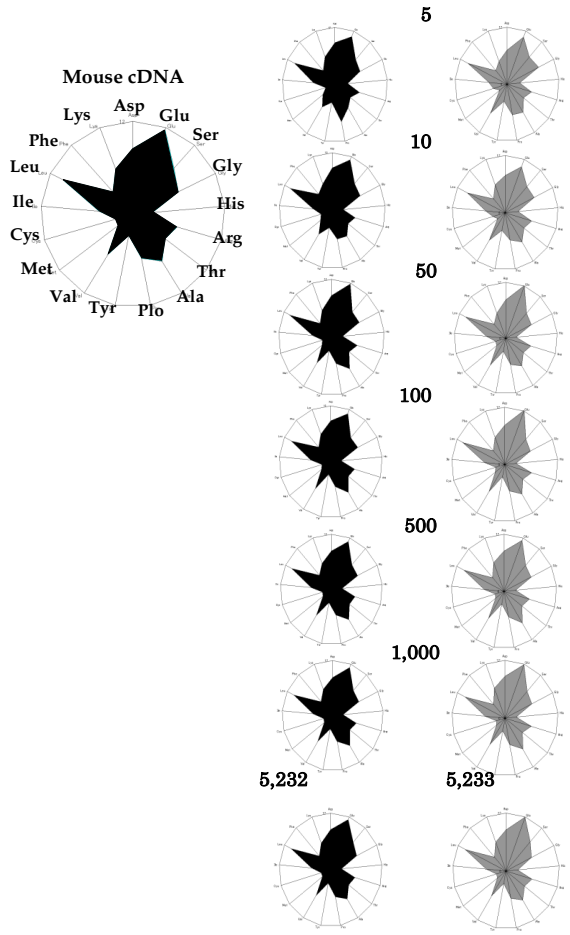

Fig. (1). The translated amino acid sequences (10,465) of FANTOM clones were divided into two equal parts; onehalf from the 5' end of the mouse genome (black) and the other half from the 3' end of the genome (gray). The amino acid compositions encoded by the first 5, 10, 50, 100, 500, and 1,000 genes in each half were analyzed. There were 5,232 and 5, 233 genes in the first and second halves, respectively. The value of each amino acid was expressed as the percentage of the total number of amino acids and all values were visualized on a radar graph. In our previous studies [11, 26], cell homogenates were hydrolyzed in 6N HCl at 105°C for 24 h, and then the cellular amino acids were analyzed using a Hitachi L8500A amino acid analyzer (Hitachi, Tokyo, Japan). In the previous analysis, the glutamine and asparagine residues were converted to glutamic acid and aspartic acid, respectively; thus, the values for the acidic forms represent the sums of both forms. In the present study, the values that were used represent the sums of both values. Tryptophan was omitted in our previous [11, 26] and present study-because this amino acid decomposed during hydrolysis and its value was 1% of the total codon frequency, based on 35 human genes [42]. This figure first appeared in *Natural Science*, **2010**, 2 (10), 1104-1112, and is reproduced with permission.

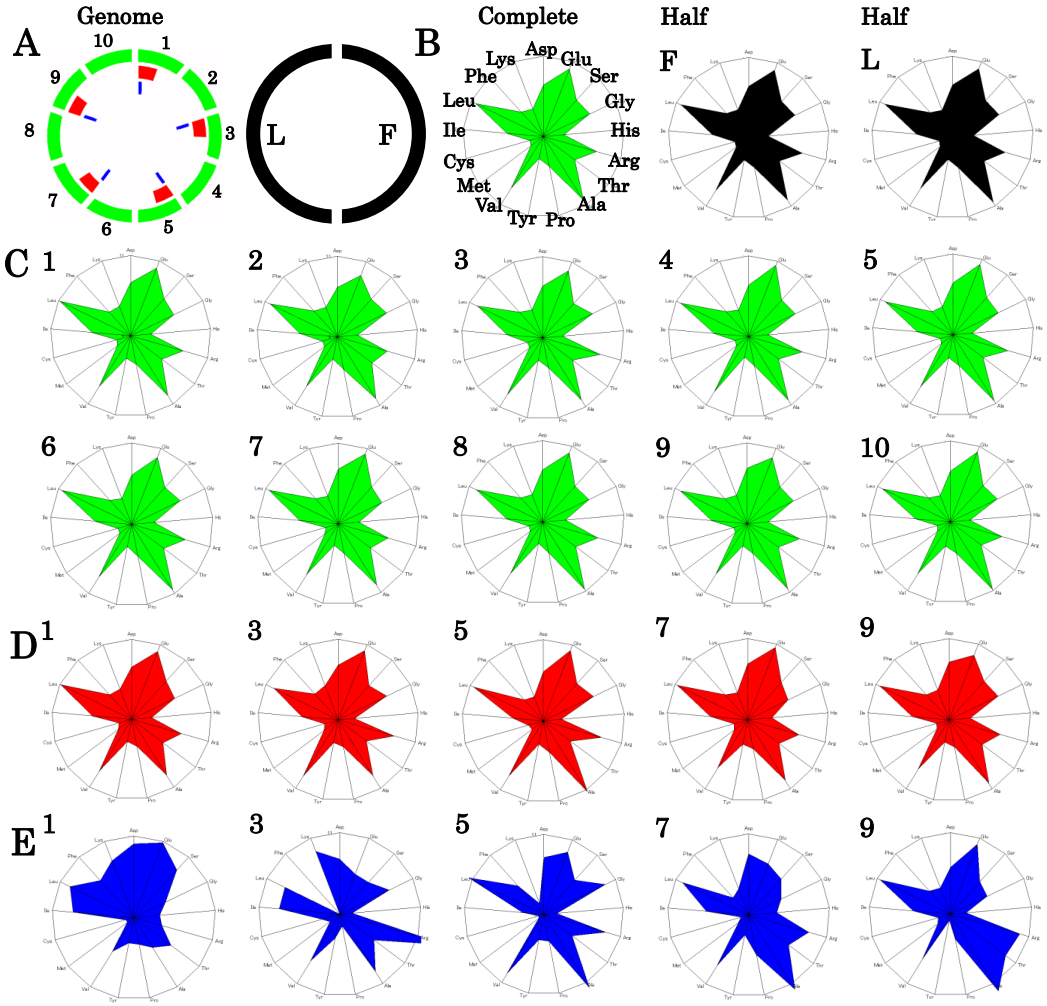

Fig. (2). Radar charts of amino acid compositions calculated from various units of the complete genome of *Treponema pallidum*. **A.** The complete genome, which consisted of 1,031 protein-coding genes and 51 RNA genes [19], was divided into two equal halves (black), and 10 (green) and 20 (red) groups. **B.** Radar charts of the amino acid compositions of the two halves of the genome (F, 515 and L, 516 genes). **C**. Radar charts of the amino acid compositions of the 10 units; 9 with 103 and 1 with 104 genes. **D.** Representative radar charts of the amino acid compositions of the 20 units (52 genes). **E.** Radar charts of the amino acid compositions of the first single gene in each of the 10 units (Fig. **2C**).

Normalization of Nucleotide and Amino Acid Contents

Normalized nucleotide and predicted amino acid contents of complete genomes can be visualized on radar charts where the proportion of each of the four nucleotides or 20 amino acids is plotted on one of the radii or on bar graphs. The normalized values for complete genome, polynucleotide, and polypeptide sequences are always different for different species. Thus, organisms can be compared based on the shapes of the resultant polygons in the radar charts, the values of the 4 and 20 angle points, and their combinations, which can be used as traits in Ward's clustering analyses [22]. A neighbor-joining method [23] yielded consistent results using amino acid and nucleotide contents as traits [24, 25]. The normalized value is indicated by the length of the bar in the graphs. It is also possible to compare organisms with only one single point value, where each single value represents the whole organism, using normalized nucleotide or amino acid contents calculated from the complete genome. Indeed, we found that the purine content of complete mitochondria sequences was enough to classify vertebrates into two groups, aquatic and terrestrial, although the pyrimidine content did not produce the same result [unpublished data]. Consistent classifications have been obtained from Ward's clustering analysis using normalized amino acid or nucleotide contents as well as from a neighbor-joining method using 16S rRNA gene sequences [24, 25].

Primitive Life Forms

It is impossible to evaluate the characteristics of primitive organisms that are now extinct. However, because natural rules are universal, some of their characteristics can be predicted based on rules obtained from extant organisms. Bacteria that are found as fossils seem to be most closely related to primitive life forms [11, 26]. The normalized amino acid composition of *Escherichia coli* shows a characteristic pattern which resembles a "starfish shape" on the radar chart (Fig. **3**). The protist, *Monosiga brevicollis*, which is thought to be close to the origin of multi-cellular organisms shows a similar amino acid composition pattern as

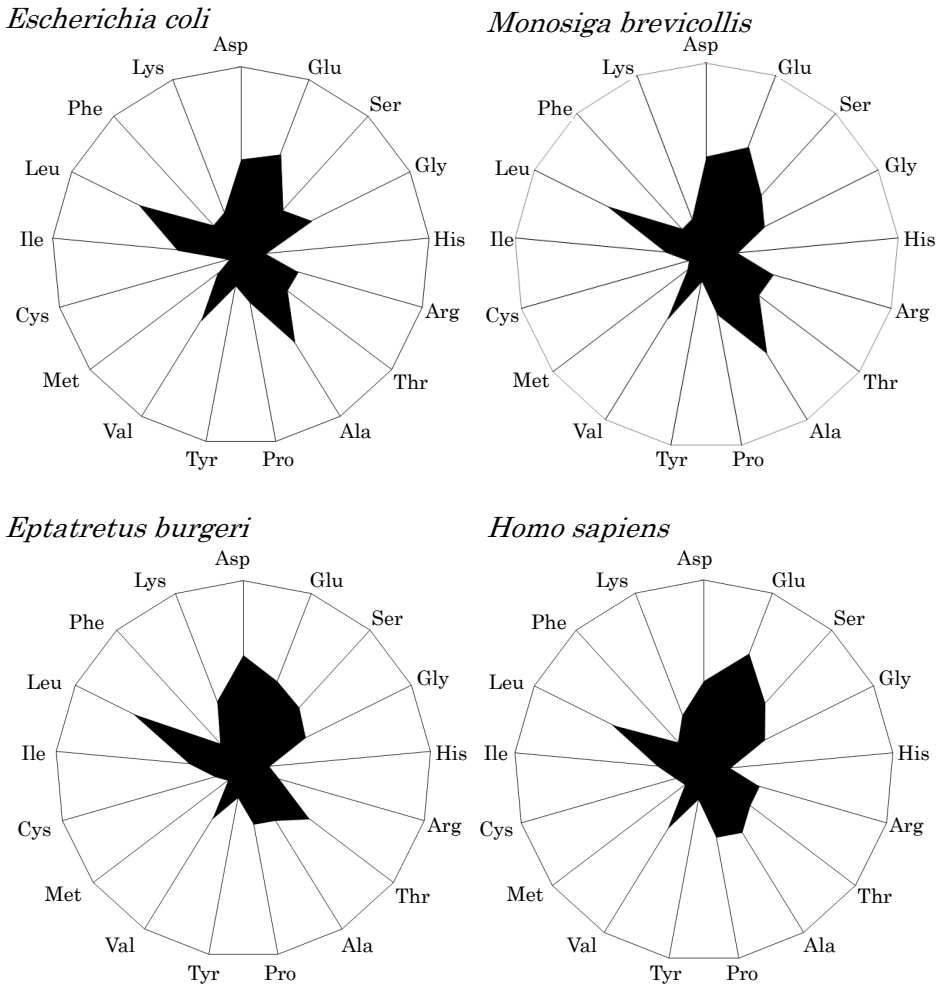

Fig. (3). Radar charts of the amino acid compositions of four complete genomes.

E. coli on the radar chart (Fig. **3**). The normalized amino acid composition pattern of hagfish (*Eptatretus burgeri*), which may be close to the origin of vertebrates [27], also resembles that of *E. coli*. The amino acid composition pattern of human (*Homo sapiens*) also resembles that of *E. coli*, as shown in (Fig. **3**). These patterns from the complete genomes imply that the encoded amino acid composition of genomes is similar from bacteria to eukaryotes. The amino acid compositions of 11 Gram-positive and 12 Gram-negative bacteria were classified into two groups, "S-type" represented by *Staphylococcus aureus* and "E-type" represented by *Escherichia coli*, based on their patterns of amino acid compositions predicted from the complete genome [28]. The amino acid composition based on the plasmid resembled that based on the parent complete genome [28]. This two group classification was independent of Gram staining [28]. The cellular amino acid composition of bacteria was first analyzed using cell hydrolysates by Sueoka [29]. We also analyzed independently the cellular amino acid compositions of cell hydrolysates obtained from bacteria, archaea, and eukaryotes [26]. The basic "starfish shape" pattern was obtained from all these cell hydrolysates [26]. Differences in the "starfish shape" reflect the evolutionary changes that have occurred in genomes. It is curious that the amino acid composition of complete genomes resembles the amino acid composition obtained from cell hydrolysates, because both values are based on different methods and different origins. However, we subsequently found that this coincidence is a result of the homogeneous genome structure. In cell hydrolysates, because differences in gene expression levels are channeled within putative small units that show similar amino acid compositions after normalization, the total cellular amino acid composition is independent of the expression and transcription levels of each gene. Thus, both cellular and genome-normalized amino acid compositions were almost equal. It can be speculated that primitive organisms may have similar amino acid compositions as extant organisms, because all extant organisms, including bacteria, archaea, and eukaryotes, have similar normalized amino acid compositions [11, 26]. This conclusion was obtained only after the nucleotide or amino acid contents of sequences were normalized across whole genomes. This conclusion is independent of whether cell hydrolysates or complete genomes are used.

Rules Governing Genome Evolution

We can speculate about past and future evolutionary phenomena based on the rules that govern the evolution of present-day organisms. By plotting the normalized nucleotide contents separately for the four nucleotides, it was found that the nucleotide contents for three nucleotides could be expressed by the

nucleotide content of the fourth using three linear regression equations [30]. This rule was applicable to complete genome sequences and coding and non-coding regions, using chromosomal [30], chloroplast [31], or plant mitochondrial genomes [31]. The nucleotide relationships were heteroskedastic in animal mitochondria [31], while the relationships between homonucleotides and their analogues, and between heteronucleotides and their analogues were linear and heteroskedastic, respectively, in animal mitochondria. These were divided into two groups, high G/C and low G/Ccontent [32]. When the nucleotide contents of coding or non-coding regions were plotted against the nucleotide contents of complete genomes, two regression lines based on chloroplast and plant mitochondria sequences were found to be closed at the edge point, because the maximum nucleotide contents in either the coding or non-coding regions against the complete genome nucleotide contents were equal in the two organelles [31].

Normalization of nucleotides was first carried out by Chargaff [33] to characterize cellular DNA consisting of double strands, as Chargaff's first parity rule [G = C, T = A, and (G + A) = (C + T)][33]. Subsequently, the rule was expanded to single-stranded DNA forming double-stranded DNA, as Chargaff's second parity rule [34]. The first rule was discovered before the Watson and Crick DNA model was proposed [35]. Based on the double-stranded DNA structure, Chargaff's first parity rule is easily understood. However, the second parity rule, based on similar nucleotide relationships in single strand DNA, has been puzzle in molecular biology, because it is impossible to imagine how pairs of G and C, and A and T formed in the single DNA strand. Using normalization of nucleotide contents, this historic puzzle was solved mathematically based on homogeneous genome structure [36]. Now, the results obtained from a huge genome dataset based on interspecies characteristics have been found to be consistent with Chargaff's rules [30, 37]. When the cellular nucleotide content of one nucleotide was fixed, the nucleotide contents of the other three nucleotides were naturally determined in chromosomes, chloroplasts, and plant mitochondria. Furthermore, not only the codon distributions but also the amino acid compositions were subsequently determined [30, 31]. In all organisms and cellular organelles except animal mitochondria, nucleotide contents can be expressed using dot plots with two duplicate points on the diagonal lines of a 0.5 square, which has been referred to as the diagonal genome universe [38].

Origin of Life

All phylogenetic trees start from a single origin. However, this feature is derived from the algorithms that are used to calculate similarities in the nucleotide or amino acid sequences of organisms. Charles Darwin described evolution and natural selection in "On the Origin of Species by Means of Natural Selection, or, the Preservation of Favoured Races in the Struggle for Life", published in 1859. Since then, the concepts of "a single origin" and "a common ancestor" of species has been generally accepted. Phylogenetic trees tend to represent the single origin of species.

We have shown that all organisms started from a single origin based on the result that nucleotide relationship lines, which represent not only the divergence of organisms but also cell organelles, closed at a single point [32]. In addition, it has been shown that vertebrates diverged from the low G/C invertebrate group [39]. These findings based on nucleotide content relationship lines have been confirmed by phylogenetic trees based on Ward's clustering analysis using 20 amino acid contents as traits [25, 40].

Phylogenetic Trees

Nucleotide and amino acid sequences reflect evolutionary divergence of organisms [1-5], and the changes in these sequences have been used to construct phylogenetic trees. Morphological changes in fossils have also been used to construct phylogenetic trees. Similarly, because normalized nucleotide and amino acid contents can be visualized in the form of radar charts, the resultant patterns can be compared computationally by multivariate analyses, using the four nucleotide contents or predicted 20 amino acid contents of chromosomal DNA as traits. The genome sequences of the organisms examined were classified into two groups, GC-rich and AT-rich [41]. It is possible to change the number of traits used; however, it should be noted that while increasing the number of traits yields better results, increasing the number of samples yields worse results because the probability of coincidence increases [25]. Using amino acid compositions predicted from complete mitochondrial genomes to draw the trees separated vertebrates into aquatic and terrestrial groups, whereas some exceptions were observed when nucleotide contents were used to draw the trees [24]. When 16S rRNA nucleotide sequences were used, the results were consistent with those obtained from mitochondrial amino acid compositions, with some minor differences [24, 25].

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 (Fig. **4**). 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, 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 [40]. Protists(*Monosiga brevicollis*) and cephalochordate (*Branchiostoma belcheri*) were classified into the low G/C and high G/C content invertebrate groups, respectively [32].

CONCLUSION

The normalized nucleotide and amino acid contents calculated from complete genomes can be accurately presented not only as numbers but also as shapes in radar charts. Normalization avoids the introduction of experimental errors and normalized nucleotide and amino acid contents have been used successfully to characterize genomes. For example,

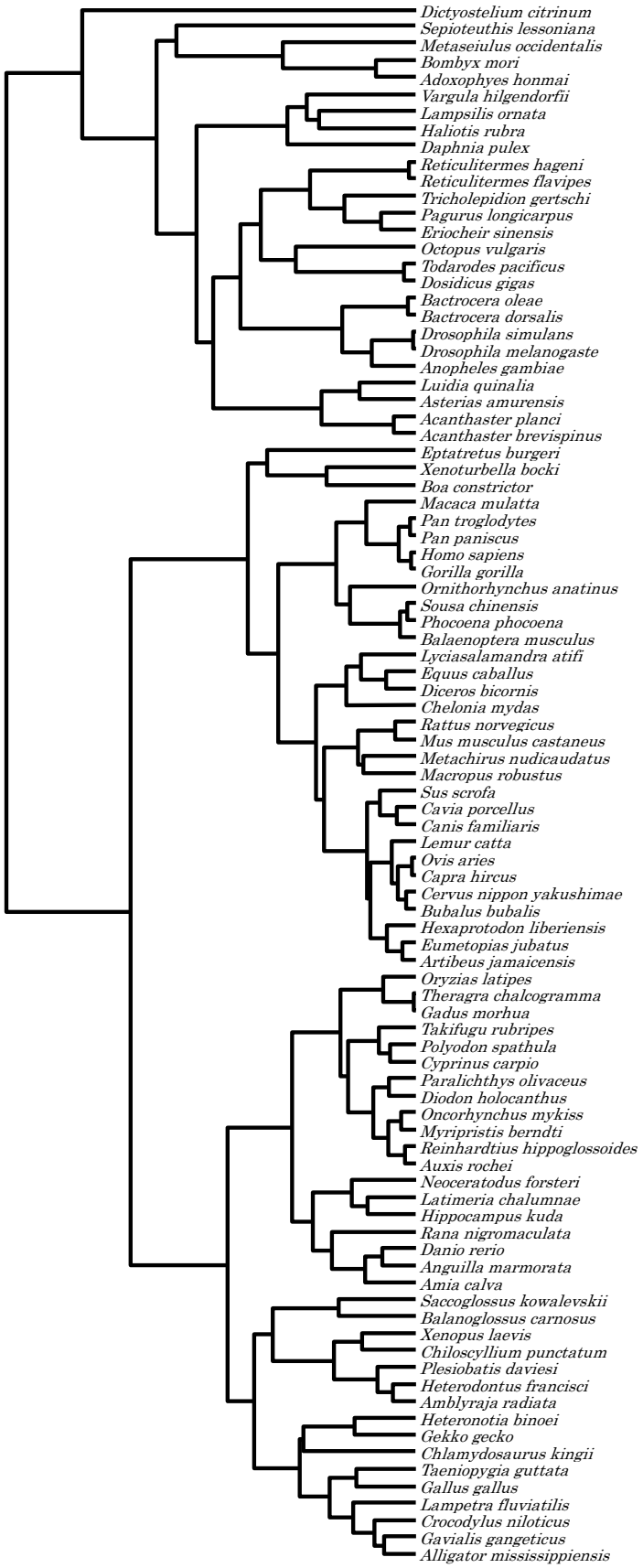

Fig. (4). Phylogenetic tree generated using Ward's cluster analysis method [22] from the predicted amino acid composition of the complete mitochondrial genomes of 58 invertebrates and 63 vertebrates. This figure first appeared in *Intl. J. Biol.*, **2014,** *6*, 82-94, and is reproduced with permission.

the concept of genome homogeneity based on putative small units was recognized based on the normalization of genome characteristics. In addition, phylogenetic trees that were constructed with the normalized nucleotide or amino acid contents predicted from complete genomes were found to give completely reasonable results. Thus, normalization of nucleotide or amino acid contents is a useful way of characterizing whole genomes consisting of numerous nucleotides and different genes.

CONFLICT OF INTEREST

The author(s) confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Dayhoff, M.O.; Park, C.M.; McLaughlin, P.J. Building a phylogenetic trees: cytochrome C. *In: Atlas of Protein Sequence and Structure*; Dayhoff, M.O.; Ed.; National Biomedical Foundation, Washington, D.C., **1977**; 5, pp.7-16.

[2] DePouplana, L.; Turner, R.J.; Steer, B.A.; Schimmel, P. Genetic code origins: tRNAs older than their synthetases? *Proc. Natl. Acad. Sci.USA*, **1998**, *95*, 11295-11300.

[3] Doolittle, W.F.; Brown, J.R. Tempo, mode, the progenote, and the universal root. *Proc. Natl. Acad. Sci. USA*, **1994**, *91*, 6721-6728.

[4] Maizels, N; Weiner, A.M. Phylogeny from function: evidence from the molecular fossil record that tRNA originated in replication, not translation. *Proc. Natl. Acad. Sci. USA*, **1994**, *91*, 6729-6734.

[5] Weisburg, W.G.; Barns, S.M.; Pelletier, D.A.; Lane, D.J. 16S ribosomal DNA amplification for phylogenetic study. *J. Bacteriol.*, **1991**, *173*, 697-703.

[6] Fleischmann, R.D.; Adams, M.D.; White, O.; Clayton, R.A.; Kirkness, E.F.; Kerlavage, A.R.; Bult, C.J.; Tomb, J.F.; Dougherty, B.A.; Merrick, J.M. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **1995**, *269*, 496-512.

[7] Lander, E.S.; Linton, L.M.; Birren, B.; Nusbaum, C.; Zody, M.C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; Funke, R.; Gage, D.; Harris, K.; Heaford, A.; Howland, J.; Kann, L.; Lehoczky, J.; LeVine, R.; McEwan, P,; McKernan, K.; Meldrim, J.; Mesirov, J.P.; Miranda, C.; Morris, W.; Naylor, J.; Raymond, C.; Rosetti, M.; Santos, R.; Sheridan, A.; Sougnez, C.; Stange-Thomann, N.; Stojanovic, N.; Subramanian, A.; Wyman, D.; Rogers, J.; Sulston, J.; Ainscough, R.; Beck, S.; Bentley, D.; Burton, J.; Clee, C.; Carter, N.; Coulson, A.; Deadman, R.; Deloukas, P.; Dunham, A;, Dunham, I.; Durbin, R.; French, L.; Grafham, D.; Gregory, S.; Hubbard, T.; Humphray, S.; Hunt, A.; Jones, M.; Lloyd, C.; McMurray, A.; Matthews, L.; Mercer, S.; Milne, S.; Mullikin, J.C.; Mungal, A.; Plumb, R.; Ross, M.; Shownkeen, R.; Sim, S.; Waterston, R.H.; Wilson, R.K.; Hillie, L.W.; McPherson J.D.; Marra, M.A.; Mardis, E.R.; Fulton, L.A.; Chinwalla, A.T.; Pepin, K.H.; Gish, W.R.; Chissoe, S.L.; Wendl, M.C.; Delehaunty, K.D.; Miner, T.L.; Delehaunty, A.; Kramer, J.B.; Cook, L.L.; Fulton, R.S.; Johnson, D.L.; Minx, P.J.; Clifton, S.W.; Hawkins, T.; Branscomb, E.; Predki, P.; Richardson, P.; Wenning, S.; Slezak, T.; Doggett, N.; Chen,g J.F., Olsen, A., Lucas, S., Elkin, C., Uberbache, E., Frazier, M. Gibbs, R.A.; Muzny, D.M.; Scherer, S.E.; Bouck, J.B.; Sodergren, E.J.; Worley, K.C.; Rives, C.M.; Gorrell, J.H.; Metzker, M.L.; Naylor, S.L.; Kucherlapati, R.S.; Nelson, D.L.; Weinstoc, G.M.; Sakaki, Y.; Fujiyama, A.; Hattori, M.; Yada, T.; Toyoda, A.; Itoh, T.; Kawagoe, C.; Watanabe, H.; Totoki, Y.; Taylor, T.; Weissenbach, J.; Heili, R.; Saurin, W.; Artiguenave, F., Brottier, P., Bruls, .T, Pelletier, E., Robert, C., Wincker, P., Smith, D.R.; Doucette-Stamm, L.; Rubenfield, M.; Weinstock, K.; Lee, H.M.; Dubois, J.; Rosenthal, A.; Platze, M.; Nyakatura, G.; Taudie, S.; Rump, A.; Yang, H.; Yu, J.; Wang, J.; Huang, G.; Gu, J.; Hood, .L,; Rowen, L.; Madan, A.; Qin, S.; Davis, R.W.; Federspiel, N..; Abola, A.P.; Proctor, M.J.; Myers, R.M.; Schmutz, J.; Dickson, M.; Grimwood, J.; Cox, D.R.; Olson, M.V.; Kaul, R.; Raymond, C.; Shimizu, N.; Kawasaki, K.; Minoshima, S.; Evans, G.A.; Athanasiou, M.; Schultz, R.; Roe, B.A.; Chen, F.; Pan, H.; Ramser ,J.; Lehrach, H.; Reinhardt, R.; McCombie, W.R.; de la Bastide, M.; Dedhia, N.; Blöcker, H.; Hornischer, K.; Nordsiek, G.; Agarwala, R;, Aravind, L.; Bailey, J.A.; Bateman, A.; Batzoglou, S.; Birney, E.; Bork, P.; Brown, D.G.; Burge, C.B.; Cerutti, L.; Chen, H.C.; Church, D.; Clamp, M.; Copley, R.R.; Doerks, T.; Eddy, S.R.; Eichler, E.E.; Furey, T.S.; Galagan, J.; Gilbert, J.G.; Harmon, C.; Hayashizi, Y.; Haussler, D;, Hermjakob, H.; Hokamp, K.; Jang, W.; Johnson, L.S.; Jones, T.A.; Kasif, S.; Kaspryzk, A.; Kennedy, S.; Kent, W.J.; Kitts, P.; Koonin, E.V.; Korf, I.; Kulp, D.; Lancet, D.; Lowe, T.M.; McLysaght, A.; Mikkelsen, T.; Moran, J.V.; Mulder, N.; Pollara, V.J.; Pontin, C.P.; Schuler, G.; Schultz, J.; Slater, G.; Smit, A.F.; Stupka, E.; Szustakowski, J.; Thierry-Mieg, D.; Thierry-Mieg, J.; Wagner, L.; Wallis J.; Wheeler, R.; Williams, A.; Wolf, Y.I.; Wolfe, K.H.; Yang, S.P.; Yeh, R.F.; Collins, F.; Guyer, M.S.; Peterson, J.; Felsenfeld, A.; Wetterstrand, K.A.; Patrinos, A.; Morgan, M.J.; de Jong, P.; Catanese, J.J.; Osoegawa, K.; Shizuya, H.; Choi, S.; Chen, Y.J.; International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, **2001**, *409*, 860-921.

[8] Venter, J.C.; Adams, M.D.; Myers, E.W.; Li, .W.; Mural, R.J.; Sutton, G.G.; Smith, H.O.; Yandell, M.; Evans, C.A.; Holt, R.A.; Gocayne, J.D.; Amanatides, P.; Ballew, R.M.; Huson, D.H.; Wortman, J.R.; Zhang, Q.; Kodira, C.D.; Zheng, X.H.; Chen, L.; Skupski, M.; Subramanian, G.; Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G.; Nadeau, J.; McKusick, V.A.; Zinder, N.; Levine, A.J.; Roberts, R.J.; Simon, M.; Slayman, C.; Hunkapiller, M.; Bolanos, R.; Delcher, A.; Dew, I.; Fasulo, D.; Flanigan, M.; Florea, L.; Halpern, A.; Hannenhall, S.; Kravitz, S.; Levy, S.; Mobarry, C.; Reinert, K.; Remington, K.; Abu-Threideh, J.; Beasley, E.; Biddick, K.; Bonazzi, V.; Brandon, R.; Cargill, M.; Chandramouliswaran, I.; Charlab, R.; Chaturvedi, K.; Deng, Z.; Di Francesco, V.; Dunn, P.; Eilbeck, ., Evangelista, C.; Gabrielian, A.E.; Gan, W.; Ge, W.; Gong, F.; Gu, Z.; Guan, P.; Heiman, .TJ.; Higgins, M.E.; Ji, R.R.; Ke, Z.; Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, .J, Liang, Y., Lin, X., Lum F.,Merkulov, G.V.; Milshina, N.; Moore, H.M.; Naik, A.K.; Narayan, V.A.; Neelam, B.; Nusskern, D.; Rusch, D.B.; Salzberg, S.; Shao, W.; Shue, B.; Sun, J.; Wang, Z.; Wang, A.; Wang, X.; Wang, J.; Wei, M.; Wides, R.; Xiao, C.; Yan, C.; Yao, A.; Ye, J.; Zhan, M.; Zhang, W.; Zhang, H. Zhao, Q.; Zheng, L.; Zhong, F.; Zhong, W.; Zhu, S.; Zhao, S.; Gilbert, D.; Baumhueter, S.; Spier, G.; Carter, C.; Cravchik, A.; Woodage, T.; Ali, F.; An, H; Awe, A.; Baldwin, D.; Baden, H.; Barnstead, M.; Barrow, .I; Beeson, K.; Busam, D.; Carver, A., Center, A.; Cheng, M.L.; Curry, L.; Danaher, S.; Davenport, L.; Desilets, R.; Dietz, S.; Dodson, K.; Dou, L.; Ferriera, S.; Garg, N.; Gluecksmann, A.; Hart, B.; Haynes, J.; Haynes, C.; Heiner, C.; Hladun, S.; Hostin, D.; Houc, J.; Howland, T.; Ibegwam, C.; Johnson, J;, Kalush, F.; Kline, L;, Koduru, S,; Love, A.; Mann, F.; May, D.; McCawley, S.; McIntosh, T.; McMullen, I.; Moy, M.; Moy, L.; Murphy, B.; Nelson, K.; Pfannkoch, C.; Pratts, E.; Puri, V.; Qureshi, H.; Reardon, M.; Rodriguez, R.; Rogers, Y.H.; Romblad, D.; Ruhfe,l B.; Scott, R.; Sitter, C.; Smallwood, M.; Stewart, E.; Strong, R.; Suh, E.; Thomas, R.; Tint, N.N.; Tse, S.; Vech, C.;Wang, G.; Wetter, J.; Williams, S.; Williams, M.; Windsor, S.; Winn-Deen, E.; Wolfe, K.; Zaveri, J.; Zaveri, K.; Abril, J.F.; Guigó, R.; Campbell, M.J.; Sjolander, K.V.; Karlak, B.; Kejariwal, A.; Mi, H.; Lazareva, B.; Hatton, T.; Narechania, A.; Diemer, K.; Muruganujan, A.; Guo, N.; Sato, S.; Bafna, V.; Istrail, S.; Lippert, R.; Schwartz, R.; Walenz, B.; Yooseph, S.; Allen, D.; Basu, A.; Baxendale, J.; Blick, L.; Caminha, M.; Carnes-Stine, J.; Caulk, P.; Chiang, Y.H.; Coyne, M.; Dahlke, C.; Mays, A.; Dombroski, M.; Donnelly, M.; Ely, D.; Esparham, S.; Fosler, C.; Gire, H.; Glanowski, S,; Glasser, K.; Glodek, A.; Gorokhov, M.; Graham, K.; Gropman, B.; Harris, M.; Heil, .J; Henderson, S.; Hoover, J.; Jennings, D.; Jordan, C.; Jordan, J.; Kasha, J.; Kagan, L.; Kraft, C.; Levitsky, A.; Lewis, M.; Liu, X.; Lopez, J.; Ma, D.; Majoros, W.; McDaniel, J.; Murphy, S.; Newman, M.; Nguyen, T.; Nguyen, N.; Nodell, M.; Pan, S.; Peck, J.; Peterson, M.; Rowe, W.; Sanders, R.; Scott, J.; Simpson, M.; Smith, T.; Sprague, A.; Stockwell, T.; Turner, R.; Venter, E.; Wang, M.; Wen, M.; Wu, D.; Wu, M.; Xia, A.; Zandieh, A.; Zhu, X.The sequence of the human genome. *Science*, **2001**, *291*, 1304-

1351.

[9] Paixao-Cortes, V.R.; Viscardi, L.H.; Salzno, F.M.; Hunemeier, T.; Bortolini, M.C. *Homo sapiens*, *Homo neanderthalensis* and the Denisova specimen: New insights on their evolutionary histories using whole-genome comparisons. *Genet. Mol. Biol.*, **2012**, *35*, 4(suppl), 904-911.

[10] Dillies, M-A.; Rau, A.; Aubert, J.; Hennequet-Antier, C.; Jeanmougin, M.; Servant, N.; Keime, G.; Marot, G.; Castel, D.; Estelle, J.; Guernec, G.; Jagla, B.; Jouneau, L.; Laloe, D.; Le Gall, C.; Schaeffer, B.; Le Crom, S.; Guedj, M.; Jaffrezic, F. and on behalf of The French StatOmique Consortium. A comprehensive evaluation of normalization methods for illumine high-throughput RNA sequencing data analysis. *Brief. Bioinform.*, **2012**, *14*, 671-683.

[11] Sorimachi, K.; Itoh, T,; Kawarabayasi, Y.; Okayasu, T;, Akimoto, K.; Niwa, A. (2001) 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.

[12] Chou, K-C. Review: Applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady state systems. *Biophys. Chem.*, **1990**, *35*, 1-24.

[13] Qi, X.Q.; Wen, J.; Qi, Z.H. New 3D graphical representation of DNA sequence based on dual nucleotides. *J. Theoret. Biol.*, **2007**, *249*, 681-690.

[14] Kumar, A.; Kamaraj, B.; Sethumadhavan, R.; Purohit, R. Evolution drived structural changes in CENP-E motor domain. *Interdiscpl. Sci.*, **2013**, *5*, 102-111.

[15] Purohit, R.; Rajendran, V.; Sethumadhavan, R. Studies on adaptability of binding residues flap region of TMC-114 resistance HIV-1 protease mutants. *J. Biomol. Struct. Dynam.*, **2011**, *29*, 137-152.

[16] Kumar, A.; Purohit, R. Use of long term molecular dynamics simulation in predicting cancer associated SNPs. *PLOS Comput. Biol.*, **2014**, *10*, e1003318.

[17] Rajendran, V.; Purohit, R.; Sethumadhavan, R. In silico investigation of molecular mechanism of laminopathy caused by appoint mutation (R482W) in lamin A/C protein. *Amino Acids*, **2012**, *43*, 603-615.

[18] The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM consortium. *Nature*, **2001**, *409*, 685-689.

[19] Fraser, C.M.; Norris, S.J.; Weinstock, G.M.; White, O.; Sutton, G.G.; Dodson, R.; Gwinn, M.; Hickey, E.K.; Clayton, R.; Ketchum, K.A.; Sodergren, E.; Hardham, J.M.; McLeod, M.P.; Salzberg, S.; Peterson, J.; Khalak, H.; Richardson, D.; Howell, J.K.; Chidambaram, M.; Utterback, T.; McDonald, L.; Artiach, P.; Bowman, C.; Cotton, M.D.; Fujii, C.; Garland, S.; Hatch, B.; Horst, K.; Roberts, K.; Sandusky, M.; Weidman, J.; Smith, H.O.; Venter, J.C. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science*, **1998**, *281*, 375-388.

[20] 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.

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

[22] Ward, J.H. Hierarchic grouping to optimize an objective function. *J. Amer. Statistic. Assoc.*, **1963**, *58*, 236-244.

[23] Saitou, N.; Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **1987**, *4*, 406-425.

[24] 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 (*Eptatretus burgeri*). *Natural Science*, **2013**, *5*, 221-227.

[25] Sorimachi, K.; Okayasu, T. Phylogenetic tree construction based on amino acid composition and nucleotide content of complete mitochondrial genomes. *IOSR J. Pharm.*, **2013**, *3*, 51-60.

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

[27] Janvier, P. Micro RNAs revive old views about jawless vertebrate divergence and evolution. *Proc. Natl. Acad. Sci. USA*, **2010**, *107*, 19137-19138.

[28] Sorimach, K.; Okayasu, T. Classification of eubacteria based on their complete genome: where does Mycoplasmataceae belong? *Proc. R. Soc. Lond. B* (Suppl.), **2004**, *271*, S127-S130.

[29] Sueoka, N. Correlation between base composition of deoxyribonucleic acid and amino acid composition in proteins. *Proc. Natl. Acad. Sci. USA*, **1961**, *47*, 1141-1149.

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

[31] Sorimachi, K.; Okayasu, T. Universal rules governing genome evolution expressed by linear formulas. *Open Genom. J.*, **2008**, *1*, 33-43.

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

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

[34] Rundner, R.; Karkas, J.D.; Chargaff, E. Separation of *B. subtilis* DNA into complementary strands. 3. Direct analysis. *Proc. Natl. Acad. Sci. USA*, **1968**, *60*, 921-922.

[35] Watson, J.D.; Crick, F.H.C. Genetical implications of the structure of deoxyribonucleic acid. *Nature*, **1953**, *171*, 964-967.

[36] Sorimachi, K. A proposed solution to the historic puzzle of Chargaff's second parity rule. *Open Genom. J.*, **2009**, *2*, 12-14.

[37] Mitchell, D.; Bridge, R. A test of Chargaff's second rule. *Biochem. Biophys. Res. Commun.*, **2006**, *340*, 90-94.

[38] Sorimachi, K. Evolution based on genome structure: the "diagonal genome universe". *Natural Science*, **2010**, *2*, 1104-1112.

[39] 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.

[40] 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. *Intl. J. Biol.*, **2014**, *6*, 82-94.

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

[42] Alff-Steinberger, C. (1987) "Codon usage in Homo sapiens: evidence for a coding pattern on the non-coding strand and evolutionary implications of dinucleotide discrimination", *J. Theor. Biol.*, **1987**, *124*, 89-95.

Received on: November 19, 2014 **Revised on: January 09, 2015** **Accepted on: January 19, 2015**