category: system-level analysis of biological systems

Analysis of DNA coding regions

Fumihiko Takeuchi[*†]
fumi@is.s.u-tokyo.ac.jp

Kenji Yamamoto[*]
backen@ri.imcj.go.jp

Hiroshi Yoshikura[*]
yoshikura@ri.imcj.go.jp

[*] *Research Institute, International Medical Center of Japan,* 162-8655, *Japan,* +81(3)32027181
[†] *Dept. of Information Science, Univ. of Tokyo,* 113-0033, *Japan*

Abstract

The most fundamental information of biological systems is encoded in their DNA or RNA sequences. Though these sequences contain the basic program for development, functions, sex, etc., the four nucleotides t (or u), c, a, g in the sequences seem to be aligned fairly random. To what extent are they random? This is the question we consider in this paper. The understanding of this randomness must be basic for system-level comprehension of biological systems.

We approach this problem by analyzing two kinds of proportions of amino acids in coding regions of DNA sequences. The two are the real and theoretical proportions. The real proportion in a coding region is the usual proportion after translation. The theoretical proportion, introduced by King & Jukes, is the expected proportion calculated from the proportions of nucleotides in the coding region. If the nucleotides t, c, a, g in coding regions were aligned uniformly random, the two figures should match. However, this is not the case. We will analyze the tendencies these proportions have. We begin by verifying the results in King & Jukes, and then proceed to much extensive analysis, such as classification of amino acids according to their distribution of these proportions.

1 Introduction

Vast amount of information is embedded in genetic materials, DNA or RNA, of living organisms. Information for development, body function, sex, etc. is all in DNA or RNA. Yet, if we look at the genomic sequences which consist of the four nucleotides thymine (t) (or uracil (u)), cytosine (c), adenine (a) and guanine (g), they are apparently indistinguishable from sequences obtained by random alignment of the four bases. To what extent are these genetic sequences random? This was the motivation for the research in this paper.

We approach this problem by analyzing the proportion of the 21 amino acids (regarding the stop codon as one) in coding regions. We consider two types of proportions of amino acids. The *real* proportion is the usual proportion after translation. The *theoretical* proportion is the mean proportion computed as follows: First, compute the proportions of t, c, a, g in a coding region. Let them be $p_{\mathrm{t}}, p_{\mathrm{c}}, p_{\mathrm{a}}, p_{\mathrm{g}}$. And then, calculate the expected proportion of the amino acids using these values and the codon table. For example, for alanine (A) translated from gct, gcc, gca, gcg, the proportion is $p_{\mathrm{g}}p_{\mathrm{c}}p_{\mathrm{t}} + p_{\mathrm{g}}p_{\mathrm{c}}p_{\mathrm{c}} + p_{\mathrm{g}}p_{\mathrm{c}}p_{\mathrm{a}} + p_{\mathrm{g}}p_{\mathrm{c}}p_{\mathrm{g}}$. This becomes the proportion of the amino acid for a randomly rearranged nucleotide sequence with the same nucleotide composition. If DNA sequences were uniformly random, the real and the theoretical proportions should not differ much. However, we show discrepancy between these proportions, which indicates non-randomness from this viewpoint. Amino acids have particular distribution of proportions. It might be interesting

to study further the biochemical properties of the amino acids showing peculiar behaviors in our study.

Our calculations are based on 270 samples each of coding regions of *Pyrococcus abyssi*, *E. coli*, and *Saccharomyces cervisiae*. We can find some tendencies common among these three different species.

2 Preceding works

The consideration of theoretical proportion first appeared in [9] [10]. Compared to the time, much more genomic information has become available today. We begin by verifying their results and then proceed to further analysis. Their analysis was based on 53 samples from mammalian, but our experiments are done on 270 samples each from the three species *Pyrococcus abyssi*, *E. coli*, and *Saccharomyces cervisiae*. Their definitions of theoretical proportions were slightly different from ours. As for the proportions of nucleotides used to calculate theoretical proportions, 1/4 each was used in [9], and the average proportions calculated over all coding regions was used in [10].

Study of the proportion of a nucleotide or a pair of neighboring nucleotides (dinucleotide) in DNA sequences have been done extensively, and it has been known that they represent some signature for each species [1] [3] [8] [7] [12]. Codon usage also has been studied intensively [2] [4] [5] [6] [11]. Our subject looks similar, but is not relevant directly to these two subjects.

3 Real and theoretical proportions of amino acids

Means. In [9] [10], they analyzed the mean of the real or theoretical proportion for each amino acid. We begin by giving the plot of the means of each amino acid over all coding regions. **(Fig. 1)** As mentioned in Section 2, we can use different definitions for theoretical proportions, but the fit to the diagonal becomes slightly worse. When measured by mean squared distance from the diagonal, the definition in [9] gives $3.72, 1.98, 2.27 \times 10^{-4}$, [10] gives $2.86, 1.98, 1.73 \times 10^{-4}$, compared to ours $2.80, 1.96, 1.68 \times 10^{-4}$, for *Pyrococcus abyssi*, *E. coli*, and *Saccharomyces cervisiae*, respectively. This means the match between the actual amino acid proportion and that expected from random nucleotide alignment becomes better when nucleotide composition is that of the gene encoding the protein. In the following analysis, we use our definition of theoretical proportion defined in Section 1. In other words, we are using as random sequences the ones derived from the nucleotide pool whose composition are same as in the coding sequence of the actual protein.

Our study with current databases confirms the observations in [9] [10] that amino acid proportions in proteins are close to those expected by re-aligning the four nucleotides at random. This is observed for all of the three species, but as mentioned in the previous paragraph, the best match is observed in *Saccharomyces cervisiae*, and then in *E. coli*, and then in *Pyrococcus abyssi*. The pattern of amino acid proportions are almost the same for these species. A closer look shows that

- Cysteine (C), Arginine (R), Histidine (H), Tryptophan (W) have smaller real proportion than theoretical, and
- Glutamic acid (E), Aspartic acid (D), Lysine (L), Alanine (A), Phenylalanine (F) have larger real proportion than theoretical.

The smaller real proportion of Cysteine, Arginine, Histidine matches the description in [9] [10]. They explain that Cysteine and Histidine are deficient because these amino acids have special functions. Also, the larger real proportion of Glutamic acid and Aspartic acid is explained by charge neutrality. Alanine is explained to be abundant because of its function as a "filler".

Raw plots. Taking a closer look and plotting, for each amino acid, its real and theoretical proportions in the coding regions is interesting, but for space limitation, we omit the plots for this version. For amino acids with real proportion zero, we change this value of proportion to 10^{-3} to enable analyses by logarithm. The calculations for the stop codon $(*)$ is also indicated, but they sometimes behave exceptionally. Because, this amino acid appears only once for each coding region, and the real proportion becomes one to the number of amino acids in the coding region. In the next table, we show the mean and the standard deviation of the common logarithm of real and theoretical proportions for each amino acid. The correlation between the logarithm of real and theoretical proportions is also indicated. **(Table 1)**

Scatters. The means of these distributions were given in Fig. 1. To visualize the shape of the "scatters" of the distributions, we give a plot showing for each amino acid the difference of the standard deviation of the common logarithm of real and theoretical proportions, and the correlation of the logarithm of real and theoretical proportions of the coding regions, using the numerical data in Table 1. **(Fig. 2)**

Roughly speaking, there are three kinds of amino acids in the omitted figure of raw plots made for each amino acid in various proteins:

- Type (I). The plots are found along the diagonal line. They are amino acids E, F, K and L in *Pyrococcus abyssi.*
- Type (II). The plots are diffusely distributed in the area around a point in the diagonal line. They are G, I and P in *Pyrococcus abyssi.*
- Type (III). The plots are distributed horizontally. They are $*$, A, C, D, H, M, N, Q, R, S, T, V, W and Y in *Pyrococcus abyssi.*

In Fig. 2, the coordinates of amino acids of type (I) should be like $(0, +)$, type (II) like $(0, 0)$, and type (III) like $(+, ?)$. Amino acids of type (I) or (II) are those with small x value, thus have larger theoretical scatter, since the real scatter does not vary much between different amino acids (see below). Also, they tend to be the amino acids having larger real mean proportion than theoretical in Fig. 1.

In the three species, the amino acids are behaving similarly with regard to the above three types. However, there are deviations. For example, though Glutamic acid (E) is type (III) in *E. coli*, it is like type (I) in the other two species; though Glycine (G) is type (II) in *Pyrococcus abyssi* and *E. coli*, it is like type (I) in the other one species; though Glutamine (Q) or Valine (V) are typically type (III) in *Pyrococcus abyssi*, they are more like type (I) or (II) in *Saccharomyces cervisiae*. In general the amino acids are tending to be more like type (I) or (II) in the order *Pyrococcus abyssi*, *E. coli*, *Saccharomyces cervisiae*. This might explain the convergence to the diagonal in Fig. 1 discussed in the beginning of this section.

As a whole, the points in Fig. 2 are showing negative correlation. The mean points are moving from bottom-right toward top-left with coordinates changing as $(0.517, 0, 334)$, $(0.515, 0.332)$, $(0.456, 0.389)$ in the order *Pyrococcus abyssi*, *E. coli*, *Saccharomyces cervisiae*. Also, the convergence to the mean point is becoming stronger with the sums of squared distance $0.117, 0.088, 0.085$ in this order. Aspartic acid (D) is apart in the top-right in any of the plots, but is becoming closer according to this order. Some of type (II), Glycine (G) and Isoleucine (I) in *Pyrococcus abyssi*, and Isoleucine (I) in *E. coli* are apart in the bottom-left, but the number of those are also decreasing according to the order.

Codon usage. One might think, for an amino acid, the similarity or difference of its distribution of real and theoretical proportions between different species might have relation with the similarity or difference of its codon usage between the species. But, basically they are not correlated. For pairs of species, we calculated, for each amino acid, the distance of the corresponding points in the plot of the means of the distributions in Fig. 1 or the plot of the scatters in Fig. 2. We also calculated the difference of codon usage on our data. This difference was measured by taking the square root of sum of squares of the differences of the usage of the triplets corresponding to that amino acid. The distance in plots of means and the difference in codon usage showed weak

correlation ($0.3 \sim 0.5$). The distance in plots of scatters and the difference in codon usage showed no correlation ($-0.2 \sim 0.1$).

Real and theoretical scatters. Let us now focus on the scatters of each of the real or theoretical proportions. We can observe that, among different amino acids, the scatters of the real proportions do not vary much, but the theoretical proportions vary larger. More precisely, we measure the scatters of real or theoretical proportions by the standard deviations of the common logarithm of the proportions under consideration. The standard deviation of the logarithm of real proportions is $0.12 \sim 0.49$. We can see

- Cysteine (C) and Tryptophan (W) have wide scatters of real proportions.

For the theoretical proportions, the standard deviation of the common logarithm is $0.023 \sim 0.16$. We can observe that

- Aspartic acid (D) and Serine (S) have small scatters of theoretical proportions, and
- Phenylalanine (F), Lysine (K) and Proline (P) have wide scatters of theoretical proportions.

Similarity among species. The figures and tables which have appeared look similar for the three species. We can indeed conclude that these distributions of real and theoretical proportions are similar, by comparing the correlation of these proportions between different species against the same correlation but after random relabeling of amino acids.

References

[1] Campbell, A., Mrázek, J. & Karlin, S. (1999). Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc. Natl. Acad. Sci. USA* **96**, 9184–9189.

[2] Grantham, R., Gautier, C., Gouy, M., Mercier, R. & Pavé, A. (1980). Codon catalog usage and the genome hypothesis. *Nucleic Acids Res.* **8**, r49–r62.

[3] Hanai, R. & Wada, A. (1990). Doublet preference and gene evolution. *J. Mol. Evol.* **30**, 109–115.

[4] Ikemura, T. (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. *J. Mol. Biol.* **146**, 1–21.

[5] Ikemura, T. (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J. Mol. Biol.* **151**, 389–409.

[6] Ikemura, T. (1982). Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes: Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. *J. Mol. Biol.* **158**, 573–597.

[7] Josse, J., Kaiser, A.D. & Kornberg, A. (1961). Enzymatic synthesis of deoxyribonucleic acid. *J. Biol. Chem.* **236**, 864–875.

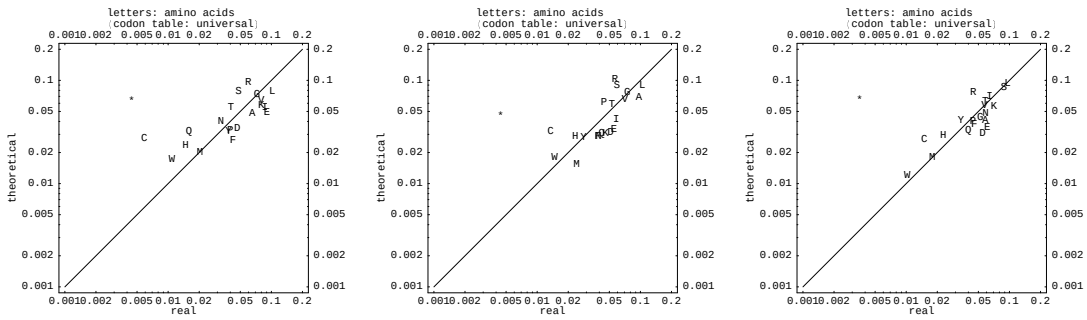
[8] Jukes, T.H. (1978). Codons and nearest-neighbor nucleotide pairs in mammalian messenger RNA. *J. Mol. Evol.* **11**, 121–127.

[9] Jukes, T.H., Holmquist, R. & Moise, H. (1975). Amino acid compositions of proteins: selection against the genetic code. *Science* **189**, 50–51.

[10] King, J.L. & Jukes, T.H. (1969). Non-Darwinian evolution. *Science* **164**, 788–798.

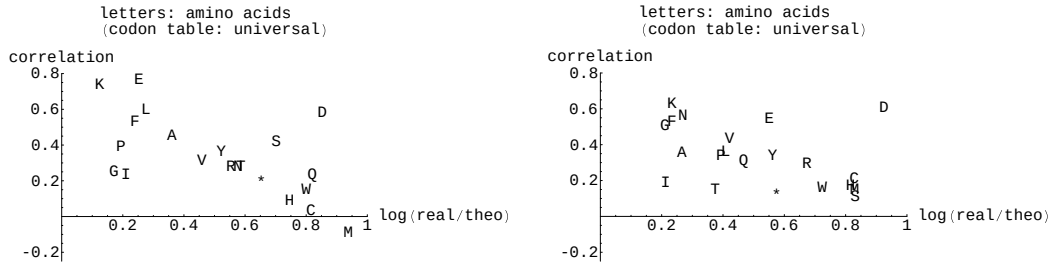
[11] Post, L.E., Strycharz, G.D., Nomura, M., Lewis, H. & Dennis, P.P. (1979). Nucleotide sequence of the ribosomal protein gene cluster adjacent to the gene for RNA polymerase subunit β in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **76**, 1697–1701.

[12] Russell, G.J., Walker, P.M.B., Elton, R.A. & Subak-Sharpe, J.H. (1976). Doublet frequency analysis of fractionated vertebate nuclear DNA. *J. Mol. Biol.* **108**, 1–23.

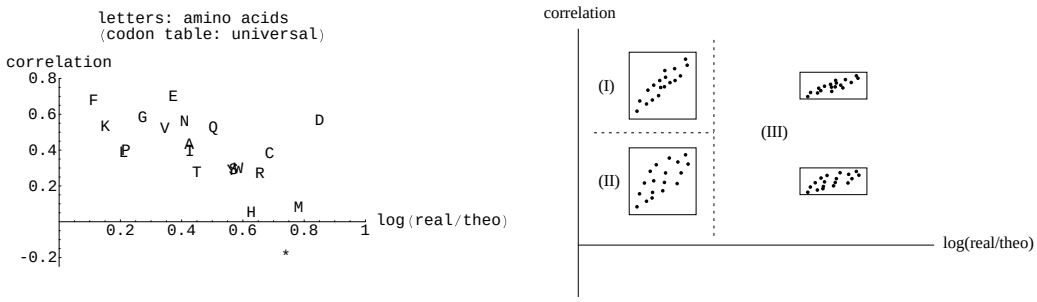


(a) *Pyrococcus abyssi* (b) *E. coli* (c) *Saccharomyces cervisiae*

FIG. 1. A plot of the real and theoretical proportions of the 21 amino acids averaged over all coding regions. The calculations are based on the universal codon table.



(a) *Pyrococcus abyssi* (b) *E. coli*



(c) *Saccharomyces cervisiae* (d) illustration of types

FIG. 2. A plot showing the type of distribution for each amino acid. The x-value is the difference of the common logarithm of the standard deviation for the real and theoretical proportions. Amino acids with wider real scatter in the omitted figure of raw plots have positive x-value and those with wider theoretical scatter have negative x-value. The y-value is the correlation between the logarithm of real and theoretical proportions. See Table 1. The calculations are based on the universal codon table.

amino acid		*	A	C	D	E	F	G	H	I	K
real	μ	−2.42	−1.22	−2.51	−1.36	−1.07	−1.43	−1.16	−1.94	−1.08	−1.12
	σ	.236	.176	.477	.183	.180	.277	.144	.375	.118	.159
theoretical	μ	−1.20	−1.33	−1.57	−1.46	−1.33	−1.61	−1.15	−1.64	−1.27	−1.26
	σ	.052	.077	.073	.026	.101	.160	.097	.067	.073	.120
correlation		.185	.450	.032	.580	.762	.529	.245	.085	.232	.733

L	M	N	P	Q	R	S	T	V	W	Y
−1.01	−1.79	−1.54	−1.44	−1.93	−1.26	−1.36	−1.43	−1.11	−2.13	−1.47
.151	.347	.251	.202	.395	.217	.204	.193	.126	.438	.270
−1.12	−1.70	−1.40	−1.50	−1.50	−1.02	−1.11	−1.26	−1.19	−1.77	−1.51
.080	.040	.067	.129	.060	.061	.040	.050	.044	.069	.081
.595	−.094	.277	.385	.233	.278	.414	.273	.311	.148	.361

(a) *Pyrococcus abyssi*

amino acid		*	A	C	D	E	F	G	H	I	K
real	μ	−2.44	−1.04	−2.04	−1.33	−1.31	−1.45	−1.15	−1.74	−1.25	−1.39
	σ	.261	.138	.442	.225	.259	.195	.147	.365	.144	.247
theoretical	μ	−1.35	−1.18	−1.50	−1.50	−1.48	−1.56	−1.13	−1.55	−1.39	−1.54
	σ	.069	.075	.066	.027	.073	.114	.090	.056	.088	.144
correlation		.110	.352	.205	.601	.539	.528	.502	.170	.184	.626

L	M	N	P	Q	R	S	T	V	W	Y
−1.00	−1.69	−1.45	−1.40	−1.41	−1.29	−1.25	−1.30	−1.17	−2.00	−1.62
.129	.326	.212	.244	.204	.228	.159	.138	.142	.454	.268
−1.06	−1.81	−1.56	−1.22	−1.52	−.997	−1.05	−1.23	−1.19	−1.75	−1.56
.050	.048	.114	.098	.069	.048	.023	.058	.054	.086	.073
.359	.146	.557	.333	.307	.291	.107	.143	.429	.158	.335

(b) *E. coli*

amino acid		*	A	C	D	E	F	G	H	I	K
real	μ	−2.56	−1.28	−1.99	−1.32	−1.27	−1.39	−1.33	−1.71	−1.22	−1.19
	σ	.319	.240	.428	.270	.278	.204	.248	.304	.185	.229
theoretical	μ	−1.20	−1.40	−1.58	−1.51	−1.47	−1.45	−1.38	−1.54	−1.17	−1.28
	σ	.062	.082	.092	.048	.121	.157	.140	.075	.069	.143
correlation		−.216	.459	.352	.535	.744	.662	.664	.095	.388	.534

L	M	N	P	Q	R	S	T	V	W	Y
−1.03	−1.82	−1.27	−1.39	−1.47	−1.42	−1.08	−1.27	−1.27	−2.16	−1.53
.118	.320	.220	.216	.269	.315	.145	.193	.155	.439	.267
−1.03	−1.75	−1.33	−1.42	−1.49	−1.13	−1.07	−1.21	−1.25	−1.94	−1.39
.073	.066	.086	.135	.084	.075	.040	.070	.077	.133	.073
.392	.085	.557	.415	.522	.271	.298	.283	.538	.309	.285

(c) *Saccharomyces cervisiae*

TABLE 1. For each amino acid, the mean μ and the standard deviation σ for the common logarithm of their real and theoretical proportions in coding regions are denoted. The correlation between the logarithm of the real and theoretical proportions is denoted in the final row. The calculations are based on the universal codon table.