Proc. Natl. Acad. Sci. USA
Vol. 81, pp. 2421–2425, April 1984
Evolution

Birth of a unique enzyme from an alternative reading frame of the preexisted, internally repetitious coding sequence

(oligomeric repeats/nylon oligomer degrading enzymes)

SUSUMU OHNO

Beckman Research Institute of The City of Hope, Duarte, CA 91010

Contributed by Susumu Ohno, January 4, 1984

ABSTRACT **The mechanism of gene duplication as the means to acquire new genes with previously nonexistent functions is inherently self limiting in that the function possessed by a new protein, in reality, is but a mere variation of the preexisted theme. As the source of a truly unique protein, I suggest an unused open reading frame of the existing coding sequence. Only those coding sequences that started from oligomeric repeats are likely to retain alternative long open reading frames. Analysis of the published base sequence residing in the pOAD2 plasmid of *Flavobacterium* Sp. KI72 indicated that the 392-amino acid-residue-long bacterial enzyme 6-aminohexanoic acid linear oligomer hydrolase involved in degradation of nylon oligomers is specified by an alternative open reading frame of the preexisted coding sequence that originally specified a 472-residue-long arginine-rich protein.**

In analogy with the immortal dictum "*omnis cellula a cellula*" of Rudolph Virchow (1821–1902), it might be said that all of the new genes arose from redundant copies of the preexisted genes (1). The mechanism of gene duplication, however, is inherently self limiting in that a new protein arising by this mechanism invariably retains substantial amino acid sequence homology and, therefore, functional relatedness with its immediate ancestor. Thus, one wonders if this mechanism alone sufficed at the very beginning of life when a large variety of polypeptide chains with divergent functions had to be created almost simultaneously. The same can be said of this 20th century when a variety of microorganisms suddenly found themselves facing an onslaught of man-made artificial compounds. It has recently occurred to me that the gene started from oligomeric repeats at its certain stage of degeneracy (base sequence diversification) can specify a truly unique protein from its alternative open reading frame. In this paper, I shall show that this mechanism can be invoked to explain the sudden birth of a plasmid-encoded bacterial enzyme, 6-aminohexanoic acid linear oligomer hydrolase (6-AHA LOH) that degrades nylon oligomers (2–4).

Three Virtues of Oligomeric Repeats as the Primordial Coding Sequence

Recently, a number of investigators independently arrived at the conclusion that all polypeptide chains were originally endowed with short periodicities, thus implying original internal repetitiousness in all coding sequences (5–7). There are three virtues that uniquely qualify oligomeric repeats as the primordial coding sequence (8). The first is their translatability to polypeptide chains of substantial lengths. Under the universal coding system with three chain-terminating base triplets, a fraction of the n-base-long randomly generated base sequences having n-base-long open reading frames is represented by $(^{61}/_{64})^{n/3}$. Thus, only 0.819% of the 4^{300} variety of 300-base-long randomly generated base sequence shall have 100-codon-length open reading frame with regard to one particular phase, while the other two reading frames shall be loaded with their customary share of chain-terminating codons. Furthermore, a chance of a chain initiator A-T-G occurring in phase with and near the 5′ terminus of that one particular open reading frame is not particularly good. The situation is far more favorable with regard to repeats of base oligomers. Provided that the number of bases in the oligomeric unit is not a multiple of 3, three consecutive copies of it translated in three different reading frames constitutes the translation unit of such oligomeric repeats. Accordingly, $(^{61}/_{64})^{3n/3}$, simplified to $(^{61}/_{64})^{n}$, the fraction of the repeats of n-base-long randomly generated base oligomers, shall have not one, but all three, open reading frames which equal the total length of repeats—e.g., 59.14% of the monodecameric repeats shall have all three reading frames open for indefinite length.

The second virtue of oligomeric repeats as the primordial coding sequence is found in the periodicity they give to polypeptide chains they specify. Such periodicities are quite conducive to the formation of either α-helical or β-sheet secondary structure.

The third and probably the most important virtue of oligomeric repeats as the primordial coding sequence is found in their inherent imperviousness to randomly sustained base substitutions, deletions, and insertions. Missense base substitutions are no concern to them, as each merely disturbs one of the many identical periodicities. Furthermore, provided that the number of bases in the oligomeric unit is not a multiple of 3, they should be totally impervious to deletions and insertions. Customarily, deletions or insertions of bases that are not multiples of 3 are very deleterious to the coding sequence as they cause frame-shifts, thereby completely altering downstream amino acid sequences and usually ending up in premature chain terminations. The very fact that three consecutive copies of the oligomeric unit of these repeats are translated in three different reading frames insures that such deletions and insertions shall cause only a local perturbation, the original periodicity resuming shortly thereafter. Customarily, rarest but most destructive of the base substitutions is the type that changes an amino acid specifying codon in the midst of coding sequences to a chain terminator, the expected ratio between missense, samesense, and chain-terminating base substitutions being 2.77:1:0.15. Inasmuch as these oligomeric repeats are open for all three reading frames to yield a set of three polypeptide chains with the originally identical periodicity, even this most destructive base substitution can damage only one-third of their coding potentials.

What if the modern coding sequence ultimately derived from oligomeric repeats still retained a sufficient degree of internal repetitiousness and, therefore, an alternative open

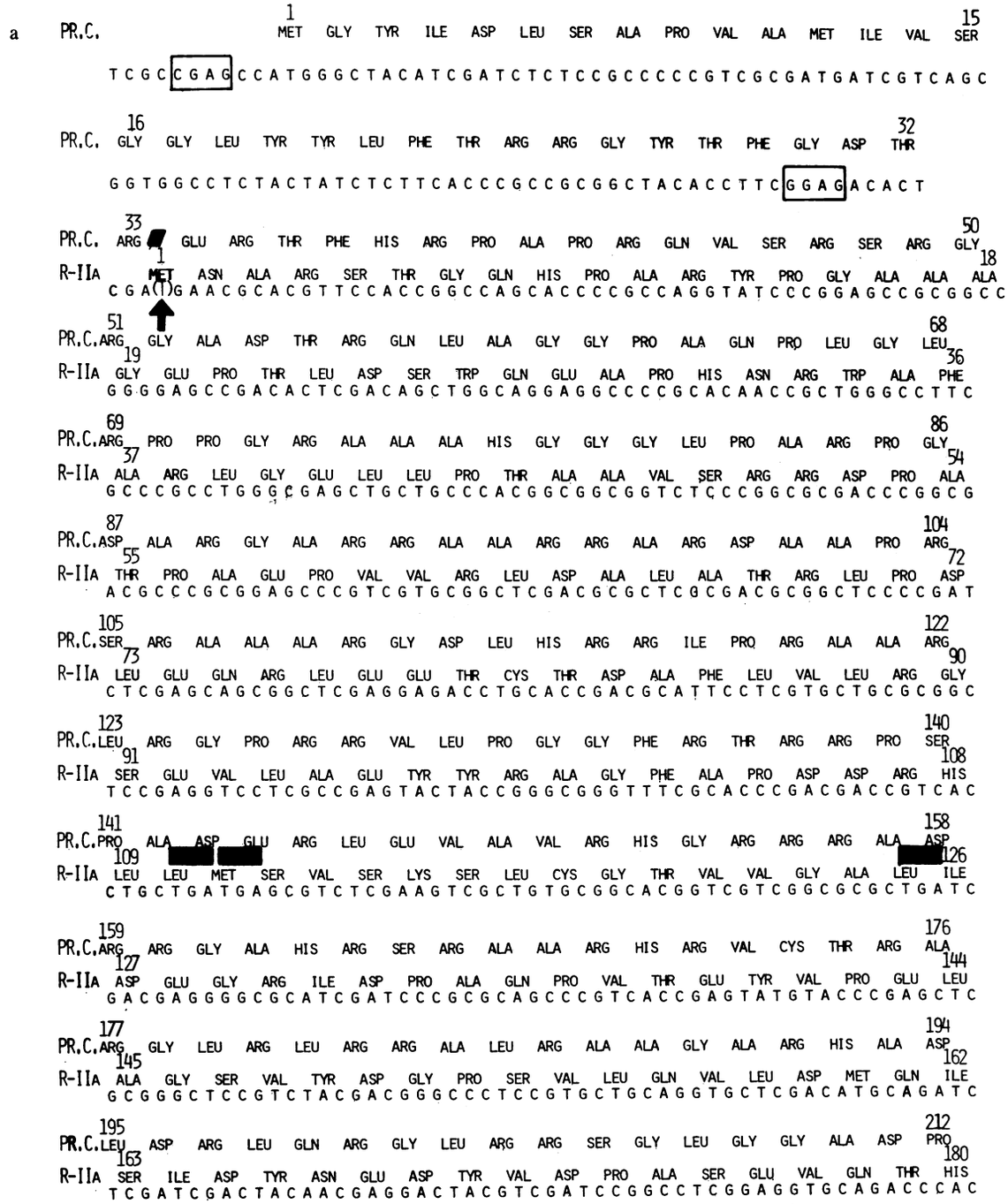
The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: 6-AHA CDH, 6-aminohexanoic acid cyclic dimer hydrolase; 6-AHA LOH, 6-aminohexanoic acid linear oligomer hydrolase.

reading frame or frames? A simple base change might silence the original coding potential, while giving a chain initiator to that alternative open reading frame, thereby, in a single step, producing a new enzyme *sensu stricto* with a hitherto nonexistent substrate specificity. I contend that this was how *Flavobacterium* Sp. KI72, formerly known as *Acromobacter guttatus* Sp. KI72 (2), has acquired two plasmid-encoded enzymes for sequential degradation of nylon oligomers (2–4).

R-II$_A$ Coding Sequence for 6-AHA LOH Embodies an Alternative, Longer Open Reading Frame That Might Have Been the Original Coding Sequence

Waste water from nylon factories contains ε-caprolactum, 6-aminohexanoic acid, 6-aminohexanoic acid cyclic dimer, and 6-aminohexanoic acid oligomers. In spite of the fact that nylon synthesis began only several decades ago, it was found, as early as 1975, that *Flavobacterium* Sp. KI72 could grow in a culture medium containing 6-aminohexanoic acid cyclic dimer as the sole source of carbon and nitrogen, as quoted in ref. 2. Soon, two enzymes responsible for this metabolism of 6-aminohexanoic cyclic dimer were identified as 6-aminohexanoic acid cyclic dimer hydrolase (6-AHA CDH) and 6-AHA LOH (2, 3). The swiftness with which these two enzymes have evolved is truly remarkable, for several decades are but a flash in the evolutionary time scale. More recently, a pair of coding sequences for two isozymic forms of 6-ALA LOH has been identified in pOAD2 plasmid harbored by *Flavobacterium* Sp. KI72 (4). Deduced amino acid sequences of these two isozymic forms differed from each other by 46 of the 392 residues. Inasmuch as both 6-AHA CDH and 6-AHA LOH demonstrated no inclination whatso-



(Fig. 1 continues on the next page.)

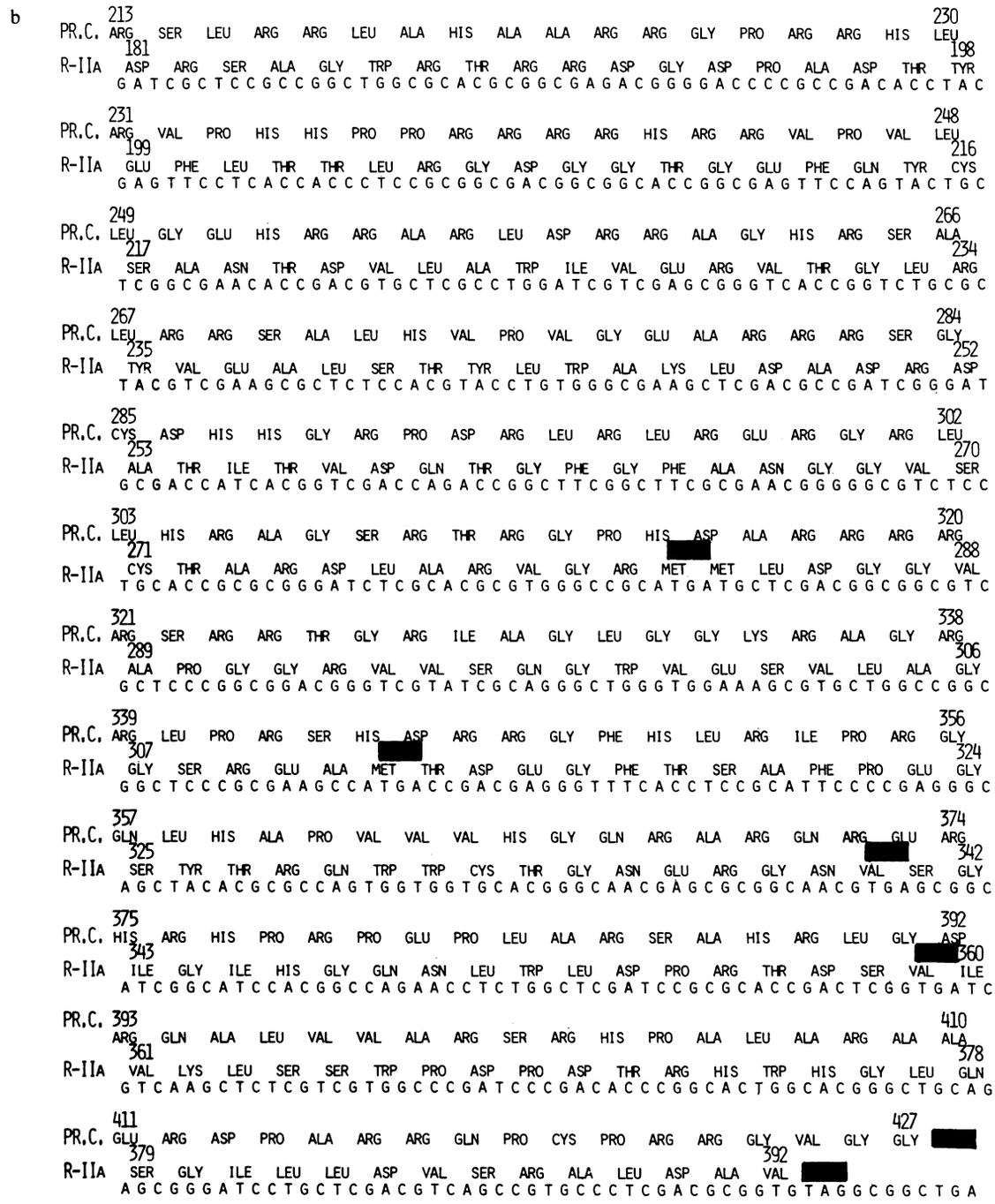


FIG. 1. (*a* and *b*) The 1295-base-long segment of pOAD2 plasmid DNA harbored by *Flavobacterium* Sp. KI72 published by Okada *et al.* (4) is shown here accompanied by two amino acid sequences specifiable by two long open reading frames. The one identified as R-II$_A$ was thought by these authors to be the coding sequence for one isozymic form of 6-AHA LOH, 392 residues long. I assume that the longer open reading frame identified as PR.C. was the original coding sequence of this stretch of plasmid DNA until several decades ago. When pOAD2 plasmids encountered nylon by-products, an insertion of T indicated by an arrow in the 3rd row of *a* proved advantageous, for this insertion silenced the PR.C. coding sequence by creating the T-G-A chain terminator; at the same time, the newly emerged A-T-G created a new coding sequence from an alternative open reading frame, which happened to specify a polypeptide chain with 6-AHA LOH activity for degradation of nylon by-products. As the G-G-A-G base tetramer recurs many times within this stretch of DNA, the convenient presence of the boxed-in Shine–Dalgarno sequence in front of the newly emerged A-T-G must also have have been accidental. The related sequence C-G-G-A in front of A-T-G of PR.C. at the 1st row of *a* is also boxed-in. The positions in respective polypeptide chains of first and last residues of each row of amino acid sequence are identified by numbers. Every chain-terminating base triplet in *a* and *b* is identified by a black box placed above it.

ever to attack any of the natural amide bonds tested (2, 3), the ancestry of neither cyclic dimer nor linear oligomer hydrolase can be readily sought among the known natural amidohydrolases. Furthermore, 11.35% amino acid sequence divergence observed between two isozymic forms of 6-ALA LOH suggests their independent origin from the same family of repeated sequences identified within pOAD2 plasmid, rather than one being ancestral to another, for so extensive an amino acid sequence divergence is not expected to occur in so short a time span—i.e., 40 years or thereabout.

As an alternative to the customary process of the birth of a unique gene from a redundant copy of the preexisted gene of a related function, I suggest that each of these unique genes for degradation of nylon by-products arose *de novo* independently from an alternative reading frame of the preexisted, internally repetitious coding sequence. In particular, I suggest

that the RS-II$_A$ base sequence was originally a coding sequence for an arginine-rich polypeptide chain 427 or so residues long in its length and that the coding sequence for one of the two isozymic forms of 6-ALA LOH arose from its alternative open reading frame.

In Fig. 1 *a* and *b*, the published coding sequence for one 6-AHA LOH isozyme is identified as R-II$_A$, the simplified version of RS-II$_A$ in the publication of Okada *et al.* (4), and it is accompanied by the amino acid sequence of a linear oligomer hydrolase it specifies. It should be noted that the base sequence shown in Fig. 1 *a* and *b* has yet another longer open reading frame a two-base-shift away from the R-II$_A$ coding sequence. It should be further noted that if T marked by an arrow in the 3rd row of Fig. 1*a* is deleted, this longer open reading frame identified as PR.C. (the abbreviation of the preexisted coding sequence) in Fig. 1 *a* and *b* starts from the first A-T-G in 1st row of Fig. 1*a* and ends in T-G-A that are the last three bases in the last row of Fig. 1*b*, thus specifying a polypeptide chain 427 amino acid residues long. I propose that this was the original coding sequence contained in the 1295-base-long stretch of pOAD2 plasmid harbored by *Flavobacterium* Sp. KI72. It is of interest to note here that this stretch of base sequence is duplicated elsewhere within the pOAD2 genome roughly 90°C away and that the coding sequence for the second isozymic form of 6-AHA LOH is found in this duplicated stretch (4). Thus, a pair of isozymic preexisted coding sequences might have given rise independently to the coding sequences for two isozymic forms of 6-AHA LOH. At any rate, an insertion of T indicated by an arrow at the position in the 3rd row of Fig. 1*a* would have silenced PR.C. by creating the T-G-A chain terminator at that position, while creating the chain initiator A-T-G, thus giving rise to a 392-codon-length coding sequence for 6-AHA LOH, identified as R-II$_A$ in Fig. 1 *a* and *b*.

As discussed at the beginning of this paper, the probability of a nonrepetitious base sequence simultaneously harboring one 427-codon-length and the other 392-codon-length open reading frames is practically nil. Furthermore, it should be noted that in spite of the presence of seven internal chain terminators (all T-G-As, three in Fig. 1*a* and four in Fig. 1*b*), the third reading frame of this 1295-base-long base sequence can still specify a 178-amino acid-residue-long polypeptide chain; beginning from A-T-G in the 3rd from the bottom row of Fig. 1*a* and ending at T-G-A in the 6th row of Fig. 1*b*. It is granted that this 1295-base-long stretch is very G+C rich, the A+T/G+C ratio being roughly 1/2.4. While this G+C richness no doubt contributed to the absolute paucity of chain-terminating base triplets within, the very fact that there are no marked ups and downs in the A+T/G+C ratios among 24 individual rows of base sequences of Fig. 1 *a* and *b* indicates that this very G+C richness was the reflection of the entire 1295-base-long sequence originating from repeats of the particular G+C-rich base oligomer. Although PR.C. and R-II$_A$ were the same base sequence decoded in two different reading frames, a longer polypeptide chain specified by PR.C. was considerably more monotonous in its amino acid composition when compared to that specified by R-II$_A$. As shown in Table 1, Arg comprised 28.33% of its 427 residues, while Ala, Gly, and Pro together made up another 33.02%. Not surprisingly, the internal repetitiousness of this 1295-base-long sequence was more clearly reflected in the PR.C. amino acid sequence than in the R-II$_A$ amino acid sequence.

As summarized in Fig. 2, numerous, interrelated tetrapep-

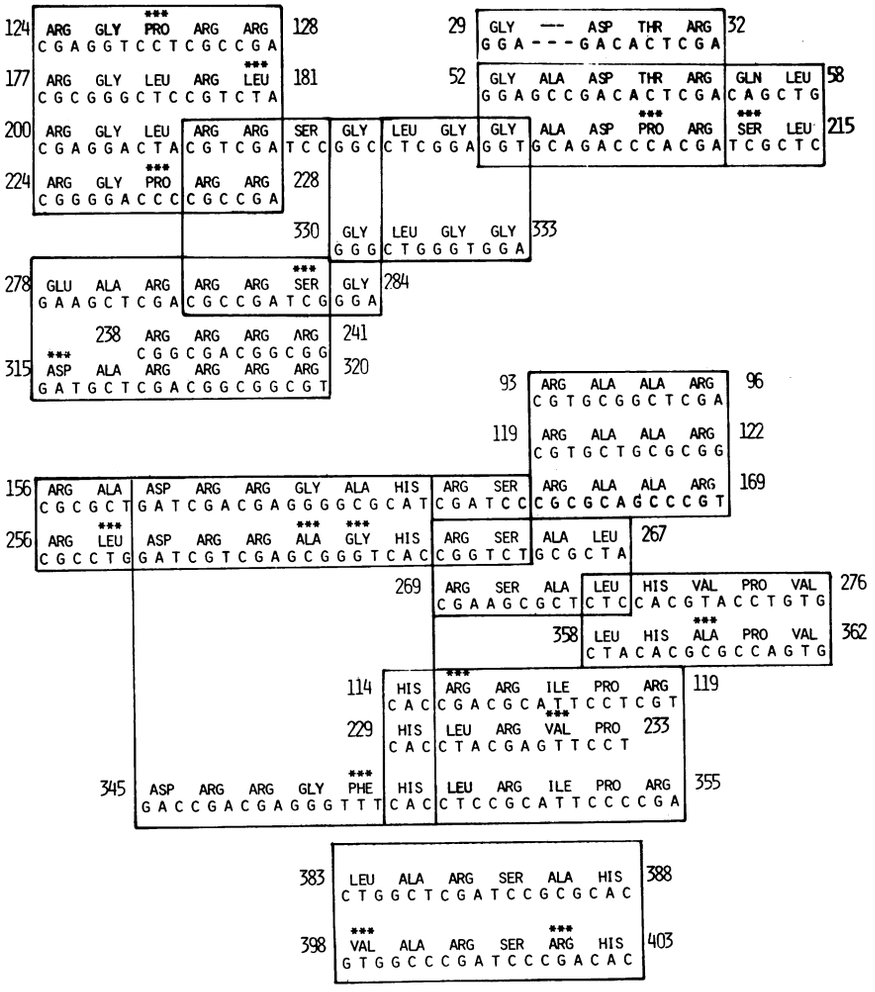


FIG. 2. Internal homology within the PR.C. coding sequence. Many, but by no means all, of the oligopeptide sequences recurring within the PR.C. amino acid sequence were selected from Fig. 1 *a* and *b* and are shown here collectively accompanied by base sequences specifying them. Each oligopeptide's position in the PR.C. polypeptide chain is identified by the number of the first and last residues. Each group of identical or homologous oligopeptides is placed in a box, and nonhomologous residues among them are marked by asterisks.

Table 1. Amino acid composition of the PR.C. and R-II$_A$ polypeptide chains

Amino acid	% composition PR.C.*	R-II$_A$†
Trp	0.00 (0)	2.81 (11)
Asn	0.00 (0)	2.04 (8)
Arg	28.33 (121)	7.65 (30)
His	6.32 (27)	1.79 (7)
Lys	0.23 (1)	0.77 (3)
Thr	2.11 (9)	6.89 (27)
Met	0.47 (2)	1.55 (6)
Tyr	0.94 (4)	3.06 (12)
Glu	2.34 (10)	5.87 (23)
Ile	1.17 (5)	2.55 (10)
Asp	4.22 (18)	8.42 (33)
Phe	1.17 (5)	2.30 (9)
Val	4.45 (19)	8.16 (32)
Ser	3.98 (17)	6.63 (26)
Gln	2.11 (9)	3.32 (13)
Pro	7.96 (34)	5.35 (21)
Cys	0.94 (4)	1.28 (5)
Ala	13.58 (58)	9.95 (39)
Gly	11.48 (49)	9.95 (39)
Leu	8.67 (37)	9.18 (36)

Actual numbers of individual residues are shown in parentheses.
*472 residues long.
†392 residues long.

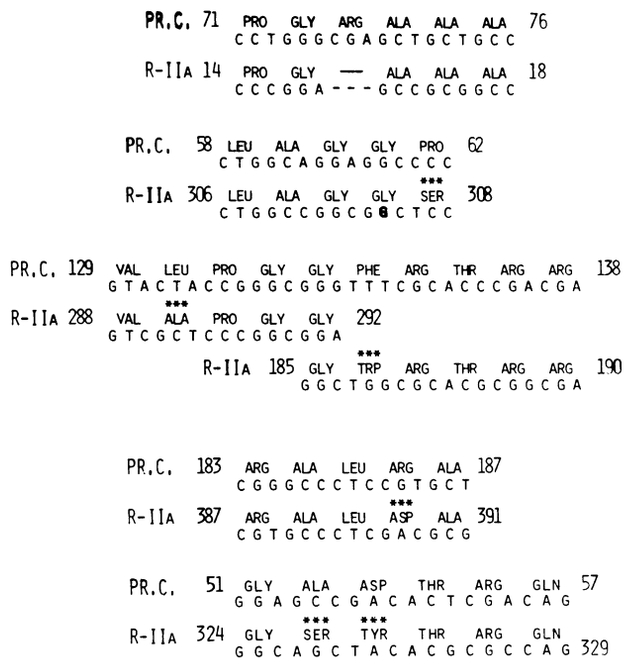


FIG. 3. Segmental homology between PR.C. and R-II$_A$. Five pairs of homologous oligopeptides found among PR.C. and R-II$_A$ polypeptide chains were singled out from Fig. 1 *a* and *b* and collectively are shown here as pairs. Numbers and asterisks are used in the same manner as they are used in Fig. 2.

tidic sequences recurred within the PR.C. polypeptide chain. For example, the tetrapeptidic Arg-Ala-Ala-Arg sequence recurred thrice, as shown in the middle right of Fig. 2. Furthermore, four more tetrapeptidic sequences recurred twice each: Arg-Arg-Ser-Gly, Gly-Leu-Gly-Gly, Ala-Arg-Arg-Arg, and Arg-Ser-Ala-Leu (top left, top center, middle left, and bottom left of Fig. 2). Although not summarized in Fig. 2, two other tetrapeptidic sequences, Arg-Ala-Ala-Ala and Arg-Arg-Arg-Arg, also recurred twice each (5th and 7th rows of Fig. 1*a*; 2nd and 6th rows of Fig. 1*b*). As to longer stretches of internal homology, a pair of decapeptidic sequences occupying the 156th to 165th and the 256th to 265th amino acid positions of the PR.C. polypeptide chain were 70% homologous with each other, reflecting underlying 66.7% base sequence homology between a pair of 30-base-long sequences (middle left of Fig. 2). Similarly, a pair of heptapeptidic sequences representing the 52nd to 58th and the 209th to 215th amino acid positions were 71.43% homologous, again reflecting underlying 66.7% base sequence homology between a pair of 21-base-long sequences (top right of Fig. 2). Fig. 2 also includes three other examples of homologous hexapeptidic sequences and two examples of homologous pentapeptidic sequences. All in all, there remains little doubt that the entire base sequence shown in Fig. 1 *a* and *b* originally arose from repeats of the G+C-rich base oligomer, such as the decamer C-G-A-C-G-C-C-G-C-T. Three consecutive copies of the above decamer should have given the decapeptidic periodicity Arg-Arg-Arg-Ser-Thr-Pro-Leu-Asp-Ala-Ala to the ancestral polypeptide chain.

Residual Amino Acid Sequence Homology Retained by Otherwise Divergent PR.C. and R-II$_A$ Polypeptide Chains

While the originally simple construction of the base sequence from oligomeric repeats was clearly reflected in repetitious monotony of the PR.C. amino acid sequence, a glance at Table 1 shows that this was not at all the case with the R-II$_A$ amino acid sequence. Befitting its 6-AHA LOH enzymatic activity, the 392-residue-long sequence contained substantial numbers of Trp, His, Tyr, and Ser residues and 30 Arg residues were balanced by 33 Asp residues. The extent of divergence between PR.C. and R-II$_A$ amino acid sequences is truly remarkable, when it is realized that these two coding sequences are but the same base sequence merely decoded in different reading frames. Yet, PR.C. and R-II$_A$ polypeptide chains are not entirely alien to each other either. Because of the underlying, abundant internal repetitiousness, it so happens that homologous base sequences here and there are translated in the same reading frame, thus producing segmental homology between noncorresponding portions of two polypeptide chains. Five pairs of such oligopeptide homologous segments were singled out from Fig. 1 *a* and *b* and collected in Fig. 3. Nevertheless, no functional relatedness is expected between PR.C. and R-II$_A$ polypeptide chains, because of such segmental homology occurring in noncorresponding portions. Indeed, the very basic former totally lacking Trp and Asn residues is not likely to function as an enzyme of any sort, whereas the latter in one swoop has acquired the capacity to degrade man-made nylon byproducts.

This work was supported in part by National Institutes of Health Grant A1 15620 and research grants from the Bixby Foundation and Wakunaga Pharmaceutical Co. of America.

1. Ohno, S. (1970) *Evolution by Gene Duplication* (Springer, Heidelberg).
2. Kinoshita, S., Negoro, S., Murayama, M., Bisaria, V. S., Sawada, S. & Okada, H. (1977) *Eur. J. Biochem.* **80**, 489–495.
3. Kinoshita, S., Terada, T., Taniguchi, T., Takene, Y., Masuda, S., Matsunaga, N. & Okada, H. (1981) *Eur. J. Biochem.* **116**, 547–551.
4. Okada, H., Negoro, S., Kimura, H. & Nakamura, S. (1983) *Nature (London)* **306**, 203–206.
5. Ohno, S. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 7657–7661.
6. Gō, M. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 1964–1968.
7. Blake, C. (1983) *Nature (London)* **306**, 535–537.
8. Ohno, S. & Epplen, J. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 3391–3395.