Systematization of Species-Specific Diversity of Genes in Codon Usage: Comparison of the Diversity Among Bacteria and Prediction of the Protein Production Levels in Cells

Shigehiko Kanaya[1]
kanaya@eie.yz.yamagata-u.ac.jp
Yoshihiro Kudo[1]
ykudo@eie.yz.yamagata-u.ac.jp
Shinya Suzuki[1]
a93619@eie.yz.yamagata-u.ac.jp
Toshimichi Ikemura[2]
tikemura@ddbj.nig.ac.jp

[1] Department of Electric and information Engineering, Faculty of Engineering, Yamagata University, Yonezawa, Yamagata-ken 992, Japan

[2] Department of Evolutionary Genetics, National Institute of Genetics, and the Graduate University for Advanced Studies, Mishima, Shizuoka-ken 411, Japan

Abstract

In the present study, we have developed the procedure for estimating species-specific heterogeneous codon usage among intraspecific genes called diversity in codon usage and for systematizing species by the species-specific diversity on the basis of principal component analysis. We tried to quantify differences of the diversity among five species, Escherichia coli (Ec), Salmonella typhimurium (St), Haemophilus influenzae (Hi), Bacillus subtilis (Bs), and Synechocystis sp. (Ss). In the five species, many of genes involved in the translation process and energy metabolism had positive values ($Z_1 > 0$) on the first principal component (PC1). In Ss, many of genes involved in photosynthetic system had also postive Z_1-values. These genes are thought to be highly expressed. By the direction of PC1, the five species were roughly classified into three categories, [Ec, St, Hi], [Ss], [Bs]. The dendrogram constructed was roughly consistent with the rRNA-based phylogeny, but interesting differences were also observed between the two phylogenic trees.

1 Introduction

The choice among synonymous codons in prokaryotic and eukaryotic genes is not random although it does not affect the nature of proteins synthesized, and the codon-usage pattern is undoubtedly useful in the search for genes in newly obtained sequences. In the genes of a particular unicellular organism, codon-choice patterns (called dialects) are similar regardless of gene function. Taxonomically related organisms have similar dialects and there is also considerable within-species heterogeneity. The extent of codon bias was related to the protein production level of individual genes[1][2][3][4]. Codon usage in genes encoding abundant proteins is almost always more dependent on the tRNA content (strong accent) than in moderately or poorly expressed genes (moderate accent), that is, highly expressed genes almost always have a strong accent but those with moderate or low expression have a moderate accent. The extent of codon bias was related to the protein production level of individual genes in *Escherichia coli*[4] and in T4 phage[5]. Correlation equations between codon-usage pattern and the protein production level for *E. coli* were developed in the previous study[6].

In the present study, we have developed the multivariate procedure based on principal component analysis (PCA) for estimating species-specific heterogeneous codon usage among intraspecific genes called species-specific diversity of genes in codon usage, and for systematizing species by the diversity. We tried to quantify differences of the species-specific diversity among five species, *E. coli*, *Salmonella typhimurium*, *Haemophilus influenzae*, *Bacillus subtilis*, and *Synechocystis sp.* and examined the potential of the measure of the diversity for estimating the protein-production levels in cells.

2 Methodology

The purpose of this study was to construct measures reflecting the heterogeneous codon usage among intraspecific genes, called the diversity of genes in codon usage in this study, and to develop a method for systematizing the species-specific diversity of genes in codon usage.

2.1 Representation of Species-specific Diversity in Codon Usage

Codon usage in a gene can be described by a vector consisting of codon-usage frequencies. To assess the species-specific diversity in codon usage, we conducted principal component analysis (PCA)[7][8], constructing axes which reflect the most heterogeneous codon usage among genes. The transformation in PCA is not object- but variable-oriented, so variations of the original variables directly influence the transformed variables (called Z'_k); that is, the variables Z'_k are affected by the multivariate representation of codon usage. This property of PCA differs from those of methods such as correspondence factor analysis based on chi-squared metric[9] and spectral map analysis based on two-way analysis of variance[10].

First, codon-usage patterns for genes were represented by *Eq.*(1).

$$x_{ij(m)} = f_{ij(m)} / [\sum_{j=1}^{M(m)} f_{ij(m)} / M(m)] \tag{1}$$

Here, in the mth amino acid, $f_{ij(m)}$ denotes the number of the jth codon for the ith gene, and M(m) denotes the synonymous codon number. With this representation it is expected that the

effect of the amino acid composition could be excluded and the resolutions of codon frequencies affected by the synonymous codon number, M(m), in each amino acid could be equivalent.

Secondly, in quantification of the diversity of genes in the 61-dimensional space representing codon frequencies (omitting the three termination codons) and to reduce the 61 variables into fewer and more fundamental variables while keeping most of the essential information about variance of the original variables, PCA was done for the reference data sets consisting of genes for each species, which have been characterized experimentally. In the present study, the codon-usage pattern represented in *Eq.*(1) was represented by the following matrix. Here, variables and elements are denoted by upper- and lower-case letters (X_j and x_{ij}, respectively).

$$\begin{array}{cccccc} X_1 & X_2 & .... & X_j & .... & X_{61} \end{array}$$

$$\begin{pmatrix} x_{11} & x_{12} & .... & x_{1j} & .... & x_{161} \\ ... & ... & .... & ... & .... & ... \\ x_{i1} & x_{i2} & .... & x_{ij} & .... & x_{i61} \\ ... & ... & .... & ... & .... & ... \\ x_{n1} & x_{n2} & .... & x_{nj} & .... & x_{n61} \end{pmatrix}$$

The vector consisting of the relative codon frequencies for the ith gene, [x_{i1}, x_{i2}, ..., x_{i61}], was transformed into a vector consisting of principal components (PCs), [z_{i1}, z_{i2}, ..., z_{i61}], according to the following conditions: (i) The correlations of the principal components between Z'_k and $Z'_{k'}$ (k < k', k = 1, 2,.., 60; k' = 2, .., 61) were zero; (ii) the first principal component, Z'_1, was the linear combination of variables X_j with the largest variance, the second principal component, Z'_2, was the linear combination of the variables with second largest variance, and so on.

$$Z'_k = \sum_{j=1}^{61} b_{kj} X_j \tag{2}$$

where

$$\sum_{j=1}^{61} b_{kj}^2 = 1.0$$

We refer to the b_{kj} as principal coefficient of the kth PC for the jth original variable, and the vector [b_{k1},..., b_{k61}] as b-vector for the kth PC. The b-vector for species represents the direction of species-specific diversity. If two species, s and s', have similar directions of the diversity, the absolute value of inner product of the two b-vectors, $\mathbf{b}_s$ and $\mathbf{b}_{s'}$ is close to one. On the other hand, the absolute value is close to zero, if two species have different directions.

There are some criteria for choosing PCs, for example, criteria by cumulative percentage of total variation, size of variances of PCs, cross validatory choice, partial correlation coefficients and so on[11]. The purpose of the present analysis was to construct measures with the widest scale in the space consisting of the 61 original variables, so we tentatively selected PCs with variance larger than the maximum of the original variables by *Eq.*(3), which corresponds to the criterion by size of variance of PCs as in the Kaiser's rule[12], because any PC with variance less than those for the original variables contain less information than the original variables, so is not worth retaining.

$$PR[Z'_k] > \max_j \{PR[X_1], PR[X_2], .., PR[X_j], .., PR[X_{61}]\} \tag{3}$$

where,

$$PR[Z'_k] = Var[Z'_k] / \sum_{j=1}^{61} Var[X_j] \tag{4}$$

$$PR[X_j] = Var[X_j] / \sum_{j=1}^{61} Var[X_j] \tag{5}$$

Here, Var[A] represents variance of variable A.

The following parameters, factor loadings and Z-parameters, were used to interpret PCs. The factor loadings r(Z'_k, X_j) denoted by *Eq.*(6) shows the contribution of the jth codon frequency to the kth principal component Z'_k.

$$r(Z'_k, X_j) = Cov[Z'_k, X_j] / (Var[Z'_k] Var[X_j])^{1/2} \tag{6}$$

where, Cov[A,B] denotes the covariance between the two variables A and B.

To standardize the scale of the PCs, Z'_k, we normalized each of those for the reference data set to zero for the average and unity for the standard deviation by *Eq.*(7).

$$Z_k = (Z'_k - Av[Z'_k]) / SD[Z'_k] \tag{7}$$

Here, Av[Z'_k] and SD[Z'_k] are average and standard deviation of Z'_k for the reference data. We refer to Z_k as Z-parameter for the kth PC. We can understand the direction of species-specific diversity of genes in codon usage by b-vector, and comprehend the specificity of codon usage for genes by scattering genes on a map consisting of the first kth Z-parameters (Z_k). The Z-parameters also provide the statistical information about the diversity of genes in codon usage. For example, according to statistical theory, 68.8% of genes are in the range -1 $< Z_1 <$ 1, 95% of genes in the range -2 $< Z_1 <$ 2, and so on.

2.2 Systematization of Species-specific Diversity in Codon Usage

As mentioned in Section 2.1, the direction of the species-specific diversity is represented by the b-vector. So, we examined to classify species by the b-vectors as follows.

First, we used Euclidean distance, D(s_k, $s'_{k'}$) in *Eq.*(8), to represent similarity of the diversity between two species, s and s', because the inner product of the two b-vectors stated above is reflected in this distance representation.

$$D(s_k, s'_{k'})^2 = \sum_{j=1}^{61} (b_{kj}^{(s)} - b_{k'j}^{(s')})^2 = 2(1 - \sum_{j=1}^{61} b_{kj}^{(s)} b_{k'j}^{(s')}) \tag{8}$$

where $b_{kj}^{(s)}$ denotes b-vector of the jth codon in the kth PC for species s.

Then, to classify species by the species-specific diversity, we applied centroid clustering method[13] to the Euclidean distance data.

3 Results and Discussion

So that differences in bacterial strains would not cause confusion, sequences annotated '*Escherichia coli K12* ', '*Salmonella typhimurium LT2*', '*Haemophilus influenzae Rd*', '*Bacillus*

Table 1: Proportions for the first four PCs. (NS and NP represent the number of genes analyzed and the number of significant PCs, respectively.)

Species	NS	NP	$\Pr[Z'_1]$	$\Pr[Z'_2]$	$\Pr[Z'_3]$	$\Pr[Z'_4]$
Ec	610	3	0.250	0.135	0.062	0.048
St	130	2	0.240	0.150	0.062	0.055
Hi	1034	1	0.176	0.095	0.083	0.049
Ss	253	4	0.159	0.074	0.070	0.060
Bs	150	4	0.126	0.114	0.082	0.069

subtilis 168', and '*Synechocystis sp. PCC6803*' were extracted from bacterial sequences in DDBJ (DNA DataBank of Japan). Hereinafter, we abbreviate *E. coli* as Ec, *S. typhimurium* as St, *H. influenzae* as Hi, *B. subtilis* as Bs, and *S. sp.* as Ss. In Ss, the number of genes which have been experimentally characterized was less than fifty thus far, so we constructed a reference data set consisting of function-known genes and genes whose sequences were reported to be homologous to those for function-known genes. So that multiple entries of one gene would be avoided, only sequences whose gene names were detected in linkage maps[14][15][16][17][18][19] for respective species were selected. The largest sequence was selected when sequences with the same gene name but different coding lengths were present. If the gene name and length were the same but some bases were different, one was selected arbitrarily. In order to construct principal component axes, we selected from them sequences longer than 500 bases. Table 1 shows proportions for the first four PCs in each species. The first four PCs for Ss and Bs, the first three PCs for Ec, the first two PCs for St, and the first PC for Hi were statistically significant axes by *Eq.*(3) and these PCs were characterized.

Relationships among the direction of the diversity of genes in codon usage were examined by calculation of inner products between all pairs of b-vectors, as shown in Table 2. The largest five products were 0.97, between b_1 for Ec and b_1 for St; 0.96, between b_2 for Ec and b_2 for St; 0.63, between b_2 for Ec and b_2 for Bs; 0.62, between b_2 for St and b_2 for Bs; and 0.62, between b_2 for Ec and b_3 for Ss. The remaining products were less than 0.60. The b-vectors were almost the same between two taxonomically related organisms Ec and St. This suggests that the present methodology is appropriate for estimating the diversity of genes in codon usage, regardless of the number of genes in reference data set. Except a pair of Ec and St, the direction of b_1-vectors were quite different between any two species. On the other hand, Ec, St, and Bs have similar tendencies in the direction of b_2-vectors. Table 3 shows the correlations between Z-parameters of the first four PCs (denoted by *Eq.*(7)) and G+C% at the codon third position. Z-parameters of PC2 for four species (Ec, St, Hi, and Bs) were highly correlated to G+C% at the codon third positions, though PC2 for Hi was not statistically significant principal component by *Eq.*(3). This indicates that similar tendencies in the direction of the diversity of PC2 for Ec, St, Hi, and Bs in Table 2 are caused by the G+C% variation at the codon third position for genes. Thus, PC1 reflects species-specific diversity of genes in codon usage, on the other hand, PC2 has more general diversity among species such as G+C% at codon third position.

To characterize PC1s, we examined factor loadings of Z_1 for the five species shown in Table

Table 2: Inner products between b-vectors for significant PCs.

	St-b_1	St-b_2	Hi-b_1	Ss-b_1	Ss-b_2	Ss-b_3	Ss-b_4	Bs-b_1	Bs-b_2	Bs-b_3	Bs-b_4
Ec-b_1	0.97	0.01	0.40	-0.59	0.25	-0.01	-0.06	-0.32	-0.43	-0.22	0.14
Ec-b_2	0.01	0.96	0.52	0.12	0.51	0.62	0.08	-0.28	0.63	-0.49	0.30
Ec-b_3	0.07	0.00	-0.05	-0.24	0.36	-0.42	0.01	-0.27	0.22	0.59	-0.30
St-b_1	1.00	0.00	0.38	-0.59	0.20	0.00	-0.14	-0.31	-0.42	-0.20	0.14
St-b_2		1.00	0.53	0.11	0.56	0.52	0.11	-0.35	0.62	-0.53	0.17
Hi-b_1			1.00	0.09	-0.13	-0.20	-0.14	-0.63	0.03	-0.31	0.28
Ss-b_1				1.00	0.00	0.00	0.00	0.13	0.35	0.09	0.01
Ss-b_2					1.00	0.00	0.00	-0.56	0.31	-0.07	0.03
Ss-b_3						1.00	0.00	0.27	0.34	-0.49	0.56
Ss-b_4							1.00	0.08	0.05	-0.03	-0.13

Table 3: Correlation coefficients between Z_k and G+C% at the codon third position.

Species	Z_1	Z_2	Z_3	Z_4
Ec	0.50	0.73	0.21	0.08
St	0.42	0.80	0.12	0.22
Hi	-0.20	0.75	0.05	0.02
Ss	0.67	0.45	0.15	0.04
Bs	-0.36	0.78	-0.00	-0.02

4. In *E. coli*, twenty of the twenty-one optimal codons (noted by '$*$' in Table 4) assigned by Ikemura[2][3][20] contributed positively to PC1. In Bs, seventeen of the eighteen major codons ('+') for each amino acid in highly expressed genes indicated by Perriere et al.[21] contributed positively to PC1. Table 5 shows genes with Z_1 larger than 2. In all species, almost all the genes in Table 5 are involved in the translation process and energy metabolism and are thought to be highly expressed. In Ec, we developed a correlation equation between Z_1 and the protein production levels in cells represented by *Eq.*(9). Details are described in [6].

$$log(Rich) = 2.44 + 0.55Z_1 \tag{9}$$

Here, log(Rich) represents the common logarithm of the amount of protein molecules per genome in cells grown in rich medium. In three purple bacteria (Ec, St, Hi), we could assign 99 genes for St to those for Hi, 317 genes for Ec to those for Hi, and 90 genes for Ec to those for St. Correlations between Z_1-parameters for these species were also observed as follows; correlation coefficients were 0.91, between Ec and St; 0.75, between St and Hi; and 0.67, between Ec and Hi. These indicate that PC1 reflects species-specific preferability in codon usage connected with protein production levels. Assuming the total protein content per cell among the three species are nearly the same level, the protein production levels for genes in St and Hi may be approximately assessed by substituting Z_1 in these species for *Eq.*(9). In common with the five species, ten variables (CGU, UCU, GGU, GUU, AUC, AAC, GAA , CAC, UUC UAC) contributed positively to PC1 and fourteen variables (CGA, AGA, AGG, UCG, AGU, GGA, GGG, GUC, AUA, AAU, GAG, CAU, UUU, UAU) contributed negatively to PC1. These suggest that the former ten are the common preferential codons for the five species. The remaining 35 variables (= 61 - 2 [Trp, Met] - 10 - 14) are especially important factors reflecting species-specific diversity of genes in codon usage.

We classified the five species using b_1-vectors because PC1 is thought to be the most important factor reflecting species-specific diversity in codon usage as mentioned above. Fig. 1 shows dendrogram of the five species. The five species were classified into three comprehensive categories, Category I [Ec, St, Hi], Category II [Bs] and Category III [Ss]. Category I is characterized as purple bacteria. The classification of the five species by the direction of PC1 is roughly consistent with the rRNA-based phylogeny, but interesting differences are also observed between the two phylogenic trees as follows. In rRNA-based phylogeny of prokaryotes[22] which reflects genome G+C% content, purple bacteria (corresponding to Category I) and low G+C Gram-positive bacteria (Bs) are merged to form the cluster [Ec, St, Hi, Bs], and then, this cluster and Cyanobacteria (Ss) are merged. In Fig. 1 which is expected to be excluded the G+C% at the codon third position, Bs and Ss are merged to form the cluster [Bs, Ss], and then, the cluster [Bs, Ss] and purple bacteria cluster [Ec, St, Hi] are merged. These differences could be explained as follows; in the former, species are classified by the general genome property similar to PC2, on the other hand, the classification by the latter reflects more species-specific diversity than that by the former.

Table 4: Factor loadings of Z_1 for five species. ('$*$' and '$+$' are explained in the text.)

	Codon	Ec		St	Hi	Ss	Bs	
Arg	CGU	0.651	*	0.512	0.860	0.263	0.676	+
	CGC	0.018	*	0.078	-0.216	0.130	0.490	
	CGA	-0.515		-0.447	-0.537	-0.435	-0.252	
	CGG	-0.504		-0.380	-0.267	0.306	-0.645	
	AGA	-0.455		-0.511	-0.399	-0.344	-0.290	
	AGG	-0.301		-0.243	-0.280	-0.297	-0.447	
Leu	UUA	-0.714		-0.628	0.597	-0.525	0.003	
	UUG	-0.519		-0.520	-0.374	0.191	-0.090	
	CUU	-0.531		-0.518	-0.011	-0.098	0.541	+
	CUC	-0.111		-0.142	-0.245	0.342	-0.232	
	CUA	-0.466		-0.474	-0.276	-0.165	0.020	
	CUG	0.898	*	0.878	-0.366	0.285	-0.387	
Ser	UCU	0.436		0.602	0.414	0.071	0.488	+
	UCC	0.513		0.529	-0.230	0.766	-0.148	
	UCA	-0.465		-0.250	0.212	-0.384	-0.024	
	UCG	-0.254		-0.260	-0.235	-0.233	-0.338	
	AGU	-0.501		-0.575	-0.379	-0.556	-0.019	
	AGC	0.040		-0.340	-0.019	-0.052	-0.131	
Ala	GCU	0.146	*	0.276	-0.092	0.108	0.213	+
	GCC	-0.185		-0.117	-0.304	0.444	-0.273	
	GCA	-0.098	*	-0.131	0.349	-0.483	-0.017	
	GCG	0.137	*	-0.015	-0.025	-0.275	0.022	
Gly	GGU	0.370	*	0.388	0.590	0.447	0.470	+
	GGC	0.328	*	0.340	-0.016	0.151	-0.004	
	GGA	-0.592		-0.560	-0.467	-0.471	-0.149	
	GGG	-0.493		-0.498	-0.382	-0.260	-0.336	
Pro	CCU	-0.360		-0.590	-0.235	-0.153	0.146	+
	CCC	-0.575		-0.461	-0.281	0.514	-0.235	
	CCA	-0.333		-0.249	0.500	-0.477	0.203	
	CCG	0.699	*	0.736	-0.202	-0.162	-0.115	
Thr	ACU	0.194	*	0.032	0.351	-0.347	0.414	+
	ACC	0.535	*	0.508	-0.075	0.677	-0.413	
	ACA	-0.525		-0.329	-0.113	-0.473	0.216	
	ACG	-0.395		-0.378	-0.207	-0.243	-0.287	
Val	GUU	0.169	*	0.128	0.065	0.035	0.356	+
	GUC	-0.271		-0.201	-0.207	-0.004	-0.356	
	GUA	0.020	*	-0.148	0.211	0.073	0.249	
	GUG	0.028	*	0.154	-0.124	-0.070	-0.230	
Ile	AUU	-0.430		-0.436	-0.115	-0.201	0.003	
	AUC	0.655	*	0.662	0.405	0.451	0.183	+
	AUA	-0.544		-0.529	-0.437	-0.456	-0.262	
Asn	AAU	-0.684		-0.712	-0.399	-0.542	-0.489	
	AAC	0.684	*	0.712	0.399	0.542	0.489	+
Asp	GAU	-0.533		-0.519	-0.155	-0.361	0.002	
	GAC	0.533		0.519	0.155	0.361	-0.002	+
Cys	UGU	-0.181		-0.129	0.089	-0.165	-0.098	
	UGC	0.181		0.129	-0.089	0.165	0.098	+
Gln	CAA	-0.511		-0.508	0.315	-0.054	0.121	+
	CAG	0.511	*	0.508	-0.315	0.054	-0.121	
Glu	GAA	0.246	*	0.311	0.217	0.175	0.369	+
	GAG	-0.246		-0.311	-0.217	-0.175	-0.369	
His	CAU	-0.601		-0.539	-0.337	-0.573	-0.332	
	CAC	0.601		0.539	0.337	0.573	0.332	+
Lys	AAA	0.086	*	-0.054	0.343	0.030	0.240	+
	AAG	-0.086		0.054	-0.343	-0.030	-0.240	
Phe	UUU	-0.686		-0.643	-0.547	-0.571	-0.285	
	UUC	0.686	*	0.643	0.547	0.571	0.285	+
Tyr	UAU	-0.525		-0.620	-0.302	-0.398	-0.291	
	UAC	0.525	*	0.620	0.302	0.398	0.291	+

Table 5: Genes with Z_1 larger than 2.

Species	Gene(Z_1)				
Ec	tufA(3.24)	tufB(3.11)	mopA(3.01)	rpsI(2.98)	rpsW(2.73)
	rplL(2.73)	rpmB(2.58)	ompA(2.52)	rplD(2.49)	fusA(2.48)
	rpoC(2.46)	atpD(2.45)	rplO(2.44)	rplQ(2.44)	atpA(2.41)
	rplK(2.40)	rplA(2.38)	rpsC(2.36)	rpmH(2.34)	rpmG(2.32)
	tpiA(2.31)	rpmA(2.30)	lpd(2.29)	rpsL(2.38)	aceE(2.26)
	pfkA(2.23)	sodA(2.20)	rpsF(2.18)	dnaK(2.18)	glyA(2.13)
	adk(2.12)	pnp(2.11)	glnA(2.10)	deoD(2.08)	purA(2.07)
	aceF(2.03)	rplB(2.03)	ppa(2.03)	hupA(2.02)	rpoB(2.00)
St	tufA(3.48)	tufB(3.42)	rpsU(2.86)	rplL(2.83)	ompA(2.78)
	nmpC(2.74)	fusA(2.58)	glyA(2.54)	hupA(2.46)	ompC(2.46)
	rpsL(2.31)	cspS(2.24)	sodA(2.21)	rpoB(2.12)	adk(2.12)
	ahpC(2.12)				
Hi	rplL(4.06)	rplA(3.65)	rpmG(3.62)	rplI(3.58)	rpsA(3.49)
	rplK(3.44)	tsf(3.43)	pal(3.41)	gapA(3.40)	rpsO(4.29)
	tufB(3.25)	eno(3.15)	tufA(3.14)	pflB(3.13)	rpmF(3.11)
	rpsO(3.08)	rplD(3.07)	rplS(2.99)	pnp(2.97)	rpmE(2.91)
	fusA(2.90)	ompA(2.85)	sodA(2.86)	rpmI(2.83)	rplV(2.83)
	rplP(2.80)	atpD(2.74)	mdh(2.70)	gmpA(2.65)	rpsI(2.62)
	pgk(2.60)	deoD(2.59)	yejX(2.58)	rpsB(2.57)	dnaK(2.56)
	rplB(2.54)	rplM(2.54)	rpsJ(2.54)	rpsK(2.53)	aceE(2.52)
	rplC(2.45)	yaeC(2.44)	glpT(2.36)	cspD(2.35)	frr(2.35)
	rpoB(2.31)	lpd(2.30)	ilvC(2.30)	fis(2.30)	adk(2.29)
	trxA(2.28)	rpmB(2.28)	secG(2.26)	yjbP(2.26)	atpA(2.26)
	rpsL(2.24)	groS(2.23)	rpsH(2.22)	rpmH(2.21)	rpsG(2.20)
	ppa(2.20)	appB(2.19)	fabG(2.19)	pckA(2.19)	glyA(2.17)
	tig(2.16)	aspA(2.16)	dapD(2.15)	pykA(2.15)	grxA(2.13)
	rplJ(2.13)	atpF(2.11)	rpoC(2.10)	rplR(2.09)	thrS(2.08)
	groL(2.08)	fabB(2.05)	metE(2.05)	tpiA(2.04)	crr(2.02)
	rplU(2.02)				
Ss	rbcL(3.45)	petF(3.10)	apcB(2.99)	psbA3(2.81)	psbA2(2.76)
	glnA(2.61)	groEL(2.46)	atpB(2.31)	atpA(2.27)	rplL(2.16)
	petB(2.12)	rplS(2.09)	apcA(2.08)		psII(11kD)(2.00)
Bs	rplB(4.49)	rpsC(4.38)	rplE(3.89)	rplD(3.70)	rplJ(3.64)
	rplJ(3.64)	gtaB(2.71)	glyC(2.57)	rpoB(2.44)	oppA(2.06)
	cysK(2.05)	purA(2.03)			

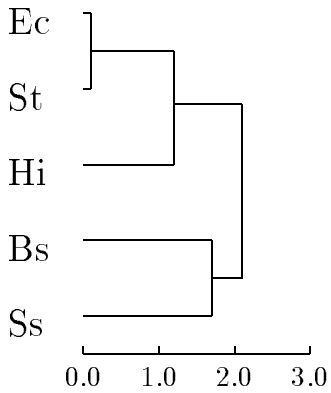

Figure 1 Dendrogram between five species.

4 Summary and Conclusions

In the present study, we have developed the multivariate procedure for estimating species-specific diversity of genes in codon usage, and could quantify differences of the diversity among five species (*Escherichia coli*, *Salmonella typhimurium*, *Haemophilus influenzae*, *Synechocystis sp.* and *Bacillus subtilis*). PC1, which is connected with the preferential codon usage and the protein production levels in cells, has more species-specific structure of the diversity in codon usage than other PCs. The classification of the five species by the direction of PC1 was roughly consistent with the rRNA-based phylogeny, but interesting differences were also observed between the two phylogenic trees. The methods developed here could be certainly applicable to prokaryotes and eukaryotes which a large number of genes have been sequenced, and could be useful for systematizing organisms from the viewpoint of the species-specific heterogeneous codon usage among genes. We could also suggest the potential of the measure developed here for estimating the protein production levels in cells.

References

[1] T. Ikemura, *J. Mol. Biol.*, Vol. 146, pp. 1-12, 1981.

[2] T. Ikemura, *J. Mol. Biol.*, Vol. 151, pp. 389-409, 1981.

[3] T. Ikemura, *J. Mol. Biol.*, Vol. 158, pp. 573-597, 1982.

[4] T. Ikemura, *Mol. Biol. Evol.*, Vol. 2, pp. 13-34, 1985.

[5] T. Kunisawa, *J. theor. Biol.*, Vol. 172, pp. 287-298, 1992.

[6] S. Kanaya, Y. Kudo, Y. Nakamura, and T. Ikemura, *CABIOS*, Vol. 12, pp. 213-225, 1996.

[7] Y. Chien, Interactive Pattern Recognition, *Electrical Engineering and Electronics III*, Marcel Dekker, Inc., New York, pp. 25-64, 1978.

[8] I. T. Jolliffe, *Principal component analysis*, Springer Series in Statistics, Springer-Verlag, New York, pp. 64-91, 1986.

[9] M. O. Hill, *Appl. Statist.*, Vol. 23, pp. 340-354, 1974.

[10] L. Lewi, *Chem. Int. Lab. Sys.*, Vol. 5, pp. 105-116, 1989.

[11] I. T. Jolliffe, *Principal component analysis*, Springer Series in Statistics, Springer-Verlag, New York, pp. 92-114, 1986.

[12] H. F. Kaiser, *Edu. Psychol. Meas.*, Vol. 20, pp. 141-151, 1960.

[13] R. A. Johnson and D. W. Wichern, Applied multivariate statistical analysis, Prentice Hall International, Inc., London, pp. 573-627, 1992.

[14] B. J. Bachmann, *Microbiol. Rev.*, Vol. 54, pp. 130-197, 1990.

[15] P. J. Piggot and J. A. Hoch, *Microbiol. Rev.*, Vol. 49, pp. 158-179, 1985.

[16] R. D. Fleischmann et al., *Science*, Vol. 269, pp. 496-612, 1995.

[17] T. Kaneko, A. Tanaka, S. Sato, H. Kotani, T. Sazuka, N. Miyajima, M. Sugiura, and S. Tabata, *DNA Res.*, Vol. 2, pp. 153-166, 1995.

[18] H. Kaneko, T. Kaneko, T. Matsubayashi, S. Sato, M. Sugiura, and S. Tabata, *DNA Res.*, Vol. 1, pp. 303-307, 1994.

[19] K. E. Sanderson, A. Hessel, and K. E. Rudd, *Microbiol. Rev.*, Vol. 59, pp. 241-303, 1995

[20] T. Ikemura, In D. L. Hatfield, B. J. Lee, R. M. Pirlte (eds), *Transfer RNA in protein synthesis*,CRC Press, London, pp. 87-111, 1992.

[21] G. Perriere, M. Gouy, and T. Gojobori, *Nucl. Acids Res.*, Vol. 22, pp. 5525-5529.

[22] G. J. Olsen, C. R. Woese, and R. Overbeek, *J. Bacteriol.*, Vol. 176, pp. 1-6, 1994.