Nucleic Acids Research, 1996, Vol. 24, No. 21 **4249–4255**

A possible origin of newly-born bacterial genes: significance of GC-rich nonstop frame on antisense strand

Kenji Ikehara*, Fumiko Amada, Shigeko Yoshida, Yuji Mikata and Akira Tanaka[1]

Department of Chemistry and [1]Department of Biological Science, Faculty of Science, Nara Women's University, Kita-uoya Nishi-machi, Nara, Nara 630, Japan

Received July 1, 1996; Revised and Accepted September 18, 1996

ABSTRACT

Base compositions were examined at every position in codons of more than 50 genes from taxonomically different bacteria and of the corresponding antisense sequences on the bacterial genes. We propose that the nonstop frame on antisense strand [NSF(a)] of GC-rich bacterial genes is the most promising sequence for newly-born genes. Reasons are: (i) NSF(a) frequently appears on the antisense strand of GC-rich bacterial genes; (ii) base compositions at three positions in the codon are nearly symmetrical between the gene having around 55% GC content and the corresponding NSF(a); (iii) amino acid compositions of actual proteins are also similar to those of hypothetical proteins from the GC-rich NSF(a); and (iv) proteins from NSF(a) of 60% or more GC content are flexible enough to adapt to various molecules encountered as novel substrates, due to the high glycine content. To support our proposition, using a computer we generated hypothetical antisense sequences with the same base compositions as of NSF(a) at each base position in the codon, and examined properties of resulting proteins encoded by the imaginary genes. It was confirmed that NSF(a) of GC-rich gene carrying about 60% GC content is competent enough for a newly-born gene.

INTRODUCTION

One of the most interesting research projects in molecular evolution may be studies on the mechanism of creation of genes. Although little is known about the mechanism as yet, two main ideas are proposed by several groups (1–6). One is the gene duplication theory (1,2), which anticipates that genes to be coding for new enzymes should be generated from the duplicated genes encoding extant enzymes catalyzing related reactions. The other is the overprinting mechanism whereupon new genetic sequences are produced from currently unused reading frames on sense or antisense strands (3–6). The latter is a more promising mechanism for newly-created bacterial genes, since the new genes produced by the overprinting could have novel functions quite different from the ancestral genes. Since different frames are used in overprinting, the original gene function is retained even after a newborn gene has been created. The typical examples of overlapping genes were found in bacteriophages ϕX174 (7–9), fd (10) and an animal virus SV40 (11,12). However, these might be specialized examples of the viral genes on small genomes. On the other hand, Yomo *et al.* originally provided the overprinting mechanism encoded by the antisense sequences of the same frame as the genes on sense strands in generation of the nylon oligomer-degrading enzymes of *Flavobacterium* and *Pseudomonas* species (5,13–16). Subsequently, we expanded the overprinting mechanism on the other genes of *Flavobacterium* sp. (6).

If newly-born genes could be produced by the overprinting, several conditions must be satisfied to develop the noncoding sequences into new coding sequences. One is that nonstop frame (NSF) must be long enough to encode several hundreds of amino acids without any interruptions, i.e. without meeting any stop codons along the new reading frame. Another is that amino acid composition of the polypeptide produced from the newly-born coding sequence must be similar to that of extant proteins. In other words, if the new polypeptide encoded by the NSF has largely different amino acid composition from proteins encoded by the active genes, it will be difficult for the former even to fold into a functional three-dimensional structure because of the unfavorable amino acid composition. Moreover, flexibility of polypeptides from the newly-born coding sequences should be ensured to adapt to novel substrates, because the newborn 'enzymes' would hardly function as mature enzymes from the beginning. Thus, as a second step, the much less functional enzymes produced from the newly-born genes must efficiently evolve in the cells into competent enzymes under appropriate conditions. In this respect, bacterial cells should be advantageous judging from the large number of population and the short time of generation. In this paper, we provide the hypothesis that NSF(a) of the GC-rich bacterial gene should be the most promising sequence for a newly-born gene and for subsequent development of the originally noncoding sequence [NSF(a)] into a newly coding gene.

* To whom correspondence should be addressed. Tel: +81 742 20 3402; Fax: +81 742 3424; Email: ikehara@cc.nara–wu.ac.jp

RESULTS AND DISCUSSION

Unusually biased base sequences were generally observed on GC-rich bacterial genes

In addition to seven genes from *Flavobacterium sp*. described in the previous paper (6), 14 GC-rich genes from eight species of bacteria distantly related to each other were arbitrarily selected from the GenBank database by using the program IDEAS-seqman (6). GC contents and codon numbers of the selected genes ranged from 63.5 to 72.8% and from 137 to 709, respectively (14 genes from the bottom of Table 1). They were all house-keeping genes, different from *Flavobacterium nyl* genes, which are possibly newly-created genes (5). As shown in Table 1, characteristics of the gene products were widely distributed, since the genes were selected arbitrarily. Base compositions at every position of codons of the genes clearly indicate that all GC-rich genes carrying >60% GC content showed unusually biased nucleotide sequences, (GNC/g)n, like the cases of *Flavobacterium sp.* genes (6). In the symbol (GNC/g), the capital letters, G and C, indicate that they are the main bases at each position, and the small letter g means that guanine is the second main at the third position. N is either of the four bases. Several common features were revealed on almost all GC-rich genes examined in this and the previous papers (6). The features are (i) guanine and thymine were observed at the first position in the codon at the maximum and minimum frequencies, respectively, except that thymine was contained slightly more than adenine in two genes, *Halobacterium sp.* H-*sod* and *Mycobacterium tuberculosis thyA* genes, (ii) at the second base position, four kinds of bases were almost equally contained, (iii) cytosine was detected most frequently at the third position of the codon, and guanine was the second most, without exception. (iv) Both adenine and thymine contents were very low at the third base position and, moreover, the scores of adenine were usually lower than those of thymine.

Dependence of base compositions at the codon positions on GC content of bacterial genes

Nucleotide sequences of 56 genes [48 given in Table 1, 8 in the previous paper (6)], most of them from taxonomically different bacteria, were collected from the GenBank database with the computer program, IDEAS-seqman, developed by Dr M. Kanehisa (Kyoto University, Institute of Chemical Research, Uji, Kyoto), as described in the previous paper (6). Base compositions of the genes at three positions in the codon were numerically scored. In the first base position of codons (Fig. 1a), guanine was most frequently detected even on AT-rich genes having ~30% GC content, although guanine gradually decreased under GC/AT pressure (17,18) as the GC content of the genes decreased, while the values of thymine were usually low even on the AT-rich genes. The values of cytosine were roughly half of the guanine across the wide range of GC content of the genes. Adenine largely increased as the GC content decreased, compensating for the decrease of guanine and cytosine.

In the second position (Fig. 1b), it is obvious that compositions of all bases were rather insensitive to the change of GC content of the genes, in comparison with the other base positions (18,19), although adenine was most, and thymine was least dependent on the GC content under the GC/AT pressure (17,18). It was also found that contents of three bases except guanine were almost equal ~60% GC content.

Table 1. Some properties of bacterial genes examined in this study[a,b]

Bacterial species	Gene	GC (%)	Codon no.	SC(a)[c]
Clostridium thermosulfurogenes	*lacA*	30.8	716	28
Vibrio cholerae	*tcpJ*	31.5	253	24
Rickettsia prowazekii	*pepA*	33.2	500	35
Clostridium acetobutylicum	*lyc*	34.7	324	24
Bacillus cereus	*sph*	35.6	333	25
Vibrio cholerae	*aldD*	36.8	506	31
Staphylococcus aureus	*glu*	37.4	357	13
Streptococcus thermophilus	*lacZ*	37.4	985	27
Staphylococcus aureus	*lip*	37.5	690	44
Flavobacterium meningosepticum	*endoF1*	39.0	339	13
Bacillus cereus	*glnA*	39.1	444	35
Flavobacterium meningosepticum	*pep*	39.2	705	18
Haemophilus influenzae	*crp*	39.3	224	16
Bacillus coagulans	*ctsA*	39.5	373	33
Treponema denticola	*pyrB*	41.4	475	15
Haemophilus influenzae	*sodC*	41.9	187	18
Acinetobacter calcoaceticus	*gld*	42.4	801	47
Bacillus subtilis	*sigG*	43.0	260	14
Proteus vulgaris	*rplA*	45.0	233	12
Bacillus subtilis	*sigF*	44.8	255	11
Moraxella sp.	*lip1*	45.9	319	14
Methanobacterium formicium	*fdhC*	46.2	280	14
Morganella morganii	*hisD*	47.4	378	10
Nitrosomonas europaea	*amoA*	48.0	275	4
Zymomonas mobilis	*adhB*	48.8	383	4
Bacillus stearothermophilus	*aml*	50.0	549	6
Chlorobium vibrioforme	*cycA*	50.8	206	4
Zymomonas mobilis	*pdc*	52.4	559	2
Chlorobium vibrioforme	*hemA*	53.9	415	9
Bacillus stearothermophilus	*ala.rac*	56.2	286	3
Treponema pallidum	*pyr.red*	56.7	260	7
Mycobacterium tuberculosis	Mn-*sod*	59.4	206	3
Mycobacterium tuberculosis	*folA*	59.7	214	3
Desulfovibrio vulgaris	*srd*	59.9	218	0
Rhodopseudomonas capsulata	*cycA*	63.5	137	0
Mycobacterium tuberculosis	*thyA*	63.8	449	6
Caulobacter crescentus	*hfaB*	64.5	230	0
Paracoccus denitrificans	*cyc*	65.2	226	0
Paracoccus denitrificans	*ndh*	65.7	431	1
Caulobacter crescentus	*hfaA*	66.0	147	0
Rhodopseudomonas capsulata	*lac*	68.7	293	0
Halobacterium sp.	H-*sod*	68.8	200	0
Mycrococcus luteus	*uvrB*	68.6	709	0
Streptomyces coelicolor	*argG*	70.6	487	0
Cellulomonas fimi	*cex*	71.1	484	0
Caulobacter crescentus	*divJ*	72.2	596	3
Cellulomonas fimi	*cenA*	72.5	449	0
Streptomyces coelicolor	*afsB*	72.8	243	0

[a]Genes are listed in the order of the increase of GC content.
[b]Eight genes of *Flavobacterium sp.* are ommited from this table for simplicity. About the genes, see ref. 6.
[c]SC(a) indicates number of stop codons on antisense strand.

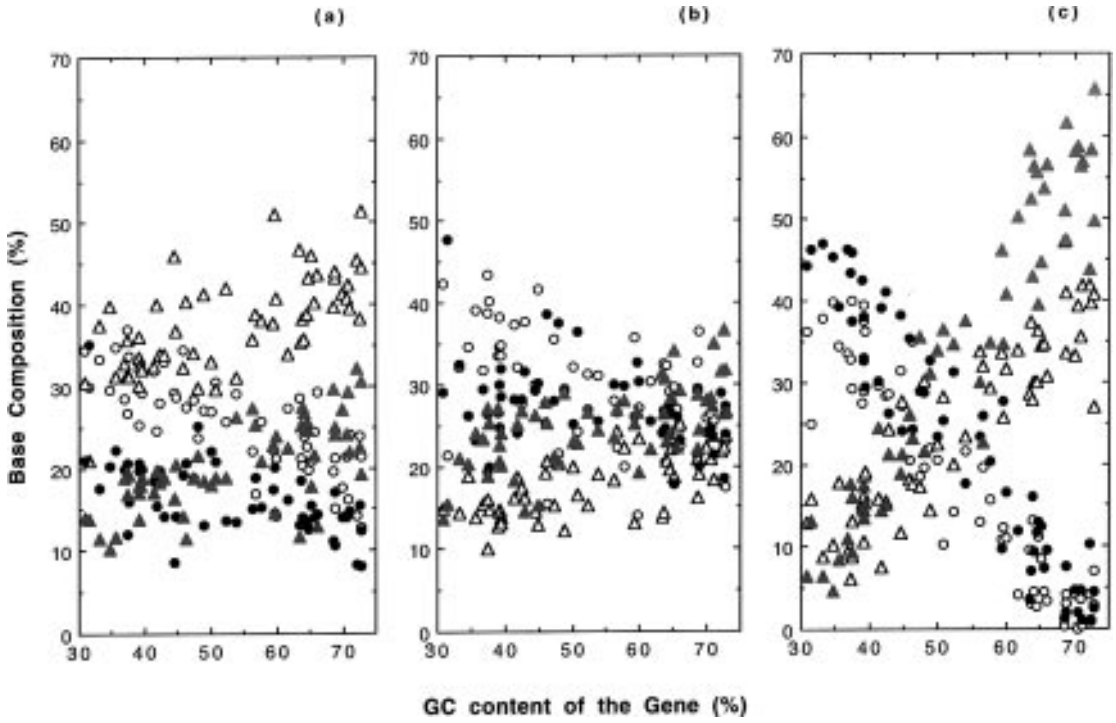

Figure 1. Base compositions are plotted against GC content of the genes at the first (**a**), second (**b**) and third (**c**) base positions in the codon. More than 50 bacterial genes were observed (see Table 1 and ref. 6). Open and filled circles, and open and filled triangles indicate adenine, thymine, guanine and cytosine, respectively.

In the third position (Fig. 1c), the contents of the four kinds of bases were much more dependent on the GC content of the genes than the other base positions, as was expected from the fact that most base substitutions at the third position in the codon do not cause amino acid exchange due to degeneracy of the genetic code at that position. A couple of interesting properties were also observed at this position. Cytosine content was much more dependent on the GC content of the genes than that of guanine. Adenine and thymine contents were extremely low in the genes ~70% GC content, whereas the contents of guanine and cytosine were not so low in the genes of ~30% GC content. The former was significantly lower than the latter, in spite of the same distances from 50% GC content and of probably similar strength of GC/AT pressure. It is also apparent that the small amounts of adenine and thymine in the high GC-rich region cannot be simply accounted for by the fact that stop codons are by definition absent from coding regions, because the amount of thymine was as low as that of adenine at the third codon position (Fig. 1c).

The data of bacterial genes shown in Figure 1 led to the following equations when approximated by the linear least-squares method.

$$G_1 = 0.289\ C_{GC} + 0.222 \qquad A_1 = -0.349\ C_{GC} + 0.442$$
$$C_1 = 0.267\ C_{GC} + 0.060 \qquad T_1 = -0.207\ C_{GC} + 0.276$$
$$G_2 = 0.207\ C_{GC} + 0.073 \qquad A_2 = -0.295\ C_{GC} + 0.452$$
$$C_2 = 0.227\ C_{GC} + 0.127 \qquad T_2 = -0.138\ C_{GC} + 0.347$$
$$G_3 = 0.726\ C_{GC} - 0.136 \qquad A_3 = -0.937\ C_{GC} + 0.681$$
$$C_3 = 1.284\ C_{GC} - 0.347 \qquad T_3 = -1.074\ C_{GC} + 0.801$$
$$(E_{obs.}1\sim12)$$

where G, A, C and T represent frequencies of bases; numerical subscripts indicate the position in the codon; and C_{GC} means GC content of the gene. These equations are used for further analyses of following sections.

Bacterial GC-rich genes produce NSF on the corresponding antisense strand at a high probability

Figure 2 shows that number of stop codons on antisense strand [SC(a)] decreases as GC content of bacterial gene increases, and that the SC(a) values approach to almost zero in the region larger than ~60% GC content. The latter fact is interesting, since long NSFs should be detected at a high probability on the antisense strands of the widely distributed bacterial GC-rich genes. The biased base sequences, (GNC/g)n, on the sense strands as well as the high GC content may offer promising regions for possible new genes on the antisense strands without stop codon.

To investigate the contribution of the biased base compositions in producing NSF(a), the observed SC(a) frequencies on actual genes were compared with the $SC(a)_{biased}$ values calculated by equation **E1** set up by taking consideration of the biased base compositions at each position in the codon.

$$SC(a)_{biased} = T_1T_2A_3 + T_1C_2A_3 + C_1T_2A_3 \qquad \textbf{E1}$$

The values of T_1, T_2, A_3, C_1 and C_2 were obtained from appropriate equations described in the above section ($E_{obs.}$ 1~12).

The observed SC(a) values were also compared with the $SC(a)_{random}$ values obtained by calculation of equation **E2**, in which random distribution of the bases on the reading frame (and on the antisense strand) was assumed.

$$\begin{aligned} SC(a)_{random} &= T_1T_2A_3 + T_1C_2A_3 + C_1T_2A_3 \\ &= T_1A_2A_3 + T_1G_2A_3 + T_1A_2G_3 \\ &= [1 - C_{GC} - (C_{GC})^2 + (C_{GC})^3]/8 \end{aligned} \qquad \textbf{E2}$$

where $T_1 = T_2 = A_2 = A_3 = (1 - C_{GC})/2$ and $C_1 = C_2 = G_2 = G_3 = C_{GC}/2$.

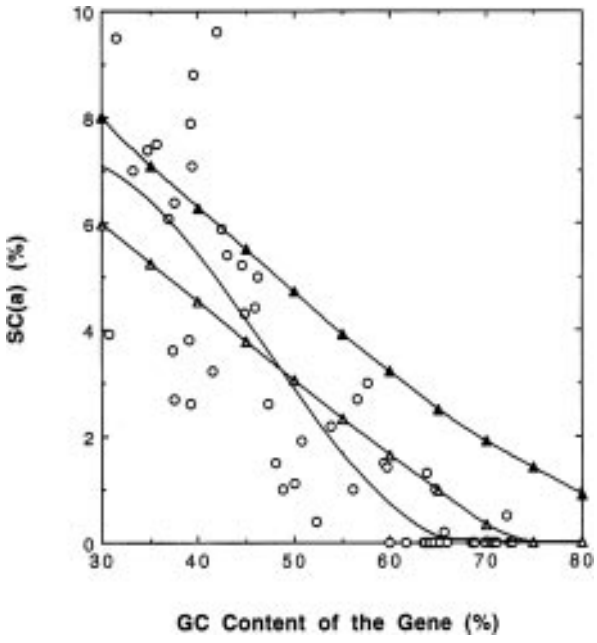

Figure 2. Percentages of SC(a) against GC content of the genes. Open circles indicate the values of observed SC(a) (Table 1). The average curve (no symbol) was obtained by approximation with a trinominal equation: $SC(a)_{obs.} = -4.44 + 1.03C_{GC} - 2.74 \times 10^{-2}(C_{GC})^2 + 1.94 \times 10^{-4}(C_{GC})^3$. Open and closed triangles show percentages of $SC(a)_{biased}$ and $SC(a)_{random}$ values calculated from equations **E1** and **E2** described in the text, respectively.

As can be seen in Figure 2, the $SC(a)_{biased}$ values were clearly lower than the $SC(a)_{random}$ values irrespective of the GC content. On the other hand, the curves of the calculated $SC(a)_{biased}$ values were located closely to the average curve of the observed SC(a) values. Both the observed SC(a) and the $SC(a)_{biased}$ curves gradually decreased as the GC content of the genes increased and approached zero in the region of high GC content. The observed SC(a) values were steadily scattered below the line of $SC(a)_{random}$ in GC-rich region, indicating that the biased base distribution in the codon lowered actual SC(a). This was true even ~50% GC content. These may be attributed to the facts that content of adenine at the third position on sense strand, or that of uracil composing a part of stop codons at the first position on antisense strand, is extremely low, and that contents of both cytosine and thymine at the first position in the codon on sense strand, which corresponds to guanine and adenine at the third position on antisense strand, are much smaller than those of other bases (Fig. 1a). Compared with the calculated $SC(a)_{biased}$ line, the observed SC(a) values tended to shift upward in the AT-rich region and downward in the GC-rich region (Fig. 2). These tendencies may result from not only the biased base compositions in the codon but also deviation of the local base sequences from the randomness, although the possibility cannot be ruled out that it is simply caused by the rough approximation of equations ($E_{obs.}$ 1~12) in the AT-rich and GC-rich regions.

The results in Figures 1 and 2 clearly show that the biased base compositions in codon positions are mainly responsible for producing NSF(a). Therefore, many bacterial genes with a GC content >60% certainly produce NSF(a)s at a high probability. In other words, it can be inferred that NSF(a)s of the GC-rich bacterial genes satisfy the primary condition for producing novel genes. At the same time, the results also indicate that it is practically difficult to make long NSF(a) on AT-rich bacterial genes. This is also of importance when the origin of AT-rich bacterial genes is discussed because AT-rich bacteria must not be able to produce new genes in their own cells and they must receive newly-born genes from others, probably GC-rich bacteria, through horizontal gene transfer.

Base compositions in the codon of bacterial genes with GC content ~55% are nearly symmetrical to those on the antisense strands

It can be seen that contents of four kinds of bases at the first position in the codon are roughly similar to those of complementary bases at the third position (Fig. 1a and c). In addition to that, at the second position the lines approximated by the linear least-squares method seem to converge at ~55% GC content except for the line of guanine (Fig. 1b). This means that the base compositions in the same frames on sense and antisense strands of the genes approaches mutual symmetry in the GC-rich region as, for example, GNC.

To confirm these tendencies quantitatively, the symmetry value in the codon [Sym(c)] was calculated by equation **E3** as an index of symmetry between corresponding codons on the sense and antisense strands.

$$Sym(c) = |A_1 - T_3| + |T_1 - A_3| + |G_1 - C_3| + |C_1 - G_3| + |A_2 - T_2| + |G_2 - C_2| \quad \textbf{E3}$$

where A, T, G and C represent frequencies of bases, and numerical subscripts indicate the position in the codon. The values of base contents were obtained from equations, ($E_{obs.}$ 1~12). In the equation, we considered the fact that contents of A, G, T and C at the first position in the codon correspond to those of the complementary bases, T, C, A and G at the third position on the antisense strand, respectively, and absolute values were used so that the smaller Sym(c) value means the higher symmetry of base compositions between the codons on sense and antisense strands. The calculated Sym(c) values were plotted against GC content of the genes in Figure 3. The base compositions in codons of bacterial genes were most symmetrical at ~55% GC content in relation to those between the sense and the corresponding antisense strands, since the Sym(c) value reached minimum at that point.

Similarity of amino acid compositions of proteins derived from sense and antisense strands

Next, it was investigated whether the GC-rich NSF(a)s can encode polypeptides composed of similar amino acids to those found in real proteins, since the symmetrical base compositions in the codons on sense and antisense sequences do not always guarantee to encode polypeptide chains composed of similar amino acid compositions. Here we compared the amino acid compositions of actual proteins encoded by seven genes [*B.stearothermophilus ala.rac*, *T.pallidum pyr.red*, *M.tuberculosis* Mn-*sod*, *folA* and *thyA*, *D.vulgaris srd* and *R.capsulata cycA* (Table 1)] having ~60% GC contents with those of possible proteins encoded by the corresponding antisense strands (Fig. 4). Similar amounts of amino acids were contained in proteins encoded by both sense (black bars) and antisense (hatched bars) sequences, although arginine showed difference. This similarity of amino acid composition suggests that proteins derived from the antisense strands could be functional enough in GC-rich region.

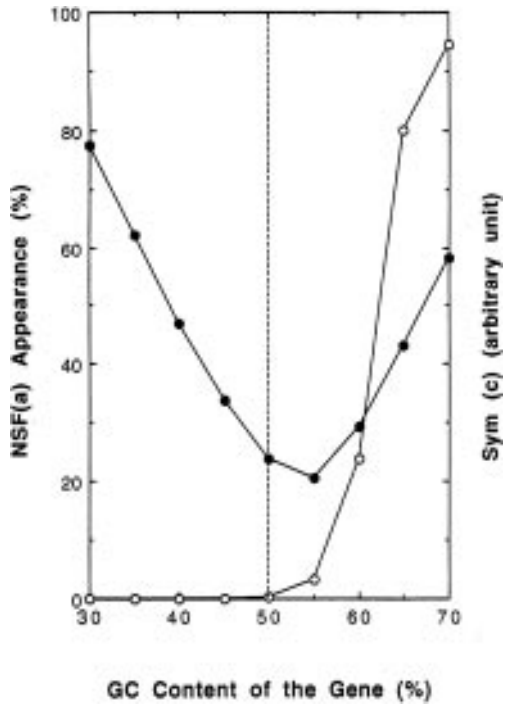

Figure 3. NSF(a) and Sym(c) against GC content of the genes. NSF(a) appearance (open circles) were estimated from the average $SC(a)_{obs}$ values observed on the actual genes described in Table 1 and Figure 2. Sym (c) values (closed circles) were calculated by equation **E3**. Dotted line shows the position of 50% GC content.

Properties of imaginary proteins encoded by computer-generated base sequences

It is important to investigate the properties of imaginary proteins encoded by computer-generated base sequences having the biased base compositions ~60% GC-rich genes, because it will make possible not only to confirm our conclusions, but also to create artificial genes from synthetic polynucleotides having random but biased base compositions in the codon. Thus, hypothetical base sequences with 60% GC content similar to the actual GC-rich genes were generated by a computer using a hand-written program and the random number table. We used the 12 straight lines ($E_{obs.}$ 1~12) described above to generate the hypothetical genes. Base compositions used at three positions were: G_1 = 39.6%, A_1 = 23.3%, C_1 = 22.2%, T_1 = 15.2%, G_2 = 19.7%, A_2 = 27.5%, C_2 = 26.4%, T_2 = 26.4%, G_3 = 30.0%, A_3 = 11.9%, C_3 = 42.9% and T_3 = 15.7%. By using the simple program, five hypothetical NSF(a) genes composed of 208 codons were generated and some properties of the computer-generated genes were examined (Table 2). First, it was ascertained that the average GC content of the hypothetical genes was expectedly ~60% (59.6) and the average SC(a) number was as small as 3.4, which was comparable with the value of seven actual genes (2.6 per 208 codons). Amino acid compositions of the hypothetical polypeptides encoded by both sense and antisense sequences on the computer-generated genes were compared with those of proteins on the seven bacterial genes having ~60% GC content (Fig. 4). Interestingly, it was found that the both amino acid compositions on the hypothetical proteins from the computer-generated base sequences were similar to those of proteins encoded by the actual bacterial genes. In other words, the amino acid compositions of functional proteins produced from the actual genes do not largely deviate from those of the hypothetical proteins from the computer-generated NSF(a) sequences. Thus, it is probable that artificial bacterial genes may be produced from the synthetic polynucleotides, which also supports our hypothesis on the origin of bacterial genes derived from GC-rich NSF(a). Actually, we are trying to produce the hypothetical, synthetic genes in our laboratory.

Antisense sequences of GC-rich bacterial genes can encode flexible polypeptide chains

As written in the above section, antisense GC-rich sequences with ~60% GC content can encode polypeptides composed of amino acid compositions similar to those of proteins encoded by the corresponding genes. In addition to that, the GC-rich NSF(a) has another advantage in creating novel genes. The possible genes

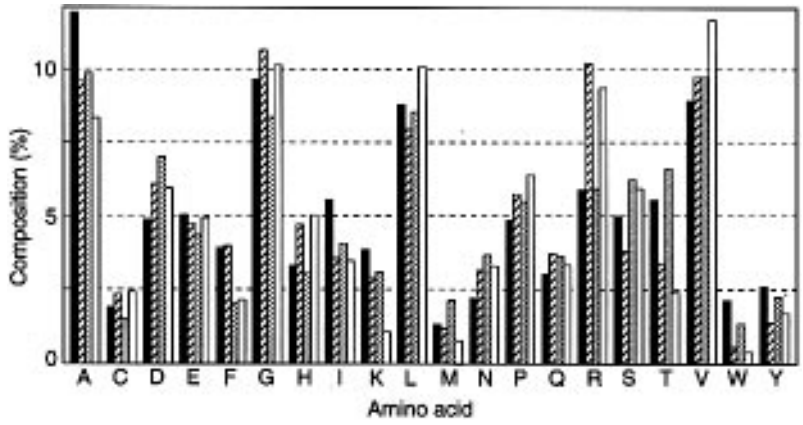

Figure 4. Comparison of average amino acid compositions of proteins encoded by bacterial genes* (black bars), by the GC-rich antisense sequences of bacterial genes (hatched bars), by sense (shaded bars) and antisense sequences (open bars) of the computer-generated base sequences with~60% GC content (Table 2), maintaining the same base compositions at each position in the codon as the actual bacterial genes having. *Seven genes selected from Table 1, having~60% GC content: *B.stearothermophilus ala.rac, T.pallidum pyr.red., M.tuberculosis* Mn-*sod, folA* and *thyA, D.vulgaris srd* and *R.capsulata cycA.*

could encode more flexible polypeptide chains than those from the extant genes, because the former polypeptides should contain more glycine, a nonbulky and slightly hydrophilic amino acid which is translated by GGN codons, than in polypeptides encoded by the corresponding GC-rich genes or by AT-rich sense and antisense sequences. Figure 5 shows that most of real proteins contained 7~9% of glycine on average, with a slight tendency to increase toward the high GC content. On the other hand, the glycine content of hypothetical proteins varied more and exceeded 10% on average in proteins from GC-rich NSF(a)s, although the values were very low in those from AT-rich NSF(a)s. Actual proteins have adequate flexibility of the structure, since they have moderate percentage of glycine, a less bulky amino acid. The large flexibility of newly-born enzymes must be important especially to accommodate three-dimensional structures of the proteins for adaptation of the structures to newly encountered substrates. Therefore, it is extremely important for a possible newborn polypeptide that GC-rich NSF(a) gives a high content of glycine, thus enough flexibility to the future enzyme.

Table 2. Some properties of the computer-generated genes having GC-content ~60%

Hypothetical gene[a]	GC (%)	Codon no.	SC(a)[b]	SC(s)[b]
Gene 115	60.7	208	2	3
Gene 376	58.8	208	5	5
Gene 753	61.9	208	3	1
Gene 826	60.6	208	3	5
Gene 952	56.1	208	4	2

[a]The hypothetical genes are arbitralily named as numbers used in generating the imaginary genes.
[b]SC(a) and SC(s) indicate numbers of stop codons on antisense and sense strands, respectively.

Proposition of a possible origin for newly created bacterial genes

Observations described in this paper are summarized as follows. (i) Probability of appearance of NSF(a) increases rapidly in the region with >55% GC content as GC content increases (Figs 2 and 3). (ii) Base compositions are nearly symmetrical between codons of the resulting polypeptides from both sense and antisense strands of the GC-rich bacterial genes ~55% GC content (Fig. 3). (iii) Amino acid compositions are nearly the same between those from sense and antisense strands ~60% GC content. This is also true in the hypothetical polypeptides derived from GC-rich sense and antisense sequences generated by a computer (Fig. 4). (iv) Hypothetical proteins from GC-rich NSF(a) contain much glycine, a less bulky amino acid, so that they may have enough flexibility for possible proteins from new genes (Fig. 5). Judging from the point of the symmetry of base compositions in the codon, the most favorable GC content may be ~55% for production of novel genes. However, the number of SC(a) and the glycine content suggest that the most promising GC content is ≥60% for creating new genes. Taking all the facts (i–iv) into consideration, we would like to propose here a working hypothesis that NSF(a) of the bacterial genes ~60% GC content are the most suitable sequences as a possible origin of newly-created bacterial genes.

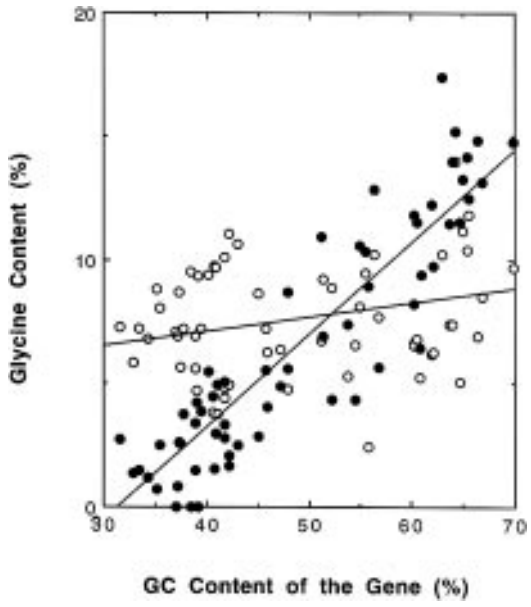

Figure 5. Glycine content against GC content of the genes. Open circles, proteins encoded by real genes. Closed circles, imaginary proteins encoded by the corresponding antisense sequences. Lines were drawn by the linear least-squares method.

We have recently found by investigation of the actual proteins encoded by functional genes and the hypothetical proteins from the possible GC-rich NSF(a) that several advantages for producing new bacterial genes are inherent in the GC-rich NSF(a) with GC content of ≥60% (Ikehara, *et al*., manuscript in preparation). The main results obtained are summarized as below. (i) Polypeptides encoded by GC-rich NSF(a) contain totally similar numbers of acidic and basic amino acids, therefore they are almost neutral. (ii) The hydrophobicity of the hypothetical protein from GC-rich NSF(a) is appropriate for folding them into soluble and globular three-dimensional structures, since total values of the hydrophobicity indices of the hypothetical proteins encoded by the GC-rich NSF(a)s were close to those of real proteins. (iii) Abilities for forming secondary structures (α-helix, β-structure and β-turn) in the hypothetical proteins were also similar to those in the actual proteins. These properties of the hypothetical proteins from GC-rich NSF(a) disclosed in further papers support our hypothesis of creating new bacterial genes described in this paper.

ACKNOWLEDGEMENT

We thank Dr Y. Takagi in Department of Biological Science, Faculty of Science, Nara Women's University for helpful discussion and suggestions in the course of this study.

REFERENCES

1 Ohno, S. (1970) *Evolution by Gene Duplication.* Springer, Heidelberg.
2 Langridge, J. (1991) *Molecular Genetics and Comparative Evolution.* Research Studies Press, Taunton, Somerset, UK, p. 216.
3 Shaw, D. C., Walker, J. E., Northrop, F. D., Barell, B. G., Godson, G. N. and Fiddes, J. C. (1978) *Nature* (London) **272**, 510–515.
4 Keese, P. K. and Gibbs, A. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 9489–9493.
5 Yomo, T., Urabe, I. and Okada, H. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 3780–3784.

6 Ikehara, K. and Okazawa, E. (1993) *Nucleic Acids Res*., **21**, 2193–2199.
7 Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., Hutchison, C. A. III, Slocombe, P. M. and Smith, A. J. (1977) *Nature*, **265**, 5596.
8 Sanger, F., Coulson, A. R., Friedman, T., Air, G. M., Barrell, B. G., Brown, N. L., Fiddes, J. C., Hutchison, C. A. III, Slocombe, P. M. and Smith, A. J. (1978) *J. Mol. Biol.*, **125**, 225–246.
9 Air, G. M., Coulson, A. R., Fiddes, J. C., Friedman, T., Hutchison, C. A. III, Sanger, F., Slocombe, P. M. and Smith, A. J. (1978) *J. Mol. Biol.*, **125**, 247–254.
10 Beck, E., Sommer, R., Auerswald, E. A., Kurz, C., Zink, B., Osterburg, G., Schaller, H., Sugimoto, K., Sugisaki, H., Okamoto, T. and Takanami, M. (1978) *Nucleic Acids Res.*, **5**, 4495–4503.
11 Fiers, W., Contreras, R., Haegemann, G., Rogiers, R., Van de Voorde, A., Van Heuverswyn, H., Van Herreweghe, J., Volkaert, G. and Ysebaert, M. (1978) *Nature*, **273**, 113–120.
12 Buchman, A. R., Burnett, L.and Berg, P. (1980) in Tooze, J. (ed.) *DNA Tumor Viruses*, 2nd edition. Cold Spring Harbor Laboratory, pp. 799–823.
13 Okada, H., Negoro, S., Kimura, H. and Nakamura, S. (1983) *Nature*, **306**, 203–206
14 Tsuchiya, K., Fukuyama, S., Kanzaki, N., Kanagawa, K., Negoro, S. and Okada, H. (1989) *J. Bacteriol.*, **171**, 3187–3191.
15 Kanagawa, K., Negoro, S., Takada, N. and Okada, H. (1989) *J. Bacteriol.*, **171**, 3181–3186.
16 Kakudo, S., Negoro, S., Urabe, I. and Okada, H. (1992) *J. Bacteriol.*, **174**, 7948–7953.
17 Sueoka, N. (1962) *Proc. Natl. Acad. Sci. USA*, **48**, 582–592.
18 Osawa, S., Jukes, T. H., Muto, A., Yamao, F., Ohama, T. and Angachi, Y. (1987) *Cold Spring Harbor Symp. Quant. Biol.*, **70**, 777–789.
19 Muto, A. and Osawa, S. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 166–169.