Vol. 15, pp. 75–85 (2018)
doi: 10.2142/biophysico.15.0_75

Biophysics and Physicobiology
https://www.jstage.jst.go.jp/browse/biophysico/

Hypothesis and Perspective

Biological meaning of “habitable zone” in nucleotide composition space

Shigeki Mitaku[1] and Ryusuke Sawada[2]

[1]*Emeritus Professor of Nagoya University, Kokubunji, Tokyo 185-0021, Japan*
[2]*Division of System Cohort, Medical Institute of Bioregulation, Kyushu University, Fukuoka 812-8582, Japan*

Received September 20, 2017; accepted March 17, 2018

Organisms generally display two contrasting properties: large biodiversity and a uniform state of “life”. In this study, we focused on the question of how genome sequences describe “life” where a large number of biomolecules are harmonized. We analyzed the whole genome sequence of 2664 organisms, paying attention to the nucleotide composition which is an intensive parameter from the genome sequence. The results showed that all organisms were plotted in narrow regions of the nucleotide composition space of the first and second letters of the codon. Since all genome sequences overlap irrespective of the living environment, it can be called a “habitable zone”. The habitable zone deviates by 500 times the standard deviation from the nucleotide composition expected from the random sequence, indicating that unexpectedly rare sequences are realized. Furthermore, we found that the habitable zones at the first and second letters of the codon serve as the background mechanisms for the functional network of biological systems. The habitable zone at the second letter of the codon controls the formation of transmembrane regions and the habitable zone at the first letter controls the formation of molecular recognition unit. These analyses showed that the habitable zone of the nucleotide composition space and the exquisite arrangement of amino acids in the codon table are conjugated to form biological systems. Finally, we discussed the evolution of the higher order of genome sequences.

Key words: genome, nucleotide composition, membrane protein, molecular recognition, habitable zone

In general, organisms exhibit two contrasting properties. On the one hand, organisms show very large diversity, and on the other hand all organisms have the common state of matter, “life”. The biological diversity of living organisms is easy to understand intuitively because the genome sequence is always diversified by random mutations. In fact, it is possible to identify species by examining, for example, difference in codon usages or GC content in genome sequences [1,2]. In contrast, it is not easy to answer the question of how different genome sequences can realize a common state of “life”. An effective approach to this type of problem is to extract common features by comparing the genome sequences of various organisms. For example, ribosomes are molecular devices present in all organisms, and ribosomal RNA sequences are the indication of the existence of “life”. However, what we are trying to address here is whether the state of “life”, ie the harmonization of all genes, can be expressed based on a small number of parameters extracted from the

Corresponding author: Shigeki Mitaku, Kokubunji, Minamicho 3-21-1-1108, Tokyo 185-0021, Japan.
e-mail: mitakus@nifty.com

◀ ***Significance*** ▶

We studied on the question of how genome sequences describe “life” where a large number of biomolecules are harmonized. We found that all organisms are plotted in a narrow region called habitable zone of the nucleotide composition space from the genome sequence. We discussed the biological significance of the habitable zone from three aspects: the great reduction of the number of possible sequences, the relationship between the second letter and the distribution of membrane proteins, and the relationship between the first letter and molecular recognition units. Hypotheses for the mechanism of the habitable zone formation were also discussed.

whole genome sequence. Total genome sequences of thousands of organisms are now available [3]. Therefore, there is a possibility to solve this problem by using actual sequence data.

The question is how we can reduce the information from the genome sequence to some parameters describing the state "life". This problem setting may seem nonsense in the biological research area, but in a wider scientific field, we have already encountered this type of problem setting. In physics, the state of a material consisting of molecules of Avogadro's number can be described with a small number of thermodynamic parameters. For example, the three states of matter can be specified by two parameters, temperature and pressure. The important point here is that the thermodynamic parameters describing the state are the intensive parameters. If the analogy is not merely superficial similarity, the possible parameters specifying the state of "life" will be the intensive parameters from the genome sequence. And the simplest intensive parameter obtained from the genome sequence is the nucleotide composition, ie the occurrence probability of nucleotides at mutations.

Before conducting research using nucleotide compositions, we must first answer the question whether the coding region or noncoding region in the genome sequence is appropriate for the analysis of the state of "life". It has already been reported how the ratio of coding and noncoding regions correlate with the complexity of organisms from prokaryotes to humans [4,5]. The complexity of the organism had a good correlation with the size of the noncoding region. For example, the noncoding region of prokaryotes is about 10% of the entire genome, whereas the proportion of the noncoding regions in the human genome is as large as 98%. Conversely, the correlation between the size of the coding region and the complexity of the organism is small. For example, the number of genes in the whole genome is similar between nematodes and humans. Therefore, it is reasonable to assume that the information describing the state of "life" is hidden in the coding region where the information of the protein which is the element of the biological system is written. In fact, in previous papers we could show that the biasing of the nucleotide composition at the second letter of the codons determines the proportion of membrane proteins in the genome [6,7]. And we proposed that the proportion of membrane proteins is one of the parameters characterizing the state of "life" [7].

In this work, we analyzed the whole genome sequences of 2664 organisms and found that all organisms are plotted in narrow regions of the nucleotide composition spaces of the first letter and the second letter of the codon. Since the whole genome sequences of prokaryotes (2551), including thermophilic bacteria (173), and eukaryotes (113) almost overlap in the narrow regions, we call them "habitable zones". The habitable zone is a kind of "life and death phase diagram" in the meaning that the organism cannot survive outside the habitable zone. We have studied the biological meaning of the habitable zone from two aspects. Large bias from the random sequence will result in a great reduction in the number of possible sequences. The first problem is to quantitatively evaluate how rare phenomenon the nucleotide composition is in the habitable zone (section 1). The features of the habitable zone at each position of the codon are different from each other and it raises the second question what kind of features of the protein are related to each position of the codon. Section 2 discusses the relationship between the nucleotide composition at the second letter of the codon and the proportion of membrane protein [7]. Section 3 explains the relationship between the nucleotide composition at the first letter of the codon and the molecular recognition unit of the protein. The habitable zone plays an important role in the formation of the biological system by conjugating with the exquisite arrangement of amino acids in the codon table. Therefore, finally, we discussed the relationship between the habitable zone clarified in our research and the codon formation [8,9] in the primordial environment. We also discussed hypotheses about the formation mechanism of the habitable zone.

1. Statistical Rareness of "Habitability zone"

Many previous studies on biased nucleotide composition seem to focus on the biodiversity [1,2,10–13]. In particular, the nucleotide composition at the third letter of the codon is widely distributed depending on the species of the organism, and it is possible to identify organisms mainly based on the information of the third letter of the codon. Conversely, the nucleotide compositions at the 1st and 2nd letters of the codon are in a relatively limited areas and are unlikely to be useful for the discussion of biodiversity. However, this controlled nucleotide compositions can conversely be regarded as a characteristic of the state of "life" that is common to all living organisms.

In Figures 1 to 3, we plotted the genome sequences of 2664 organisms in 4-dimensional spaces with the composition of 4 kinds of nucleotides as the axis. In order to represent a four-dimensional space, generally six orthogonal planes are required. Thus, Figures 1 to 3 show six two-dimensional graphs corresponding to the first to third letters of the codon, respectively. For convenience of explanation, we call the point (0.25, 0.25) of each two-dimensional graph as the "origin" because the probability of occurrence of each nucleotide expected from perfect random mutations is 0.25. We denote the composition of each nucleotide by bracket < >.

The features of the plots in Figures 1 to 3 can be summarized as follows.

(1) Figure 1 shows that 2664 organisms are plotted in a narrow nucleotide composition region in the first letter of the codon. In the graph of <G> vs. <T> (Fig. 1C), almost all organisms are plotted in the second quadrant. That is, for almost all organisms, <G> is greater than

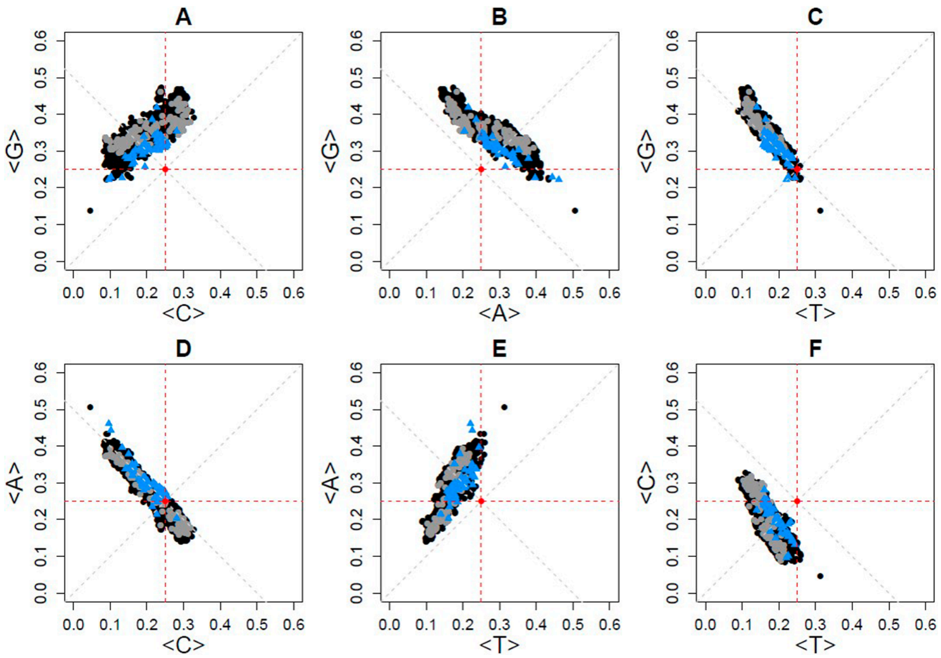


Figure 1 Plots of genomes into the nucleotide composition space at the first letter of the codon. We call the plotted area as the "habitable zone". Three types of organisms are represented by different marks: 2378 prokaryotes (black circles), 173 thermophiles containing hyper-thermophilic bacteria (gray circles) and 113 eukaryotes (blue triangles). The list of organisms is shown in the Supplementary Table 1. Since the nucleotide composition space is a four-dimensional space, six graphs are required in order to project on the two-dimensional planes, as shown in A to F. Bracket < > represents the composition of nucleotides in the entire coding region. The red dashed lines show the average value 0.25 of the nucleotide composition for perfectly random mutations. As is clear from the graph of <G> vs. <T>, <G> is greater than 0.25 and <T> is less than 0.25 for almost all organisms. Furthermore, for all creatures, <G> + <A> is greater than 0.5 and <T> + <C> is less than 0.5, as shown in the graphs of <G> vs. <A> and <T> vs. <C>, respectively. The habitable zone at the first letter of the codon is biased from the origin by 0.05 or more, as shown in Figures A, B and F.

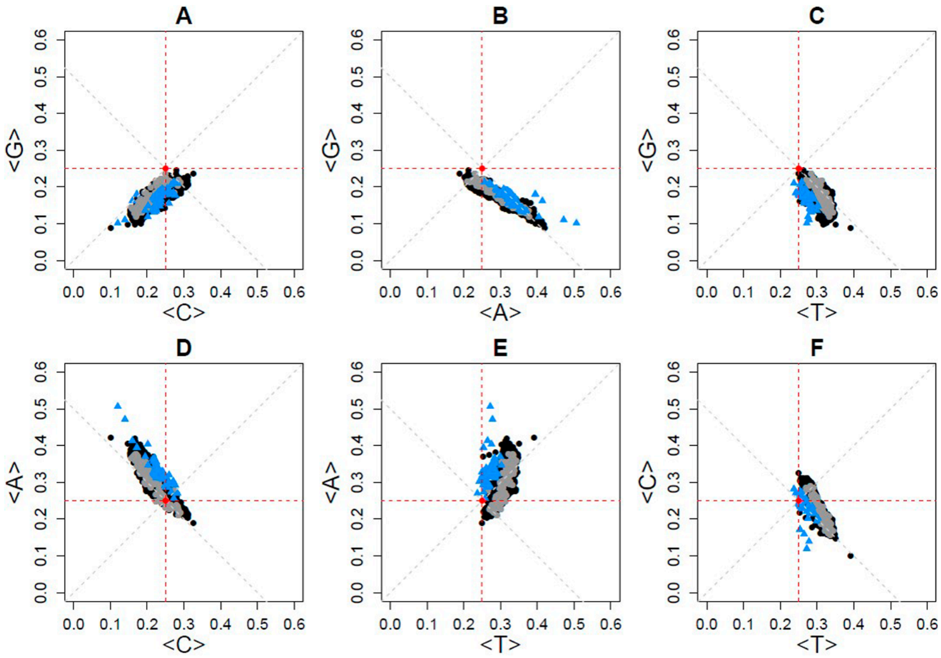


Figure 2 Plots of genomes into the nucleotide composition space at the second letter of the codon. The habitable zone for the second letter is even smaller than that of the first letter. The data set of biological genomes is the same as that in Figure 1. As is clear from the graph of <G> vs. <T>, <T> is greater than 0.25 and <G> is less than 0.25 for almost all organisms. Furthermore, for most organisms, <T> + <A> is greater than 0.5, as shown in the graph of <T> vs. <A>.

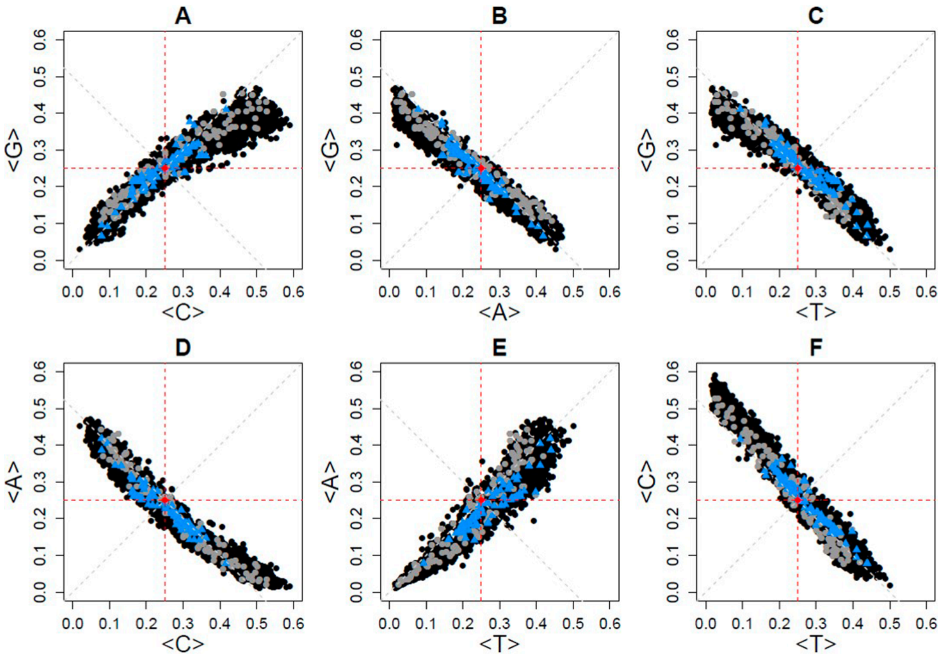


Figure 3 Plots of genomes into the nucleotide composition space at the third letter of the codon. The nucleotide composition is very widely dispersed from almost 0 to 0.5 or more. This large scattering cannot be explained by simple random mutations. So, we also call the plot of the third letter as the “habitable zone”.

0.25 and <T> is less than 0.25. Furthermore, for all creatures, <G> + <A> is greater than 0.5 and <T> + <C> is less than 0.5, as shown in Figures 1B and 1F, respectively. The habitability area in the first letter of the codon is 0.05 or more away from the origin in Figures 1A, 1B and 1F.

(2) The scatter diagrams in Figure 2 show that the habitable zone for the second letter of the codon is even narrower than that of the first letter. In other words, the mutations at the second letter must be very severely controlled. In the graph of <G> vs. <T> (Fig. 2C), almost all organisms are plotted in the fourth quadrant. That is, for almost all organisms, <T> is greater than 0.25 and <G> is less than 0.25. Furthermore, for most organisms, <T> + <A> is greater than 0.5, as shown in Figure 2E.

(3) Figure 3 shows scatter diagrams of the nucleotide composition at the third letter of the codon. The biggest feature of the graphs of the third letter is that the nucleotide composition is very widely dispersed from almost 0 to 0.5 or more. We also call the plot of the third letter as the “habitable zone” because the large scattering may be essential for various organisms to survive.

(4) Eukaryotes are plotted in narrower regions of the nucleotide composition spaces than prokaryotes. Interestingly however, plots of eukaryotes almost overlap with plots of prokaryotes.

(5) Thermophilic and hyper-thermophilic bacteria that are examples of extreme environmental organisms overlap with ordinary environmental organisms in all graphs.

When prokaryotes and eukaryotes, or extreme environmental organisms and normal environment organisms are plotted in the nucleotide composition space, they overlap each other. That is, the habitable zone is common to all living things. Therefore, the habitable zone in the nucleotide composition space can give important information on the state of “life” common to all organisms. So, we discuss the habitable zone from two aspects: the statistical constraints on the whole genome sequence, and the biological significance of the total proteins from the genome. In this section, we will examine the statistical significance of the habitable zone and discuss the biological significance of the habitable zone in Sections 2 and 3.

Figure 1A, 1B and 1F show that all the organisms are plotted in regions greatly biased from the origin (0.25, 0.25). The magnitude of bias is 0.05 or more. This indicates that the four-dimensional space of the nucleotide composition at the first letter of the codon is largely biased from the origin. Furthermore, the bias in the 12-dimensional space of the entire codon between the actual genome sequence and the origin can be evaluated by the following equations, (1) to (3).

$$d(\alpha) = \sqrt{\sum_{\beta}\{x(\alpha;\beta) - 0.25\}^2} \quad (1)$$

Here, α means 1, 2 or 3 which is the letter position of the codon, and β represents the nucleotide A, T, G or C. So, the parameter $x(\alpha;\beta)$ represents the composition of the nucleotide β at the α position of the codon. Therefore, $d(\alpha)$ indicates the

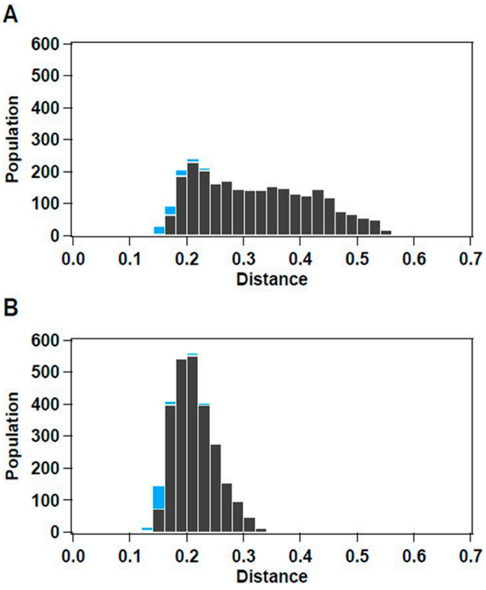


Figure 4 The nucleotide composition distance between the actual genome and the random sequence is shown in the histogram. Among 2664 species, 2551 prokaryotes are shown in black and 113 eukaryotes are shown in blue. A: Histogram of the composition distance for the three letters of the codon calculated by Eq. (2). B: Histogram of the composition distance for the first and second letters of the codon calculated by Eq. (3).

distance between the genome sequence and the origin at the α position of the codon. Then, the distance D of the bias for all positions of the codon is represented by the following equation.

$$D = \sqrt{d(1)^2 + d(2)^2 + d(3)^2} \tag{2}$$

From the viewpoint of the properties of amino acids, the first and second letters of the codon are important, and the distance D_{12} in these two letters can be expressed by the following equation.

$$D_{1,2} = \sqrt{d(1)^2 + d(2)^2} \tag{3}$$

Figures 4A and 4B show the histograms of the distances D and D_{12} between the nucleotide compositions from the actual genome sequences of 2664 species and those expected from the perfectly random mutation. In histograms for prokaryotes, the distance exceeds 0.15 and the histogram peak is at about 0.2. The histogram for eukaryotes shows a peak at a distance slightly less than the prokaryotic peak, ie about 0.15. Let us calculate how much the number of possible sequences is reduced when the nucleotide composition is deviated by 0.2 in prokaryotes and by 0.15 in eukaryotes.

The distribution of nucleotide composition from random sequences becomes a normal distribution, and the expected value of the composition for each nucleotide is 0.25. And the standard deviation becomes $\sigma = \sqrt{3/16N}$, when the number of nucleotides in the coding region of the genome sequence is N. The number N of nucleotides in the coding region of the prokaryotic genome utilized in this study is typically in the order of 10^6. Then, the standard deviation σ for $N \geq 10^6$ is less than 0.0004 for prokaryotes, and the bias of about 0.2 corresponds to about 500σ or more. In contrast, the eukaryotic genome size is much larger than that of prokaryotes. However, the ratio of the coding region to the whole genome is small in the eukaryotic genome. In fact, the number of nucleotides N in the coding region of the eukaryotic genome is typically in the order of 10^7. For example, the number of nucleotides in the coding region of the human genome is about 5×10^7. The standard deviation σ for $N \geq 10^7$ is less than 0.00014 for eukaryotes, and the bias of about 0.15 corresponds to about 1000σ or more. That is, the genome sequence within the habitable zone is a very rare phenomenon in both prokaryotic and eukaryotic organisms.

We showed that the habitable zone in the nucleotide composition space of the whole genome is largely out of the region expected from the random sequence. This can be reexamined from the viewpoint of narrowing down the individual mutation. In the case of a genome with nucleotide number N, the number of possible sequences is 4^N. Assuming that the individual mutations are narrowed down by a factor of κ ($0<\kappa<1$), the effective number of possible sequences becomes $(4\kappa)^N$. That is, assuming that the narrowing down factor κ^N due to individual mutations corresponds to the macroscopic bias in the habitable zone, the following equation is obtained.

$$\kappa^N = \frac{1}{\sqrt{2\pi}} \exp\left\{-(D/\sigma)^2/2\right\} \tag{4}$$

Since $\sigma = \sqrt{3/16N}$ as described above, the factor κ of the narrowing down of a single mutation is as follows.

$$\kappa = \exp(-8D^2/3)$$

The narrowing-down factor κ for each mutation is a function of only the degree of deviation D of the habitable zone, not depending on the size N of the whole sequence. In the case of D=0.2, the narrowing-down factor κ becomes about 0.9. In other words, relatively small biases in individual mutations can create the large macroscopic bias of the habitable zone in the nucleotide composition space of the entire genome sequence.

Another argument can be made about the factor for the maintenance of the habitable zone by examining the genomes of extreme environmental organisms. The gray marks in Figures 1 to 3 represent groups of thermophilic bacteria including hyper-thermophilic bacteria. It should be noted that the plots of extreme environmental organisms overlap with the habitable zone of ordinary environmental organisms. Since the GC pair forms three hydrogen bonds and the AT pair forms two hydrogen bonds, a high GC content makes the DNA double helix more stable and slows down the melting rate [14]. Given that the genome sequence is affected by the environment, it is expected that the genome of organisms adapted to high temperatures will contain more GC pairs, as a naïve prediction. However, this trend does not exist in the actual genome. In other words, the habitable zone is hardly affected at least by the current environment in the level of the nucleotide composition. Of course, it cannot be denied that the environment has a great influence on primordial codon formation and genome formation [8].

Another very important aspect of the habitable zone is that there should be some order in the total protein obtained from the genome sequence, which is advantageous for survival of the biological system. In the following sections, we discuss the unique mechanism behind the functional network of biological systems. Protein population formed by genome sequence in the habitable zone has almost constant appearance probabilities of membrane proteins (Section 2) and molecular recognition units (Section 3), which are advantageous for the survival of organisms.

2. Background mechanism of bio-system formation by the habitable zone at the second letter of codon

A biological system can be expressed by huge networks of protein functions. On the other hand, the genome sequence, which is the blueprint for the biological system, is made up of the accumulation of mutations. Therefore, the question of how functional networks of organisms are formed by mutation accumulation is essential for understanding organisms. In the previous papers, we reported that the proportion of membrane proteins is constant regardless of species [6,7]. Determining whether it becomes a membrane protein or a soluble protein places a strong restriction on the function of each protein. Therefore, their composition will be an important prerequisite for constructing the entire biological system. From this viewpoint, it would be useful to study the relationship between the habitable zone of the nucleotide composition space and the proportion of membrane protein in detail.

We have developed a software system SOSUI for predicting membrane proteins from amino acid sequences [15–17] and studied the proportion of membrane proteins in biological systems by analyzing the whole genome sequences. The results showed that all organisms have almost constant proportions of membrane proteins (about 23%) which are determined by the nucleotide composition of the entire coding region of the genome sequence [6,7]. Furthermore, we performed simulations of mutations to genome sequences at the DNA level [6] as well as amino acid level [18]. We introduced a large amount of mutation to the actual genome sequence on the computer and analyzed how the ratio of the membrane protein changes. When the nucleotide composition was fixed to the value of the actual genome sequence, the proportion of membrane protein reached a plateau of about 23%, which was the same value for the actual organism [6]. Similar simulations at the amino acid level also reached the same result [18].

Figure 5 summarizes the results of these researches in the form of flow charts. Regarding the second letter of the codon, the composition of thymine is large and the composition of guanine is small (Fig. 5A). As can be seen from the codon table (Fig. 5B), the hydrophobicity of amino acids is biased by the nucleotides at the second letter of the codon, which is due to the action of the translation system by ribosome and transfer RNA. Since the composition of thymine in the second letter is actually large, the frequency of hydrophobic amino acids of the proteins constituting the organism is large. Next, the membrane translocation device consisting of the translocon and signal recognition particles incorporates hydrophobic segments into the lipid membrane [19]. At this point, the membrane topology of the transmembrane helix is determined by the asymmetric sandwich structure of the hydrophobic segment and the amphiphilic amino acid segment. Figure 5C shows the distribution of hydrophobicity and amphiphilicity around the transmembrane region. The membrane protein prediction system SOSUI with high performance uses this distribution of the physical properties around the transmembrane domain [15–17]. Various intracellular devices must cooperate to make the proportion of membrane proteins constant, as shown in Figure 5D.

Here, we note that the composition of guanine and cytosine in the second letter of the codon is small, which also contributes to the formation of transmembrane helices. The amino acid groups of guanine and cytosine include secondary structure breakers, proline and glycine. We have already reported that serine and threonine clusters are also secondary structure breakers [20]. Since many amino acids in which the second letter of the codon is guanine and cytosine act as the secondary structure breakers, the occurrence probability of the secondary structure breaker is suppressed by the small composition of those nucleotides.

3. Background mechanism of bio-system formation by the habitable zone at the first letter of codon

The question in this section is what role the first letter of the codon plays in biological systems. When the first letter is guanine, most amino acids have small side chains like glycine and alanine. Since steric hindrance is small for amino acids with small side chains, the segments consisting of these amino acids become flexible. On the other hand, the group of amino acids in which the first letter of the codon is thymine contains many hydrophobic aromatic amino acids (phenylalanine, tyrosine and tryptophan). Since the bulky aromatic side chains show large steric hindrance, the segments containing those amino acids become less flexible. Furthermore, cysteine contained in the amino acid group of thymine at the first letter forms SS bonds, which reduces the degree of freedom of the whole protein. Thus, the nucleotide composition at the first letter of the codon must have a strong influence on the flexibility of segments in the protein.

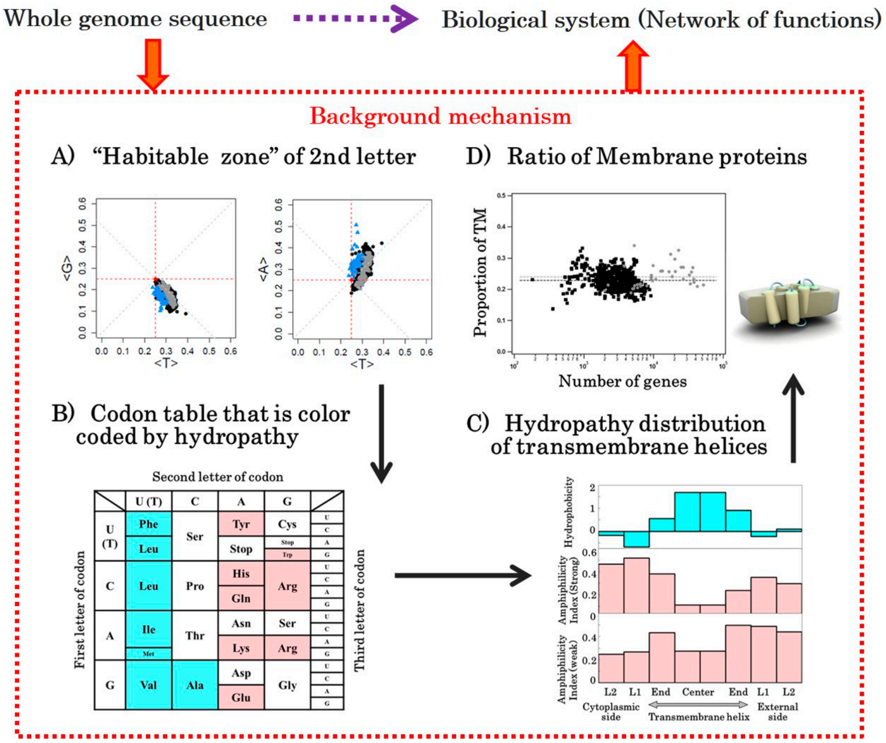
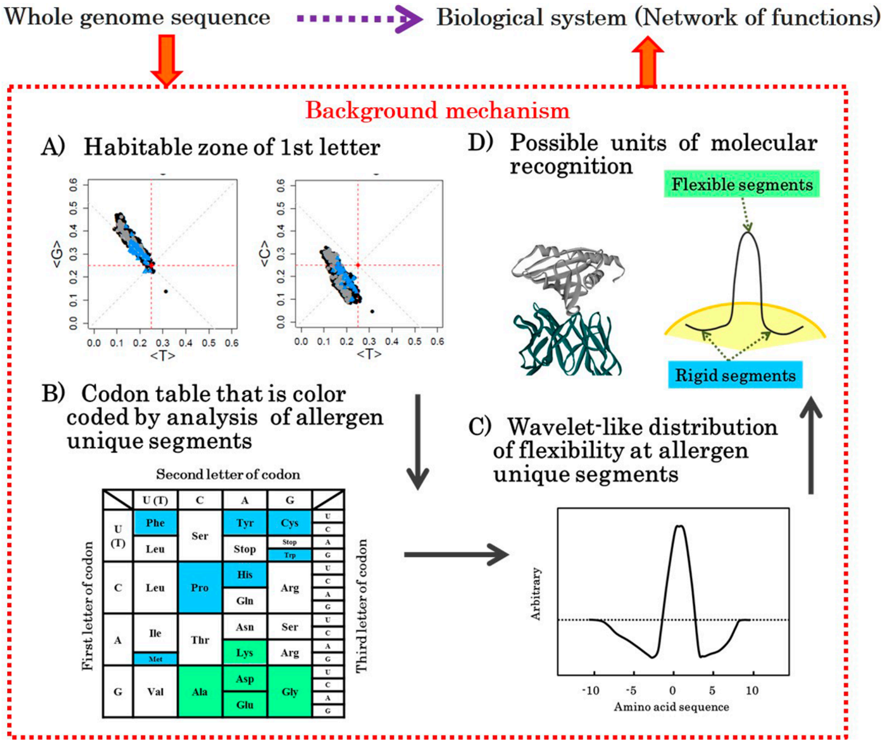


Figure 5 In the background mechanism by the habitable zone of the nucleotide composition at the second letter of the codon, the proportion of the membrane protein is controlled to become constant by biased random mutations to DNA sequences. A) The habitable zone is a narrow region in the nucleotide composition space, and greatly limits the number of possible sequences. In particular, at the second letter of the codon, the composition of thymine is large and the composition of guanine is small. B) In the codon table, hydrophobic amino acids (blue) and amphiphilic amino acids (red) are colored. The former is mainly coded by the second letter of thymine, the latter mainly coded by adenine. C) Characteristic plots of the moving average of the hydrophobicity and amphiphilicity indexes near the transmembrane helix is used in the high precision membrane protein prediction system SOSUI. These plots were obtained by analyzing amino acid sequences in the vicinity of many transmembrane helices [15–17]. D) Analysis of the total amino acid sequences from approximately 600 biological genomes showed that the proportion of membrane proteins is nearly constant regardless of species of organisms [6,7].

As shown in Figure 6A and graphs of Figure 1, the nucleotide composition of guanine in the first letter of the codon is clearly larger than 0.25 for most species, and the composition of thymine is smaller than 0.25. Furthermore, the nucleotide composition of <T> + <C> is smaller than 0.5, and the group of amino acids in which the first letter is cytosine also contains an aromatic amino acid, histidine, and an rigid imino acid, proline. Thus, the segments included in all proteins in the biological system are set to be flexible on average by the habitable zone in the first letter. However, truly important information exists in the one-dimensional fluctuation of the physical properties of amino acid sequences as discussed in the last section. The hydrophobic segment in the transmembrane helix can be regarded as a result of the fluctuation of the hydrophobicity. The question in this section is what type of segments of the protein is formed by the fluctuation in the flexibility of amino acid sequences.

Among the studies we conducted in the past, the analysis that aimed at predicting epitopes of allergens gave groupings of amino acids deeply related to the segment flexibility [21]. In this analysis, a data set of about 500 amino acid sequences was prepared for each of allergen and non-allergen. For all pairs of amino acid sequences, the homology was less than 30%. Then, homology analysis was performed in round robin, and amino acid fragments appeared only in allergens were extracted. As a result, we obtained over 10,000 fragments appearing only in allergens. Next, we examined the distribution of occurrence probabilities of amino acids in the vicinity of allergen-specific fragments. The distribution of occurrence probability of 20 kinds of amino acids could be classified into 3 types. The two were wavelet-like distributions, with peaks and valleys at the center. The third type was a distribution of white noise. The amino acids with the peak at the center were glycine, alanine, aspartic acid, glutamic acid and lysine, and the amino acids with the valley at the center were phenylalanine, tyrosine, tryptophan, cysteine, histidine, proline and methionine. The color coding of the amino acids in Figure 6B follows a grouping of amino acids obtained from the analysis of allergens. Figure 6C schematically shows the wavelet-like distribution obtained by the analysis of allergen unique fragments. When turning this graph upside down, we get a distribution with a valley at the center and small peaks on both sides away by several residues.


Whole genome sequence
Biological system (Network of functions)
Background mechanism
A) Habitable zone of 1st letter
<G>
<C>
<T>
0.0 0.1 0.2 0.3 0.4 0.5 0.6
B) Codon table that is color coded by analysis of allergen unique segments
Second letter of codon
First letter of codon
Third letter of codon
C) Wavelet-like distribution of flexibility at allergen unique segments
Arbitrary
-10 -5 0 5 10
Amino acid sequence
D) Possible units of molecular recognition
Flexible segments
Rigid segments

First letter	U (T)	C	A	G	Third letter
U (T)	Phe Leu	Ser	Tyr Stop	Cys Stop Trp	U C A G
C	Leu	Pro	His Gln	Arg	U C A G
A	Ile Met	Thr	Asn Lys	Ser Arg	U C A G
G	Val	Ala	Asp Glu	Gly	U C A G

Figure 6 In the background mechanism by the habitable zone of the nucleotide composition at the first letter of the codon, the formation of molecular recognition unit of proteins is controlled by the biased random mutations to DNA sequences. A) The habitable zone is a narrow region in the nucleotide composition space, and greatly limits the number of possible sequences. In particular, at the first letter of the codon, the composition of thymine is small and the composition of guanine is large. B) In the codon table, amino acids that make the segments flexible (green) and non-flexible (blue) are colored which were obtained from the analysis of allergen dataset [21]. The former is mainly coded by guanine of the first letter, the latter mainly coded by thymine. C) The universal distribution of amino acids that makes segments flexible has a wavelet-like shape with a peak at the center, and the distribution of amino acids making the segments non-flexible is upside down. This universal distribution and the classification of amino acids were obtained by analyzing the segment appearing only in allergen [21]. D) A schematic diagram of the flexibility distribution in the vicinity of the molecular recognition unit which is deduced from the analysis of allergen unique fragments. The molecular structure of allergen from birch pollen is also shown [22–24].

A summary of these analyses is schematically shown in Figure 6D. Allergen-specific fragments, ie, molecular recognition units of allergens, have a one-dimensional structure in which a flexible segment is sandwiched by rigid segments. The binding structure of pollen allergen and antibody is also added to Figure 6D. The epitope experimentally defined by structural analysis was in good agreement with the typical sandwich structure for allergen unique fragments by the computer [21]. Moreover, the fluctuation by B factor of the structural analysis of the allergen alone was largest at the allergen unique fragments [22–24].

The grouping of amino acids in this study was obtained by the analysis of the specific amino acid sequences of allergens, but this classification of amino acids shows good correlation with amino acid groupings by nucleotides of the first letter of the codon, as shown in Figure 6B. So, the grouping of amino acids obtained by the analysis of allergens can be attributed to the level of DNA. Therefore, it is suggested that the sandwich structure of segment flexibility can be generalized as the property of molecular recognition units.

In general, the molecular recognition of proteins has two contrasting aspects. On one hand, all proteins have molecular recognition sites that bind to substrate molecules. Then, in order to create a field of molecular recognition sites for binding the substrate, amino acid sequence fragments of some length must have physical properties that can be contacted with a substrate molecule. On the other hand, the molecular recognition of each protein is very specific, in the sense that each protein identifies a different partner. Sequence homology analysis is useful for estimating specific amino acids involved in the latter aspect of molecular recognition. When the multiple alignment analysis is performed on a group of amino acid sequences of proteins recognizing the same substrate, highly homologous amino acids contribute to the specificity of recognition. However, sequence homology is not effective for analyzing the former aspect of the molecular recognition. This is because amino acid sequence fragments that form molecular recognition sites must have common physical properties, but the homology of such sequence fragments is weak.

In material science, we often encounter problems where random noise masks the signal, and we use the "law of large numbers" as a way to extract a signal averaging large noises. The problem of extracting characteristic amino acid distribution around the molecular recognition units also has similar difficulties. The one-dimensional distribution of the physical property as a signal is masked by random mutations as noise. Therefore, "law of large numbers" is expected to be effective for solving this problem. In such an analysis, it is necessary to collect molecular recognition units for many proteins without sequence homology and to analyze the distribution of amino acids by superimposing them. If there is a distribution of physical properties common to the molecular recognition units, this averaging operation should give a characteristic distribution of occurrence probabilities of amino acids. The above work for allergen analysis was a method to accurately apply "law of large numbers" [21].

The characteristic distribution of the amino acid appearance probabilities found in allergen unique fragments may be in common with the distribution in more general molecular recognition units. Then, the distribution of amino acid appearance probability may be used for the prediction of molecular recognition units. We applied this method to various proteins, and the results suggested the possibility of the prediction of molecular recognition units (unpublished).

Here, we must discuss the exceptions of amino acid positioning in Figures 5 and 6. For example, there is a non-aromatic amino acid, serine, in the thymine group at the first letter, and relatively large valine in the guanine group. Since the average physical properties of the segment determine the transmembrane region and the molecular recognition unit, the nature of the segment of the protein can be clearly determined while permitting the exception of the positioning of amino acids in the codon table. This is an exquisite part of the background mechanism.

Furthermore, it should be noted that the mechanisms reliably control the occurrence probability of transmembrane regions and molecular recognition sites, regardless of the function of each protein. In this sense, Figures 5 and 6 can be regarded as the background mechanisms of the functional networks of biological systems. The genome sequence incorporates the highly sophisticated mechanisms in which protein-like structures as the material of biological systems are prefabricated before the formation of functional networks. Natural selection by environmental pressure probably will work very effectively against prefabricated protein-like structures.

4. Discussion

In this study, we examined the nucleotide compositions, which are intensive parameters from the genome sequence, using the data of 2664 organisms (2551 prokaryotes and 113 eukaryotes). When all organisms were plotted in the nucleotide composition space at each codon position, the plot was concentrated in a narrow region, especially at the first and second letters of the codon. Therefore, the "habitable zone" in the nucleotide composition space could be clearly defined. Since the living thing cannot survive outside, the habitable zone corresponds to the "phase diagram of life and death". Since these habitable zones are largely out of the nucleotide composition expected from random sequences, the number of possible sequences is drastically reduced. This is considered to be the basic mechanism by which living things stably survive. Also, by combining the habitable zone at the second letter of the codon with the analysis by the membrane protein prediction system, it became possible to clarify the meaning of the habitable zone at the protein level. If the occurrence probability of mutation at the second letter of codon is in the habitable zone, the proportion of membrane protein in the whole genome is controlled to be constant (about 23%) [6,7]. For the habitable zone at the first letter of the codon, the meaning at the protein level has been clarified from the analysis of the amino acid sequence of the allergen. The first letter of the codon is related to the flexibility of the amino acid and controls the formation of sequence fragments (molecular recognition units) that form the molecular recognition site of the protein.

These results further raise two questions. First, the exquisite arrangement of amino acids in the codon table is the key to the background mechanism of biological systems (Figs. 5 and 6), and the question is how did this arrangement evolve? Second, the nucleotide composition of the genome of an existing organism is kept within the narrow habitable zone, and the question is what mechanism is regulating the nucleotide composition.

M. Eigen and P. Schuster tried to answer the former question based on the concept of the realistic hypercycle [8]. At that time, no genome sequence was available at all, but they led to the evolution process of codons from the physical considerations on the stability of base pairs and amino acids. As mentioned in Sections 2 and 3, the current codon table is highly sophisticated, but the set of primordial codons had to be simpler than the current codon table. The binding energies of base pairs are much higher for GC pairs than for AU pairs, and higher GC content is more advantageous for forming longer nucleotide fragments. On the other hand, natural amino acids by prebiotic synthesis showed that amino acids with small side chains such as glycine and alanine are much more abundant than amino acids with larger side chains [25]. These physical facts show a good correlation with the fact that the codon of glycine is GGN and the codon of alanine is GCN. Here, N indicates one of four kinds of bases. Furthermore, M. Eigen and P. Schuster presumed that the evolution of codons occurred in the flow of GNC→RNY→NNN [8]. R represents purine (G or A) and Y represents pyrimidine (C or U). As Figure 1B shows, the composition of guanine and adenine is clearly large at the first letter of the codon and may be a trace of what RNY would have been the codon evolution process.

Husimi and colleagues pointed out the robustness of the mutation that amino acids with similar physicochemical properties have similar codons and thought it was a trace of codon evolution scenario [9]. The background mechanisms for the bio-system formation shown in this research are similar to the discussion by Husimi *et al.* And the codon evolution scenario by M. Eigen and P. Schuster is probably complementary to our argument in the meaning that they treat codons in the primordial environment and that we deal with the current standard codon.

The habitable zone in the nucleotide composition space shown in this study is conjugated with the exquisite arrangement of amino acids in the codon table. And the second question is what mechanism forms the habitable zone. The primordial codon is formed based on the stability of the base sequence and the composition of the natural amino acid at the time, and the problem is how the biased amino acid composition remains in the current genomic sequence as well. There are two hypotheses about this question. One hypothesis is that only organisms with nucleotide composition within the habitable zone survived, based on the characteristics of the codon table established in the process of evolution. Various theories of molecular evolution [26,27] are compatible with this hypothesis. Another possible hypothesis is the evolution of mutation management systems that control the probability of mutation within the habitability zone. Habitable zones shown in Figures 1 to 3 are conserved for all living things and may be preserved as the personality of the repair system for mutation. The advantage of this hypothesis is that it can dramatically reduce in advance the number of genomes that cannot survive. Regardless of these hypotheses, the habitable zone in the nucleotide composition space of the whole genome guarantees extremely high stability of the organism. Finally, we would like to point out that there is great potential for information analysis of whole genome sequence.

Acknowledgements

We deeply appreciate Dr. Kei Yura of Ochanomizu University for his very useful advice. We also thank very much Dr. Yuzuru Husimi of Saitama University for introduction of the concept of hypercycle.

Conflicts of Interest

All authors declare that they have no conflict of interest.

Author Contributions

S. M. directed the entire study and wrote the manuscript. R. S. performed the analysis of genome sequences.

References

[1] Abe, T., Kanaya, S., Kinouchi, M., Ichiba, Y., Kozuki, T. & Ikemura, T. Informatics for unveiling hidden genome signatures. *Genome Res.* **13**, 693–702 (2003).

[2] Forsdyke, D. R. *Evolutionary Bioinformatics*. (Springer, 2016).

[3] Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Ostell, J., Pruitt, K. D., *et al.* GenBank. *Nucleic Acids Res.* **46(D1)**, D41–D47 (2018).

[4] Taft, R. J., Pheasant, M. & Mattick, J. S. The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays* **29**, 288–299 (2007).

[5] Mattick, J. S. A new paradigm for developmental biology. *J. Exp. Biol.* **210**, 1526–1547 (2007).

[6] Sawada, R. & Mitaku, S. Biological meaning of DNA compositional biases evaluated by ratio of membrane proteins. *J. Biochem.* **151**, 189–196 (2012).

[7] Mitaku, S. & Sawada, R. What parameters characterize "life"? *Biophys. Physicobiol.* **13**, 305–310 (2016).

[8] Eigen, M. & Schuster, P. The hypercycle: A principle of natural self-organization, Part C: The realistic hypercycle. *Die Naturwissenschaften* **65**, 341–369 (1978).

[9] Aita, T., Urata, S. & Husimi, Y. From amino acid landscape to protein landscape: Analysis of genetic codes in terms of fitness landscape. *J. Mol. Evol.* **50**, 313–323 (2000).

[10] Ikemura, T. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* **2**, 13–34 (1985).

[11] Nakamura, T., Suyama, A. & Wada, A. Two types of linkage between codon usage and gene expression levels. *FEBS Lett.* **289**, 123–125 (1991).

[12] Kanaya, S., Yamada, Y., Kinouchi, M., Kudo, Y. & Ikemura, T. Codon usage and tRNA genes in eukaryotes: Correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *J. Mol. Evol.* **53**, 290–298 (2001).

[13] Fukushima, A., Ikemura, T., Kinouchie, M., Oshimae, T., Kudod, Y., Morig, H., *et al.* Periodicity in prokaryotic and eukaryotic genomes identified by power spectrum analysis. *Gene* **300**, 203–211 (2002).

[14] Cohen, R. J. & Crothers, D. M. Rate of Unwinding Small DNA. *J. Mol. Aid.* **61**, 525–542 (1971).

[15] Hirokawa. T., Boon-Chieng, S. & Mitaku, S. SOSUI: Classification and secondary structure prediction system for membrane proteins. *Bioinformatics* **14**, 378–379 (1998).

[16] Mitaku, S. & Hirokawa, T. Physicochemical factors for discriminating between soluble and membrane proteins: hydrophobicity of helical segments and protein length. *Protein Eng.* **12**, 953–957 (1999).

[17] Mitaku, S., Hirokawa, T. & Tsuji, T. Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces. *Bioinformatics* **18**, 608–617 (2002).

[18] Sawada, R., Ke, R-C., Tsuji, T., Sonoyama, M. & Mitaku, S. Ratio of membrane proteins in total proteomes of prokaryota. *Biophysics* **3**, 37–45 (2007).

[19] Walter, P. & Johnson, A. E. Signal sequence recognition and protein targeting to the endoplasmic reticulum membrane. *Annu. Rev. Cell Biol.* **10**, 87–119 (1994).

[20] Imai, K., Asakawa, N., Tsuji, T., Sonoyama, M. & Mitaku, S. Secondary structure breakers and hairpin structures in myoglobin and hemoglobin. *Chem-Bio Informatics J.* **5**, 65–77 (2005).

[21] Asakawa, N., Sakiyama, N., Teshima, R. & Mitaku, S. Characteristic amino acid distribution around segments unique to allergens. *J. Biochem.* **147**, 127–133 (2010).

[22] Gajhede, M., Osmark, P., Poulsen, F. M., Ipsen, H., Larsen, J. N., Joost van Neerven, R. J., *et al.* M.D. X-ray and NMR structure of Bet v 1, the origin of birch pollen allergy. *Nat. Struct. Biol.* **3**, 1040–1045 (1996).

[23] Mirza, O., Henriksen, A., Ipsen, H., Larsen, J. N.,Wissenbach, M., Spangfort, M. D., *et al.* Dominant epitopes and allergic cross-reactivity: complex formation between a Fab fragment of a monoclonal murine IgG antibody and the major allergen from birch pollen Bet v 1. *J. Immunol.* **165**, 331–338 (2000).

[24] Spangfort, M. D., Mirza, O., Ipsen, H., Van Neerven, R. J., Gajhede, M. & Larsen, J. N. Dominating IgE-binding epitope of Bet v 1, the major allergen of birch pollen, characterized by X-ray crystallography and site-directed mutagenesis. *J. Immunol.* **171**, 3084–3090 (2003).

[25] Miller, S. L. Production of some organic componds under possible primitive earth conditions. *J. Am. Chem. Soc.* **77**, 2351–2361 (1955).

[26] Akashi, H., Osada, N. & Ohta, T. Weak selection and protein evolution. *Genetics* **192**, 15–31 (2012).

[27] Kimura, M. & Ohta, T. Protein polymorphism as a phase of molecular evolution. *Nature* **229**, 467–469 (1971).