Statistics of trinucleotides in coding sequences and evolution

Fumihiko Takeuchi[*†]
fumi@ri.imcj.go.jp

Yasuhiro Futamura[*]
yfutamura@ri.imcj.go.jp

Hiroshi Yoshikura[‡]
yoshikura@nih.go.jp

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

[*] *Research Institute, International Medical Center of Japan,* 162-8655*, Japan*
[†] *The Organization for Pharmaceutical Safety and Research,* 100-0013*, Japan*
[‡] *National Institute of Infectious Diseases,* 162-8640*, Japan*

October 29, 2002

Corresponding author:

Fumihiko Takeuchi

Research Institute, International Medical Center of Japan,

1-21-1 Toyama, Shinjuku-ku Tokyo, 162-8655, JAPAN

phone: +81-3-3202-7181 (ext. 2856)

fax: +81-3-3202-7364

e-mail: fumi@ri.imcj.go.jp

Abstract

The aim of this paper is to give measurements indicative of evolutional stages of the species. Two types of statistics of trinucleotides in coding regions are analyzed for 27 species.

The first one is the codon space, the nucleotide ratio for each of the three codon positions. We apply principal component analysis on this space and extract two principal components faithfully describing the original distribution of the codon space. The first principal component corresponds to the GC content. The second principal component classifies the species into three evolutional groups, *Archaea*, *Bacteria* and *Eukaryota*.

The second statistics is the real and theoretical frequency of amino acids. The *real* frequency of an amino acid in a coding sequence is its frequency in the translated protein. The *theoretical* frequency is the expected frequency calculated from the ratio of nucleotides. We introduce the discrepancy between these two frequencies as an index of *nonrandomness* of nucleotides in the sequence. This index of nonrandomness divides the species into two groups: eukaryotes having smaller nonrandomness (i.e., being more random) and prokaryotes having higher nonrandomness.

1 Overview

Various investigations have been made on statistics of nucleotide or amino acid sequences (see, e.g., (Guigó & Fickett, 1995)). Fundamental information of living organisms is encoded in their genomes. The portion of information extracted from the statistics is partial, but sometimes represents features of species or genes. The studies can be classified by the length of successive nucleotides they use: one, two, three or six.

The most basic statistics are the ratio of nucleotides in genomes or coding sequences. Among them, GC content is the best known and used.

Dinucleotide frequency, the frequency of the $4 \times 4 = 16$ possible successions of two nucleotides, also has been investigated (Hanai & Wada, 1990), (Josse *et al.*, 1961), (Jukes, 1978), (Russell *et al.*, 1976). This frequency is know to be different between species but rather consistent among the genes of a species (Campbell *et al.*, 1999).

Trinucleotide frequency, the frequency of the $4 \times 4 \times 4 = 64$ possible successions of three nucleotides, also has been investigated (Fickett, 1982), (Shepherd, 1981), (Tiwari *et al.*, 1997). However, since this high dimension of 64 makes analysis difficult, the data is usually transformed to lower dimensional data for study. We study two ways to project the data to lower dimensions in this paper.

The codon space (Rowe *et al.*, 1984) is the nucleotide frequencies for the three codon positions, which become $4 \times 3 = 12$ dimensional. Cluster analysis on this data showed that most primitive organisms have the highest A or T content in the second and third codon position (Rowe *et al.*, 1984). Cluster analysis on

the lower dimensional transformation of this data by multi-dimensional scaling showed that this statistical approach was able to classify 332 coding sequences into eukaryotes, prokaryotes, viruses and phages (Rowe, 1985).

The former part of our paper studies this codon space more thoroughly. By principal component analysis, a method to compress high dimensional data to lower dimensional, we provide a projection and its visualization of the codon space of 27 species to two dimensional space without loosing much information. Interestingly, this automatic process of principal component analysis gives classification of coding sequences into *Archaea*, *Bacteria* and *Eukaryota*. Furthermore, our analysis yields an explicit measurement dividing these three groups, which was not possible by the cluster analysis above. The relation between the codon space of the genes of one species and the codon space of the number of tRNA genes, the number of copies of tRNA genes in the genome of a species, is also studied.

Codon usage is another transformation of the 64 dimensional trinucleotide frequency, classifying the 64 codons into 21 subgroups according to the amino acids they code. Extensive studies on codon usage (Grantham, Gautier & Gouy, 1980), (Grantham, Gautier, Gouy, Jacobzone *et al.*, 1980), (Grantham, Gautier, Gouy, Mercier *et al.*, 1980), (Karlin *et al.*, 1998) and their relation to the abundance of expressed tRNA or the number of tRNA genes (Dong *et al.*, 1996), (Ikemura, 1981*a*), (Ikemura, 1981*b*), (Ikemura, 1982), (International Human Genome Sequencing Consortium, 2001), (Lowe, 2000) have been done. Studies on codon usage and studies on codon space both project the trinucleotide frequency to lower dimensional space, but

in different manners.

Trinucleotides in a coding sequence are translated to amino acids in a protein. In the latter part of this paper, we discuss two kinds of amino acid frequencies. The *real* frequency of an amino acid in a coding sequence is its frequency in the translated protein. The *theoretical* frequency (Jukes *et al.*, 1975), (King & Jukes, 1969) of an amino acid in a coding sequence is the expected frequency calculated from the ratio of nucleotides. Note that the real and theoretical amino acids together make $20+20 = 40$ dimensional data extracted from the 64 dimensional trinucleotide frequency.

This part also analyzes the randomness of nucleotide sequences measured by the difference between these two kinds of amino acid frequencies. Nucleotides in genome sequences appear to be ordered randomly at a first glance. However, the nucleotides must not be completely "random", because they are encoding a vast amount of information (information for development, body function, sex, etc.), and each coding sequence is encoding proteins with specific functions. Indeed, several statistical regularities, for example, repetition of periodicity three (see (Fickett, 1982)), are known.

Our purpose here is to compare this randomness of the nucleotides among species from a viewpoint of evolution. We show that eukaryotes have a smaller nonrandomness (i.e., are more random) compared to prokaryotes. However, among prokaryotes, the two subgroups *Archaea* and *Bacteria* seemed indistinguishable by this nonrandomness.

Finally, the 4^6 dimensional hexanucleotide frequency (or hexamer frequency),

whose translation becomes the diamino acid frequency, also has been studied. This quantity is powerful in discriminating coding sequences from non-coding sequences in genomes, and has been exploited in gene finding (Salzberg *et al.*, 1998).

2 Principal component analysis of codon space separates species

2.1 Means and standard deviations

For 27 species appearing in coding usage tabulated from the international DNA sequence database (Nakamura *et al.*, 2000) (**Table 1**) we computed the nucleotide ratio for each of the three codon positions, i.e., the codon space (Rowe *et al.*, 1984), for all or some portion of its coding sequences. These are for the nuclear genome, and not for mitochondrial or chloroplastic genomes. Plots for seven species among them are shown in **Fig. 1**. For codon space, it is known, for example, that guanine favors the first codon position (Shepherd, 1981). Guanine appears least in the third codon position, adenine the least in the third, and thymine the least in the first. This skewness can be explained partially by the absence of stop codons in coding sequences.

The mean values of the 12 dimensional codon space for these 27 species are

	T	C	A	G
first	0.174	0.199	0.291	0.336
second	0.290	0.223	0.314	0.173
third	0.246	0.272	0.220	0.262

,

and the standard deviations are

	T	C	A	G
first	0.031	0.056	0.060	0.044
second	0.028	0.036	0.048	0.027
third	0.097	0.116	0.098	0.078

.

Nucleotide ratios for the third nucleotide in the codon correlate with codon usage, and have largest standard deviations. Those for the first and second nucleotides in the codons correlate with amino acid frequencies in the coding sequences. Their standard deviations are less than for the third nucleotide, but not significantly smaller.

Though, among the seven species listed in Fig. 1, the nucleotide ratios for the second codon are similar in value (only with ± 0.014 difference from the means above) with small standard deviations $0.020 \sim 0.028$, this is not true for the all 27 species we examined, where the standard deviations are $0.027 \sim 0.048$.

These data indicate that there is a quite large inter-species variation in the nucleotide ratio in each position of the codons.

2.2 Principal component analysis

We performed principal component analysis for the codon space of the 27 species above. Principal component analysis is a tool to extract a few characteristic variables from a high dimensional distribution. First, the codon space of each of the 27 species is normalized, by subtracting the mean and dividing by the standard deviation of values in subsection 2.1. To represent the distribution of these normalized $3 \times 4 = 12$ (three positions in a code $\times$ four possible nucleotides) dimensional data for the 27 species, the direction (a 12 dimensional vector) giving the longest width of the distribution is selected as the *coefficient vector for the first principal component*. Then, the direction (another 12 dimensional vector) perpendicular to the previous one(s) giving the longest width of the distribution is selected as the *coefficient vector for the second principal component*. The repetition of this procedure automatically selects vectors representing the scatter of the distribution from major ones to minor ones. The coefficient vectors for the first and second principal components are given in the two tables below. The coefficients correspond to the ratio of four possible nucleotides in three positions in a codon.

Now, the inner product of the normalized codon space for a species with the coefficient vector for the first (or second) principal component becomes the *first (or second) principal component* of that species (**Fig. 2**). The faithfulness of the transformation became 0.798 for the first principal component, and 0.912 with the second principal component together.

The coefficient vector for the first principal component (horizontal axis) is

	T	C	A	G
first	−0.290	0.293	−0.311	0.256
second	−0.160	0.282	−0.288	0.300
third	−0.308	0.314	−0.316	0.311

The correspondence with the GC content is observed by the coefficients for G or C being approximately 0.3, and T or A being −0.3.

The coefficients vector for the second principal component (vertical axis) is

	T	C	A	G
first	0.282	0.283	−0.106	−0.410
second	−0.705	0.292	0.196	−0.007
third	0.155	−0.003	−0.058	−0.115

Interestingly, the 27 *species separated into the three evolutional groups Archaea, Bacteria, Eukaryota by this second principal component derived by the automatic process of principal component analysis* (see Fig. 2). The biological explanation of these coefficients of the second principal component is an interesting issue. We could observe a strong negative −0.891 correlation between the second principal component and the sum of frequencies of hydrophobic amino acids (F, L, I, M, V, A), though this sum of frequencies separates the species only to prokaryotes and eukaryotes but not to the three groups (**Fig. 3**). Prokaryotes having more hydrophobic amino acids than eukaryotes might reflect the environment such as the temperature during the evolution of these species.

Points representing codon space for some mitochondria for *Eukaryota* under similar normalization and linear transformation (their data obtained from

(Nakamura *et al.*, 2000)) had first principal components 0.87∼7.30, and second principal components −3.29∼0.58, and scattered in the bottom-right part of Fig. 2, the higher GC content parts of *Archaea* and *Bacteria*. Points for some chloroplasts had first principal components 1.08∼4.35, and second principal components −0.10∼0.95, scattering in the region of *Bacteria* to *Eukaryota*. (Data not shown.)

Similar division into three groups can be obtained by principal component analysis of the 20 dimensional amino acid frequencies, but this division becomes less sharp.

2.3 Coding sequences of each species in the codon space

In Fig. 2, we showed that the principal component analysis of the codon space classifies species into *Archaea*, *Bacteria* and *Eukaryota*. The data of codon space used for this analysis were means for many coding sequences (whole or part of nuclear genome). Here, we show that *each* coding sequence has similar value with the *mean*, by giving similar plots for all coding sequences of *Pyrococcus abyssi*, *Escherichia coli* and *Saccharomyces cerevisiae*. As in Fig. 2, for each coding sequence, the inner product of its normalized codon space with the coefficient vector for the first (or second) principal component is plotted as the first (or second) principal component (**Fig. 4**). *For each species, points corresponding to coding sequences are converging around their mean point in Fig. 2, somewhat separated from those of other species.* (The original 12 dimensional points of codon space are separated in this 2 dimensional projection. How-

ever, remark that this transformation is not a principal component analysis for the coding sequences of each species, thus the 2 dimensions are not necessarily describing the 12 dimensional scatter faithfully.)

We also compared genes highly preserved among species. The codon space of glyceraldehyde 3-phosphate dehydrogenase (GAPDH) genes for the seven species in Fig. 1 (with 37.4% amino acids homology) were transformed as in Fig. 2. Though GAPDH genes are reported to have translated horizontally across prokaryotes and eukaryotes, still the points corresponding to GAPDH genes are found around their corresponding points in Fig. 2 (**Fig. 5**). This again shows that principal component analysis of codon space discriminates genes of each species, indicating the power of our approach.

2.4 Codon spaces of the number of tRNA genes in nuclear genome and of coding sequences

Now, we consider the relation between trinucleotide frequency of two kinds. Other than the frequency summed over coding sequences, which we have discussed, we also consider the frequency corresponding to the *the number of tRNA genes*: the number of coded genes of tRNA in the nuclear genome summed according to their *anticodons*, which gives $64-3+1 = 62$ dimensional trinucleotide frequency. The $3 \times 4 = 12$ dimensional reduction of these two kinds of trinucleotide frequencies according to the nucleotide ratio in the three codon positions become the codon space of the number of tRNA genes or of the trinucleotides of coding sequences.

For *Caenorhabditis elegans*, correlation between the the number of tRNA genes and the trinucleotide frequency in the codon sequences, both from the whole nuclear genome, are reported (Lowe, 2000). For *Homo sapiens*, only *rough correlation* is observed between these two
(International Human Genome Sequencing Consortium, 2001).

Codon usage is the relative frequency of nucleotide triplets among those coding the same amino acid. It gives another way to interpret the 4^3 dimensional trinucleotide frequency. Codon usage can be defined for (1) the abundance of expressed tRNA, (2) trinucleotide frequency of coding sequences, or (3) the number of tRNA genes. The first two, namely, the codon usage of the abundance of expressed tRNA and the codon usage of the coding triplets of highly expressed genes are known to have correlation in several species (Dong *et al.*, 1996), (Ikemura, 1981*a*), (Ikemura, 1981*b*), (Ikemura, 1982). It is though that this abundance of tRNA relates the two factors in the previous paragraph, the number of tRNA genes and the trinucleotide frequency of coding sequences.

We tried to confirm these results from the viewpoint of codon space (not trinucleotides or codon usage) dealing with multiple species together. First, the plot of the codon space of the the number of tRNA genes in nuclear genome is given (**Fig. 6**).

To check if these codon spaces and the codon spaces for coding regions in Fig. 1 are similar, we computed the first and second principal components of these tRNAs according to the transformation used in Fig. 2 (**Fig. 7**).

In this figure, for each species, the point for the coding sequences (e.g.,

labeled "Stap") and the point for the tRNAs (e.g., labeled "Stap tRNA") are not far apart but not very close. *Thus, the codon space of coding sequences and the codon space of the number of tRNA genes is only roughly similar.* This dissimilarity indicates dissimilarity in the level of trinucleotides or codon usage. Comparing Figs. 1 and 6, we can see the third codon position is not agreeing. This disagreement was not explained even if the wobble (the simple Crick's wobble law assuming that anticodon A changes to inosine and pairs U, C, A whenever possible, and that G and U can pair) was taken into consideration: for each of the 64 codons, we took the sum of the number of tRNA genes which can pair according to the wobble above, computed the 12 dimensional codon space from this 64 dimensional data, and plotted as in Fig. 7, but these points came no less closer to their encoding sequence counterparts.

3 Randomness of nucleotides increases with evolution

In this section, we analyze *randomness* of trinucleotides and its relation to evolution. The measurement of randomness we take is by the difference of two kinds of amino acid frequencies. As mentioned in section 1, the *real* frequency of an amino acid in a coding sequence is its frequency in the translated protein. The *theoretical* frequency (Jukes *et al.*, 1975) (King & Jukes, 1969) (see appendix A.2) of an amino acid in a coding sequence is the expected frequency calculated from the ratio of nucleotides. First, we show the real and theoretical frequencies

of three species from *Archaea*, *Bacteria* and *Eukaryota* **(Fig. 8)**.

If the nucleotides were random in the coding sequences, the real and theoretical frequency should coincide for each amino acid, and the points representing amino acids in the plot should be on the diagonal line. The *nonrandomness*, or the discrepancy between real and theoretical frequencies, appears as the *divergence from the diagonal.* The *randomness*, or the coincidence between real and theoretical frequencies, appears as the *concentration around the diagonal.* It can be observed that randomness appears increasing in the order *Pyrococcus abyssi*, *Escherichia coli*, *Homo sapiens* in Fig. 8.

We have analyzed the real and theoretical frequencies for 27 species, and computed an index of nonrandomness (see appendix A.3) as the divergence from the diagonal **(Table 1)**. A plot of this nonrandomness and the GC content for those species is given **(Fig. 9)**.

It can be observed that eukaryotes have smaller nonrandomness (i.e., are more random), compared to prokaryotes. Indeed, the hypothesis that the two groups have the same mean value of nonrandomness would be rejected, because Welch's approximate t-test gives p-value of 3.16×10^{-6}. However, among prokaryotes, the two subgroups *Archaea* and *Bacteria* appeared indistinguishable. The hypothesis that these two subgroups have the same mean value of nonrandomness cannot be rejected, because Welch's approximate t-test gives p-value of 0.444. It can also be seen that GC content is not affecting nonrandomness. We have also checked that the discrepancy between real and theoretical frequencies does not occur in regions not coding protein, such as rRNAs or

introns.

Our conclusion is that the randomness of nucleotides in coding sequences, measured by the coincidence between real and theoretical amino acid frequencies, is high in eukaryotes while it is low in prokaryotes.

In a sense, this is contradictory to our intuition. Because, this increase of randomness, or loss of information, is happening as the living organism is becoming more complex. However, one explanation is that higher organisms have accumulated random mutations. Another matter to consider is the environment. If the environment such as temperature is severe, there might be stricter restriction on amino acid frequency, leading to nonrandomness. However, complex organisms with many gene repertoire can move on to better conditions or keep their own optimum inner environment, and thus might have less of this restriction and more freedom enabling randomness. Analysis of evolutional time based on this model of explanation might be of interest.

4 Discussions

The most basic data of coding sequences must be the one dimensional GC content and the four (or three) dimensional ratio of nucleotides. However, they are not necessarily fine enough to classify different species. Much finer data can be obtained by the 64 dimensional trinucleotide frequencies or their reduction to the 12 dimensional codon space. However, these are still of high dimension making the visualization and analysis difficult. In this paper, we analyzed their

reduction to dimension two: the principal component analysis of the codon space. This measurement is faithful enough to discriminate genes coming from different species. The first principal component was indicating GC content, and the second divided species into *Archaea*, *Bacteria* and *Eukaryota*. The second measure we analyzed was the difference between real and theoretical amino acid frequencies. This also gave distinction of species into two groups prokaryotes and eukaryotes. These measurements are novel and interesting in that they give statistical indices of evolution possibly for the whole collection of living species. Applying principal component analysis to other sequence statistics must also be of interest. Our two measurements, in a sense, indicate the deviation of coding sequences from randomly ordered sequences. The investigation of the evolutionary mechanism causing these statistics is also an important topic.

A Appendices

A.1 Codon space

The fact that nucleotide ratios of the three codon positions differ in *coding sequences* has been known and applied to gene finding (Tiwari *et al.*, 1997). However, the reason or mechanism of this phenomena is not resolved. This disagreement does not happen in regions not coding proteins, such as ribosomal RNAs or introns. Even in coding regions, if the nucleotides positions are divided not by the multiple of three, the nucleotide ratios coincide among different positions.

A.2 Theoretical frequency

The *theoretical* frequency was first discussed by Jukes et al. (Jukes *et al.*, 1975) (King & Jukes, 1969). They analyzed the real and theoretical frequencies for a concatenation of 53 coding sequences assembled from several species. We analyzed the frequencies for each individual species and compared them from a viewpoint of evolution. Our analysis exploits the huge amount of genome data available today.

Our calculations are based on the Kazusa database (Nakamura *et al.*, 2000). We selected 27 major species with enough amount of genome data (Table 1). The nucleotide ratio for each species, used to calculate the theoretical frequencies, is the ratio in the whole collection of codons of the available coding sequences. In other words, our calculations are not averaged over coding sequences, but over the whole collection of codons. Though, the calculation in the other way yields similar results.

The expected frequency of amino acids computed from the codon space (instead of nucleotide ratio) gives an alternative definition of "theoretical" amino acid frequency. This one agrees better with the real amino acid frequency, but still some significant discrepancy remains.

A.3 Index of nonrandomness

For the index of nonrandomness, we calculated the difference of common logarithm of real and theoretical frequencies for each amino acid, and took the

square root of sum of squares of this difference over all amino acids:

$$\sqrt{(\log_{10} r_{\mathrm{A}}/t_{\mathrm{A}})^2 + (\log_{10} r_{\mathrm{C}}/t_{\mathrm{C}})^2 + \cdots + (\log_{10} r_{\mathrm{Y}}/t_{\mathrm{Y}})^2},$$

where $r_{\mathrm{A}}, r_{\mathrm{C}}, \ldots, r_{\mathrm{Y}}$ and $t_{\mathrm{A}}, t_{\mathrm{C}}, \ldots, t_{\mathrm{Y}}$ are real and theoretical frequencies of amino acids $\mathrm{A}, \mathrm{C}, \ldots, \mathrm{Y}$. Larger amount of this value means large divergence from the diagonal line, thus the nonrandomness of the nucleotides. For each species, the real and theoretical amino acid frequencies for its coding sequences is concentrating enough around their mean, and the number of coding sequences used (indicated in Table 1) is not affecting the index of nonrandomness.

Acknowledgments

We thank Wayne Dawson for discussions and comments. K. Y. was partly supported by the grants of the ministry of health, labour, and welfare (H14-nano-004 and H13-SD-01).

References

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

Dong, H, Nilsson, L, & Kurland, C.-G. (1996) Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. *J. Mol. Biol.* **260**, 649–663.

Fickett, J.-W. (1982) Recognition of protein coding regions in DNA sequences *Nucleic Acids Research* **10**, 5303–5318.

Grantham, R., Gautier, C., & Gouy, M. (1980) Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. *Nucleic Acids Res.* **8**, 1893–1912.

Grantham, R., Gautier, C., Gouy, M., Jacobzone M., & Mercier, R. (1980) Codon catalog usage as a genome strategy modulated for gene expressivity. *Nucleic Acids Res.* **9**, r43–r74.

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

Guigó, R, & Fickett, J.-W. (1995) Distinctive sequence features in protein coding genic non-coding, and intergenic human DNA. *J. Mol. Biol.* **253**, 51–60.

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

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.

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.

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.

International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.

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

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

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

Karlin, S, Mrázek, J, & Campbell, A.-M. (1998) Codon usages in different gene classes of the *Escherichia coli* genome, *Molecular Microbiology* **29**, 1341–1355.

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

Lowe, T.-M. (2000) Combining New Computational and Traditional Experimental Methods to Identify tRNA and

snoRNA Gene Families. Ph.D thesis, Washington University. `http://www-genome.stanford.edu/~lowe/thesis/node1.html`

Nakamura, Y., Gojobori, T., & Ikemura, T. (2000) Codon usage tabulated from the international DNA sequence databases: status for the year 2000. *Nucl. Acids Res.* **28**, 292 `http://www.kazusa.or.jp/codon/` (Source: GenBank Release 123.0 [15 April 2001]).

Rowe, G.-W. (1985) A three-dimensional representation for base composition of protein-coding DNA sequences. *J. theor. Biol.* **112**, 433–444.

Rowe, G.-W., Szabo, V.-L., & Trainor, L.-E.-H. (1984) Cluster analysis of genes in codon space. *J. Mol. Evol.* **20**, 167–174.

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.

Salzberg, S., Delcher, A., Fasman, K., & Henderson, J. (1998) A Decision Tree System for Finding Genes in DNA. *Journal of Computational Biology* **5**, 667-680.

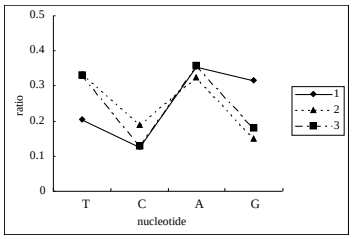
Shepherd, J.-C.-W. (1981) Periodic correlations in DNA sequences and evidence suggesting their evolutionary origin in a commaless genetic code. *J. Mol. Evol.* **17**, 94–102.

Tiwari, S., Ramachandran, S., Bhattacharya, A., Bhattacharya, S., & Ramaswamy, R. (1997) Prediction of probable genes by Fourier analysis of ge-

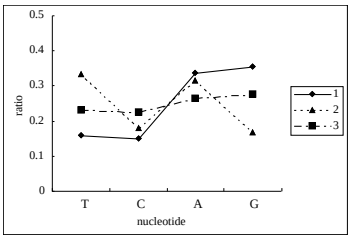
nomic sequences. *Comput. Appl. Biosci.* **13**, 263–270.

species	taxonomy	abbre-viation	nonran-domness	# of CDS
Aeropyrum pernix	Archaea; Crenarchaeota; Desulfurococcales	Aero	0.835	2702
Sulfolobus solfataricus	Archaea; Crenarchaeota; Sulfolobales	Sulf	0.889	591
Archaeoglobus fulgidus	Archaea; Euryarchaeota; Archaeoglobales	Arch	0.901	2418
Halobacterium sp. NRC-1	Archaea; Euryarchaeota; Halobacteriales	Halo	1.257	2605
Methanococcus jannaschii	Archaea; Euryarchaeota; Methanococcales	Meth	0.901	1717
Pyrococcus abyssi	Archaea; Euryarchaeota; Thermococcales	Pyra	1.017	1769
Pyrococcus horikoshii	Archaea; Euryarchaeota; Thermococcales	Pyrh	0.941	2059
Thermoplasma acidophilum	Archaea; Euryarchaeota; Thermoplasmales	Ther	1.007	1517
Streptomyces coelicolor A3	Bacteria; Firmicutes; Actinobacteria	Stre	1.295	6377
Bacillus subtilis	Bacteria; Firmicutes; Bacillus/Clostridium group	Baci	0.850	9005
Staphylococcus aureus	Bacteria; Firmicutes; Bacillus/Clostridium group	Stap	0.934	955
Pseudomonas aeruginosa	Bacteria; Proteobacteria; gamma subdivision	Pseu	1.205	6690
Escherichia coli K12	Bacteria; Proteobacteria; gamma subdivision	Ecol	0.793	10136
Helicobacter pylori J99	Bacteria; Proteobacteria; epsilon subdivision	Heli	0.821	1491
Leishmania major	Eukaryota; Euglenozoa; Kinetoplastida	Leis	0.830	1323
Plasmodium falciparum	Eukaryota; Alveolata; Apicomplexa	Plas	0.800	1042
Saccharomyces cerevisiae	Eukaryota; Fungi; Ascomycota	Sacc	0.636	11644
Schizosaccharomyces pombe (fission yeast)	Eukaryota; Fungi; Ascomycota	Schi	0.669	5737
Zea mays (corn)	Eukaryota; Viridiplantae; Streptophyta	Zea	0.770	988
Oryza sativa (rice)	Eukaryota; Viridiplantae; Streptophyta	Oryz	0.667	3940
Arabidopsis thaliana (thale cress)	Eukaryota; Viridiplantae; Streptophyta	Arab	0.571	33661
Caenorhabditis elegans	Eukaryota; Metazoa; Nematoda	Cele	0.550	19947
Drosophila melanogaster (fruit fly)	Eukaryota; Metazoa; Arthropoda	Dros	0.725	19565
Danio rerio (zebrafish)	Eukaryota; Metazoa; Chordata	Dani	0.604	863
Rattus norvegicus (Norway rat)	Eukaryota; Metazoa; Chordata	Rat	0.665	5936
Mus musculus (house mouse)	Eukaryota; Metazoa; Chordata	Mus	0.622	11484
Homo sapiens (human)	Eukaryota; Metazoa; Chordata	Homo	0.621	27143

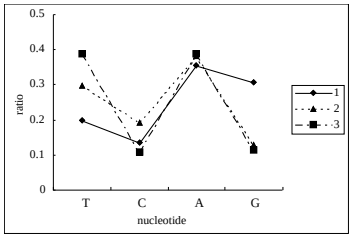
Table 1: The name of the species, its taxonomy, the abbreviation in Figs. 2, 3, 5 7, 9, its nonrandomness, and the number of coding sequences (CDS) the calculation is based on are listed for 27 species from (Nakamura *et al.*, 2000).



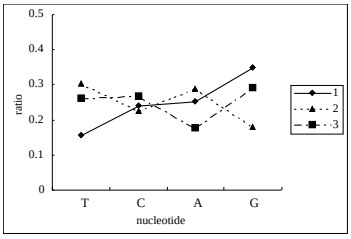
(a) *Sulfolobus solfataricus*



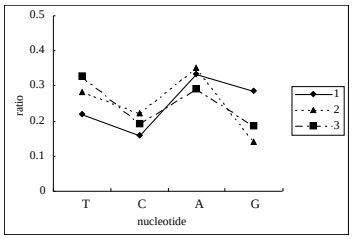
(b) *Pyrococcus abyssi*



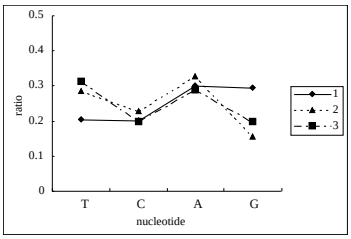
(c) *Staphylococcus aureus*



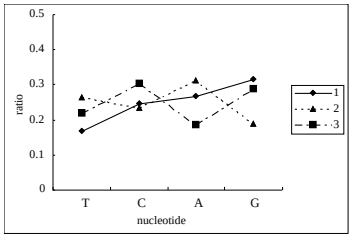
(d) *Escherichia coli* K12



(e) *Saccharomyces cerevisiae*



(f) *Caenorhabditis elegans*



(g) *Homo sapiens*

Figure 1: Codon space, the nucleotide frequencies for each of the three codon positions. Computations based on (Nakamura *et al.*, 2000).

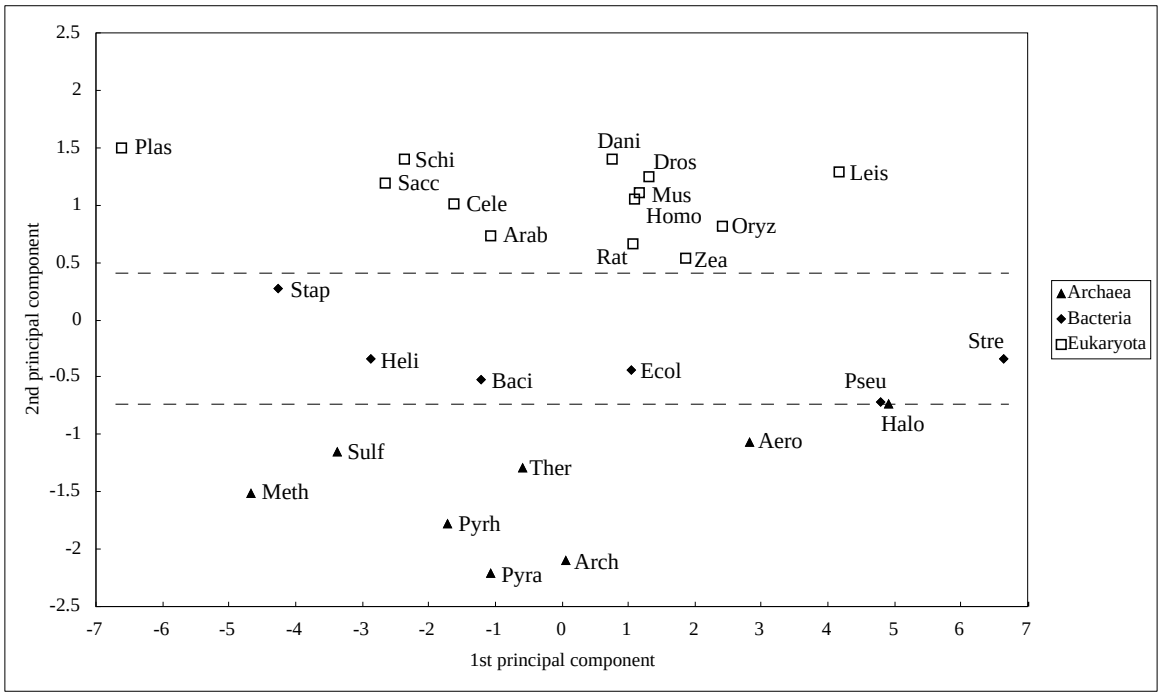


Figure 2: Plot of the first and second principal component for the codon space.

For the 27 species in Table 1 from (Nakamura *et al.*, 2000).

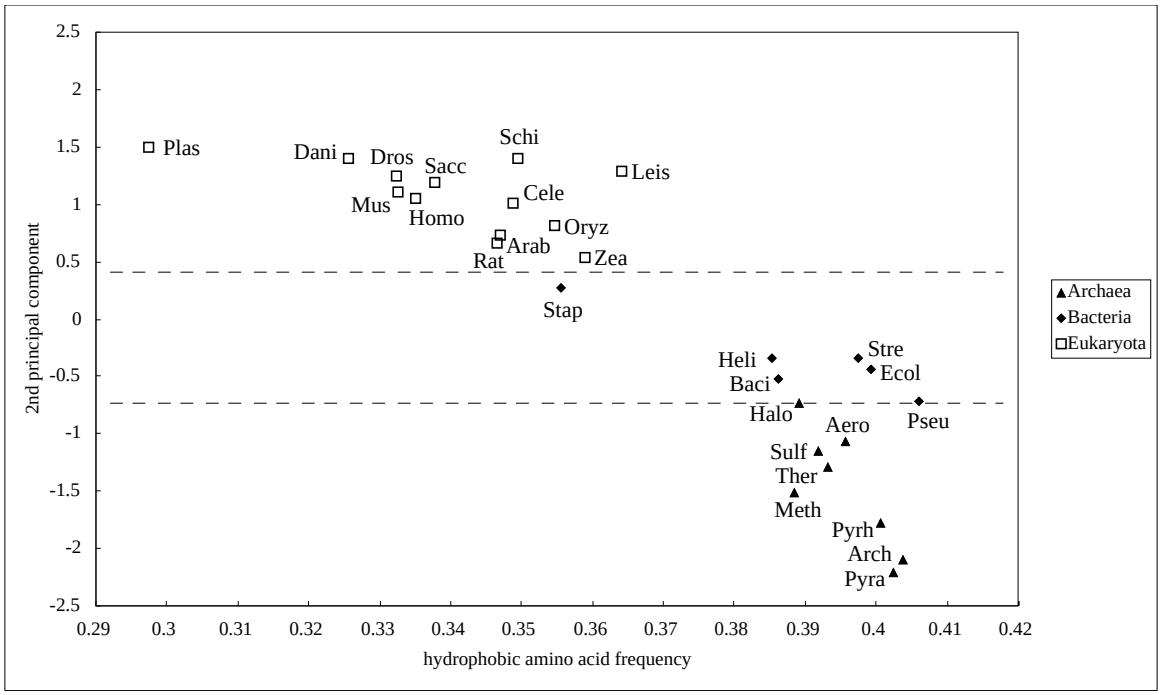
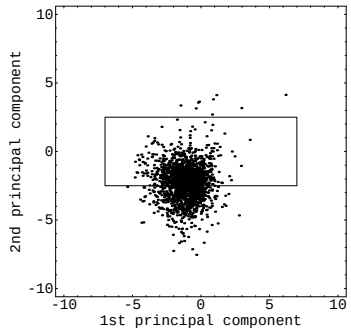
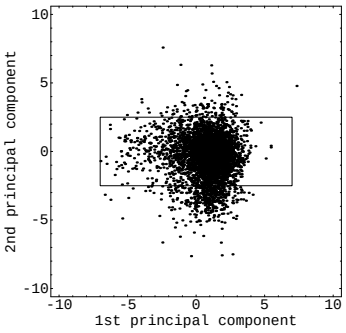


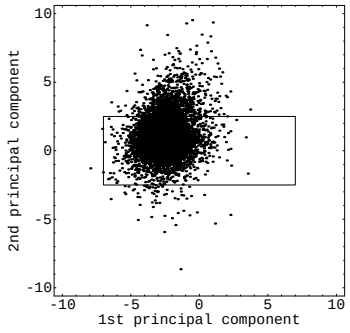
Figure 3: Plot of the sum of frequencies of hydrophobic amino acids and the second principal component of codon space as in Fig. 2. For the 27 species in Table 1 from (Nakamura *et al.*, 2000).



(a) The whole 1756 CDSs from the nuclear genome of *Pyrococcus abyssi*.



(b) The whole 4288 CDSs from the nuclear genome of *Escherichia coli* K12.



(c) The whole 6267 CDSs from the nuclear genome of *Saccharomyces cerevisiae*.

Figure 4: Codon space of each coding sequence transformed as in Fig. 2. The plot range of Fig. 2 is shown as a rectangle.

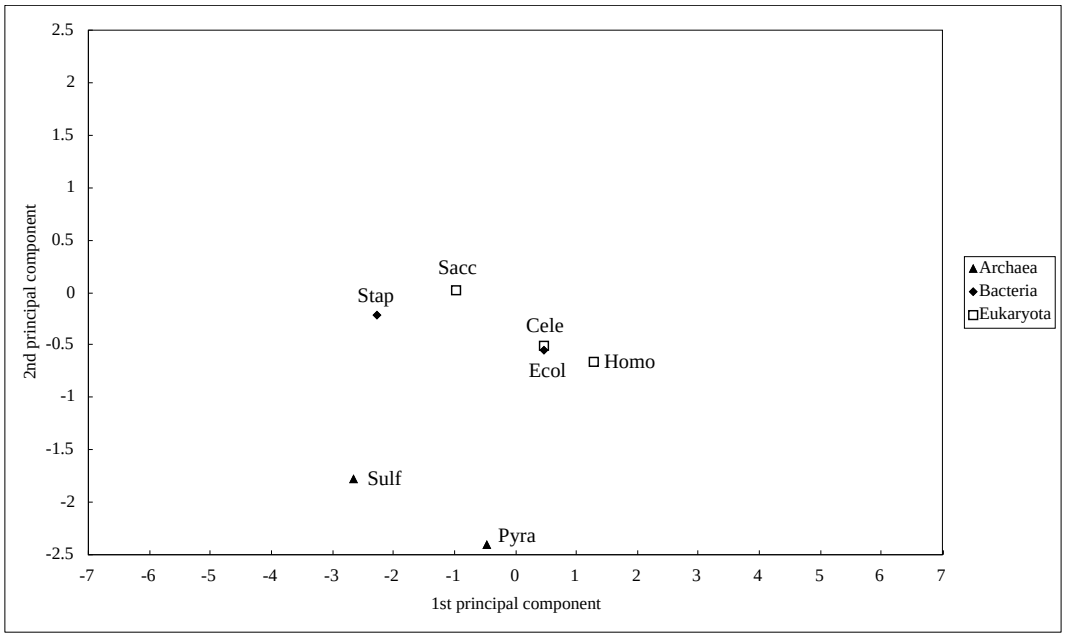
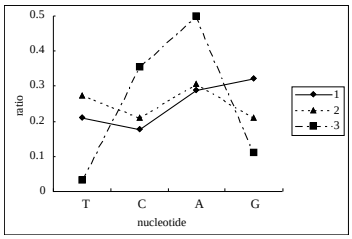
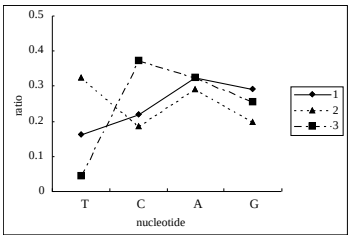


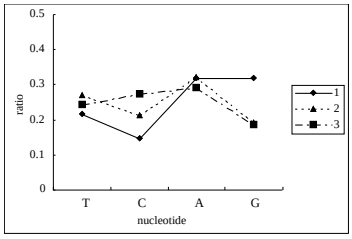
Figure 5: Codon spaces of glyceraldehyde 3-phosphate dehydrogenase (GAPDH) genes for the seven species in Fig. 1 transformed as in Fig. 2.



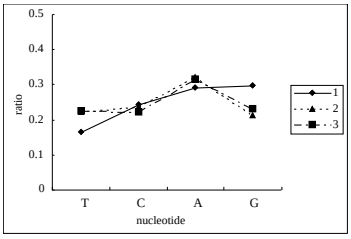
(a) The whole 62 tRNA from the nuclear genome of *Staphylococcus aureus*.



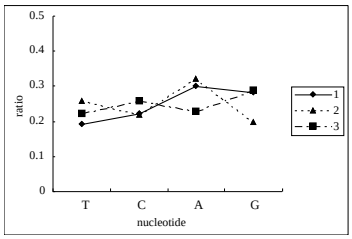
(b) The whole 86 tRNA from the nuclear genome of *Escherichia coli* K12.



(c) The whole 273 tRNA from the nuclear genome of *Saccharomyces cerevisiae*.



(d) The whole 584 tRNA from the nuclear genome of *Caenorhabditis elegans* (Lowe, 2000).



(e) The whole 496 tRNA from the nuclear genome of *Homo sapiens* (International Human Genome Sequencing Consortium, 2001).

Figure 6: Codon spaces computed from the number of tRNA genes in the nuclear genome.

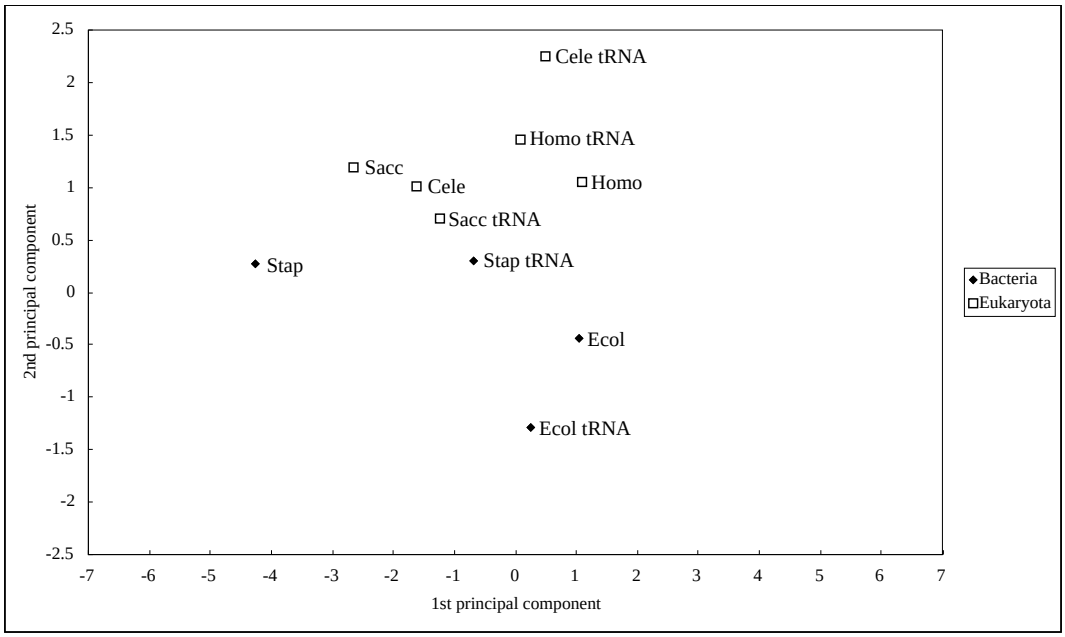
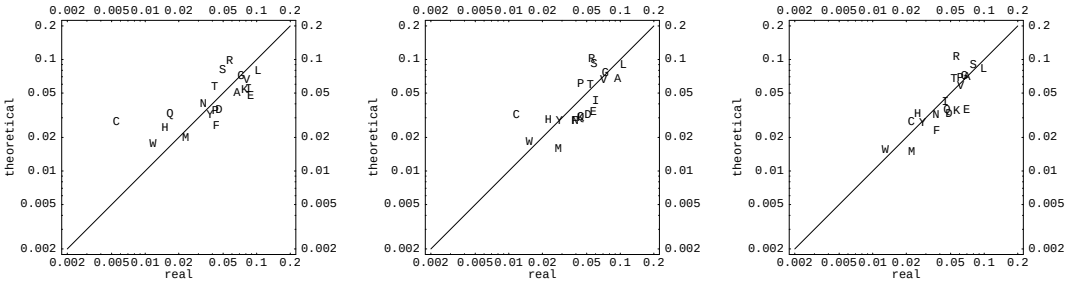


Figure 7: Plot of the first and second principal components of the codon space of the number of tRNA genes in nuclear genomes computed as in Fig. 2. The points corresponding to the coding sequences of these 5 species from Fig. 2 are also shown for reference.



(a) *Pyrococcus abyssi* (b) *Escherichia coli* (c) *Homo sapiens*

Figure 8: The 20 amino acids are plotted with real frequency as the horizontal coordinate and the theoretical frequency as the vertical coordinate for *Pyrococcus abyssi*, *Escherichia coli* and *Homo sapiens*.

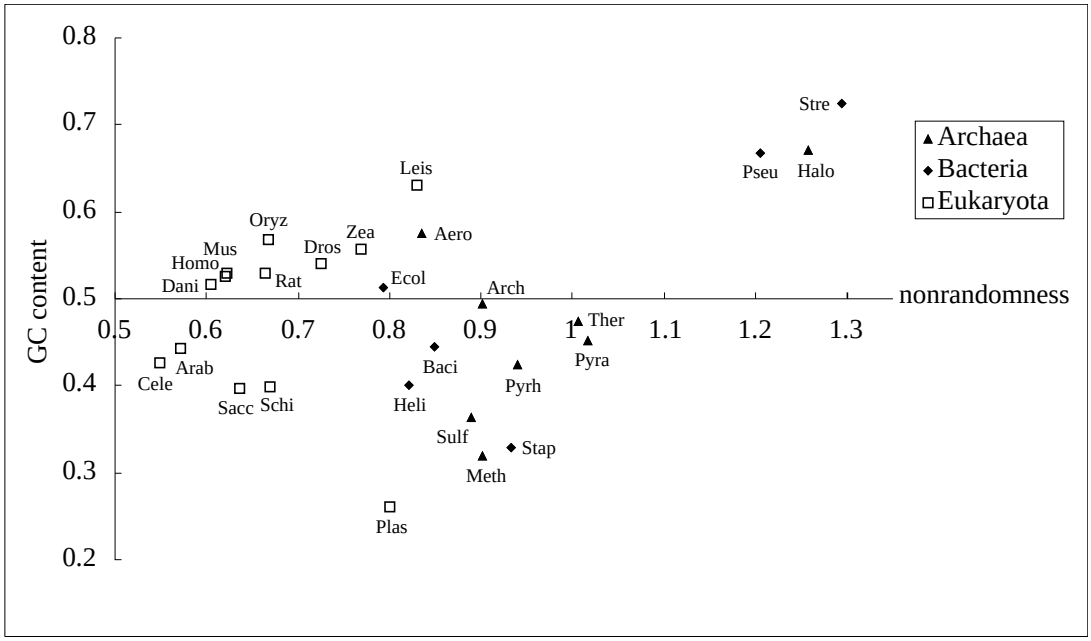


Figure 9: The 27 species from *Archaea*, *Bacteria* and *Eukaryota* are plotted with nonrandomness as the horizontal coordinate and GC content as the vertical coordinate. See Table 1 for the abbreviations of the names.