Nucleic Acids Research, 1993, Vol. 21, No. 9 2193–2199

Unusually biased nucleotide sequences on sense strands of *Flavobacterium sp.* genes produce nonstop frames on the corresponding antisense strands

Kenji Ikehara and Eriko Okazawa
Department of Chemistry, Faculty of Science, Nara Women's University, Kita-uoya-nishi-machi, Nara, Nara 630, Japan

Received December 1, 1992; Revised and Accepted March 25, 1993

ABSTRACT

From investigation of eight *Flavobacterium sp.* genes encoding enzyme proteins, it was found that six genes had nonstop frames (NSFs) on the antisense strands, and base sequences of the genes are mainly composed of repeating triplet sequence(s), 5'-GNC-3' (where G and C are guanine and cytosine, and N is either of the four bases), in the reading frames. Thus, we concluded that the biased nucleotide sequences on the sense strands produce NSFs on the corresponding antisense strands. Furthermore, from the precise alignments of both nucleotide and amino acid sequences of two related *Flavobacterium sp.* genes, *nylB* and *nylB'*, it was found that base replacements might have occurred symmetrically in the codons. That is, transversions between G and C were observed at high frequencies at the first and third positions of codons, but not at the second positions. At the first position, AG base transitions were observed much more than similar CT transitions, wheras CT transitions were found at the third positions at a relatively high frequency. These suggest that symmetrical base replacements in codons might be the main contribution to evolution in *Flavobacterium sp.* genes.

INTRODUCTION

It is generally believed that genes metabolizing by-products exhausted from modern manufacture are now evolving from some kind of ancestral genes, and several new genes encoding enzymes for modifying man-made products have actually been found, such as in *Flavobacterium* and *Pseudomonas* (1–4). Therefore, it is considered that mechanisms must exist for creation of the ancestral genes into new functional enzymes. But, little is known about the mechanism, how and from what kind of ancestral genes, new functional genes are created. In relation to this, gene duplication has been previously proposed to be a general mechanism for evolution of new genes encoding enzyme proteins (5). According to this theory, genes coding for new enzymes should be generated from the duplicated genes encoding enzymes catalyzing related reactions to new ones.

Contrary to this, Yomo et al. recently reported a novel attractive hypothesis that special mechanisms for producing and/or protecting nonstop frames (NSFs) on both the sense and the antisense strands may enable the nonfunctional NSF to evolve directly into new functional genes, as follows (6); The four genes for nylon oligomer-degrading enzymes (three related genes F-*nylB*, F-*nylB'* and P-*nylB*, from *Flavobacterium* and *Pseudomonas* species, and one *Flavobacterium* gene phylogenically independent of the *nylB* gene family), all contained a long stretch of sequence without chain-terminating base triplets (or NSF) on the antisense strand, and only one coding frame was open for both the sense and the antisense strands. A calculation of the probability of the presence of these NSFs on the antisense strands indicates that the presence of these NSFs are very rare and improbable. That is, the probability is very small (0.0001–0.0018), when the calculation was carried out on the assumption that four kinds of bases composing the *nyl* genes are randomly distributed on the DNA strands. In addition, according to another calculation, probability of the original NSF persisting in one of its descendants today is only 0.007. Thus, they suggested that there is some special mechanism for protecting these NSFs from mutation that generate stop codons, and that such a mechanism may enable NSFs to evolve into new functional genes. This may be a basic mechanism for the birth of new enzymes.

In order to inquire whether the new hypothesis provided by Yomo et al. is correct and to find a clue about how NSFs have been produced on the antisense strands of nylon oligomer-degrading genes and/or about what kind of forces have protected from random base replacements making stop codons on the antisense strands, we examined nucleotide sequences of one *Flavobacterium sp. nylA* gene (2) and eight *Flavobacterium* genes which were obtained from the most recent release of the GenBank data base (July, 1992) as full gene sequences encoding active enzymes. From the results, we unexpectedly found that six out of nine genes have NSFs on the antisense strands of not only

nylon oligomer-degrading genes (*nylB* and *nylB'*) but also housekeeping genes coding for, such as N-acyl-D-mannosamine dehydrogenase, and creatinase. It was also found that, except *endoF1* gene from *F. meningosepticum*, eight *Flavobacterium sp.* genes have had more or less the unusual and biased base composition at each position of codons in the reading frames, suggesting that the coding sequences of probably many *Flavobacterium sp.* genes are mainly composed of repeats of rather symmetrical triplet sequence(s), 5'-GNC/g-3' (where the capital letters, G and C, indicate the main bases at each position, and the small letter g means that guanine base is always successively less than cytosine at third position), typically 5'-GNC-3'. Thus, we would like to conclude that the unknown forces or mechanisms maintaining the unusually biased base compositions at the first and the third positions of the codons cause production and/or preservation of NSFs on the antisense strands of many *Flavobacterium* genes, not only of newly created genes but also of housekeeping genes. Furthermore, to investigate which kind of base replacements have occurred at each position of the codons, both nucleotide and amino acid sequences of two related genes, *nylB* and *nylB'*, were aligned with a computer. From the results, it was concluded that GC transversions occur at a high rate at the first and third positions, but not at the second position, and that CT transitions are observed relatively frequently at the third positions compared to first and second positions, although AG transitions were almost equally scored at three positions. Contrary to this, the second bases were well conserved during the gene evolution. This might imply that symmetrical base substitutions at the first and the third positions are the main motive force of the gene evolution in rather symmetrical *Flavobacterium sp.* genes. Based upon these results, a hypothesis for evolutional pathway of *Flavobacterium sp.* genes is proposed.

NUCLEOTIDE SEQUENCE ANALYSIS WITH COMPUTER

Collection of nucleotide sequences of *Flavobacterium* genes

As given in Table I, nine nucleotide sequences of *Flavobacterium* genes including one rRNA gene were collected from the recent GenBank data base (July, 1992) with the computer program, IDEAS-seqman, developed by Dr M.Kanehisa (Kyoto University, Institute for Chemical Research, Gokasho, Uji, Kyoto 606). One nylon oligomer-degrading gene, *nylA*, encoding 6-aminohexanoate-cyclic-dimer hydrolase was obtained from a recent paper (2). The collected genes are eight *Flavobacterium sp.* genes encoding creatinase (*cre*) (7), N-acyl-D-mannosamine dehydrogenase (*nam*) (8), isopenicillin-N-synthase (*pcbC*) (9), pentachlorophenol-induced periplasmic protein (*pcpA*) (3), parathion hydrolase (*opd*) (4) and nylon-oligomer-degrading genes (*nylB* (1), *nylB'* (1) and *nylA* (2)), one gene of *F. meningosepticum* encoding endoglycosidase F (*endoF1*) (10) and one 16S ribosomal RNA gene of *F. heparium* (11), as a control. The gene coding for creatinase is designated as *cre* for convenience in this study. Base composition at each position of codons in the genes was scored numerically by hand.

Alignment of both nucleotide and amino acid sequences of *Flavobacterium nylB* and *nylB'* genes

Both nucleotide and amino acid sequences of *Flavobacterium sp. nylB* and *nylB'* genes were separately aligned with the computer program, IDEAS- seqa and seqap, developed by Dr M.Kanehisa (Kyoto University). The aligned nucleotide sequences of the two genes were precisely related to the aligned amino acid sequences of the genes by hand. Thereafter, base replacements at every base position of the codons were numerically counted.

RESULTS AND DISCUSSION

Many *Flavobacterium sp.* genes have NSFs on the anti-coding strands

Flavobacterium are gram-negative, facultatively anaerobic bacteria and the primary characteristic of the genus is pigmentation. Some strains of the genus have the activities degrading synthetic pesticides exhausted from modern chemical industries. It is well known that the base composition of chromosomal DNA extracted from one subgroup of *Flavobacterium* is highly GC-rich (63 ~ 70%) and is different from the other subgroup, the base composition of which is highly AT-rich (30 ~ 42%).

The collected nine genes of *Flavobacterium* genes from the GenBank data base and one gene (*nylA*) from a paper are shown in Table I. Four genes (*cre*, *nam*, *pcbC* and *endoF1*) are housekeeping or ordinary genes, three *nyl* genes are probably new born genes for degrading nylon oligomers, as claimed by Yomo et al. (6) and the other two (*pcpA* and *opd*) are genes for metabolizing man-made chemical compounds other than nylon oligomers. As can be seen in the table, base compositions of eight genes from *Flavobacterium sp.* are all GC-rich and enzymes produced from nine genes are all middle-sized ranging from 271 to 493 amino acids. When the number of stop codons in the reading frame on the antisense strand (NSC(a)) was numerically scored, it was found that, surprisingly, six *Flavobacterium sp.* genes, including three housekeeping genes and two nylon oligomer-degrading genes (*nylB* and *nylB'*), have NSFs on the antisense strand, whereas unlike the *nylB* and *nylB'* genes, five stop codons were observed on the antisense strand of *nylA* gene. This means that NSFs are present on the antisense strands, of not only new functional genes (*nylB* and *nylB'*), but also ordinary genes (*cre*, *nam* and *pcbC*) produced probably in ancient days and, conversely, that newly evolved genes, such as *nylA* and *opd* genes, do not always have NSFs on the antisense strands. Therefore, it is conceivable that many genes in *Flavobacterium sp.* must have NSFs on their antisense strands.

Many *Flavobacterium sp.* genes have unusual base sequences composed of characteristic repeated triplets

To know the reason why so many *Flavobacterium sp.* genes have NSFs on the antisense strands, we investigated the base composition at every position of codons of all genes listed in Table I. As examples, partial nucleotide sequences of three genes (*nam*, *pcpA* and *nylB* genes) are shown in Fig. 1. As can be seen in the figure, many guanine bases are apparently observed at the first positions of codons in comparison to other bases, and much more cytosine bases are found at the third positions than the other three bases, especially adenine and thymine. Thus, we scored numerically each kind of base at three base positions of codons of ten *Flavobacterium* genes including a 16S rRNA gene, of which base compositions were examined over 1,500 base sequence (hypothetical 500 codons) at each position in three hypothetical reading frames. The results are summarized in Table II. Consequently, it was clear that every *Flavobacterium sp.* gene has exhibited unusual characteristics in base composition at the positions of codons, especially remarkable in the cases of six genes having NSFs on the antisense strands. Namely, in the eight

Table I. *Flavobacterium* genes examined in this study

Gene	No. of codons	GC (%)	NSC(a) (observed)	GeneBank Locus	Source
cre	378	64.1	0	FVBCRE	F. sp. U-188 (7)
nam	271	72.8	0	FVBNAM	F. sp. 141−8 (8)
pcbC	326	63.7	0	FVBPCBC	F. sp. (−) (9)
pcpA	271	61.6	0	FVBPCPA	F. sp. (−) (3)
opd	365	57.7	11	EVBOPD	F. sp.(ATCC27551) (4)
nylB	392	70.8	0	PO2RSA	F. sp. K172 (1)
nylB'	392	70.0	0	PO2RSB	F. sp. K172 (1)
nylA	493	64.7	5	(paper)	F. sp. K172 (2)
endoF1	339	39.0	13	FVBENDOF1A	F. meningosepticum (ATCC33958) (10)
rRNA	(500)	51.0	26 (1st) 25 (2nd) 19 (3rd)	FVBRGDA	F. heparium (11)

NSC(a)s on rRNA gene were counted in three hypothetical reading frames composed of 500 codons (1,500 bases from the first nucleotide position). Reference numbers are written in parentheses.

genes except *endoF1* gene from *F. meningosepticum* and 16 S rRNA gene from *F. heparium*, the base observed most plentifully at the first positions was always guanine, and the number of bases detected at the first positions, decreased in the order of cytosine or adenine and thymine, and contrary to that, the base observed the most frequently at the third position was always cytosine and scores of bases decreased in the order of guanine, thymine and adenine. On the other hand, the four bases are almost equally found at the second position. Particularly, in the case of six genes having NSF on the antisense strands, adenine (and thymine) content at the third position are as small as 3 to 4.5%, and cytosine content is as much as about 50 to 66% (Fig. 2). Of the other two genes (*opd* and *nylA*), similar characteristics about base compositions are also observed, though biases are less prominent than for the six genes above. The results described in Table II indicate that at least the *Flavobacterium sp.* genes examined in this study are mainly constructed from repeats of rather symmetrical triplet(s), 5'-GNC/g-3', typically 5'-GNC-3', on the reading frames, and inversely much less 5'-TNA/t-3' sequences are present than other sequences, and that the excess occurrences of 5'-GNC/g-3' triplets compensate for the reduced occurrences of AT-rich triplets.

Amino acid composition of proteins encoded by the *Flavobacterium* genes is given in Table III. As can be seen in the table, the protein composition is not unusually biased, except a considerably large amount of alanine and glycine found in many *Flavobacterium* proteins. Since all codons of two acidic amino acids, aspartic acid and glutamic acid, are composed of guanine-start codons, acidity of proteins must be one good criterion for judgement of bias of the protein composition. Judging from the ratios of acidic amino acids to basic ones (Table III), it can be also concluded that amino acid composition of *Flavobacterium* proteins is not unusually biased. These indicate that the unusual DNA composition can not be simply explained away as a function of the unusual amino acid composition of the proteins and that, in *Flavobacterium*, there is a mechanism maintaining the unusual DNA base composition in the codons.

In order to know whether the unusual base sequences described above are universal in *Flavobacterium* genes, or are limited to the genes coding for enzymes or functional proteins, we investigated the base sequence of a gene coding for *Flavobacterium* 16S ribosomal RNA (11), similarly as above.

Gene	Base Position	Nucleotide Base Sequences
nam		79 94 109 124
	1st	A G C G A G G G T C C G G G A G
	2nd	T G G C C T A C A T G A G C G T
	3rd	C C C C C C G C T T C G C C C G
		81 93 108 123
pcpA		325 340 355 370
	1st	G G C G C G T G A G A T G T A A
	2nd	G A C C C T C C G C T G T C G T
	3rd	C G G A C T G C C T G G T C C C
		327 339 354 369
nylB		379 394 409 424
	1st	G G G C A G C G C C G A G T G C
	2nd	A A G G T A C C A C T C A A T C
	3rd	C C G C C T C G G C C C G T A C
		381 393 408 423

Figure 1. Partial nucleotide sequences of three representative *Flavobacterium sp.* genes (*nam*, *pcpA* and *nylB*) described separately at each base position of codons. Numbers over and below the base sequences indicate nucleotide positions from the first bases of the genes.

From the results, it was clear that, unlike the cases of genes coding for proteins, the ribosomal RNA gene had almost equally distributed base sequences, resulting in three hypothetical reading frames containing from 19 to 26 stop codons on the antisense strand, as expected. As shown in Table IV, all of them are approximately the same as expected NSC(a) values calculated according to the equation (1) below.

From the above results, it could be expected that the unusual features of base sequences of *Flavobacterium sp.* genes have produced NSFs on the antisense strands of those genes, because the termination codons, TAA, TAG and TGA, on the antisense strand arise from the codons TTA, CTA and TCA on the sense strands, respectively. In other words, adenine base is always required at the third position on the sense strand to make stop codons on the antisense strand, but, as explained above, in the genes having NSFs on the antisense strands, the adenine bases are present only 3 to 4.5% at that position. To confirm this further arithmetically, we estimated expected the NSC(a)s by using the biased sequence data in the following procedure.

Now, consider that an ORF consists of N codons. And when the base composition of each kind of base at the nth position (n=1, 2 or 3) of codon on the sense strand is A_n, for an

Table II. Number of bases at each position of codon of *Flavobacterium* genes.

Gene	Base	Base Numbers Position of Codon 1st	2nd	3rd	Gene	Base	Base Numbers Position of Codon 1st	2nd	3rd
cre	A	90	122	17	*opd*	A	94	73	57
(378)	G	135	78	113	(365)	G	139	81	107
	T	52	91	35		T	56	109	74
	C	101	87	213		C	76	102	127
nam	A	58	47	8	*nylB*	A	63	99	14
(271)	G	139	64	73	(392)	G	166	93	139
	T	22	74	12		T	56	93	18
	C	52	86	178		C	107	107	221
pcbC	A	69	106	10	*nylB'*	A	69	101	16
(326)	G	115	47	122	(392)	G	162	87	130
	T	60	87	23		T	55	94	18
	C	82	86	171		C	106	110	228
pcpA	A	74	82	11	*nylA*	A	97	114	55
(271)	G	92	60	92	(493)	G	213	109	179
	T	44	69	32		T	62	130	64
	C	61	60	136		C	121	140	195
endoF1	A	112	114	93	*rRNA*	A	138	149	135
(339)	G	102	43	52	(500)	G	151	145	142
	T	67	90	144		T	104	99	110
	C	58	92	50		C	107	107	113

The numbers in parentheses show total amino acid numbers of the gene products. In the case of rRNA, number of hypothetical codons on 1,500 nucleotides is written in the parenthesis.

example in the case of adenine, the expected NSC(a) is expressed as follows.

$$NSC(a) = N\ (T_1T_2A_3 + T_1C_2A_3 + C_1T_2A_3) \qquad (1)$$

where T, C, and A in the parenthesis at the right side are thymine, cytosine and adenine, respectively. As can be seen in Table IV, the expected NSC(a) of the six genes are extremely small, from 0.82 to 2.21, compared to the values of 4.3 to 10 which are obtained by calculation, on assumption that the bases constructing genes are randomly distributed on the sense strand.

On the other hand, Yomo et al. have considered that the presence of NSFs on the antisense strands of four *nyl* genes is rare, based on their calculations of the probability of the presence of the NSFs with n or more codons on the three frames of the antisense sequence, ZN,n (6). But, they have assumed that four bases are randomly distributed on the reading frames without considering the biased base sequences which we found and describe above (Table II). So, taking the biased sequence data into consideration, we recalculated the probabilities of NSF appearance, ZN,N, of the ten genes according to equations 7 and 8 in their report (6). As given in Table IV, the probabilities (ZN,N) of the two nylon genes (*nylB* and *nylB'*) are not so small. Namely, we obtained the values 0.375 and 0.290, which are about 200 fold larger than the values calculated without the sequence data (0.0022 and 0.0016). This clearly mean that *nyl* genes as well as several other ordinary genes must have NSFs on the antisense strands as consequences of the unusually biased base sequences in the reading frames.

Similar conclusions can be also obtained from the results in Fig. 3 and Fig. 4. When the observed numerical scores of stop codons on the antisense strands and the expected number calculated by equation (1) are plotted on the vertical and horizontal axes, respectively, every spot fill around the diagonal line (Fig. 3), indicating that observed NSC(a)s on all genes appear

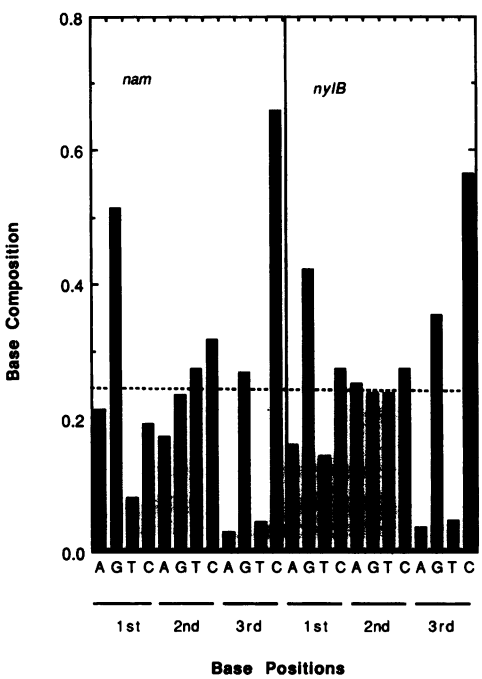

Figure 2. Base compositions at three positions of codons of two arbitrary selected *nam* and *nylB* genes. Horizontal broken line in the figure indicates the average base compositions at the base positions.

in proportion to the expected NSC(a)s. When the expected NSC(a)s with consideration of the biased sequence data were plotted against the expected NSC(a)s without consideration of the sequence data, spots of nine genes were located in the region upward from the diagonal line (Fig. 4). Particularly, six spots of the genes having NSFs were located on the line having a slope

Table III. Amino acid composition of *Flavobacterium* genes.

Amino Acid	*Flavobacterium* genes								
	cre	*nam*	*pcbC*	*pcpA*	*opd*	*nylB*	*nylB'*	*nylA*	*endoF1*
Ala	43	51	36	14	46	39	40	67	35
Cys	5	2	3	3	3	5	4	5	2
Asp	26	17	23	16	19	35	30	27	18
Glu	18	13	18	20	16	23	20	30	8
Phe	12	6	18	17	15	9	8	12	18
Gly	28	37	21	27	33	37	38	48	19
His	20	1	14	13	7	7	12	7	4
Ile	21	15	18	20	26	10	10	24	13
Lys	12	4	10	9	9	3	4	12	21
Leu	27	21	25	11	37	36	34	43	30
Met	11	11	8	8	6	6	8	9	7
Asn	17	6	17	7	6	8	10	15	35
Pro	10	10	23	13	14	21	21	30	17
Gln	13	1	6	7	11	13	12	16	10
Arg	33	21	17	17	28	31	28	32	10
Ser	19	11	19	16	28	26	25	34	29
Thr	21	17	11	24	27	28	31	24	21
Val	20	21	18	14	25	32	34	40	22
Trp	6	1	3	6	4	11	11	9	2
Tyr	16	5	18	9	5	12	12	9	18
Total No.	378	271	326	271	365	392	392	493	339
(Acidic)	44	30	41	36	35	58	50	57	26
(Basic)	65	26	41	39	44	41	44	51	35

(Acidic) and (Basic) mean sums of aspartic acid and glutamic acid, and histidine, lysine and arginine, respectively.

Table IV. Expected number of stop codons (NSC(a)) and probability of NSF (ZN,N) on antisense strands calculated with and without consideration of the biased sequence data described in Table II.

Gene	Biased Sequence Data					
	with Consideration			without Consideration		
	F	Expected NSC(a)	ZN,N	F	Expected NSC(a)	ZN,N
cre	0.0058	2.2	0.295	0.0264	10.0	0.0001
nam	0.0030	0.8	0.826	0.0160	4.3	0.0373
pcbC	0.0051	1.7	0.463	0.0270	8.8	0.0004
pcpA	0.0055	1.5	0.533	0.0298	8.1	0.0008
opd	0.0235	8.6	0.0005	0.0353	12.9	n.d.
nylB	0.0049	1.9	0.375	0.0182	7.1	0.0022
nylB'	0.0057	2.2	0.290	0.0191	7.5	0.0016
nylA	0.0149	7.4	0.0018	0.0257	12.7	n.d.
endoF1	0.0415	14.1	n.d.	0.0647	21.9	n.d.
rRNA	0.0345	17.3	n.d. (1st)	0.0379	19.0	n.d.
	0.0374	18.7	n.d. (2nd)			
	0.0416	20.8	n.d. (3rd)			

F is the average frequency of the appearance of one of the stop codons on the antisense strand (6) and n.d. means that calculated values are too small to describe in the columns.

Table V. Numbers of base replacements between *Flavobacterium sp. nylB* and *nylB'* genes at each base position of codon

Base Position	No. of Base Replacement						Total No.	Frequencies (%)
	Transition		Transversion					
	AG	CT	AC	GT	GC	AT		
1st	9	1	7	0	13	4	34	8.7
2nd	9	2	5	1	2	0	19	4.8
3rd	14	15	7	4	35	1	76	19.4
Total No.	32	18	19	5	50	5	129	11.0

Base replacements are not investigated in a region of nucleotide number from 710–735, because of frameshifting in the two genes.

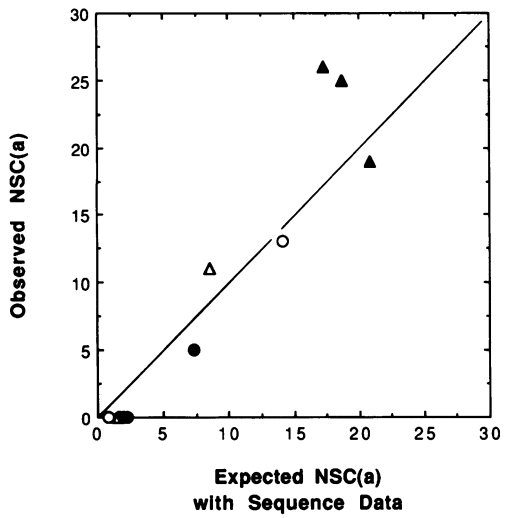

Figure 3. Plots of the observed NSC(a) against the expected NSC(a) calculated by equation (1) after consideration of the biased sequence data described in Table II. Open and closed circles indicate plots for the housekeeping genes and nylon oligomer-degrading genes, respectively. Open and closed triangles show genes for metabolizing man-made chemical compounds other than nylon oligomer and plots for hypothetical three reading frames on rRNA gene, respectively.

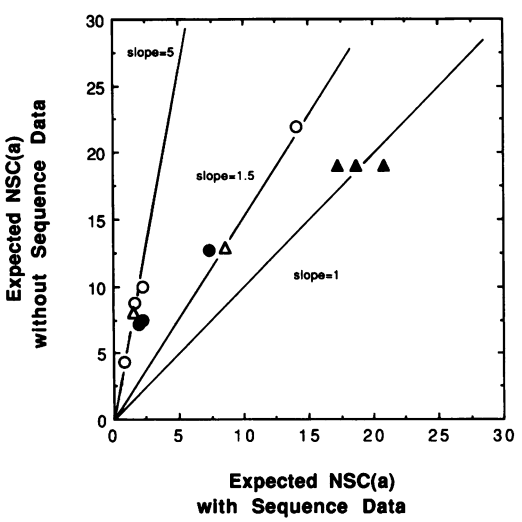

Figure 4. Plots of the expected NSC(a) calculated without consideration of sequence data described in Table II against the expected NSC(a) with consideration of the data. Symbols drawn in the figure are the same as those described in the legend of Fig. 3.

as large as 5. This means that base compositions in the reading frames of the six genes are similarly biased from random distribution. Therefore, unlike the conclusions described previously by Yomo et al. (6), we concluded that the existence of the NSFs on the antisense strands is not always limited to the newly created genes, such as *nylB* and *nylB'*, but many other genes also have NSFs, caused by the unusually biased base sequences on codons of *Flavobacterium sp.* genes. In other words, NSFs on the antisense strands covering a wide range of *Flavobacterium sp.* genes are not preserved by a (unknown) pressure that is protecting nonessential NSFs from random mutations, but by unknown forces that are maintaining unusual

Gene	Amino Acid and Nucleotide Sequences
nylB	7 G Q H P A R Y P G A A A 18 19 GGC CAG CAC CCC GCC AGG TAT CCC GGA GCC GCG GCC 55
nylB'	19 GGC TCG CAC CCC GCC CGA TAC CCC AGC GCC GCC GCG 55 7 G S H P A R Y P S A A A 18
nylB	19 G E P T L D S W Q E A P 30 56 GGG GAG CCG ACA CTC GAC AGC TGG CAG GAG GCC CCG 90
nylB'	56 GGT GAA CCG ACG CTC GAC AGC TGG CAG GAG CCC CCG 90 19 G E P T L D S W Q E P P 30

Figure 5. Alignments of partial amino acid and nucleotide sequences of *Flavobacterium sp. nylB* and *nylB'* genes. Numbers written at both sides of amino acid and nucleotide sequences are given to show amino acid positions from N-terminal of the gene products and nucleotide positions from the first base of the genes, respectively. Underlined letters indicate amino acid or nucleotide replacements.

repeating triplets, 5'-GNC/g-3', in the reading frame. If this is the case, as a matter of course, distributions of base dimers in *Flavobacterium sp.* genes should not agree with the universal rule for coding sequence construction stating that TA/CG base dimers are deficient and TG/CT dimers in excess, as reported by Ohno (12), because CG dimers must be rich in *Flavobacterium sp.* genes, since successive 5'-GNC-3' sequences make CG dimers between the triplets. In fact, excess of CG dimers were clearly observed in nylon oligomer-degrading genes (data not shown).

Characterization of base replacements between *Flavobacterium sp. nylB* and *nylB'* genes

To search for the causes of the unusual base sequences detected in *Flavobacterium sp.* genes and to know how the genes have changed their sequences and evolved from the ancestor genes, we aligned two related nylon oligomer-degrading genes (*nylB* and *nylB'*). The two genes are favorable to the above purposes, because they are not too distantly related to each other to align two entire sequences with the computer, nor too closely related to score appropriate number of base replacements between the genes. In addition to this, the two genes only were obtained from the GenBank as *Flavobacterium* genes related to each other. Both nucleotide and amino acid sequences of the two related genes could be aligned precisely using a computer program, IDEAS, seqa and seqap. The aligned partial nucleotide and amino acid sequences of the *nylB* and *nylB'* genes are shown in Fig. 5. So, one can now know which bases are different from each other and have been replaced from the common ancestor gene. Scores of all base replacements are summarized at every position of codons as described in Table V. From the results, several characteristics of the base substitutions emerge, as follows.

First, base replacements at the third position of codon have occurred the most frequently (19.4%), and GC transversions and AG, CT transitions are preferentially observed more than other base replacements at this position. This can be easily interpreted, as almost all base replacements at the third position, especially base transitions, are silent for amino acid substitutions and, thus, are neutral against natural selection. Second, substitutions at the first position (8.7%) are about two times as frequent as those at the second position (4.8%). This result may be attributed to more frequent GC transversions at the first position than at the second position. Third, interestingly, at the first and the second

positions, the AG base transitions were observed more often than CT transitions, in spite of the fact that both AG and CT transitions are almost equally scored at the third position. This may suggest that the AG transitions at the first and second positions contribute to the force of gene evolution in *nylB* and *nylB'* genes. Fourth, it is noteworthy that bases at the second position, where four kinds of bases are almost equally distributed (Table II), are replaced the least frequently (Table V), and that those at the third position, where the most biased base compositions were observed (Table II), are substituted the most frequently (Table V). These facts indicate that the second bases are certainly well conserved, since probabilities of accidental coincidence of base upon substitutions from an ancestor gene should be much larger at the third positions than at the second positions. Fifth, in spite of the fact that, in the genes of *Flavobacterium sp.*, adenine bases are rarely found at the third positions (Table II), only 2 out of 13 adenine bases (15.4%) in *nylB* gene were coincident with the same base, A, on the corresponding position in the *nylB'* gene. Contrary to this, 92 out of 98 adenine bases (93.9%) corresponded to the same base, A, at the second position. These also mean that second bases are well conserved during an evolutionary period and that a large part of adenine bases at the third positions should arise from the other bases, probably from guanine and cytosine bases. Sixth, it is considered that both *nylB* and *nylB'* genes have evolved from the common ancestor gene at a similar rate, since base replacements between two genes are relatively symmetrical (data not shown).

On evolution of *Flavobacterium sp.* genes

Before discussing a hypothetical pathway for evolution of genes in *Flavobacterium sp.*, we would like to summarize the results obtained in this study, briefly. (i) Many *Flavobacterium sp.* genes have NSFs on the antisense strands. (ii) Many *Flavobacterium sp.* genes coding for enzymes are composed of repeating rather symmetrical triplet sequences, 5'-GNC/g-3', typically as 5'-GNC-3'. Therefore, the presence of NSFs on the antisense strands should be attributed to the extremely biased base sequences on the sense strands. (iii) Transversions between G and C occur much more frequently at the third and the first positions than at the second position. (iv) CT transitions are observed relatively frequently only at the third position, but AG transitions are detected at similar frequencies at every position of the reading frame. (v) The second bases remain relatively unchanged during the evolutionary period. (vi) Except for CT transitions at the third position, base exchanges between thymine and the other bases are seldom observed.

Based on these findings, two models for evolutionary pathways of *Flavobacterium sp.* genes are considered *a priori*. One is a divergent model, in which a NSF or an ancestral gene having repeats of triplet sequences, as (5'-GNC-3')n, evolves into an active gene consisting of variable triplet sequences. Another is a convergent model, in which a random sequence changes into a new functional gene having repeats of triplet sequences, mainly composed of trimers as 5'-GNC/g-3'. But, for the following reasons, we propose the former model as a possible evolutionary pathway of *Flavobacterium sp.* genes. (i) If the latter model is correct, base substitutions between T and G should be prefered at the first position, since the most frequent and the least bases detected at the first position of codons are guanine and thymine, respectively, but this is not the case (Table V). (ii) Similarly, main base replacements at the third positions should be from A to C or G, and from T to C or G, but AC and TG transversions score relatively low (Table V). (iii) By the convergent model, it is difficult to explain the reason why so many GC transversions are observed at the first and the third positions of codons.

Contrary to this, the evolutionary pathway can be considered by the divergent model without these apparent contradictions, as follows. That is, (i) At first, *Flavobacterium* created double-stranded DNA composed of repeating symmetrical triplet sequences, as (5'-GNC-3')n, on a region of chromosomal or plasmid DNA. (ii) Occurrence of GC transversions at the first and third positions in triplets makes (5'-G/cNC/g-3')n sequence. (iii) In parallel, a small number of transitions from G to A and from C to T occurs symmetrically at the first and the third positions, respectively. Thus, until this stage, rather symmetrical nucleotide sequences in frame are maintained and, as a matter of course, NSFs are inevitably preserved on both strands in one coding frame. (iv) Meanwhile, base replacements at the second positions are repressed at a low level. (v) At a final stage, various bases are substituted at three positions to produce a new functional gene.

Recently, the complete nucleotide sequences of two *Escherichia coli* genes, *relA* and *spoT*, whose products function in the synthesis and degradation of guanosine 3', 5'-bispyrophosphate during the stringent response to amino acid starvation, have been determined (13, 14). In addition, it was reported that the two genes are extensively interrelated both with regard to amino acid and nucleotide sequences (15). So, we analyzed base composition of codons in the reading frames and aligned both the two sequences in order to score base replacements, as described above. Consequently, it was confirmed that guanine and cytosine bases were scored somewhat more than the other bases at the first and third positions of codons, respectively. When number of one base replacements in codons was scored, it was found that base replacements have occurred at the first positions about five times more than the second position and the replacements at the first and the third positions were mainly AG and CT transitions, respectively (Ikehara, K. et al., manuscript in preparation), like as in the case of *Flavobacterium nylB* and *nylB'* genes. The results of the *E.coli* genes suggest that the conclusions from *Flavobacterium sp.* genes might be also applicable to other bacterial genes.

REFERENCES

1. Okada, H., Negoro, S., Kimura, H. and Nakamura, S. (1983) *Nature*, **306**, 203–206.
2. Kanagawa, K., Negoro, S., Takada, N. and Okada, H. (1989) *J. Bacteriol.*, **171**, 3181–3186.
3. Xun, L. and Orser, C. S. (1991) *J. Bacteriol.*, **173**, 2920–2926.
4. Mulbry, W. W. and Karns, J. S. (1989) *J. Bacteriol.*, **171**, 6740–6746.
5. Ohno, S. (1970) *Evolution by Gene Duplication* (Springer, Heidelberg).
6. Yomo, T., Urabe, I. and Okada, H. (1992) *Proc. Natl. Acad. Sci.*, **89**, 3780–3784.
7. Koyama, Y., Kitao, S., Otake, H., Suzuki, M. and Nakano, E. (1990) *Agric. Biol. Chem.*, **54**, 1453–1457.
8. Yamamoto-Otake, H., Koyama, Y., Horiuchi, T. and Nakano, E. (1991) *Appl. Environ. Microbiol.*, **57**, 1418–1422.
9. Shiffman, D., Cohen, G., Aharonowitz, Y., von Dohren, H., Kleinkauf, H. and Mevarech, M. (1990) *Nucleic Acids Res.*, **18**, 660-660.
10. Tarentino, A. L., Quinones, G., Schrader, W. P., Changchien, L. M. and Pulummer, T. H. (1991) *J. Biol. Chem.*, **267**, 3868–3872.
11. Weisburg, W. G., Oyaizu, Y., Oyaizu, H. and Woese, C. R. (1985) *J. Bacteriol.*, **164**, 230–236.
12. Ohno, S. (1988) *Proc. Natl. Acad. Sci.*, **85**, 9630–9634.
13. Metzger, S., Ben-Dror, I., Aizenman, E., Schreiber, G., Toone, M., Friesen, J. D., Cashel, M. and Glaser, G. (1988) *J. Biol. Chem.*, **247**, 15699–15704.
14. Sarubbi, E., Rudd, K. E., Xiao, H., Ikehara, K., Kalman, M. and Cashel, M. (1989) *J. Biol. Chem.*, **264**, 15074–15082.
15. Metzger, S., Sarubbi, E., Glaser, G. and Cashel, M. (1989) *J. Biol. Chem.*, **264**, 9122–9125.