DNA RESEARCH 5, 251–259 (1998)

Genes from Nine Genomes Are Separated into Their Organisms in the Dinucleotide Composition Space

Hiroshi NAKASHIMA,1* Motonori OTA,2 Ken NISHIKAWA,2 and Tatsuo OOI3

School of Health Sciences, Faculty of Medicine, Kanazawa University, 5-11-80 Kodatsuno, Kanazawa 920-0942, Japan, 1 Center for Information Biology, National Institute of Genetics, Yata 1111, Mishima, Shizuoka 411-8540, Japan, 2 and Kyoto Women’s University, Kitahiyoshi-cho 35, Higashiyama-ku, Kyoto 605, Japan3

(Received 2 September 1998)

Abstract

A set of 16 kinds of dinucleotide compositions was used to analyze the protein-encoding nucleotide sequences in nine complete genomes: Escherichia coli, Haemophilus influenzae, Helicobacter pylori, Mycoplasma genitalium, Mycoplasma pneumoniae, Synechocystis sp., Methanococcus jannaschii, Archaeoglobus fulgidus, and Saccharomyces cerevisiae. The dinucleotide composition was significantly different between the organisms. The distribution of genes from an organism was clustered around its center in the dinucleotide composition space. The genes from closely related organisms such as Gram-negative bacteria, mycoplasma species and eukaryotes showed some overlap in the space. The genes from nine complete genomes together with those from human were discriminated into respective clusters with 80% accuracy using the dinucleotide composition alone. The composition data estimated from a whole genome was close to that obtained from genes, indicating that the characteristic feature of dinucleotides holds not only for protein coding regions but also noncoding regions. When a dendrogram was constructed from the disposition of the clusters in the dinucleotide space, it resembled the real phylogenetic tree. Thus, the distinct feature observed in the dinucleotide composition may reflect the phylogenetic relationship of organisms.

Key words: separation of genes; dinucleotide frequency; phylogenetic tree

1. Introduction

The complete genome sequences of several free-living organisms are now available, and this provides an opportunity to perform comprehensive comparison of genes between different organisms. The nucleotide sequences of various sources have been analyzed by many investigators1–10 in terms of the frequency of occurrence of oligomers such as dinucleotides, trinucleotides, and tetranucleotides. The general analytical method is based around the odds ratio between observed and expected values calculated from base composition. In particular, the importance of dinucleotide patterns has been pointed out by Nussinov.1 Each organism has a characteristic odds ratio between the observed and expected frequencies of dinucleotides, and clear differences between prokaryotes and eukaryotes have been reported.1,2,5 The statistical difference in the dinucleotide frequency prevails throughout a long DNA sequence (more than 1 mbp), containing both protein-coding and noncoding regions. The nucleotide sequences have been compared especially on the relative abundance of dinucleotides as a unique signature by Karlin’s group.4–10 Comparative analysis of the complete genome sequences has led to a new genome-based biology.11,12

Recently, we have reported that protein coding nucleotide sequences of human, yeast (Saccharomyces cerevisiae) and Escherichia coli have different frequencies of dinucleotides.13 The genes are distributed in a cluster around its average when each gene is expressed as a vector of the log-odds ratio of 16 components of dinucleotides. This result holds for the genes which encodes homologous proteins. For example, the human gene which encodes H+-transporting ATP synthase α-chain, which has 72% amino acid identity with that of E. coli, is separated from the corresponding E. coli gene in the dinucleotide space. This result indicated that a nucleotide sequence of a gene encodes information not only for an amino acid sequence but expresses characteristic features of an organism. As an extension of our previous study, the genes from nine complete genomes have been investigated in this work.
In our previous studies, it was shown that proteins having a folding type such as $\alpha$, $\beta$, $\alpha/\beta$ or $\alpha+\beta$ types are located in a cluster in the amino acid composition space and that they are distributed roughly in a separate group, when each protein is expressed as a vector of 20 components of amino acids. Furthermore, the distribution of enzyme versus non-enzyme proteins, and that of intracellular versus extracellular proteins are roughly separated in the amino acid composition space. Therefore, proteins which share the same structures or functions are assumed to cluster in the amino acid composition space. In the present study, we examined the degree of separation of proteins according to the sources in the amino acid composition space, and compare the results with the separation of corresponding genes in the dinucleotide space.

2. Data and Methods

2.1. Sequence data

Nucleotide sequences of nine complete genomes and those of human were obtained from the DNA Data Bank of Japan. The nine complete genomes employed are three Gram-negative bacteria ($E. coli$, $Hae. influenzae$, and $Helicobacter pylori$), two Gram-positive bacteria ($Mycoplasma genitalium$ and $M. pneumoniae$), one cyanobacterium ($Synechocystis$), two archaea ($Methanococcus jannaschii$ and $Archaeoglobus fulgidus$) and one eukaryote ($S. cerevisiae$). Nucleotide sequences encoding proteins were extracted according to the feature table of the data. Protein-coding sequences longer than 50 residues were used. Complete protein-coding sequences longer than 150 bases were employed, and the start and stop codons were excluded from the analysis. The 489 genes and 108 introns from humans used here were the same as used in the previous work. Amino acid sequence data for proteins longer than 50 residues were used.

2.2. Analytical method

The same method to analyze proteins in terms of amino acid composition was used to investigate genes in terms of dinucleotide composition. The number of 16 kinds of dinucleotides in a nucleotide sequence was counted and the dinucleotide composition is expressed in a normalized scale by the following equation:

$$V_{i,k} = (CDN_{i,k} - AV_k)/SD_k$$ (1)

where $V_{i,k}$ and $CDN_{i,k}$ are the normalized and real compositions of dinucleotides for the $k$-th component in a gene of nucleotide sequence $i$, respectively. $AV_k$ is the average composition and $SD_k$ is the standard deviation for the data set. The distance between two nucleotide sequences, $i$ and $j$, was calculated as:

$$D_{i,j} = [\Sigma(V_{i,k} - V_{j,k})^2]^{1/2}$$ (2)

The distance in the space was used to classify the genes. The assignment of genes into an organism was done as follows: the distance from a gene to each one of the ten mean points was calculated in the normalized dinucleotide composition space, and then the gene was assigned to the organism with the shortest distance.

A protein is expressed as a vector of 20 components of normalized amino acid composition. The normalization was done by:

$$C_{i,k} = (CAA_{i,k} - AV_k)/SD_k$$ (3)

where $C_{i,k}$ and $CAA_{i,k}$ are the normalized and real composition of amino acid residues for the $k$-th component in a protein of amino acid sequence $i$, respectively. $AV_k$ is the average composition and $SD_k$ is the standard deviation. The assignment of proteins into an organism was done in the same way as done for gene classification by the shortest distance in the normalized amino acid composition space.

In order to illustrate the distribution of points in the multidimensional space, points were projected on a two-dimensional (2D) plane. A 2D plane can be defined by three points on the plane, and arbitrary $x$ and $y$ axes can be drawn on the plane. The $x$ coordinate of a point projected on the 2D plane is calculated as a scalar product between the unit vector of the $x$ axis and the vector $p$. The $y$ coordinate is the scalar product of the unit vector of the $y$ axis and the vector $p$.

3. Results

The mean compositions of dinucleotides for the nine organisms are shown in Table 1. The average composition ($AV$) and standard deviation ($SD$) of the nine organisms in Table 1 were used for the normalization of dinucleotide compositions using Eq. (1). The mean composition of dinucleotides is considerably different among organisms; this is evident in the star diagrams in Fig. 1. The dinucleotides of $E. coli$ show a more or less even composition, while other species indicate more deformed shapes. A prolonged shape in the up-and-down direction, seen for $Hae. influenzae$, $Helicobacter pylori$, $Mycoplasma genitalium$, and so on, reflects $A+T$ richness, while a flat shape as seen most in humans implies G + C richness. The percentage of A + T in the whole genome is 68.5% in $Myc. jannaschii$, 68.3% in $Myc. genitalium$, 61.8% in $Hae. influenzae$, 61.7% in $S. cerevisiae$, 61.1% in $Hel. pylori$, 60.0% in $Myc. pneumoniae$, 52.3% in $Synechocystis$, 51.4% in $A. fulgidus$ and 49.2% in $E. coli$, respectively. Extreme examples are rare compositions of CpG for $Met. jannaschii$ (0.7%) and $Myc. genitalium$ (1.0%).

Frequent dinucleotides are $ApA > TpT > ApT$, and less frequent ones are $CpG < CpC < TpC$. The SDS of both $ApA$ and $TpT$ are larger than the SDS of the other dinucleotides. There are six pairs of dinucleotides
Figure 1. Star-diagrams presenting the dinucleotide composition. The mean compositions (%) over all genes (shaded) and the entire genome (non-shaded) are plotted. The data for human are exceptions (see Sequence data section), depicted here as references. The radial axes of 16 dinucleotides are allotted so that the complementary dinucleotides, AA/TT, AC/GT, etc., occupy counter positions along the circle. Complementary pairs should have equivalent amounts in the total composition over a genome. Note that the scale is different for each diagram. The innermost broken circle indicates the 5% level.
such as ApA-TpT, ApC-GpT and so on, which are complementary counterparts to each other. For the pairs of complementary dinucleotides, the frequency of occurrence is generally in the order ApA > TpT, GpG > CpC, GpA > TpC, ApG > CpT and TpG > CpA (Table 1). The frequency of ApC is equal to that of GpT. This result indicates that the purine-purine dinucleotides are more favorable than the pyrimidine-pyrimidine dinucleotides in the coding sequences. The dinucleotides ApT, TpA, CpG and GpC have identical complementary counterparts, and the frequencies are ApT > TpA and GpC > CpG. That is, the dinucleotides of purine-pyrimidine are more favorable than pyrimidine-purine.

Genes can be classified into groups by comparing their dinucleotide composition with the mean compositions in Table 1: a gene is assigned to an organism closest in distance as described above. All genes from the nine complete genomes and those from humans were mixed together and classified into organisms. Table 2 shows the comparison between the real and assigned classification of genes for the ten organisms. For instance, 86% of 4248 E. coli genes were correctly assigned as E. coli, and 5% of them were incorrectly assigned as Hae. influenzae. The genes from the ten organisms were correctly assigned with accuracy as high as 79.7%. The broken line in Table 2 divides the organisms into the following five categories: Gram-negative species, Gram-positive species, cyanobacterium, archaea and eukaryotes. The discrimination went up to 85.5% in the five-group classification.

The genes of two mycoplasma species overlap to some extent in the distribution, while the genes of two archaeal genomes have almost no overlap.

The distribution of genes can be visualized by projecting points (genes) on a two-dimensional plane. Figure 2 shows all genes of the three species, E. coli, Met. jan-

Table 1. Mean composition (%) of dinucleotides of genes in E. coli (Ec), Hae. influenzae (Hi), Hel. pylori (Hp), Myc. genitalium (Mg), Myc. pneumoniae (Mp), Synechocystis sp. (Sc), Met. jannaschii (Mj), A. fulgidus (Af), S. cerevisiae (Sc) and H. sapiens (Hs). AV and SD stand for the average and standard deviation of dinucleotide compositions for the data set, respectively.

<table>
<thead>
<tr>
<th>Dinucleotide</th>
<th>Ec</th>
<th>Hi</th>
<th>Hp</th>
<th>Mg</th>
<th>Mp</th>
<th>Sc</th>
<th>Hs</th>
<th>AV</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>7.72</td>
<td>12.57</td>
<td>13.49</td>
<td>15.95</td>
<td>12.80</td>
<td>8.73</td>
<td>15.62</td>
<td>8.49</td>
<td>11.42</td>
</tr>
<tr>
<td>TT</td>
<td>6.66</td>
<td>11.18</td>
<td>11.83</td>
<td>12.99</td>
<td>10.44</td>
<td>8.97</td>
<td>10.82</td>
<td>7.23</td>
<td>8.91</td>
</tr>
<tr>
<td>AC</td>
<td>5.40</td>
<td>4.78</td>
<td>4.03</td>
<td>5.28</td>
<td>6.56</td>
<td>4.90</td>
<td>4.30</td>
<td>4.71</td>
<td>5.80</td>
</tr>
<tr>
<td>GT</td>
<td>5.67</td>
<td>5.33</td>
<td>3.92</td>
<td>5.10</td>
<td>5.76</td>
<td>4.99</td>
<td>4.33</td>
<td>4.78</td>
<td>4.97</td>
</tr>
<tr>
<td>AT</td>
<td>6.52</td>
<td>8.95</td>
<td>7.88</td>
<td>8.83</td>
<td>6.23</td>
<td>6.71</td>
<td>10.81</td>
<td>5.50</td>
<td>8.25</td>
</tr>
<tr>
<td>TA</td>
<td>4.30</td>
<td>6.90</td>
<td>6.56</td>
<td>8.64</td>
<td>6.74</td>
<td>4.87</td>
<td>9.33</td>
<td>3.91</td>
<td>6.43</td>
</tr>
<tr>
<td>CA</td>
<td>6.11</td>
<td>6.26</td>
<td>5.57</td>
<td>5.98</td>
<td>6.54</td>
<td>6.22</td>
<td>4.68</td>
<td>5.56</td>
<td>6.93</td>
</tr>
<tr>
<td>TG</td>
<td>8.01</td>
<td>7.27</td>
<td>6.12</td>
<td>6.64</td>
<td>6.53</td>
<td>7.02</td>
<td>6.70</td>
<td>6.99</td>
<td>6.58</td>
</tr>
<tr>
<td>CC</td>
<td>5.16</td>
<td>2.99</td>
<td>3.70</td>
<td>2.58</td>
<td>4.52</td>
<td>7.82</td>
<td>2.01</td>
<td>4.18</td>
<td>4.10</td>
</tr>
<tr>
<td>GG</td>
<td>6.58</td>
<td>4.29</td>
<td>5.14</td>
<td>2.99</td>
<td>4.52</td>
<td>7.88</td>
<td>4.73</td>
<td>8.02</td>
<td>4.48</td>
</tr>
<tr>
<td>TC</td>
<td>5.26</td>
<td>4.53</td>
<td>3.98</td>
<td>3.81</td>
<td>4.08</td>
<td>4.95</td>
<td>2.84</td>
<td>5.33</td>
<td>5.67</td>
</tr>
<tr>
<td>GA</td>
<td>6.32</td>
<td>5.83</td>
<td>6.49</td>
<td>5.71</td>
<td>5.67</td>
<td>5.92</td>
<td>8.80</td>
<td>9.77</td>
<td>7.27</td>
</tr>
<tr>
<td>CG</td>
<td>7.58</td>
<td>4.03</td>
<td>3.61</td>
<td>0.97</td>
<td>3.38</td>
<td>4.43</td>
<td>0.69</td>
<td>4.61</td>
<td>3.11</td>
</tr>
<tr>
<td>GC</td>
<td>8.45</td>
<td>5.40</td>
<td>6.06</td>
<td>3.08</td>
<td>4.67</td>
<td>6.00</td>
<td>2.89</td>
<td>6.04</td>
<td>4.04</td>
</tr>
<tr>
<td>CT</td>
<td>5.41</td>
<td>4.43</td>
<td>4.90</td>
<td>5.20</td>
<td>5.39</td>
<td>5.17</td>
<td>3.76</td>
<td>5.94</td>
<td>5.47</td>
</tr>
<tr>
<td>AG</td>
<td>4.85</td>
<td>5.26</td>
<td>6.72</td>
<td>6.25</td>
<td>6.17</td>
<td>5.42</td>
<td>8.59</td>
<td>8.94</td>
<td>6.57</td>
</tr>
</tbody>
</table>

Composition of dinucleotides was calculated for individual genes and the mean was computed over all the genes considered for an organism. AV and SD were calculated as the average of mean compositions and that of standard deviations of the nine organisms.

Table 2. Comparison between the real and assigned classification (in %) of genes from E. coli (Ec), Hae. influenzae (Hi), Hel. pylori (Hp), Myc. genitalium (Mg), Myc. pneumoniae (Mp), Synechocystis sp. (Sc), Met. jannaschii (Mj), A. fulgidus (Af), S. cerevisiae (Sc) and H. sapiens (Hs).

<table>
<thead>
<tr>
<th>organism (genes)</th>
<th>assigned as</th>
<th>Ec</th>
<th>Hi</th>
<th>Hp</th>
<th>Mg</th>
<th>Mp</th>
<th>Sc</th>
<th>Hs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ec (4248)</td>
<td></td>
<td>86.0</td>
<td>5.0</td>
<td>0.0</td>
<td>1.0</td>
<td>1.0</td>
<td>2.0</td>
<td>3.0</td>
</tr>
<tr>
<td>Hi (1677)</td>
<td></td>
<td>75.0</td>
<td>2.1</td>
<td>7.1</td>
<td>3.4</td>
<td>5.0</td>
<td>2.0</td>
<td>3.0</td>
</tr>
<tr>
<td>Hp (1512)</td>
<td></td>
<td>75.0</td>
<td>1.0</td>
<td>7.1</td>
<td>5.0</td>
<td>2.0</td>
<td>3.0</td>
<td>3.0</td>
</tr>
<tr>
<td>Mg (465)</td>
<td></td>
<td>58.0</td>
<td>0.0</td>
<td>0.0</td>
<td>89.0</td>
<td>7.0</td>
<td>0.0</td>
<td>3.0</td>
</tr>
<tr>
<td>My (573)</td>
<td></td>
<td>1.0</td>
<td>3.0</td>
<td>1.0</td>
<td>3.0</td>
<td>1.0</td>
<td>4.0</td>
<td></td>
</tr>
<tr>
<td>Sc (3143)</td>
<td></td>
<td>1.0</td>
<td>2.0</td>
<td>3.0</td>
<td>2.0</td>
<td>2.0</td>
<td>5.0</td>
<td></td>
</tr>
<tr>
<td>Mj (1673)</td>
<td></td>
<td>1.0</td>
<td>3.0</td>
<td>0.0</td>
<td>1.0</td>
<td>3.0</td>
<td>1.0</td>
<td>79.0</td>
</tr>
<tr>
<td>Af (2348)</td>
<td></td>
<td>1.0</td>
<td>3.0</td>
<td>0.0</td>
<td>1.0</td>
<td>3.0</td>
<td>1.0</td>
<td>6.0</td>
</tr>
<tr>
<td>Sc (5840)</td>
<td></td>
<td>1.0</td>
<td>3.0</td>
<td>0.0</td>
<td>4.0</td>
<td>1.0</td>
<td>3.0</td>
<td>77.0</td>
</tr>
<tr>
<td>Hs (489)</td>
<td></td>
<td>1.0</td>
<td>3.0</td>
<td>0.0</td>
<td>4.0</td>
<td>1.0</td>
<td>3.0</td>
<td>74.0</td>
</tr>
</tbody>
</table>

The sum along the row gives a total of 100%. The average accuracy of classification of genes into ten organisms was 79.7%.
Figure 2. Gene distributions in the dinucleotide space. Plots of 4248 genes encoded on the E. coli genome are shown in red, 5840 genes of S. cerevisiae in blue, and 1673 genes of Met. jannaschii in yellow. Each point representing a gene is expressed in the 16-dimensional dinucleotide space. All points are then projected on a plane: the origin of the coordinate system is the point of the total average composition (i.e., given as AV in Table 1). The direction of the x-axis was defined as pointing from the mean point of E. coli to that of S. cerevisiae, and the y-axis was chosen perpendicular to the x-axis in the plane involving the mean point of Met. jannaschii.

The genes of each organism are densely distributed around the respective center, or the mean point, so that genes of different colors are almost completely separated from each other. Note that the actual points are distributed in the 16-dimensional space and therefore the separation by species is much better than in Fig. 2. This is also why genes of only three species, instead of ten, are shown here: genes of the ten organisms would appear heavily overlapped on a plane, while they are actually well separated, as shown in Table 2. Some genes in Fig. 2 seem to deviate far from the mean point of the organism to which they belong. Consulting the annotation in the database, they are genes for ribosomal proteins, ferredoxins, histones, ATP synthase F0 subunit c, etc., although most of the genes remain unidentified. The genes which deviated largely from their means are relatively short in length. A data set of genes, excluding those shorter than 300 bases, was prepared and used to classify the genes into ten organisms. It gave an average accuracy of 81.3%, slightly better than that in Table 2.

The degree of separation between clusters may be given by center-to-center distances. The distances among the mean points of genes of the ten organisms in the normalized dinucleotide composition space are shown in Table 3. Similarly calculated distances from the origin and the radius of the distribution are also listed in the last two columns of Table 3. By definition, the origin of the space is the average dinucleotide composition. Hel. pylori showed the shortest and human showed the largest distance from the origin.

In our previous analysis, E. coli noncoding sequences are similar to E. coli genes in terms of dinucleotide frequency. In order to examine the relation between genes and noncoding sequences, whole genomes were analyzed in a similar way. In this case, the dinucleotide composition becomes symmetrical for complementary dinucleotides (ApA/TpT, etc.) due to the double-stranded nature of DNA. The composition obtained from the complete genome is plotted in Fig. 1, in comparison with that of genes. For human, intron sequences were used in place of the genomic DNA. Figure 1 shows that the dinucleotide composition is almost the same between genes and the entire genome for a given organism, although remarkably different among organisms. Since coding regions (genes) occupy a large portion of a bacterial genome, it would be natural that the difference between genes and a whole genome is small. However, the same trend of common composition is seen even in yeast and human. In the latter the comparison is made between

naschii and S. cerevisiae, plotted on one plane. All the genes of each organism are densely distributed around the respective center, or the mean point, so that genes of different colors are almost completely separated from each other. Note that the actual points are distributed in the 16-dimensional space and therefore the separation by species is much better than in Fig. 2. This is also why genes of only three species, instead of ten, are shown here: genes of the ten organisms would appear heavily overlapped on a plane, while they are actually well separated, as shown in Table 2. Some genes in Fig. 2 seem to deviate far from the mean point of the organism to which they belong. Consulting the annotation in the database, they are genes for ribosomal proteins, ferredoxins, histones, ATP synthase F0 subunit c, etc., although most of the genes remain unidentified. The genes which deviated largely from their means are relatively short in length. A data set of genes, excluding those shorter than 300 bases, was prepared and used to classify the genes into ten organisms. It gave an average accuracy of 81.3%, slightly better than that in Table 2.

The degree of separation between clusters may be given by center-to-center distances. The distances among the mean points of genes of the ten organisms in the normalized dinucleotide composition space are shown in Table 3. Similarly calculated distances from the origin and the radius of the distribution are also listed in the last two columns of Table 3. By definition, the origin of the space is the average dinucleotide composition. Hel. pylori showed the shortest and human showed the largest distance from the origin.

In our previous analysis, E. coli noncoding sequences are similar to E. coli genes in terms of dinucleotide frequency. In order to examine the relation between genes and noncoding sequences, whole genomes were analyzed in a similar way. In this case, the dinucleotide composition becomes symmetrical for complementary dinucleotides (ApA/TpT, etc.) due to the double-stranded nature of DNA. The composition obtained from the complete genome is plotted in Fig. 1, in comparison with that of genes. For human, intron sequences were used in place of the genomic DNA. Figure 1 shows that the dinucleotide composition is almost the same between genes and the entire genome for a given organism, although remarkably different among organisms. Since coding regions (genes) occupy a large portion of a bacterial genome, it would be natural that the difference between genes and a whole genome is small. However, the same trend of common composition is seen even in yeast and human. In the latter the comparison is made between
the coding and noncoding (intron) regions. These results imply that the same characteristic composition of dinucleotides prevails in the entire genome, including not only coding but also noncoding regions of a particular organism.

It is interesting to compare these results with those analyzed at the level of gene products, i.e., proteins. Table 4 shows the mean amino acid composition of proteins in the ten organisms, corresponding to the gene data in Table 1. Some characteristic features were observed in the mean amino acid composition for the different organisms. For example, large differences between organisms exist in the mean composition of Ala (9.4% in E. coli versus 5.5% in Met. jannaschii) or of Lys (4.7% in E. coli and 5.5% in Met. jannaschii).
Figure 3. A dendrogram derived from the distances (upper part in Table 3) between the mean points (expressed by Eq. (4) in the text) of the gene clusters, and an unrooted tree was produced by the neighbor-joining method. Each branch length indicated was measured using the same scale.

versus 10.6% in \textit{M. jannaschii}). Similar to the analysis of genes, individual proteins can be plotted as points in the 20-dimensional amino acid composition space, and they are classified into groups depending on the distribution in the composition space.\textsuperscript{14–17} All proteins from the nine complete genomes and those from humans were classified into organisms by the distance in the normalized amino acid composition space. The accuracy of classification of proteins by the amino acid compositions was 54.4%. This figure, although significant in comparison with the random guess (ca. 10%), is much lower than that for genes classified by the dinucleotide composition. This suggests that the variation at the protein level seems to be a secondary effect.

4. Discussion

The clustering nature of DNA sequences in the dinucleotide space, depending on the organism, is remarkable. Genes from different organisms are densely clustered around respective mean points (Fig. 2). This trend is not limited to genes or to protein-coding regions of DNA, but prevails throughout the entire genome itself. The shape of the star-diagrams in Fig. 1 shows a mean dinucleotide composition characteristic of each organism. However, when comparing the diagrams, it is peculiar to find no systematic similarities in their shapes. For instance, the diagram of yeast (\textit{S. cerevisiae}) is closer in shape to that of one of the micoplasma (\textit{Myc. pneumoniae}) than to that of human, another eukaryote. The shape of \textit{E. coli} is very different from those of the other two Gram-negative bacteria, \textit{Hae. influenzae} and \textit{Hel. pylori}. These relations are quantitatively given as distances among mean compositions in the lower half of the diagonal part in Table 3. Using these distances, it is possible to construct a dendrogram. The resulting dendrogram (data not shown) was completely different from the usual phylogenetic tree for these organisms.

We then tried to alter the definition of the measure of distance in the 16-dimensional space, by replacing Eq. (1) with the following equation.

\[ V_{i,k} = CDN_{i,k} - \text{EXP}_{i,k} \]  

(4)

Here, \text{EXP} is a vector expressing the expected dinucleotide composition calculated from its mononucleotide composition by multiplication. The origin of the new coordinate system is the point of the expected dinucleotide composition of each organism, and the new vector \text{V} implies the net deviation from the expected composition at the dinucleotide level. Using this new vector as a variable, the distance among the mean dinucleotide composition of various organisms was calculated (upper half of Table 3), and the dendrogram in Fig. 3 was obtained by the neighbor-joining method.\textsuperscript{27} The relationship among various organisms is more or less similar to the real phylogenetic relation among them, except that eukaryotes and archaea are collected in one group. The position of human in the phylogenetic tree might change, as it was derived using an incomplete genome. It was used here as a reference. As a whole, the dendrogram diverges into
several branches almost according to the five subgroups of the organisms, as indicated in Table 2. The dinucleotide composition is thought to consist of two parts. One part is attributed to mononucleotide composition, or more simply the G + C content, and the other is a deviation from the expectation given by the multiplication of mononucleotide contents. The present results suggest that only the second part, deviation from the expected composition, is correlated with the phylogenetic relationship of these organisms. If this is the case, this correlation is interesting and of importance in analyzing the DNA data of genomes in comparative studies. Also, it would provide a clue to the question of what kind of factors brings about species-specific dinucleotide composition of genes as well as the genome. However, before proceeding further it would be necessary to confirm the above correlation by using more DNA data available from a larger number of organisms.

Our main aim in the present study was to examine to what extent genes of various organisms could be separated into clusters in the dinucleotide space, as suggested in a previous study.\textsuperscript{13} The degree of separation depends on the detailed definition of the scaling. In our previous work, the log-odds ratio of dinucleotide frequencies in a sequence was employed. Applying this measure to the ten organisms in the present study, the correct assignment obtained was 77.0%, a little less than that in Table 2. If the absolute dinucleotide composition per se was used instead, the accuracy of assignment was 76.4%. The best separation (79.7% accuracy) was obtained with the normalized composition as shown in Table 2. If equation (4) is employed, the accuracy drops to 73.8%. These results indicate that the two factors mentioned above, i.e., the G + C content and the deviation from the expected value, affect the separation of gene distributions. Our question then is, what causes the deviation of a whole genome at the dinucleotide level that characterizes individual organisms by 80% confidence?

The present study confirmed the main results of our previous analysis done for three species, \textit{E. coli}, yeast and human, and suggest a possible cause of gene separation: the scarcity of dinucleotide CpG and excess of TpG/CpA in humans is due to the methylation-deamination mutation that converts CpG to TpG. The scarcity of TpA in humans might be related to the mRNA stability due to the UpA-selective ribonucleases.\textsuperscript{28} However, the reason why individual organisms have unique dinucleotide compositions seems beyond these explanations.

Acknowledgments: We thank Dr. Kaoru Fukami-Kobayashi for her help in drawing the dendrograms.

References


