J. Biochem. 2012;151(2):189–196 doi:10.1093/jb/mvr132

JB THE JOURNAL OF BIOCHEMISTRY

Biological meaning of DNA compositional biases evaluated by ratio of membrane proteins

Received July 8, 2011; accepted October 18, 2011; published online October 28, 2011

Ryusuke Sawada* and Shigeki Mitaku

Department of Applied Physics, Graduate School of Engineering, Nagoya University, Furocho, Chikusa-ku, Nagoya 464-8606, Japan

*Ryusuke Sawada, Department of Applied Physics, Graduate School of Engineering, Nagoya University, Furocho, Chikusa-ku, Nagoya 464-8603, Japan. Tel: +81-(0)52-789-4465; Fax: +81-(0)52-789-4465; email: sawada@bp.nuap.nagoya-u.ac.jp

Membrane spanning regions can be used as markers for studying the robustness of biologically important units of proteins against evolutionary change (R. Sawada and S. Mitaku, *Genes to Cells*, 2010). We carried out computational experiments of extensive DNA mutations on the assumption of constant GC content or constant codon positional nucleotide biases. Randomized sequences were evaluated by membrane protein prediction systems SOSUI and SOSUIsignal. When all amino acid sequences from the total real genomes of 538 prokaryotes were analysed, ratios of membrane proteins to all genes in the total genomes were almost constant around a ratio of 22% with a standard deviation of 1.56. When the nucleotide sequences were randomized, keeping only the GC contents constant, the ratios of membrane proteins became highly diverse with a standard deviation of 10.1. When the codon positional nucleotide biases were taken into account; however, the diverse ratios of membrane proteins converged to a value of ∼25% with a standard deviation of 3.55. These results suggest that codon compositional biases play an important role in the evolution of prokaryotes for maintaining a constant ratio of membrane proteins. Further detailed analysis suggested that non-uniform nucleotide compositional biases at the terminal regions are the reason for the small but significant deviation.

Keywords: GC content/membrane protein/nucleotide compositional bias/nucleotide sequence simulation/ prokaryote genome.

Many reports have described biases in nucleotide compositions and non-random codon usages (*1–6*). It has been found that the four nucleotide bases are not evenly distributed among the three codon positions, and that this biased nucleotide composition leads to more convergent amino acid substitutions despite the different evolutionary histories of various biological organisms (*7*, *8*). Therefore, amino acid compositions are probably controlled by nucleotide compositional biases, which were formed by many mutations in the evolutionary process (*1–6*). However, the relationship between the average amino acid composition of all proteins in a biological genome and the characteristics of the organism is not yet understood.

To understand the biological meaning of the amino acid composition in a genome, we recently carried out simulations of random mutations of all amino acid sequences from a set of total genomes and estimated the ratio of membrane proteins by using the high performance membrane protein predictors SOSUI and SOSUIsignal (*9*, *10*). The ratio of membrane proteins was used to evaluate the meaning of the amino acid compositions. The results indicated that only if the real amino acid compositions were maintained, the ratio of membrane proteins was kept almost constant even after extensive mutations (*11*), suggesting that the distribution of transmembrane regions in amino acid sequences is strongly influenced by the amino acid composition. Therefore, the control of the amino acid composition by the biased nucleotide composition, and the control of the membrane protein ratio by the amino acid composition, leads to the control of the ratio of membrane proteins by the biased nucleotide composition. However, to what degree the ratio of membrane proteins is controlled by this mechanism is an open question, because the matching of the two processes and natural selection may create deviations in the ratios of membrane proteins.

Recently, we investigated the effect of the exon–intron structure on the formation of membrane proteins in whole genomes, using transmembrane regions as markers in exons (*12*). The population of all genes in a genome showed a single exponential decay against the number of exons per gene, suggesting that the number of exons per gene changes dynamically, and its distribution has already reached equilibrium. The question of how exons move or how introns are inserted and eliminated in the process of biological evolution was studied by measuring the change in positional distribution of exons containing transmembrane regions among prokaryotes, early eukaryotes and higher eukaryotes. In this way, a high-performance prediction system for membrane proteins can be a powerful tool for investigating the relationship between evolutionary changes in sequences and the formation of types of proteins.

In the present work, we carried out sequence simulations at the level of DNA in which all the nucleotide sequences in whole genomes of prokaryotes were randomized. When the nucleotide compositional biases at the three codon positions were maintained, the ratios of membrane proteins converged to ∼20% for the simulated genomes. In contrast, in the sequence simulation, ignoring the nucleotide biases at the three codon positions led to a very large divergence of the ratio of membrane proteins.

Materials and Methods

Genome data

We used data from 538 prokaryotic genomes that were obtained from NCBI (ftp://ftp.ncbi.nih.gov/genome/Bacteria/), RefSeq 24. The data set consisted of 43 archaea and 495 eubacteria. The general codon table of the genetic code was used for translation. All information about regions of ORFs was obtained from descriptions in the database.

Random sequence simulations at the level of nucleotide sequences

To investigate the effect of mutation of DNA sequences in the 538 prokaryotic genomes, all nucleotide sequences in the coding regions were mutated at the rate of a single nucleotide mutation per 100 bases per step of the simulation. A new nucleotide was selected according to the given nucleotide compositional biases. We used two kinds of nucleotide compositional biases: (i) The GC content of the genome was given without any biases at the codon positions and (ii) According to the model by Knight *et al.* (*7*), the nucleotide compositional bias at each codon position was given. If the terminal codon emerged during the simulation, a new nucleotide was randomly selected until a codon for an amino acid was selected on the assumption that the mutations producing terminal codons are so deleterious. The simulation was performed for 500 simulation steps. After 500 steps at which the sequences were completely randomized under the compositional biases, the nucleotide sequences were translated into amino acid sequences. Then, the analysis of the amino acid compositions and the membrane proteins predictions were carried out for evaluating the effect of the nucleotide compositional biases.

Prediction of membrane proteins

The numbers of membrane proteins were estimated by using SOUSI and SOSUIsignal, which predict membrane proteins and signal peptides, respectively (*9*, *10*). The accuracy of the prediction systems is ~95 and 90% for SOSUI and SOSUIsignal, respectively. The accuracy of SOSUI was very high; however, this system had a weak point in which membrane proteins predicted by SOSUI included a fraction of secretory proteins. The false-positive prediction by SOSUI was corrected by the SOSUIsignal system, which accurately predicts the existence of signal peptides.

Deviation of membrane protein ratio from the average value

The results of the simulations were evaluated by calculating the number of membrane proteins and comparing the average and the deviation of the membrane protein ratio between the simulated and the real genomes. Since the total number of open reading frames (ORFs) in biological genomes is different among biological species, the deviation of the membrane protein ratio was normalized by the square root of the total number of ORFs.

$$X = \frac{M_{\text{total}} - \text{AN}}{\sqrt{N}} \tag{1}$$

where M_{total}, N and A represent the total number of membrane proteins in a simulated genome, the total number of ORFs in the corresponding genome and the average value for the membrane protein ratio for all real genomes: 0.2197. Genomes with different total numbers of ORFs could be compared by this normalization procedure.

Averaged difference of codon compositional biases in sequence divisions

For all proteins, sequences were divided into sequence divisions with a length of 10 amino acids from the N- and C-terminal ends. Compositional biases were calculated at each division. The averaged differences of compositional biases of nucleotide k on the codon letter position j at the sequence division i between a real genome and simulated genome were calculated by the following equation;

$$D(i,j,k) = \frac{\sum_{n}^{N} [C_{\text{sim}}(i,j,k,n) - C_{\text{real}}(i,j,k,n)]}{N} \tag{2}$$

where $C_{\text{sim}}(i,j,k,n)$ and $C_{\text{real}}(i,j,k,n)$ indicate the composition of nucleotide k on codon letter position j at the sequence division i of organism n for simulated and real genomes. N indicates the total number of organisms used in this study.

Results

Dependence of membrane protein ratio on GC contents in prokaryotes

Since amino acids are encoded by three nucleotides, *i.e.* codons, the nucleotide composition in the genome sequence is one of the essential factors for determining the amino acid composition. Figure 1 shows a histogram of GC contents in the coding regions of 538 prokaryotic genomes in which the GC content scattered from ~0.3 to 0.7. The broken line represents the combination of three Gaussian distributions whose medium values are 0.32, 0.45 and 0.64, and standard deviations are 0.040, 0.101 and 0.030, respectively. The agreement between the histogram of the GC contents and the Gaussian distributions indicates the possibility of some random processes directed to three target GC contents.

When a nucleotide sequence is translated to an amino acid sequence according to the genetic code, the GC content has to influence the characteristics of the resulting proteins, because codons with fewer Gs or Cs code for more hydrophobic amino acids. In Fig. 2, the numbers of membrane proteins were plotted as a function of the number of ORFs in the total genomes of 538 prokaryotes. Despite of the great diversity in GC content, the ratio of membrane proteins seems to be a constant value of ~22%. Previously, we reported a very similar relationship for a much smaller number of biological genomes, and the present result indicates that the constant ratio of membrane proteins is a general tendency for prokaryotic genomes. We also studied the dependence of the membrane protein ratios on GC contents as shown by the grey scale. Although

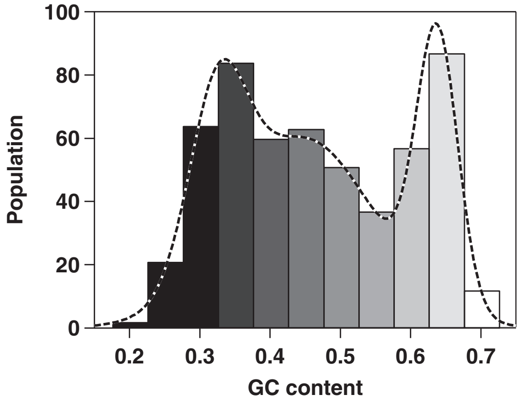

Fig. 1 Histogram of genomic GC content in 538 prokaryotes. The broken line represents the combination of three Gaussian distributions whose medium values are 0.32, 0.45 and 0.64, and standard deviations are 0.040, 0.101 and 0.030, respectively.

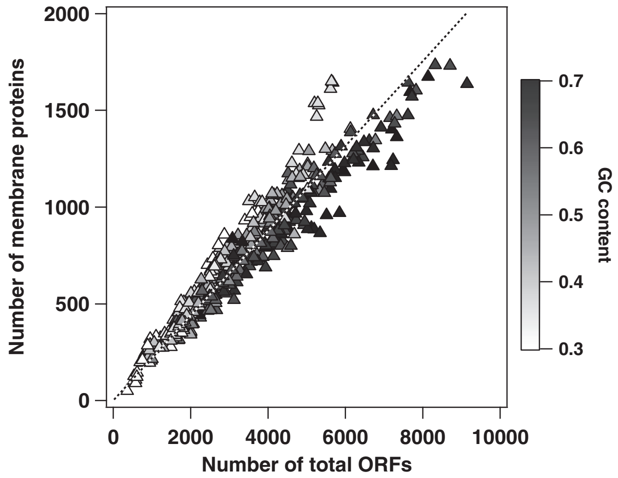

Fig. 2 Number of membrane proteins in 538 prokaryote genomes. The dotted line was obtained by least square deviation analysis; $y = 0.2192x$ with an R^2-value of 0.9151.

there is a slight tendency for a higher ratio of membrane proteins for genomes with smaller GC contents, the convergence of the ratio of membrane proteins raises the question of how the convergence of the ratio of membrane proteins was attained during evolution.

Membrane proteins in random nucleotide genomes

It is considered that real genomes were formed by a series of mutations. Therefore, the control for real genomes should be completely randomized sequences. Thus, we carried out randomization of all nucleotide sequences while maintaining the GC content of each genome. After the randomized simulation of genomes, the numbers of membrane proteins were estimated from amino acid sequences translated from simulated sequences by using SOSUI and SOSUIsignal. Figure 3A shows examples of the time course of the ratio of membrane proteins for three organisms: *Staphylococcus epidermidis*, *Escherichia coli* and *Xanthobacter autotrophicus*, whose genomes have GC contents of 0.330, 0.521 and 0.679, respectively. The ratio of membrane proteins strongly depended on the GC content, and the ratio became almost constant after 300 steps, indicating that the ratio of membrane proteins reaches equilibrium by extensive mutations. In Fig. 3B, the numbers of membrane proteins for the equilibrium values of simulated genomes are plotted as a function of total ORFs indicated as rectangles coloured by grey scale of genomic GC contents. The numbers of membrane proteins from real genomes are indicated by triangles as the reference. The ratios of membrane proteins highly deviated from almost 0–83% as shown in Fig. 3B. The ratio of membrane proteins systematically changed from almost 0% at GC contents of 0.7 to almost 80% at GC contents of 0.3. Figure 3C shows the deviations of the ratios of membrane proteins from the average ratio, indicating that the distribution of deviation for simulated genomes was completely different from that for real genomes. This relationship between the ratio of membrane proteins and the GC contents is quite reasonable, because many hydrophobic amino acids are encoded by codons with low GC contents. For the prediction of transmembrane helix segment, the SOSUI system mainly use the information about clustering of hydrophobic amino acid residues and other properties such as periodicity of particular amino acids are not used. When the amino acid sequences are determined by randomized nucleotide sequences, the probability of membrane spanning helices occurring will be influenced by the contents of hydrophobic amino acids, and a comparison between Figs 2 and 3 strongly suggest that the real nucleotide sequences have to be adjusted so that there is some mechanism for converging the ratio of membrane proteins.

Nucleotide biases within codon positions of first, second and third letters

It is well known that there are biases in the nucleotide occurrence at the three positions of codons. If nucleotides occur randomly, the propensity of each nucleotide at the first, second and third positions of codons will be the same with each other. However, the propensities of nucleotides are in fact very different among the three codon positions. Figure 4 shows the ratio of the four nucleotides as a function of the GC content. If there is no skewness of A/T and G/C, the nucleotide compositions for many genomes should form a straight line with slopes of 0.25 for G and C and −0.25 for A and T, as shown by the broken lines. In fact, the average of the ratios of the four nucleotides in whole genomes agreed well with the line corresponding to a random occurrence of nucleotides. However, when the nucleotide compositions at the three codon positions were analysed for the coding sequences, a significant deviation from the random occurrence lines was observed (Fig. 4), as was reported previously (*7*, *8*). It should be noted that the trends of the four nucleotides at the three codon positions showed a significantly high correlation throughout more than 500 genomes, including tens of extremophile genomes that are in extreme environments. The very high correlation in the trends of the four nucleotides independent of their environments suggests the existence of some mechanism to control nucleotide compositional biases other than natural selection.

The deviations from the random distribution (broken lines) were different among the three positions. A large skewness was observed for the first codon position: the ratio of thymine (T) was much smaller than that of adenine (A), whose distribution agreed well with the random distribution, while the ratio of guanine (G) was much larger than that of cytosine (C), whose distribution also agreed with the random distribution. The skewness of nucleotides at the second and third positions was small, but the slopes of the compositions of four nucleotides against the GC content were smaller at the second position and larger at the third position than the random distributions. The complementary slopes of the compositions at the second and the third positions apparently lead to a random distribution for the whole genomes.

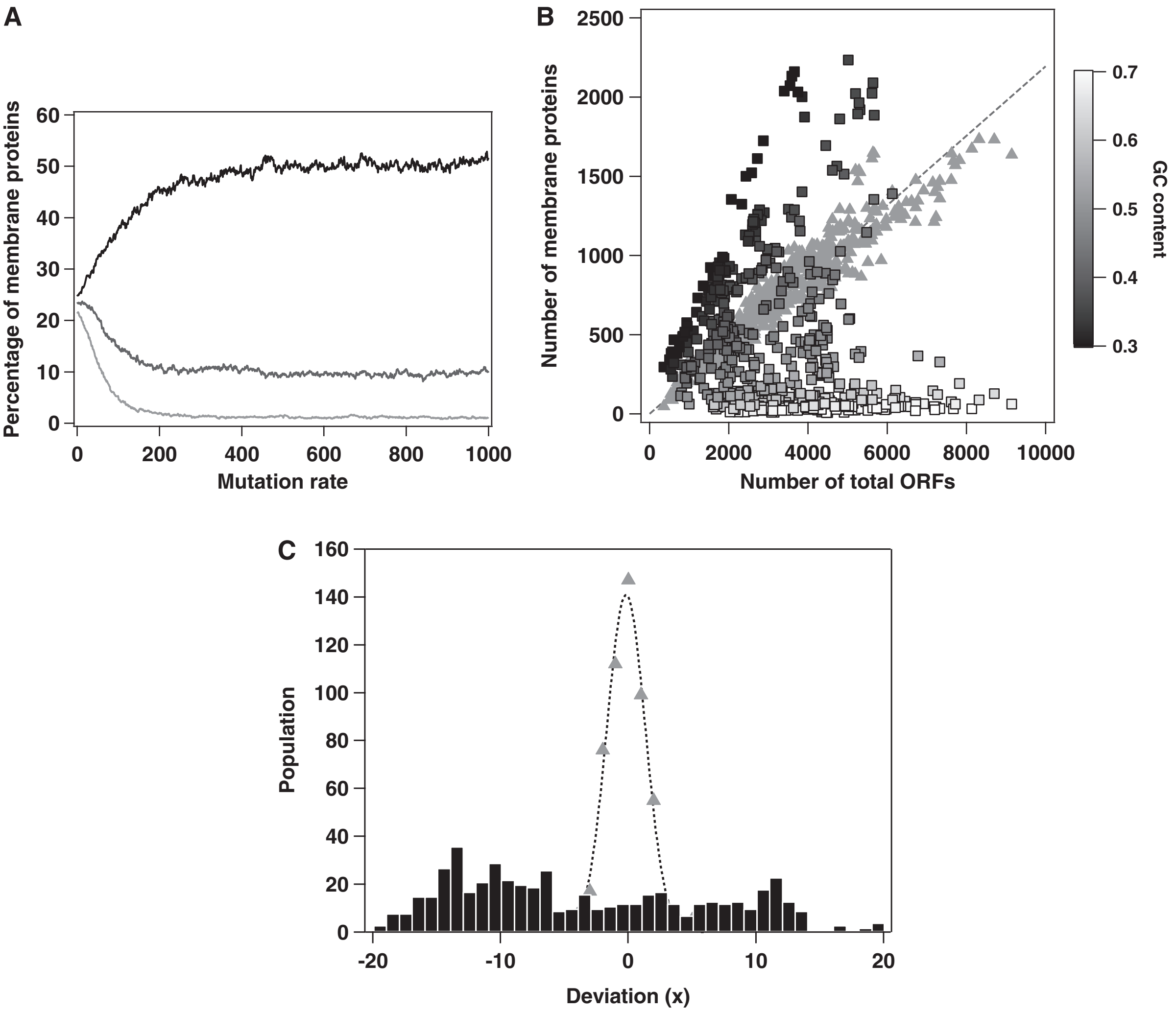

Fig. 3 Ratios of membrane proteins are highly deviated in the simulated proteomes with random nucleotide sequences. (A) Solid lines of black, dark grey and light grey represent the variation in the percentages of membrane proteins for *S. epidermidis*, *E. coli* and *X. autotrophicus*, respectively, plotted as a function of the randomized simulation up to the 1000th step. (B) Numbers of membrane proteins are plotted as a function of total ORFs for 538 prokaryotes. Rectangles indicate the number of membrane proteins in the random sequence genomes. All nucleotide sequences of protein-coding regions for 538 prokaryote proteomes were randomized according to the genomic GC content. For comparison, the number of membrane proteins and the linear approximation for real genomes are indicated as grey triangles and the dotted line. (C) A histogram of the deviations of membrane protein ratios of simulated genomes from the constant ratio of membrane proteins in real genomes is indicated by the closed bars. Deviation was calculated by using Eq. 1. The grey triangles indicate the deviations of membrane protein ratios in real genomes from the constant ratio of membrane proteins. A Gaussian distribution fitted to the data points is represented as a dotted line. Skewness, kurtosis and standard deviation of distribution are 0.117, 3.588 and 1.574, respectively (*15*). Light closed triangles and the dotted line indicate the results for real genomes for comparison.

These results for more than 500 genomes are consistent with previous works (*7*, *8*).

Effect of nucleotide biases on the ratio of membrane proteins

We investigated how the ratios of membrane proteins change when nucleotide sequences are randomized with the nucleotide compositional biases maintained. Figure 5A shows three examples of a time course of the ratio of membrane proteins for simulations of the same organisms as in Fig. 3A. The initial decay of the membrane protein ratio is much smaller in the simulations when assuming nucleotide compositional biases than in simulations without any biases. Small initial decays were in fact observed, but the levels of the decays were almost the same as the fluctuations. Figure 5B shows the number of membrane proteins in the amino acid sequences obtained from simulated nucleotide sequences of genomes estimated by SOSUI and SOSUIsignal. The ratio of membrane proteins for simulated genomes was 24.6%, which was almost the same as that for real genomes (grey triangle). The deviations of the membrane protein ratios are indicated for the real and simulated genomes in Fig. 5C. The distribution for both kinds of genomes fit a Gaussian distribution well (lines). The standard deviation of the distribution for simulated genomes was 3.55 and that for real genomes was 1.574. It is clear that both values of the average ratio and standard deviations were greatly improved when the codon compositional biases were taken into account. These results indicate that codon compositional biases play an important

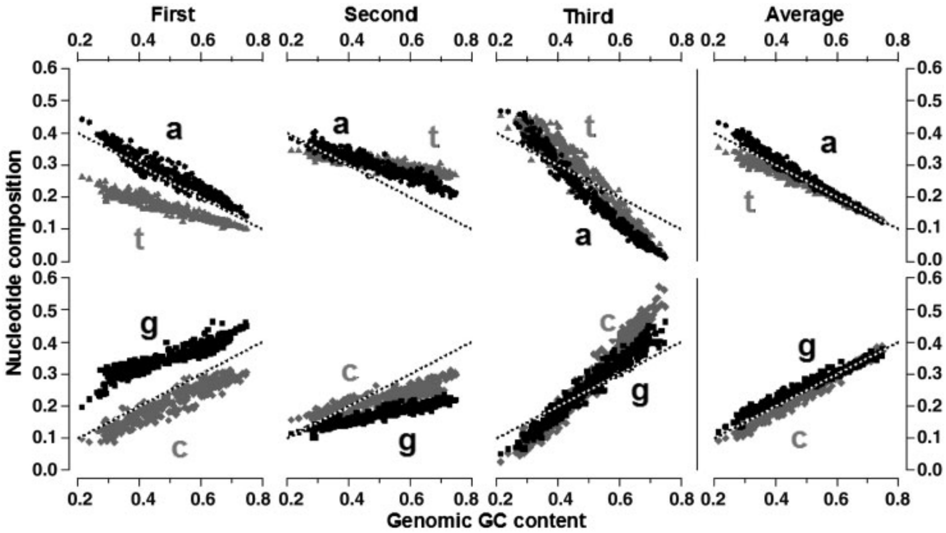

Fig. 4 Nucleotide compositions of the first, second and third codon positions in 538 prokaryotes are plotted as a function of the corresponding genomic GC content. Closed dark circles, closed light triangles, closed dark rectangles and closed light diamonds indicate nucleotides A, T, G and C, respectively. Dashed lines indicate the logical composition of random nucleotide sequence genomes.

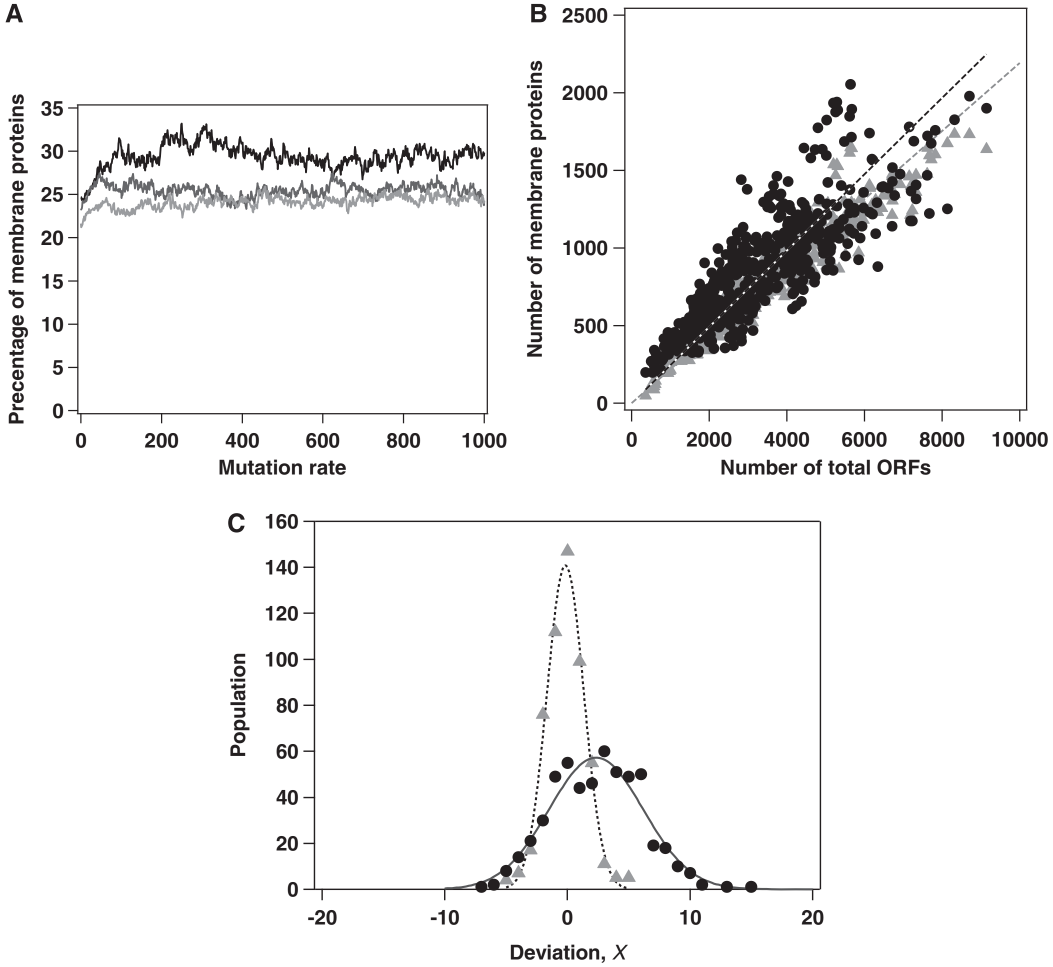

Fig. 5 The number of membrane proteins in simulated nucleotide sequences of genomes. All nucleotide sequences of 538 prokaryote genomes were randomized while maintaining the nucleotide compositions within the codon positions. (A) Solid lines of black, dark grey and light grey represents the variation in percentages of membrane proteins for *S. epidermidis*, *E. coli* and *X. autotrophicus*, respectively, plotted as a function of the randomized simulation up to the 1000th step. (B) Numbers of membrane proteins at the 500th mutational step are plotted as a function of the numbers of all proteins coded in the total genomes (closed circle). Linear approximation was performed by using a least square deviation analysis: $y = 0.2459x$, with an R^2-value of 0.6133 (dark dotted line). Grey closed triangles and dotted lines indicate the number of membrane proteins and the linear approximation in real genomes for comparison. (C) The distribution of the deviation of simulated genomes from the constant ratio among real genomes is shown as closed circles. A Gaussian distribution fitted to the data points is represented as a black line. Skewness, kurtosis and standard deviation of distribution are 0.070, 2.715 and 3.541, respectively (*15*). Light closed triangles and the dotted line indicate the result for real genomes for comparison.

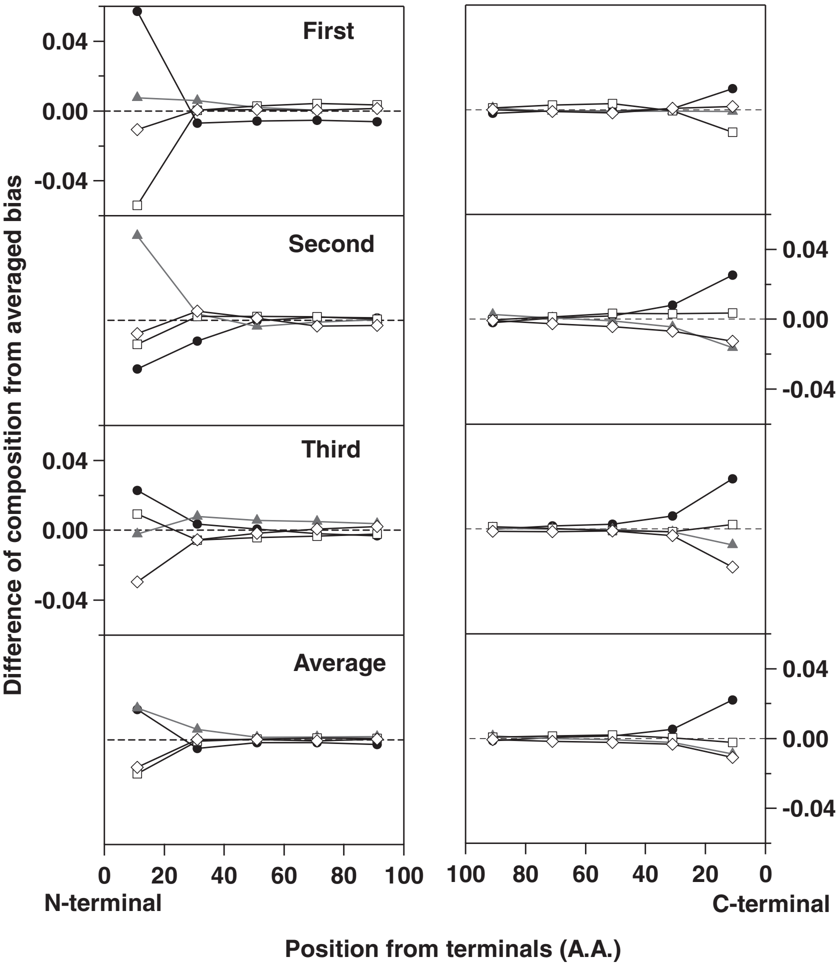

Fig. 6 Averaged differences of compositional biases of nucleotides in the three codon positions and their averages. Averaged differences in each division were calculated by using Eq. 2 from both N- and C-termini. Dark circles, light triangles, open rectangles and open diamonds represent nucleotides A, T, G and C, respectively.

role in the evolution of biological systems for maintaining a constant ratio of membrane proteins. However, the standard deviation of membrane protein ratios is still two times larger than that of real genomes. Another evolutionary factor needs to be considered for the sequence simulation to reproduce a constant ratio of membrane proteins in real genomes.

Deviation from the average of nucleotide compositional biases at terminal regions

Recently, it was reported that transmembrane regions are not evenly distributed throughout sequences and are statistically located at the end regions (*12*). Thus, we investigated the deviation of nucleotide compositional biases at the end regions from the average values at the central regions of sequences. Figure 6 shows the positional distribution of the deviation of the codon composition at the N-terminal and C-terminal ends from the average at the central regions. We found that the compositional biases significantly deviated at terminal divisions of 20 residues long from the average. At the first codon position, the propensity of adenine significantly increased and guanine decreased for the 20 residues of the N-terminal end. Higher nucleotide compositional biases were observed for thymine at the second codon position and adenine at the third position, and lower biases were observed for adenine at the second position and cytosine at the third. As for the C-terminal end, the deviations were mostly smaller than those for the N-terminal end. The propensity of adenine was larger than the average at all codon positions. This would cause an increase in hydrophobic amino acids at the terminal divisions, which would result in an increase of transmembrane regions (*12*).

Discussion

This work can be summarized in two points. (1) When a simulation of random mutations in genome sequences was carried out, the introduction of nucleotide compositional biases led to a convergence of the membrane protein ratios for 538 prokaryotic genomes to a single value of ~22%, which is similar to that of the real genomes. (2) However, the deviation of the ratio of membrane proteins for the simulated genomes was somewhat larger than for the real genomes, indicating that the average nucleotide compositional biases cannot regenerate the distribution of the ratio of membrane proteins. When the nucleotide compositional biases were analysed in detail, particularly at the amino- and carboxyl-terminal regions, the nucleotide compositional biases showed an anomalous distribution, which is probably the main reason for the difference in the dispersion of the ratio of membrane proteins between the real and simulated genomes.

Relevance of genome sequence simulation ignoring natural selection

Before discussing the biological meaning of the current work, we should explain its evolutionary relevance. In our simulation of random mutations, we did not care about the function of proteins. Therefore, discussing evolutionary processes on the basis of the present simulation seems nonsensical for functional proteins. However, there are two reasons why we performed the current work. First, statistical analyses of the nucleotide compositions in real genomes indicates significant biases at the three codon positions, which was revealed >40 years ago (*8*). This fact has been explained by directional mutation pressure rather than by the pressure of natural selection under various environments (*13*). Therefore, if the influence of the nucleotide compositional biases on the distribution of various proteins can be analysed independently of the functional effect, the contribution of directional mutation pressure can be elucidated. The simulation of extensive mutations is a good control experiment against real evolutionary processes.

The second reason is related to the general characteristics of proteins. It is known that a pair of proteins whose structures are similar do not necessarily show high sequence homology. A possible explanation is that, if a pair of proteins with low sequence homology shows a similar distribution of physical properties along their sequences, the proteins will form similar molecular structures. It is well known that an amino acid sequence becomes a transmembrane region if the hydrophobicity of the region is high enough, and if there are clusters of amphiphilic residues at the ends of the hydrophobic region. Based on knowledge about the physical properties of transmembrane helices, we developed the membrane protein prediction system, SOSUI. This system uses only physicochemical parameters for high-performance prediction of membrane proteins. Taking advantage of this system, which is based on the physical properties alone, we studied the effect of extensive random mutations in DNA sequences on the ratio of membrane proteins.

In this work, we have analysed all amino acid sequences from 538 prokaryotic genomes, including more than 70 genomes of the extremophiles by the membrane protein prediction system SOSUI. When membrane proteins were predicted for all amino acid sequences of real genomes, the ratio of membrane proteins to all amino acid sequences was almost constant at ~23%, independent of the environment of the organisms. If this constant ratio of membrane proteins is due to the directional mutation of DNA sequences, the simple introduction of nucleotide compositional biases will lead to an average ratio of membrane proteins in every prokaryotic genome.

Evolutionary mechanism for constant ratio of membrane proteins

We analysed two kinds of simulated genomes in comparison with the real genomes: In one simulation, the nucleotide compositional biases were kept the same as in the real genomes, while biases were ignored in the other simulations. The results were completely different between the two kinds of simulated genomes. The genomes with assumed nucleotide compositional biases had a membrane protein ratio of ~22%, which is very similar to the real genomes, whereas the randomized genomes without nucleotide compositional biases showed a very large dispersion of membrane protein ratios. It should be pointed out that no functional information was considered in these simulations, and that the difference between the two kinds of simulation arose only from the nucleotide compositional biases. This fact indicated that a systematic change in the distribution of various types of proteins may occur by the introduction of the nucleotide compositional biases into DNA sequences without any natural selection due to environmental factors.

This work implies that the membrane protein ratio is stable against random mutations in the genome only when the nucleotide compositional biases are maintained. Another important aspect of the simulation of extensive mutations is that the membrane proteins show decay curves with a characteristic number of about 100 steps, as shown in Figs 2A and 5A. In other words, the ratio of membrane proteins reaches an equilibrium after 300 simulation steps. The question of whether the real genomes of prokaryotes are in equilibrium cannot be answered, but the simulated genomes, assuming the nucleotide compositional biases as shown in Fig. 5, are certainly in equilibrium. The agreement of the ratio of membrane proteins between these simulated genomes and the real genomes suggest that the equilibrium has already been reached in the real genomes.

The simulation assuming a simple mechanism of codon positional nucleotide compositional biases showed good convergence of constant membrane protein ratios compared to that of a randomized nucleotide sequence. However, there was a small but observable difference in the dispersion of the ratio of membrane proteins between the simulated genomes and real genomes. This fact implies that nucleotide biases are an essential factor for maintaining a constant ratio of membrane proteins, but that another mechanism is also functioning to maintain the constant ratio. A clue to understanding this mechanism was obtained by studying the positional distribution of the nucleotide compositional biases shown in Fig. 6, which showed that the compositional biases at the terminal regions varied from the central regions. Recently, we reported from an analysis of all amino acid sequences from whole genomes that the transmembrane regions of membrane proteins are abundant at the terminal regions, particularly at the N-terminal region in membrane proteins of prokaryotes (*12*). Although the mechanism of the variation of the nucleotide compositional biases is as yet unknown, the slight difference between the simulated and real genomes shown in Fig. 6 may be due to the variation of the nucleotide compositional biases at the terminal regions.

Finally, we point out that the development of a high-performance prediction system of proteins can be very useful for studying biological evolution. When the system is based on sequence homology, the evolutionary distance between biological organisms can be estimated using pairs of homologous proteins. However, many amino acid sequences are orphans in sequence homology analysis. In contrast, when the system is based on physicochemical parameters, all amino acid sequences from whole genomes can be equally analysed, and the constitution of various types of proteins is obtained. In fact, the families of proteins responsible for big events in biological evolution can be revealed by analyzing the electric charge distribution (*14*). In this work, we applied this approach to simulated genomes and compared the ratio of membrane proteins between the simulated and the real genomes. Since the simulated genomes correspond to virtual evolution in which natural selection does not occur, the difference between the real genomes and the simulated gnomes will be partly due to natural selection.

Conflict of interest

None declared.

References

1. Francino, M.P., Chao, L., Riley, M.A., and Ochman, H. (1996) Asymmetries generated by transcription-coupled repair in enterobacterial genes. *Science* **272**, 107–109
2. Kano-Sueoka, T., Lobry, J.R., and Sueoka, N. (1999) Intra-strand biases in bacteriophage T4 genome. *Gene* **238**, 59–64
3. Ochman, H. (2003) Neutral mutations and neutral substitutions in bacterial genomes. *Mol. Biol. Evol.* **20**, 2091–2096
4. 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
5. Grigoriev, A., Freeman, J.M., Plasterer, T.N., Smith, T.F., and Mohr, S.C. (1998) Genome arithmetic. *Science* **281**, 1923a
6. Sueoka, N. (1988) Directional mutation pressure and neutral molecular evolution. *Proc. Natl. Acad. Sci. USA* **85**, 2653–2657
7. Knight, R.D., Freeland, S.J., and Landweber, L.F. (2001) A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol.* **2**, RESEARCH0010
8. Sueoka, N. (1961) Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. *Proc. Natl. Acad. Sci. USA* **47**, 1141–1149
9. Gomi, M., Sonoyama, M., and Mitaku, S. (2004) High performance system for signal peptide prediction: SOSUIsignal. *Chem-Bio Inf. J.* **4**, 142–147
10. Hirokawa, T., Boon-Chieng, S., and Mitaku, S. (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics* **14**, 378–379
11. Sawada, R., Ke, R., Tsuji, T., Sonoyama, M., and Mitaku, S. (2007) Ratio of membrane proteins in total proteomes of prokaryota. *Biophysics* **3**, 37–45
12. Sawada, R. and Mitaku, S. (2011) How are exons encoding transmembrane sequences distributed in the exon–intron structure of genes? *Genes Cells* **16**, 115–121
13. Sueoka, N. (1962) On the genetic basis of variation and heterogeneity of dna base composition. *Proc. Natl Acad. Sci. USA* **48**, 582–592
14. Ke, R., Sakiyama, N., Sawada, R., Sonoyama, M., and Mitaku, S. (2008) Vertebrate Genomes Code Excess Proteins with Charge Periodicity of 28 Residues. *J. Biochem.* **143**, 661–665
15. Pearson, E.S. and Hartley, H.O. (1954) *Biometrika Tables for Statisticians*. University Press, Cambridge