# PSEUDOGENE IN THE GENOME OF BACTERIOPHAGE LAMBDA?

Jaroslav Kypr and Jan Mrázek

Institute of Biophysics, Czechoslovak Academy of Sciences,
612 65 Brno, Czechoslovakia

Received April 17, 1987

We find a region in the non-coding part of bacteriophage lambda genome that codes for the conserved fold which repressors and other proteins use for specific DNA binding. The region is involved in a long open reading frame exceeding one kilobase and is read in the same frame as gene A in the opposite strand. The putative translation product of this open reading frame has a highly ordered secondary structure with a predominance of alpha helices, which is typical of repressors. In addition, codon usage in this frame suggests a protein-coding region. However, there is a TGA stop codon located between the putative gene start point and the region coding for the DNA binding fold. It thus appears that bacteriophage lambda had one more DNA binding protein, perhaps repressor, in the past that was inactivated by a mutation. © 1987 Academic Press, Inc.

There is a number of significant differences between prokaryotes and eukaryotes, one of them being a reversed relative proportion of protein-coding and non-coding DNA. In eukaryotic genomes protein-coding pieces represent not more than one or two per cent of total genomic DNA while the rest serves other as yet largely unknown purposes. The non-coding DNA includes pseudogenes - originally protein-coding sequences that were inactivated by one or more defects which prevent from their proper expression (for a review, see, for example, ref. 1). On the other hand, prokaryotes and bacteriophages in particular take use of 90 or even more per cent of their DNA for coding purposes. It is fairly frequent that their genes overlap. It is thus apparently unreasonable to search for pseudogenes in bacteriophage genomes. Yet, bacteriophage lambda is suspect to contain a pseudogene as will be shown in this communication.

## MATERIAL AND METHODS

Nucleotide sequence of the lambda genome and identification of its genes have been taken from literature (2). The se-

0006-291X/87 $1.50
*Copyright © 1987 by Academic Press, Inc.*
*All rights of reproduction in any form reserved.*

quence coding for the conserved DNA binding fold of repressors was identified using our computer program written in Fortran for an ICL 2950/10 computer. The program works using a scoring system incorporating the knowledge of sequences known to form the fold (3-11). The program finds one DNA binding fold in a random sequence of 100,000 amino acids. Protein secondary structure prediction was performed by our program JAMSEK combining several variants of the statistical algorithms of Chou and Fasman with hydrophobicity profile and helical wheel representation of the sequence into a single algorithm. JAMSEK works reliably with repressors (12) and is useful in estimation of the total amount of alpha helices in proteins. Codon usage was analysed using the approach of Macchiato and Tramontano based on the concept of codon information value (13).

## RESULTS AND DISCUSSION

In a previous work, we analysed bacteriophage lambda genes by our program and found four proteins containing the conserved DNA binding fold of repressors (14). Here we extend this work to non-coding parts of the genome, including both strands and all reading frames and, to our surprise, find an additional copy of the fold. It occurs in the complementary strand of gene A and both messages are read in phase. The starting nucleotide of the region coding for the fold is at position 1574 in the genome map (2) and the last at position 1509.

The amino acid sequence of the fold is presented in Table 1 along with sequences of the folds occurring in other proteins. The fold contains the key residues Gly 11, Ala 7 and Ile 17. In addition, its ten more residues also occur in the respective positions in other folds. The remaining nine amino acids are unique for the fold in the complementary strand of lambda gene A but in all these positions one can find at least five different amino acids in other folds (Table 2) to indicate that nature of the amino acid in these positions is not to a certain limit crucial for the helix-turn-helix formation.

Gaining a suspicion that the region in the complementary strand of gene A codes for such an important protein property as specific DNA binding we searched for open reading frames in its neighbourhood. The nearest initiation and termination codons spanning the putative repressor DNA binding motif were found at positions 1862 and 771, respectively, so that this part of the genome can code for a polypeptide chain containing 363 amino acid residues including the starting Met. The amino acid sequence of this polypeptide is given in Table 3. However there is a stop codon TGA shortly preceding the fold so that

Table 1. Amino acid sequence of the DNA binding motif found in the complementary strand of lambda gene A and its comparison to sequences of the motifs occurring in various repressors and related proteins that form (as shown in crystals) or may form (as anticipated from sequence homologies) the conserved helix-turn-helix fold

| Protein | Position in the fold | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 |
| Complementary strand of lambda gene A | Lys | Val | Gln | Leu | Leu | Leu | Ala | Asp | Asp | Ala | Gly | Ile | Met | Leu | Ala | Glu | Ile | Lys | His | Ala | Gly | Gly |
| Crystals | | | | | | | | | | | | | | | | | | | | | | |
| lambda cro | Phe | Gly | Gln | Thr | Lys | Thr | Ala | Lys | Asp | Leu | Gly | Val | Tyr | Gln | Ser | Ala | Ile | Asn | Lys | Ala | Ile | His |
| lambda cI | Leu | Ser | Gln | Glu | Ser | Val | Ala | Asp | Lys | Met | Gly | Met | Gly | Gln | Ser | Gly | Val | Gly | Ala | Leu | Phe | Asn |
| 434 cI | Leu | Asn | Gln | Ala | Glu | Leu | Ala | Gln | Lys | Val | Gly | Thr | Thr | Gln | Gln | Ser | Ile | Glu | Gln | Leu | Glu | Asn |
| E. coli CAP | Ile | Thr | Arg | Gln | Glu | Ile | Gly | Gln | Ile | Val | Gly | Cys | Ser | Arg | Glu | Thr | Val | Gly | Arg | Ile | Leu | Lys |
| E. coli trp | Met | Ser | Gln | Arg | Glu | Leu | Lys | Asn | Glu | Leu | Gly | Ala | Gly | Ile | Ala | Thr | Ile | Thr | Arg | Gly | Ser | Asn |
| Homologies | | | | | | | | | | | | | | | | | | | | | | |
| E. coli lac | Val | Thr | Leu | Tyr | Asp | Val | Ala | Glu | Tyr | Ala | Gly | Val | Ser | Tyr | Gln | Thr | Val | Ser | Arg | Val | Val | Asn |
| E. coli gal | Ala | Thr | Ile | Lys | Asp | Val | Ala | Arg | Leu | Ala | Gly | Val | Ser | Val | Ala | Thr | Val | Ser | Arg | Val | Ile | Asn |
| Mat al | Lys | Glu | Lys | Glu | Glu | Val | Ala | Lys | Lys | Cys | Gly | Ile | Thr | Pro | Leu | Gln | Val | Arg | Val | Trp | Cys | Asn |
| 434 cro | Met | Thr | Gln | Thr | Glu | Leu | Ala | Thr | Lys | Ala | Gly | Val | Lys | Gln | Gln | Ser | Ile | Gln | Leu | Ile | Glu | Ala |
| P 22 repressor | Ile | Arg | Gln | Ala | Ala | Leu | Gly | Lys | Met | Val | Gly | Val | Ser | Asn | Val | Ala | Ile | Ser | Gln | Trp | Gln | Arg |
| P 22 cI | Arg | Gly | Gln | Arg | Lys | Val | Ala | Asp | Ala | Leu | Gly | Ile | Asn | Glu | Ser | Gln | Ile | Ser | Arg | Trp | Lys | Gly |
| P 22 cro | Gly | Thr | Gln | Arg | Ala | Val | Ala | Lys | Ala | Leu | Gly | Ile | Ser | Asp | Ala | Ala | Val | Ser | Gln | Trp | Lys | Glu |

The above amino acid sequences were identified as DNA binding folds in the following papers: lambda cro (4), lambda cI (5), 434 cI (10), E. coli CAP (3), E. coli trp (11), E. coli lac and gal (7), Mat al (8), 434 cro, P 22 repressor, cI, and cro (6).

Table 2. Amino acids in the non-conservative positions of the DNA binding fold of repressors

| Position in the fold | Predominant amino acid | Also occurring amino acids | Amino acids in the fold encoded by the complementary strand of lambda gene A |
|---|---|---|---|
| 2 | Thr | Gly, Ser, Asn, Glu, Arg | Val |
| 4 | - | Thr, Glu, Ala, Gln, Arg Tyr, Lys | Leu |
| 5 | Glu | Lys, Ser, Asp, Ala | Leu |
| 13 | Ser | Tyr, Gly, Thr, Lys, Asn | Met |
| 14 | Gln | Arg, Ile, Tyr, Val, Pro Asn, Asp | Leu |
| 16 | Thr | Ala, Gly, Ser, Gln | Glu |
| 18 | Ser | Asn, Gly, Glu, Thr, Arg Gln | Lys |
| 19 | Arg | Lys, Ala, Gln, Val, Leu | His |
| 21 | - | Ile, Phe, Glu, Leu, Ser Lys, Val, Cys, Gln | Gly |

it is spanned by two termination codons and probably is not expressed. Yet, was this protein synthesized in the past? It is a difficult task to find a conclusive answer to this question but there is a possibility to look for some characteristic properties of the inactivated gene.

We first used an algorithm of Tramontano and Macchiato (13) to distinguish between coding and non-coding nucleotide sequences which in fact relies on a determination of how the potential protein spatial structure is resistant to mutations. This criterion says that the open reading frame containing the DNA binding fold is coding (information value 2.24). Another criterion to hint whether bacteriophage lambda had one more protein in the past is its secondary structure. This we predicted using our computer program JAMSEK that was used in some previous studies (12,15). It predicted 35-40% of the polypeptide chain to form alpha helices and 25-30% beta sheets. This high degree of spatial order of the polypeptide chain and predominance of alpha helices is typical of repressors. It should be pointed out that JAMSEK only predicts 12% of residues in random sequences to be involved in alpha helices

Table 3. Amino acid sequence of the inactivated "repressor" of phage lambda. Amino acids constituting the DNA binding fold are underlined Note the boxed termination codon TGA

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Met | Leu | Phe | Pro | Leu | Cys | His | His | Phe | Ser | Ile | Arg | Thr | Phe | Ala |
| Asn | Phe | Arg | Leu | Pro | Arg | Leu | Thr | Glu | Arg | Gly | Val | Tyr | Glu | Gly |
| Phe | Thr | Phe | Ser | Arg | Ile | Pro | Phe | Arg | Phe | His | Pro | Val | Phe | Asp |
| Asn | Leu | His | Pro | Gly | Gly | Glu | Arg | Ala | Val | Arg | Cys | Pro | Asp | Val |
| Lys | Gly | His | Thr | Val | Arg | Trp | Leu | Asn | Leu | Phe | Thr | Gly | $\boxed{\text{TGA}}$ | Arg |
| Lys | Pro | Glu | Asn | Ala | Ile | Thr | Gly | Pro | Asp | Pro | Gly | Leu | Phe | Ala |
| Asp | Ile | Thr | Gly | Ile | Ser | Lys | Val | Gln | Leu | Leu | Leu | Ala | Asp | Asp |
| Ala | Gly | Ile | Met | Leu | Ala | Glu | Ile | Lys | His | Ala | Gly | Gly | Val | Ile |
| Arg | Arg | Pro | Phe | Glu | Ala | Lys | Arg | Arg | Leu | Phe | Val | Ala | Lys | Phe |
| Lys | Ile | Leu | Leu | Leu | Pro | Ala | Met | Arg | Ala | Gly | Asn | Met | Lys | Thr |
| His | Lys | Met | Arg | Gly | Phe | Thr | Gly | Cys | Thr | Leu | Asn | Leu | Thr | Gly |
| Ala | Ser | His | Phe | Trp | Arg | Gly | Ala | Thr | Asp | Gly | Leu | Trp | Pro | Asp |
| Arg | Ala | Phe | Asn | Thr | Leu | Val | Thr | Gln | Glu | Arg | Arg | Arg | Ala | Phe |
| Leu | Phe | Asn | Ile | Ile | Ile | Lys | Ser | Ser | Lys | Phe | Ile | Ile | Thr | Arg |
| His | Ile | His | Arg | Leu | Phe | Thr | Val | Val | PHe | Cys | Arg | Phe | Thr | Ala |
| Gln | Ala | Pro | Glu | Ala | Thr | Pro | Ile | Ser | Glu | Thr | Leu | His | Gly | Glu |
| Arg | Val | Ile | Pro | Val | Leu | Phe | Ala | Ile | Pro | Arg | Gly | Gln | Arg | Gln |
| Gln | Arg | Arg | Asn | Ile | Thr | Asn | Ser | Arg | Leu | Asn | Val | Gly | Phe | His |
| Lys | Val | Leu | Gly | Ile | Thr | Ile | Arg | Arg | Gln | Pro | Asp | Lys | Gly | Val |
| Ala | Leu | Leu | Met | Leu | Tyr | Lys | Val | Gly | Ile | Asn | Thr | Gln | Gln | His |
| Phe | Gly | Ile | Thr | Asp | Thr | Gly | Arg | Leu | His | His | Ile | His | Leu | Thr |
| Asp | Val | Val | Ala | Ala | His | Arg | Ile | His | Asp | Gly | Pro | Leu | Lys | Gly |
| Gln | Cys | Phe | Pro | Ala | Pro | Phe | Leu | Val | Cys | Gly | Phe | Phe | Arg | Glu |
| Ile | Val | Ile | Ser | Ile | Arg | Pro | Phe | Asn | Gly | Gly | Leu | Trp | Leu | Arg |
| Pro | Glu | Gln | | | | | | | | | | | | |

(15). As for the signal sequences often preceding genes (16), no wonder they are not properly developed prior to the open reading frame which might code for one more lambda repressor but one finds five consecutive purines starting at position 1880 and an (A+T) rich block of eight nucleotides at position 1997, reminding of Shine-Dalgarno and Pribnow boxes, respectively.

The presence of a relatively long open reading frame involving the conserved repressor fold for DNA binding, typical secondary structure and characteristic choice of codons make likely a possibility that the complementary strand of gene A coded for a repressor in the past and that it was inactivated by a mutation, perhaps to improve functional properties of the gene A protein product which is coded in the opposite strand. This mutation resulted in the appearance of the central serine in the tripeptide -Ser-Ser-Ser- in protein A which belongs among the most frequent tripeptides in proteins (17). The idea that lambda phage had originally two repressors to maintain the lysogenic way of life is not as much surprising because a related phage P22 of Salmonella typhimurium has also two repressors for this purpose of which one has no counterpart in today's bacteriophage lambda.

## ACKNOWLEDGEMENT

We greatly appreciate the interest of Dr. J. Reich in this work, many stimulating discussions and attracting our attention to the presence of two repressors in phage P22. Manuscript preparation by Mrs. Irina Hebelková is gratefully acknowledged.

## REFERENCES

1. Jeffreys, A.J. and Harris, S. (1984) BioEssays 1, 253-258.
2. Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F. and Petersen, G.B. (1982) J. Mol. Biol. 162, 729-773.
3. McKay, D.B. and Steitz, T.A. (1981) Nature 290, 744-749.
4. Anderson, W.F., Ohlendorf, D.H., Takeda, Y. and Matthews, B.W. (1981) Nature 290, 754-758.
5. Pabo, C.O. and Lewis, M. (1982) Nature 298, 443-447.
6. Sauer, R.T., Yocum, R.R., Doolittle, R.F., Lewis, M. and Pabo, C.O. (1982) Nature 298, 447-451.
7. Weber, I.T., McKay, D.B. and Steitz, T.A. (1982) Nucl. Acids Res. 10, 5085-5102.
8. Ohlendorf, D.H., Anderson, W.F. and Matthews, B.M. (1983) J. Mol. Evol. 19, 109-114.
9. Takeda, Y., Ohlendorf, D.H., Anderson, W.F. and Matthews, B.M. (1983) Science 221, 1020-1026.
10. Anderson, J.E., Ptashne, M. and Harrison, S.C. (1985) Nature 316, 596-601.
11. Schevitz, R.W., Otwinowski, Z., Joachimiak, A., Lawson, C.L. and Sigler, P.B. (1985) Nature 317, 782-786.
12. Kypr, J. and Mrázek, J. (1985) Biochem. Biophys. Res. Commun. 131, 780-785.
13. Tramontano, A. and Macchiato, M.F. (1986) Nucl. Acids Res. 14, 127-135.
14. Kypr, J. and Mrázek, J. (1986) J. Mol. Biol. 191, 139-140.
15. Kypr, J. and Mrázek, J. (1986) Int. J. Biol. Macromol. 9, 49-53.
16. Staden, R. (1984) Nucl. Acids Res. 12, 505-519.