No stop codons in the antisense strands of the genes for nylon oligomer degradation
(newborn gene/nonstop frame/gene lifetime)

TETSUYA YOMO, ITARU URABE*, AND HIROSUKE OKADA†

Department of Biotechnology, Faculty of Engineering, Osaka University, 2-1 Yamadaoka, Suita-shi, Osaka 565, Japan

Communicated by Susumu Ohno, January 9, 1992

ABSTRACT Genes for nylon oligomer-degrading enzymes are unique in the sense that the enzymes encoded by them are found not to have any appropriate substrates during most of the period of their evolution. Furthermore, these nylon genes form a family not related to any other known gene families. The base sequences of these genes were examined and a common characteristic was found: a long stretch of sequence without chain-terminating base triplets, defined as a nonstop frame (NSF), is being maintained on the antisense strand. Moreover, a certain coding frame is open for both the sense and the antisense sequences, while the other frames have many stop codons. The probability of the presence of these NSFs on the antisense strand of a gene is very small (<0.0001–0.0018). In addition, another gene for nylon oligomer degradation was found to have a NSF on its antisense strand, and this gene is phylogenically independent of the nylon genes. Therefore, the presence of these NSFs is very rare and improbable. Even if the common ancestral gene of the nylon family was originally endowed with a NSF on its antisense strand, the probability of this original NSF persisting in one of its descendants of today is only 0.007. Unless an unknown force was maintaining the NSF, it would have quickly disappeared by random emergences of chain terminators. Therefore, the presence of such rare NSFs on all three antisense strands of the nylon gene family suggests that there is some special mechanism for protecting these NSFs from mutations that generate the stop codons. Such a mechanism may enable NSFs to evolve into new functional genes and hence seems to be a basic mechanism for the birth of new enzymes.

Classification of living organisms into genera, families, orders, classes, and phyla is thought to reflect the propinquity of their descendents. As a rule, genes too can be classified into families and superfamilies. Congruently, it has been thought that the prime driving force of molecular evolution lies in the mechanism of gene duplication (1). However, as this mechanism does not always rationalize the manifestation of the origin of a gene family, another mechanism is also likely to exist for the creation of the ancestral gene. Therefore, in search of this mechanism, genes harbored in the plasmids obtained from Flavobacterium sp. K172 and Pseudomonas sp. NK87 were analyzed.

Flavobacterium sp. K172 metabolizes by-products of nylon manufacture by using newly evolved enzymes, 6-aminohexanoate-cyclic-dimer hydrolase [EC 3.5.2.12] (F-EI) and 6-aminohexanoate-cyclic-dimer hydrolase [EC 3.5.1.46] (F-II) (2–4). The F-EI gene (F-nylA) and the F-EII gene (F-nylB) are on the plasmid pOAD2 harbored by this bacterium (5, 6). In addition, the F-EII gene (F-nylB)*, which is homologous to the F-EII gene, is present on the same plasmid (6, 7). The primary structures of the F-EII and F-EII' enzymes are very similar (88% identity), but the activity of F-EII' toward 6-aminohexanooate dimer is <1% of that of F-EII (7). Thus these two genes seem to be the descendants that diverged through the duplication of a common ancestral gene (7).

It is surprising that neither F-EII (2, 8) nor F-EII' (unpublished results) shows any activity toward natural amide compounds so far tested. This means that these enzymes had no appropriate substrates until several decades ago, when the manufacturing of nylon had begun. Furthermore, curing of pOAD2 does not inhibit the growth of Flavobacterium sp. K172 on a minimal medium supplemented with 6-aminohexanoate as the sole carbon and nitrogen source (5), thereby indicating that genes on pOAD2 are not essential for the growth of this bacterium. Therefore, it is likely that F-nylB and F-nylB' were nonessential genes until the accumulation of the by-products of nylon manufacture.

Pseudomonas sp. NK87 is another bacterial strain that grows on the by-products of nylon factories (9). This bacterium also has the same two enzymes, P-EI and P-EII, but the P-EI gene (P-nylA) and the P-EII gene (P-nylB) are on different plasmids (9).

In this paper, we show that these nylon genes have a long stretch of sequence without stop codons (a nonstop frame, NSF) on their antisense strands. This feature might lead us to understand the mechanism behind the emergence of new genes.

RESULTS AND DISCUSSION

Genes for Nylon Oligomer-Degrading Enzymes Have NSFs on Their Antisense Strands. The nucleotide sequences of F-nylB and F-nylB' (7) and P-nylB were examined. As described in the next section, the homology between the amino acid sequences of F-EII and P-EII is 37% and is much lower than the homology between F-EII and F-EII'. Interestingly, it was found that the antisense strands of these genes bear NSFs (anti-F-nylB, anti-F-nylB', and anti-P-nylB, respectively). The regions wherein these NSFs are located just coincide with those of the corresponding open reading frames (ORFs) on the sense strands, shown as arrows in Fig. 1. Furthermore, the reading frames of these NSFs (the first lines in Fig. 1a–c) share the same triplets as those for the ORFs. Thus, the presence of anti-P-nylB in addition to anti-F-nylB and anti-F-nylB' strongly suggests that these nylon genes might have evolved from a special region of the DNA.

Another nylon oligomer-degrading enzyme, EIII, is also encoded on pOAD2. The ORF corresponding to EIII was found in the upstream region of F-nylB (Fig. 1a). Although no

Abbreviations: ORF, open reading frame; NSF, nonstop frame.
*To whom reprint requests should be addressed.
†Present address: Department of Applied Microbial Technology, The Kumamoto Institute of Technology, 4-22-1 Ikeda, Kumamoto 860, Japan.
‡The sequences reported in this paper have been deposited in the GenBank data base (accession numbers D10678 [P-nylB] and D10686 [EII gene]).

3780
homology was found between the deduced amino acid sequences of the EII and EIII enzymes, similar results as for the nylB genes were obtained upon analyzing the nucleotide sequence of the EIII gene: There is a NSF on the antisense strand of the EIII gene, the location of the NSF just coincides with the ORF for EIII, and the reading frame for the NSF shares the same triplet as that for the ORF (Fig. 1a). These results imply that there may be some unknown mechanism behind the evolution of these genes for nylon oligomer-degrading enzymes.

The amino acid sequences deduced from the nucleotide sequences of anti-F-nylB, anti-F-nylB', anti-P-nylB, and the NSF on the antisense strand of the EIII gene show that the codon usages of these NSFs are similar to those of F-nylB, F-nylB', P-nylB, and the EIII gene, respectively (Fig. 2). The codon usages of these ORFs and NSFs follow the symmetry rule on base sequences; i.e., the degree of excesses and deficiencies of the base oligomers remained very similar between two complementary strands (10).

**Phylogeny of nylB Family and Anti-nylB Family.** The homology between the deduced amino acid sequences of F-nylB, F-nylB', and P-nylB is shown in Fig. 3a (numbers above the diagonal). Though the homology between P-nylB and F-nylB (or F-nylB') is not high, it was confirmed to be significant by the method of Pearson and Lipman (11), indicating that these nylB genes are evolutionarily related. This strongly suggests that the nylB gene family has diverged from a common ancestral gene. In addition, the evolutionary distances \(d_k\) between F-nylB, F-nylB', and P-nylB were calculated as the number of amino acid substitutions per site by the method of Kimura (12) (Fig. 3a, numbers below the diagonal). The distance between P-nylB and F-nylB (or F-nylB') is much larger than that between F-nylB and F-nylB'. The time of the divergence of F-nylB and P-nylB is estimated to be at least 1.4 \(\times 10^8\) years ago, using a very high amino acid-substitution rate of 9 \(\times 10^{-9}\) per site per year for the fibrinopeptide (13). Therefore, most of the amino acid substitutions from the ancestor of the nylB gene family to its descendants of today might have occurred before the beginning of nylon manufacture.

Fig. 3b shows the homology and the evolutionary distances between anti-F-nylB, anti-F-nylB', and anti-P-nylB obtained from their deduced amino acid sequences. Like the nylB gene family described above, these anti-nylB genes are also likely

---

**Fig. 1.** Distribution of the stop codons on the antisense strands of the nylB gene family. The three horizontal lines (from 5' to 3') for each antisense strand of the F-nylB and EIII gene (a), F-nylB' (b), or P-nylB (c) including flanking regions show the three kinds of coding frames of each sequence: the top lines correspond to the native coding frames of the nylB family and the EIII gene, and the middle and the bottom lines correspond to frames shifted 1 base forward and backward, respectively, from the top frame. The vertical lines on these horizontal lines show the position of stop codons. Arrows show the position of the ORFs on the sense strands for F-EII and EIII (a), F-EII' (b), and P-EII (c). The lengths of the genes are as follows: anti-F-nylB, 392 codons; anti-F-nylB', 392 codons; anti-P-nylB, 394 codons; and the NSF of the EIII gene, 355 codons.

---

**Fig. 2.** Symmetry of codon usage. The number of occurrences of a codon observed in all the ORFs of F-nylB, F-nylB', P-nylB, and the EIII gene (total, 1535 codons) is plotted against the number of occurrences of the same codon observed in all the NSFs on the antisense strands of the above genes (total, 1533 codons). Stop codons were omitted.

---

**Fig. 3.** Amino acid substitution matrix of nylB (a) and anti-nylB (b) families. Numbers above the diagonal are homologies (H, % identity) between amino acid sequences deduced from the nucleotide sequences, and the numbers below the diagonal are evolutionary distances \(d_k\). Amino acid sequences were aligned by finding the local similarities and were confirmed to be statistically significant by the improved version of programs in the FASTA program package (11). For the calculation of H, deletions and insertions were excluded. The number of amino acids compared are shown in parentheses. The \(d_k\) values were calculated as the number of amino acid substitutions per site from the following equation (12): \(d_k = -\ln[H - 0.2(1 - H)]\).
to have a common ancestral gene and form a family. Moreover, the evolutionary distances between the members of the anti-nylB family are significantly larger than the corresponding distances for the nylB family. These results suggest that the evolutionary rate of the anti-nylB family is faster than that of the nylB family, as the time of divergence is the same for both families. This may be due to the fact that the products of the nylB family have a positive function of degrading nylon oligomers at present.

When protein sequence data banks (Protein Identification Resource Release 29.0 and Swiss-Prot Release 18.0) were searched, no proteins homologous to the expected product of anti-nylB genes and the NSF on the antisense strand of the EIII gene were found. Likewise, F-EII, F-EII', P-EII, and EIII enzymes also showed no homology to the proteins cited in the data bank. Therefore, we conclude that the nylB gene family and the anti-nylB gene family are not related to any known gene families and might have evolved from a common ancestral gene.

**Probability of the Presence of a NSF on the Antisense Strand.**

The presence of a long NSF in the antisense strand seems to be a rare case, but it may be due to the unusual characteristics of the genes or plasmids for nylon oligomer degradation. One of the possible candidates that may increase the chance for the presence of a long NSF is the high G+C content of these genes (see Table 1). We estimated the probability of the presence of such long NSFs by using the sequence data of these genes in the following procedure.

Let us consider that an ORF consists of \( N \) codons. There are several NSFs with different length on its antisense strand, and the longest one consists of \( n \) codons. When \( n \) is larger than \( N/2 \), the probability for the longest NSF to be present on a frame of the antisense sequence is expressed as follows. For \((N - 1)/2 < n \leq N - 1\),

\[
P_{N,n} = 2F(1 - F)^n + F^2(1 - F)^n(N - 1 - n),
\]

and for \( n = N\),

\[
P_{N,N} = (1 - F)^N,
\]

where \( F \) is the average frequency of the appearance of one of the stop codons (TAA, TAG, and TGA) on the antisense strand, the first term of the right side of Eq. 1 is the probability for the NSFs delimited by one edge of the ORF and one stop codon (i.e., they are located at the ends), and the second term of the right side of Eq. 1 is the probability for the NSFs delimited by two stop codons. From Eq. 1,

\[
P_{N,n+1} - (1 - F)P_{N,n} = -F^2(1 - F)^{n+1}.
\]

This equation becomes

\[
\sum_{i=1}^{N-n} P_{N,n+i} - (1 - F) \sum_{i=0}^{N-n} P_{N,n+i}
= -F^2(1 - F)^n \sum_{i=1}^{N-n} (1 - F)^i
= -F(1 - F)^{n+1}[1 - (1 - F)^{N-1-n}].
\]

On the other hand, the sum of the probabilities for the NSFs to consist of \( n \) or more codons is defined as

\[
S_{N,n} = \sum_{i=n}^{N} P_{N,i}.
\]

Transforming the left side of Eq. 4 using Eq. 5,

\[
S_{N,n} = P_{N,n} - P_{N,n+1} - (1 - F)(S_{N,n} - P_{N,n+1} - P_{N,n})
= -F(1 - F)^{n+1}[1 - (1 - F)^{N-1-n}] + 1.
\]

Rearranging Eq. 6,

\[
S_{N,n} = (1 - F)^n[1 + n - (1 - F)(N - n)].
\]

Thus, the probability \((Z_{N,n})\) of the presence of the NSFs with \( n \) or more codons on the three frames of the antisense sequence can be estimated from Eqs. 7 and 8.

\[
Z_{N,n} = 1 - (1 - S_{N,n})^3
\]

The values of the average frequency \((F)\) of one of the stop codons to appear in the antisense sequences of the genes for F-EII, F-EII', P-EII, and EIII were calculated by using the base composition of each antisense strand (Table 1). We also estimated the frequency from the sequence data of a 12.144-base-pair fragment of pOAD2 (unpublished results); the F-EII gene is located on this fragment. This estimation was done because there may be such a selection bias in the composition of base triplets in the nucleotide sequence of this plasmid as to increase the frequency. In this calculation, the total number of occurrences of the three termination triplets (TAA, TAG, and TGA) in all the six triplet frames of the fragment was divided by the total number of all the triplets examined. The value of the frequency thus obtained is 0.020 per codon and is almost the same as those obtained from the base composition (Table 1). This means that there is no special bias in the composition of the base triplets in the sequence of this pOAD2 fragment. Therefore, we used the \( F \) values listed in Table 1 for the estimation of the \( Z_{N,n} \) values for the antisense strands of the four ORFs from Eqs. 7 and 8, and the results are shown in Table 1. These values indicate the rareness of the NSFs observed above. On the other hand, the expected number of stop codons in a sequence with a given length and an \( F \) value can also be calculated; the value for the antisense strand of each ORF is included in Table 1. These values, except for that of EIII, coincide well with the average numbers of stop codons observed in the two frames not having the NSFs (see Table 1). This means that the estimated \( F \) values shown in Table 1 are reasonable. These results clearly show that the presence of NSFs such as those found on the antisense strands of the three nylB genes and the EIII gene is rare. In addition, the nylB genes and the EIII gene are concerned with nylon oligomer degradation but are not phylogenetically related, as described above. Therefore, the events that occurred on the EIII gene and one of the nylB genes are independent, and the presence of all these NSFs is very rare and improbable. Accordingly, the actual existence of these NSFs leads us to speculate that some special mechanism exists in the regions of these genes.

The rareness of the observed NSFs can also be shown from an evolutionary point of view. It has been assumed that ORFs encoding essential proteins are usually conserved but that nonessential ones are not. If nonessential ORFs are not

<table>
<thead>
<tr>
<th>G+C, %</th>
<th>ORF</th>
<th>( N )</th>
<th>( n )</th>
<th>( F )</th>
<th>( Z_{N,n} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F-EII</td>
<td>392</td>
<td>392</td>
<td>0.019</td>
<td>0.0018</td>
</tr>
<tr>
<td></td>
<td>F-EII'</td>
<td>392</td>
<td>392</td>
<td>0.020</td>
<td>0.0010</td>
</tr>
<tr>
<td></td>
<td>P-EII</td>
<td>396</td>
<td>394</td>
<td>0.025</td>
<td>0.0003</td>
</tr>
<tr>
<td></td>
<td>EIII</td>
<td>355</td>
<td>355</td>
<td>0.017</td>
<td>0.0080</td>
</tr>
</tbody>
</table>

Values of \( Z_{N,n} \) were calculated from Eqs. 7 and 8 by using the corresponding average frequency \((F)\) of a stop codon.
conserved, random mutation will lead to the occurrence of the stop codons in the ORFs. However, in the case of the nylB and anti-nylB families, their ORFs and NSFs are kept open, even though none of the genes on pOAD2 have been found to be essential for the growth of the host strain, as described earlier. The improbability of the presence of these NSFs (and ORFs) in all the nylB gene family members was evaluated by estimating the lifetime of a nonessential NSF with n codons under random mutation.

After suffering m mutations (nucleotide substitutions) in the NSF, the probability \( P_{m,i} \) of the presence of i stop codons in the frame is expressed as

\[
P_{m,i} = [(n - i + 1)R/n]P_{m-1,i-1} + [(n - i)(1 - R)/n + i(1 - Q)/n]P_{m-1,i} + [(i + 1)Q/n]P_{m-1,i+1} \quad (1 \leq m, 0 \leq i \leq n), \tag{10}
\]

where \( Q \) is the probability for the change of a stop codon to a nonstop codon by one mutation, and \( R \) is the probability for the change of a nonstop codon into a stop codon by one mutation. The value of \( Q \) is calculated to be 23/27. When the distribution of the number of stop codons is at an equilibrium state, then \( R \) is estimated from the following equation:

\[
R(1 - F) = Q \cdot F, \tag{11}
\]

where \( F \) is the average frequency of appearance of one of the stop codons.

The fate of a nonessential NSF on the antisense strand of a gene with 392 codons \( (n = 392) \) was calculated from Eq. 10 under the conditions that the NSF is suffering random mutation but the corresponding ORF of the gene is kept open during the mutation. In this case, the \( F \) value depends on the codon usage of the genome in which the ORF is present. As the termination codons, TAA, TAG, and TGA, in the antisense sequence arise from the codons TTA, CTG, and TCA, respectively, in the sense sequence of the gene, then the average frequency that one of the chain terminators on the antisense strand will emerge in the same coding frame as the ORF of the gene is equivalent to the average frequency of the appearance of one of the codons (TTA, CTG, and TCA) in the sense sequence of the gene. From Table 2 of codon usage (14), the average frequency of one of the codons (TTA, CTG, and TCA) to appear in the genes of Escherichia coli, Salmonella typhimurium, and Bacillus subtilis is 0.018, 0.021, and 0.039 per codon, respectively. As the value of 0.018 is smaller than the others, and as it is almost the same as the \( F \) values estimated from the base composition (Table 1), we used this value for the calculation of \( R \). Thus, the \( R \) value for a NSF on the antisense strand of a gene is estimated to be 0.016 from Eq. 11. Fig. 4 shows the fate of the NSF thus obtained. The number of mutation per codon \( (m/n) \) on the abscissa is related to the number of amino acid substitutions per site \( (d_{a}) \) in Fig. 3 by the equation

\[
m/n = d_{a}/(0.71 \times 2), \tag{12}
\]

where 0.71 is the frequency of nonsynonymous substitution in a codon by random mutation (15). Fig. 4 shows that the probability of the presence of the remaining NSF decreases sharply, frames having one stop codon appear transiently, and finally almost all the frames appear to have two or more stop codons. Fig. 4 also shows that half of the NSFs initially present disappear after 0.12 mutation per codon, and the probability of maintaining the frame open until one present descendant from the common ancestral NSF of anti-F-nylB and anti-F-nylB' is 0.29. Furthermore, if the common ancestral gene of the nylB family was endowed with a NSF on its antisense strand, the probability for its persistence in one of the descendants of today is only 0.007. Of course, the probability is much less if the ancestral gene did not have the NSF.

Fig. 4 also shows the fate of an NSF \( (n = 392) \) that is not on the antisense strand of a gene. In this case, an \( F \) value of 0.253 \( \times 3 \) was used for the calculation because there is no limitation for the codon usage in the complementary sequence. The lifetime of such a NSF is much less under random mutation due to the absence of the limitation for the codon usage.

These results indicate that the lifetime of a nonessential NSF is very short, and it is impossible for such a NSF to persist for a long period of evolution. Therefore, we strongly suggest that the existence of the NSFs on all the three antisense strands of the nylB gene family points to an unknown force that is preserving these nonessential NSFs; otherwise, they would have quickly disappeared by random emergences of chain terminators.

**Keeping NSFs Open Is a Basic Mechanism of Enzyme Evolution.** It has been shown that the apparently nonessential NSFs on the antisense strands of the nylB gene family have been protected for a long time from mutations that generate the stop codons. The same mechanism seems to have worked on the sense strand of the nylB gene family, because these genes appear to have become important only after the accumulation of the by-products of nylon manufacture; before that time, they, as well as their antisense NSFs, seem to have been nonessential genes. Though the time at which their gene products obtained the new enzyme activity is still unknown, a functional but not necessarily important ancestral nylB gene is found to exist on a NSF that had been protected for a long time against heavy mutation.

We suppose that the mechanisms described above are working not only on the sense and antisense strands of the nylB gene family but also on the regions where new genes are going to be born. In fact, it will take a long time until an NSF has evolved into a new gene. As shown in Fig. 4, nonessential NSFs are apt to disappear during mutations, which are the actual steps of evolution. Therefore, the mechanism of preserving nonessential NSFs always open is essential for the birth of new enzymes.

Gene duplication has been proposed to be a general mechanism for evolution of new enzymes (1). Indeed, this is one of the important mechanisms for making nonessential NSFs, which are the starting points for new enzymes. However, without the mechanism of maintaining the duplicated gene open, it seems improbable for the gene to become a new gene.
with quite a different sequence. For example, when the homology of the sequences between the duplicated gene and its original one with 392 codons becomes 50%, the probability of the duplicated gene still occurring is only $6.8 \times 10^{-6}$ (Fig. 4). Therefore, we conclude that preparing new NSFs and keeping the NSFs open are basic mechanisms for the occurrence of new enzymes.

In nature, there may be many nonessential but persisting NSFs, whether or not they are expressed as nonfunctional or functional proteins. The number of examples of such nonessential NSFs will increase greatly as the nucleotide sequences of the whole genomes of *E. coli*, rice, and humans are read. Although it is considered that not all of the NSFs in the present sequences have been or will be kept for a long time, the presence of a DNA region where both strands have NSFs, as shown in this study, must be good evidence that a special mechanism of preserving NSFs is working on the DNA region. Therefore, it is possible that such a double-NSF region may lead to the making of a new enzyme; consider that the *nylB* genes appeared on such double NSFs. Furthermore, the anti-*nylB* family may come to encode new enzymes in the near future, or the *nylB* family itself might be born from NSFs on the antisense strands of functional genes similar to the anti-*nylB* family. It follows that the NSFs on the antisense strands are sources of new enzymes. Transcription of the antisense DNA strands of coding regions has been reported in eukaryotic cells (16–19). If these NSFs are also transcribed and then translated, the proteins may have functions, such as catalysis, recognition of molecules, and so on (20, 21).

We thank Dr. Susumu Ohno for helpful discussion. We thank Mrs. Elizabeth Ko Mitamura for correcting English.