The *rIIA* Gene of Bacteriophage T4. I. Its DNA Sequence and Discovery of a New Open Reading Frame Between Genes *60* and *rIIA*

Patrick Daegelen[1] and Edward Brody

Centre de Génétique Moléculalre du C.N.R.S., Laboratoire propre associé à l'université Pierre et Marie Curie, 91190 Gif-sur-Yvette, France

Manuscript received October 30, 1989
Accepted for publication February 16, 1990

ABSTRACT

We have determined the DNA sequence of the *rIIA* gene and have discovered a small open reading frame, *rIIA.1*, between genes *60* and *rIIA*. The predicted molecular weights of these proteins are 82,840 for rIIA and 8,124 for rIIA.1. The rIIA protein has a repeated motif which suggests that the gene has evolved by duplication. It also has a motif which suggests that it belongs to a group of ompR-like proteins that control regulation of gene expression in response to changes in the external environment. We have sequenced three different missense mutants whose mutations lie in the *Ala* segment of the *rIIA* genetic map. All three changes are found within the first 35 bp of the *rIIA* coding sequence. The region of control of protein synthesis is identical in the *rIIA* gene and in gene *44* of T4. We relate this finding to the high sensitivity of both RNAs to translational repression by the T4 *regA* gene product.

BENZER's (1959, 1961) analyses of the *rII* loci of bacteriophage T4 are one of the cornerstones of modern molecular genetics. The *rIIA* and *rIIB* genes have served as tools for numerous studies since that time, including the nature of the genetic code (CRICK *et al.* 1961), mechanisms of gene expression [see SINGER, SHINEDLING and GOLD 1983 for a review] and the mechanism of frameshift mutagenesis induced by acridine dyes (RIPLEY *et al.* 1988) . The *rII* mutations serve to define the *rex* genes of bacteriophage λ (MATZ, SCHMANDT and GUSSIN 1982), although the nature of this interaction remains unknown. Both the rIIA and rIIB proteins copurify with membranes from T4-infected *Escherichia coli*, partitioning with the true inner membrane rather than into the cell wall (ENNIS and KIEVITT 1973; TAKACS and ROSENBUSCH 1975; WEINTRAUB and FRANKEL 1972). In addition, the rII proteins have been reported to be DNA binding proteins (HUANG and BUCHANAN 1974) and to be part of a membrane-free DNA-protein complex containing newly replicated DNA (MANOIL, SINHA and ALBERTS 1977). The rII proteins are bound to such a DNA-protein complex even when parental DNA cannot replicate (UZAN *et al.* 1985).

The *rIIB* gene has been sequenced (PRIBNOW *et al.* 1981; HUANG 1986) and has been the subject of recent studies on translation (SHINEDLING *et al.* 1987a,b) and transcription (SHINEDLING, WALKER and GOLD 1986). The wealth of genetic information concerning the *rIIA* gene has not yet been fully exploited. The DNA sequence of a carboxy-terminal region of the *rIIA* gene has been published (PRIBNOW *et al.* 1981; SUGINO and DRAKE 1984), leaving two-thirds of the gene unsequenced. We have taken a first step in remedying this situation by sequencing the region between gene *60* and the carboxy-terminal portion of *rIIA*. This completes the DNA sequence of the *rIIA* gene. We have found, as well, a small open reading frame (ORF) between genes *60* and *rIIA*. In a companion study (DAEGELEN and BRODY 1990) we analyze the transcriptional control of *rIIA* gene expression.

MATERIALS AND METHODS

Enzymes and biochemicals: Restriction endonucleases were purchased from New England Biolabs and Bethesda Research Laboratories (BRL). T4 DNA ligase and *E. coli* DNA polymerase I Klenow fragment were purchased from Amersham or BRL. Avian myeloblastosis virus (AMV) reverse transcriptase came from Genofit (Geneva). T4 polynucleotide kinase, deoxy- and dideoxynucleotides, 17-mer M13 primer, desoxyadenosine 5'-[α-^{35}S]thiotriphosphate ([α^{35}S]dATP), and adenosine 5'-[γ-^{32}P]triphosphate ([^{32}P] ATP) were all obtained from Amersham.

Plasmids: DNA of plasmid pTB101 was a gift from R. H. EPSTEIN (Geneva) . Plasmid pTB101 is a derivative of pBR313 (BOLIVAR *et al.* 1977); it contains a 2-kb *Eco*RI-*Hind*III fragment of T4 DNA. This fragment, which we shall call *60-A*, includes the distal portion of gene *60* and the proximal two-thirds of gene *rIIA* (SELZER *et al.* 1978, 1981) . We shall use the name p60A for this plasmid.

Bacteriophages: We have used T4D wild-type phage and three mutants in the Ala region of the *rIIA* gene of phage T4B originally isolated by BENZER (1961). Two were iso-

[1] Present address: Groupe Systèmes Parallèles et Biologie, Institut de Biotechnologie, I.N.R.A. Domaine de Vilvert, 78350 Jouy en Josas, France.

TABLE 1

T4-specific oligodeoxynucleotides

Name	Length	Sequence	Location[a]
r2A1	20-mer	5′-GATAACTTGATGCACGGCTG-3′	1703–1723
r2A2	19-mer	5′-GAACCATTACCAAGAATTG-3′	572–590
r2A3	23-mer	5′-CATTAAGTGCATGAGCATCAATC-3′	702–724
r2A6	20-mer	5′-GCAGCTTTAGGACGAGGAGC-3′	2014–2033

[a] The location refers to the sequence shown in Figure 2.

lated after 2-aminopurine treatment (*AP80, AP129*) and the other (*F120*) is a spontaneous mutant. Phage M13 derivatives mp10 and mp11 (Messing and Vieira 1982) and mp18 and mp19 (Norrander, Kempe and Messing 1983) were obtained from Amersham.

Bacteria: The p60A plasmid (Selzer *et al.* 1978, 1981) was transferred into the *E. coli* coli strain MC1061 (F^- $araO_{139}$ *del* $(araABOIC\text{-}leu)_{7679}$ *del* $(lac)_{x74}$ *galK rpsL* hsr^- hsm^+; Casadaban and Cohen 1980) . *E. coli* JM101 and JM105 strains obtained from Amersham were used as the recipient for M13 phage DNA transformations and as the hosts for the propagation of M13 phages. *E. coli* B^E (su^-) was the host for the growth of T4B and T4D phages.

T4 phage infection and RNA purification: *E. coli* BE was grown in M9 medium supplemented with 1-casamino acids (Daegelen and Brody 1976) at 30°. When bacteria reached 5 × 10^8 cells per ml, they were infected with T4 phages at a multiplicity of infection of five. Five minutes later, 10 ml of infected cells were added to 2 ml of lysis buffer containing 5% (wt/vol) $NaDodSO^4$ and 1.5 M sodium acetate (pH 5.2), in a boiling water bath (adapted from Uzan, Favre and Brody 1988). The samples were held at 100° for 2–3 min, then an equal volume of phenol saturated with 0.25 M sodium acetate (pH 5.2) was added; RNA was then extracted at 65°. Two and in some cases three subsequent extractions with phenol and chloroform were carried out before ethanol precipitation of the RNA (in the presence of 7% acetic acid to solubilize M9 salts). Finally the RNA pellet was washed once with 70% ethanol, dried, and resuspended in distilled water.

DNA sequencing: All DNA sequence determinations were done by the chain-termination method (Sanger, Nicklen and Coulson 1977) according to the procedures described in the DNA sequencing handbook from Amersham. Recombinant M13 phage DNAs carrying T4 inserts were used as single-stranded DNA templates, for chain-extension and termination with the Klenow fragment of DNA polymerase I, and 17-mer M13 primer. In some cases we have used, in place of this primer, T4-specific synthetic oligodeoxynucleotides (see Table 1). The annealing and sequencing reactions were done at room temperature. Chain termination products were analyzed on 5 or 6% polyacrylamide, 7 M urea sequencing gels. The gels, fixed and washed with a mixture of 7% acetic acid, 10% methanol in water, were dried prior to autoradiography.

Primer-extension reactions: The r2A1, r2A2, r2A3 and r2A6 oligodeoxynucleotides were 5′-end labeled by T4 polynucleotide kinase with [γ-^{32}P]ATP and purified on 20% polyacrylamide, 7 M urea sequencing gels. The annealing reaction was performed with 50 μg of RNA in a total volume of 10 μl, containing 1 pM of the labeled oligonucleotide and 2 μl of the 5× concentrated annealing buffer (250 mM Tris-HCl (pH 8.3), 300 mM NaCl, 50 mM dithiothreitol) . The mixture was heated for 3 min at 85° and then rapidly frozen in a solid CO_2/ethanol bath. After thawing of the samples on ice, 2 μl of the annealing reaction were distributed in each of four tubes; then we added 1 μl of a solution of all four deoxynucleotide triphosphates (each at 2 mM in 1× annealing buffer) plus 1 μl of one of the dideoxynucleotide triphosphates (at 250 μM in 1× annealing buffer). Afterward 1 μl of AMV reverse transcriptase mixture (containing, in 25 μl, 25 units of AMV reverse transcriptase and 5 μl of the 5x reverse transcriptase reaction buffer: 250 mM Tris-HCl (pH 8.3), 300 mM NaCl, 50 mM dithiothreitol and 150 mM magnesium acetate) was added. Incubation was carried out for 30 min at 37°, after which reaction mixtures were frozen before loading onto sequencing gels (McPheeters *et al.* 1986)

***In vitro* synthesis of T4 RNA:** T4 RNA was synthesized *in vitro* by using T4*D*+ DNA at 28 μg/ml and *E. coli* RNA polymerase holoenzyme at 15 μg/ml under standard reaction conditions (Brody, Rabussay and Hall 1983) at 0.2 M NaCl. Incubation was for 2 hr at 37°. T4 RNA was then purified by standard phenol extraction methods, precipitated with ethanol, and resuspended in water.

Restriction site mapping of *60-rIIA*: First, 1 μg of p60A DNA was completely digested with *Eco*RI nuclease. Each liberated extremity was labeled with 10 pM of [α-^{35}S]dATP and 5 units of Klenow fragment (for 30 min using 100 μmol of dCTP, dGTP and dTTP in the same buffer used for the digestion by *Eco*RI nuclease). The linearized vector was phenol-extracted, precipitated with ethanol, and resuspended in the *Hin*dIII restriction nuclease buffer. After total digestion with this enzyme, the two fragments generated were separated by electrophoresis in a 0.7% agarose gel. After sufficient separation, the gel was stained with ethidium bromide; then the bands were visualized under UV-irradiation. A small piece of DEAE-paper (Schleicher and Schuell, NA45, pretreated by soaking 10 min in 10 mM EDTA (pH 8.0), then 5 min in 0.5 M NaOH, and finally washed extensively with water until pH 7.0), was inserted in front of each DNA band. DNA fragments were then trapped on DEAE paper by further electrophoresis (Dretzen *et al.* 1981). This labeled DNA was then redigested partially and completely by a variety of restriction endonucleases. Products of these digestions were analyzed on polyacrylamide sequencing gels as described above.

Computer analysis: Most of the analyses were done using the Sequence Analysis Software Package of the University of Wisconsin Genetics Computer Group (UWGCG; Devereux, Haeberli and Smithies 1984) running on a VAX 11/750 under VMS V5. In particular, we have used these programs extensively for sequence comparisons, data-base searching, pattern recognition and protein sequence analysis. Nucleic acid or protein sequences were aligned with BESTFIT (algorithm of Smith and Waterman 1981) or GAP (algorithm of Needleman and Wunsch 1970) programs from UWGCG. The search for similarities between our sequences and entire databases (GENBANK, EMBL, NBRF, PASTEUR) was done using an implementation of FASTP (Lipman and Pearson 1985) on the VAX. These three last algorithms were used with the scoring matrix of

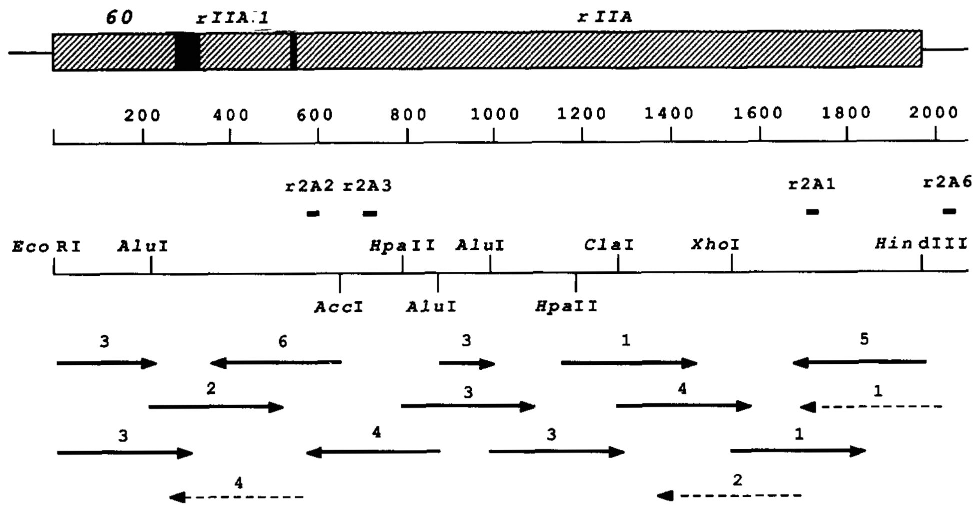

FIGURE 1.—Restriction map of the *Eco*RI-*Hin*dIII *60-rIIA* fragment of T4 DNA and the sequencing strategy of the region. The distal part of gene *60*, the small ORF *rIIA.1*, and the proximal part of gene *rIIA* are shown as shaded areas above the nucleotide scale. Intercistronic regions between genes *60* and *rIIA.1*, and between genes *rIIA.1* and *rIIA* are shown as black boxes. The restriction map of the 1.97 kb shows the sites used for subcloning in M13 during the sequencing project. Recombinant M13 phage DNA carrying inserts of T4 DNA were used as single-stranded templates for the SANGER dideoxy chain-termination reactions (SANGER, NICKLEN and COULSON 1977). Sequences obtained from this method are shown with filled arrows. In some cases, we have used in place of the 17-mer M13 primer the oligodeoxynucleotides r2A1, r2A2, r2A3 and r2A6 complementary to T4-specific regions of the DNA genome; they are shown as small bars on the map. R2A1, r2A2, r2A3 and r2A6 have been also used as primers for AMV reverse transcriptase using T4 RNA as templates. In all cases, the digits shown above each arrow represent the number of times each T4 region was sequenced.

DAYHOFF, BARKER and HUNT (1983) or RISLER *et al.* (1988). In some cases we have used equivalently the computing facilities of the "Base Informatique sur les Séquences d'Acides Nucléiques pour les Chercheurs Européens" at the CITI2 (Centre Inter-Universitaire d'Informatique à Orientation Biomédicale; Paris). Secondary-structure predictions of rIIA and rIIA.1 proteins were determined by the team of J. GARNIER according their most recent published work (BIOU *et al.* 1988). Analysis of hydrophobic clusters was done using the program of GABORIAUD *et al.* (1987).

RESULTS

We have sequenced the 2-kb *Eco*RI-*Hin*dIII *60-RIIA* fragment by subcloning in M13 and using the dideoxy chain termination method. The map of this fragment and the cloning strategy are outlined in Figure 1.The regions which were also sequenced using RNA from T4 infected cells and AMV reverse transcriptase are indicated by dotted arrows. The complete DNA sequence of this fragment plus the previously published sequence of the distal part of the *rIIA* gene (starting at the *Hin*dIII site at nucleotide 1970; PRIBNOW *et al.* 1981; SUGINO and DRAKE 1984) are shown in Figure 2. Between the end of gene *60* and the beginning of the *rIIA* gene is an ORF which we call *rIIA.1*. This ORF codes for a hypothetical protein (M_r = 8124) whose amino acid composition is given in Table 2. The ATG of this ORF is separated by seven nucleotides from an AGGA SHINE-DALGARNO sequence, which suggests that this ORF codes for a T4 protein. HUANG *et al.* (1988 and personal communication) have independently sequenced part of this gene and have obtained evidence that this protein is synthesized after transcription from a T7 promoter on a plasmid. The early and middle *rIIA* promoters which are analyzed in the accompanying article are found, respectively, just before and in the coding portion of this ORF. The *rIIA* gene codes for a protein (M_r = 82,840) whose amino acid composition is also given in Table 2.

We have analyzed the rIIA protein derived from our DNA sequence with respect to its structure, its possible function (or functions) and the control of its biosynthesis. The structure of the rIIA protein has been analyzed using the algorithms of BIOU *et al.* (1988) and GABORIAUD *et al.* (1987). The structure predictions for these two methods are shown in Figures 3a and 4a.

Region around the *rIIA* initiator ATG: One of the most striking features of the *rIIA* sequence is shown in Figure 5. The region controlling the translational start of the *rIIA* gene is identical over 19 nucleotides (−12 to +7, with the A of the initiator AUG serving as the +1 reference) to that of T4 gene *44*. The SHINE-DALGARNO sequence, the distance from this sequence to the AUG, and all the nucleotides between

A

```
        10                  30                  50                  70                  90
GAATTCGCTTTGTCAAAACTCCTGTAATCATCGCTCAGGTCGGTAAAAAACAAGAATGGTTTTATACAGTCGCTGAATATGAGAGTGCCA
   I  R  F  V  K  T  P  V  I  I  A  Q  V  G  K  K  Q  E  W  F  Y  T  V  A  E  Y  E  S  A  K

                 110                 130                 150                 170
AAGATGCTCTACCTAAACATAGCATCCGTTATATTAAGGGACTTGGCTCTTTGGAAAAATCTGAATATCGTGAGATGATTCAAAACCCAG
   D  A  L  P  K  H  S  I  R  Y  I  K  G  L  G  S  L  E  K  S  E  Y  R  E  M  I  Q  N  P  V

       190                 210                 230                 250                 270
TATATGATGTTGTTAAACTTCCTGAGAACTGGAAAGAGCTTTTTGAAATGCTCATGGGAGATAATGCTGACCTTCGTAAAGAATGGATGA
   Y  D  V  V  K  L  P  E  N  W  K  E  L  F  E  M  L  M  G  D  N  A  D  L  R  K  E  W  M  S
                                                                                    end of
                 290                 310                 330                 350
GCCAGTAGTTTACTTTACCACAAGGATGTGGTATAATTAATTGGGCAAATGAGGATATTGAAATGAAATCATATAAAGTAAATTTAGAAC
   Q  *                                                        M  K  S  Y  K  V  N  L  E  L
gene 60                                                       start of gene rIIA.1
       370                 390                 410                 430                 450
TTTTTGATAAAGCAGTTCATCGAGAATATAGAATCATTCAACGCTTTTTCGATATGGGAGAAGCCGAAGAATTTAAAACCCGCTTTAAAG
   F  D  K  A  V  H  R  E  Y  R  I  I  Q  R  F  F  D  M  G  E  A  E  E  F  K  T  R  F  K  D

                 470                 490                 510                 530
ATATTAGAGATAAAATTCAATCCGACACCGCAACTAAAGATGAACTACTAGAAGTTGCTGAAGTTATTAAGCGTAATATGAATTAATGAG
   I  R  D  K  I  Q  S  D  T  A  T  K  D  E  L  L  E  V  A  E  V  I  K  R  N  M  N  *  *
                                                                     end of gene  rIIA.1
       550                 570                 590                 610                 630
GAAATTATGATTATCACCACTGAAAAAGAAACAATTCTTGGTAATGGTTCTAAATCAAAAGCATTTAGCATCACAGCATCTCCTAAAGTA
       M  I  I  T  T  E  K  E  T  I  L  G  N  G  S  K  S  K  A  F  S  I  T  A  S  P  K  V
     start of gene rIIA
                 650                 670                 690                 710
TTTAAAATTCTGTCATCTGATTTGTATACAAACAAGATTCGCGCAGTAGTCCGTGAATTGATTACTAACATGATTGATGCTCATGCACTT
 F  K  I  L  S  S  D  L  Y  T  N  K  I  R  A  V  V  R  E  L  I  T  N  M  I  D  A  H  A  L

       730                 750                 770                 790                 810
AATGGAAATCCTGAAAAATTTATCATACAAGTTCCTGGACGTTTAGACCCACGATTTGTTTGTCGAGATTTTGGTCCGGGTATGAGTGAT
 N  G  N  P  E  K  F  I  I  Q  V  P  G  R  L  D  P  R  F  V  C  R  D  F  G  P  G  M  S  D

                 830                 850                 870                 890
TTTGATATTCAAGGTGATGATAATTCTCCTGGGTTGTATAATTCATACTTCAGTTCATCTAAAGCTGAATCTAATGACTTTATTGGCGGA
 F  D  I  Q  G  D  D  N  S  P  G  L  Y  N  S  Y  F  S  S  S  K  A  E  S  N  D  F  I  G  G

       910                 930                 950                 970                 990
TTTGGTTTAGGTTCTAAATCTCCGTTTAGTTATACTGATACGTTTAGTATTACTTCGTATCATAAAGGTGAAATTCGTGGTTATGTAGCT
 F  G  L  G  S  K  S  P  F  S  Y  T  D  T  F  S  I  T  S  Y  H  K  G  E  I  R  G  Y  V  A

                1010                1030                1050                1070
TACATGGATGGTGATGGTCCACAGATTAAACCTACATTCGTAAAAGAAATGGGTCCAGATGATAAAACTGGTATTGAAATCGTAGTTCCA
 Y  M  D  G  D  G  P  Q  I  K  P  T  F  V  K  E  M  G  P  D  D  K  T  G  I  E  I  V  V  P

      1090                1110                1130                1150                1170
GTTGAAGAAAAAGACTTTAGAAACTTTGCTTATGAAGTTTCTTATATCATGCGACCGTTCAAAGATTTGGCTATCATTAATGGTCTTGAC
 V  E  E  K  D  F  R  N  F  A  Y  E  V  S  Y  I  M  R  P  F  K  D  L  A  I  I  N  G  L  D

                1190                1210                1230                1250
CGCGAAATTGATTATTTTCCGGATTTTGATGACTATTACGGTGTAAATCCAGAAAGATACTGGCCTGATCGTGGTGGATTATATGCTATC
 R  E  I  D  Y  F  P  D  F  D  D  Y  Y  G  V  N  P  E  R  Y  W  P  D  R  G  G  L  Y  A  I

      1270                1290                1310                1330                1350
TACGGTGGTATTGTTTATCCTATCGATGGTGTTATTAGAGACCGTAACTGGCTAAGCATTCGCAATGAAGTGAATTACATTAAGTTTCCA
 Y  G  G  I  V  Y  P  I  D  G  V  I  R  D  R  N  W  L  S  I  R  N  E  V  N  Y  I  K  F  P

                1370                1390                1410                1430
ATGGGTTCACTTGATATTGCTCCATCTCGCGAGGCTCTTTCACTGGATGATCGCACTCGTAAAAATATTATTGAACGAGTTAAAGAACTC
 M  G  S  L  D  I  A  P  S  R  E  A  L  S  L  D  D  R  T  R  K  N  I  I  E  R  V  K  E  L
```

FIGURE 2.—The complete nucleotide sequence of gene *rIIA* and ORF *rIIA.1*, and their corresponding amino acid sequences. The sequence is numbered from the left end of the *Eco*RI-*Hin*dIII *60-rIIA* fragment which falls into the distal part of gene *60* already sequenced (HUANG *et al.* 1988) . Our sequence continues up to the first *Hin*dIII site (located at nucleotide 1970) contained in the *rIIA* gene. For convenience we also present the previously published sequence distal to this *Hin*dIII site (PRIBNOW *et al.* 1981; SUGINO and DRAKE 1984), as well as the beginning of the *rIIB* gene sequence (PRIBNOW *et al.* 1981). Based on the DNA and RNA sequences, the amino acid sequence of rIIA protein is shown under the nucleotide sequence, as is that of the *rIIA.1* protein. The *Eco*RI and *Hin*dIII sites at the ends of the *60-A* fragment, and the next *Hin*dIII site are underlined.

the SHINE-DALGARNO sequence and nucleotide +7 are identical in the two genes. Moreover, after differences between +8 and +12, the coding sequences of the two genes between +13 and +34 are identical in 18 of 22 nucleotides (20 of 22 if one allows a one-base gap once in each sequence). We find this identity particularly remarkable because the 44 protein and the rIIA protein are reported to be the T4 proteins most sensitive to the translational repression mediated by the *regA* gene of T4 (KARAM and BOWLES 1974; WIBERG and KARAM 1983; WINTER *et al.* 1983). Moreover, very recently, WEBSTER, ADARI and SPICER (1989) have shown that the region in gene *44* RNA between −11 and +9 is sufficient to specify regA recognition. RNase protection experiments show interaction between regA protein and this RNA between positions −10 and +2. Therefore we have, in Figure 5, compared the *rIIA* RNA sequence in this region to similar regions from other proteins sensitive to *regA* translational repression. They are arranged roughly in decreasing order of sensitivity to *regA* inhibition (MILLER *et al.* 1987; WINTER *et al.* 1987; WEBSTER, ADARI and SPICER 1989) . A number of correlations of sequence to sensitivity to *regA* inhibition appear; such correlations could help define what constitutes strong and weak *regA* binding sites. We note the following:

1. AAUU before the AUG and AUUA after the AUG are features only of the two strongest sites (*rIIA* and *44*).

2. There is a rough correlation between the number of identical bases (compared to the *rIIA* and *44* sequences) in the region and sensitivity to *regA* inhibition.

```
B
      1450                1470                1490                1510                1530
AGTGAGAAAGCATTTAATGAAGATGTAAAACGATTTAAAGAATCTACATCTCCTCGTCACACATATCGTGAATTGATGAAGATGGGGTAT
S  E  K  A  F  N  E  D  V  K  R  F  K  E  S  T  S  P  R  H  T  Y  R  E  L  M  K  M  G  Y

                1550                1570                1590                1610
TCTGCTCGAGATTATATGATTAGTAATTCAGTCAAATTCACGACTAAAAATCTGTCATATAAAAAGATGCAGAGCATGTTTGAACCTGAC
S  A  R  D  Y  M  I  S  N  S  V  K  F  T  T  K  N  L  S  Y  K  K  M  Q  S  M  F  E  P  D

      1630                1650                1670                1690                1710
AGTAAGTTATGCAACGCGGGAGTTGTGTATGAAGTAAATCTTGACCCTCGACTGAAGCGCATTAAGCAAAGTCATGAAACTTCAGCCGTT
S  K  L  C  N  A  G  V  V  Y  E  V  N  L  D  P  R  L  K  R  I  K  Q  S  H  E  T  S  A  V

                1730                1750                1770                1790
GCATCAAGTTATCGTCTGTTTGGTATTAATACAACAAAAATTAATATCGTTATTGATAATATTAAAAATCGTGTTAATATTGTCCGTGGA
A  S  S  Y  R  L  F  G  I  N  T  T  K  I  N  I  V  I  D  N  I  K  N  R  V  N  I  V  R  G

      1810                1830                1850                1870                1890
TTAGCACGTGCGTTAGATGATAGTGAATTTAATAACACTTTGAATATTCATCATAACGAACGTCTTCTGTTTATTAATCCAGAAGTAGAA
L  A  R  A  L  D  D  S  E  F  N  N  T  L  N  I  H  H  N  E  R  L  L  F  I  N  P  E  V  E

                1910                1930                1950                1970
TCGCAGATTGATTTGCTTCCTGATATTATGGCGATGTTTGAAAGTGATGAAGTTAACATTCATTATTTGTCAGAAATTGAAGCTTTAGTA
S  Q  I  D  L  L  P  D  I  M  A  M  F  E  S  D  E  V  N  I  H  Y  L  S  E  I  E  A  L  V

      1990                2010                2030                2050                2070
AAAAGTTATATTCCAAAGGTAGTTAAAAGTAAAGCTCCTCGTCCTAAAGCTGCTACAGCGTTTAAGTTTGAAATTAAAGACGGGCGCTGG
K  S  Y  I  P  K  V  V  K  S  K  A  P  R  P  K  A  A  T  A  F  K  F  E  I  K  D  G  R  W

                2090                2110                2130                2150
GAAAAGAGGAATTATTTACGCCTCACATCAGAAGCAGATGAAATTACTGGTTATGTAGCGTATATGCATCGTTCTGATATTTTCTCTATG
E  K  R  N  Y  L  R  L  T  S  E  A  D  E  I  T  G  Y  V  A  Y  M  H  R  S  D  I  F  S  M

      2170                2190                2210                2230                2250
GATGGTACTACATCTCTTTGTCATCCATCTATGAATATTTTGATTCGTATGGCTAATCTTATTGGCATTAATGAATTTTATGTTATTCGT
D  G  T  T  S  L  C  H  P  S  M  N  I  L  I  R  M  A  N  L  I  G  I  N  E  F  Y  V  I  R

                2270                2290                2310                2330
CCGCTTTTACAGAAAAAGGTAAAAGAACTCGGTCAGTGCCAATGTATTTTTGAAGCTTTGCGTGATTTATATGTAGATGCTTTTGATGAT
P  L  L  Q  K  K  V  K  E  L  G  Q  C  Q  C  I  F  E  A  L  R  D  L  Y  V  D  A  F  D  D

      2350                2370                2390                2410                2430
GTAGATTATGATAAGTATGTAGGTTATTCAAGTTCAGCTAAACGATATATTGATAAAATTATCAAGTATCCTGAGTTAGATTTTATGATG
V  D  Y  D  K  Y  V  G  Y  S  S  S  A  K  R  Y  I  D  K  I  I  K  Y  P  E  L  D  F  M  M

                2450                2470                2490                2510
AAGTACTTCAGTATAGATGAAGTTTCTGAAGAATATACACGACTCGCTAATATGGTTAGTTCATTACAGGGTGTATATTTTAATGGTGGA
K  Y  F  S  I  D  E  V  S  E  E  Y  T  R  L  A  N  M  V  S  S  L  Q  G  V  Y  F  N  G  G

      2530                2550                2570                2590                2610
AAAGATACCATCGGTCATGACATTTGGACAGTAACTAATCTTTTTGATGTATTATCAAATAATGCTTCAAAAAACAGTGATAAAATGGTT
K  D  T  I  G  H  D  I  W  T  V  T  N  L  F  D  V  L  S  N  N  A  S  K  N  S  D  K  M  V

                2630                2650                2670                2690
GCTGAGTTTACCAAGAAATTCCGTATTGTTTCCGACTTCATCGGTTATCGCAACTCTTTAAGTGATGATGAAGTTTCCCAAATCGCTAAA
A  E  F  T  K  K  F  R  I  V  S  D  F  I  G  Y  R  N  S  L  S  D  D  E  V  S  Q  I  A  K

      2710                2730                2750                2770                2790
ACTATGAAGGCCCTTGCGGCCTAATAAGGAAAATTATGTACAATATTAAATGCCTGACCAAAAACGAACAAGCTGAAATTGTTAAACTGT
T  M  K  A  L  A  A  *  *          M  Y  N  I  K  C  L  T  K  N  E  Q  A  E  I  V  K  L  Y
         end of gene rIIA         start of gene rIIB
                2810                2830
ATTCAAGTGGTAATTACACCCAACAGGAATTGGCTGATTGGCAA
  S  S  G  N  Y  T  Q  Q  E  L  A  D  W  Q
```

FIGURE 2.—Continued.

3. The box between −4 and +7 has no Gs or Cs for the *rIIA* and *44* sequences (disregarding the G of the initiator AUG). For the others the G + C content varies (in this box) from 1 to 3. Again there is a rough inverse correlation between the number of (G + C)s and the *regA* effect.

Structural motifs in the rIIA protein: Mutants in the *rIIA* and *rIIB* genes have similar phenotypes. The proteins seem to have similar characteristics. Could the two be structurally related? Analysis of the *rIIA* and *rIIB* nucleic acid and protein sequences shows no region of extensive similarity, certainly no more than is found when other T4 genes are compared to *rIIA*. We do, however, find a hint that the present *rIIA* gene may have evolved by gene duplication. When tyrosine 101 is aligned with tyrosine 481 (tyrosine 481 is 118 nucleotides downstream of the midpoint amino acid 363 of the *rIIA* gene), a similarity between the two halves of the molecule becomes evident (Figure 6). Although the overall similarity is only 28% (calculated with BESTFIT using the matrix of RISLER), the clustering is impressive, because there are no gaps in the sequence alignment. The predicted structures for these two zones (Figure 3, a and b) do not show extensive similarity.

We have analyzed the rIIA sequence for a variety of known sequence motifs. Starting with leucine 544 and ending with leucine 570 there is the helix-turn-helix motif shown in Figure 7. This motif shows some similarity to the helix-turn-helix motif in prokaryotic repressors of transcription; the similarity seems limited to the glycine at the turn and to the second α-helix motif. Less similarity is seen when consensus protein sequences for activators of transcription or

TABLE 2

Amino acid composition of rIIA and rIIA.1 proteins

Amino acid residue	rIIA Number	rIIA Mole percent	rIIA.1 Number	rIIA.1 Mole percent
Ala	38	5.241	4	5.970
Cys	5	0.690	0	0.000
Asp	57	7.862	6	8.955
Glu	45	6.207	8	11.940
Phe	39	5.379	5	7.463
Gly	41	5.655	1	1.493
His	10	1.379	1	1.493
Ile	63	8.690	5	7.463
Lys	54	7.448	8	11.940
Leu	48	6.621	4	5.970
Met	23	3.172	3	4.478
Asn	42	5.793	3	4.478
Pro	29	4.000	0	0.000
Gln	11	1.517	2	2.985
Arg	41	5.655	6	8.955
Ser	61	8.414	2	2.985
Thr	31	4.276	3	4.478
Val	44	6.069	4	5.970
Trp	4	0.552	0	0.000
Tyr	39	5.379	2	2.985
Isoelectric point	6.58		7.91	

for sigma factors are compared to this rIIA motif. The most striking similarity, however, is found when this rIIA motif is compared to a series of proteins, the prototype of which is ompR, which act as regulators of gene expression in response to environmental change (IKENAKA *et al.* 1988). This extended helix-turn-helix motif is thought to be the DNA-binding part of these proteins which activate gene expression in response to the external environment. Leucine 544 would correspond to the highly conserved leucine (7 out of 9) in these proteins (see Figure 4 of IKENAKA *et al.* 1988) . Also noteworthy is the conservation of isoleucine 553 which corresponds to a highly conserved hydrophobic amino acid (valine, leucine or isoleucine) and which, when mutated in the ompR protein, leads to the loss of the cell's ability to regulate porin synthesis in response to changes in the osmolarity of the culture medium. It has been known for many years that the phenotype of *rIIA* mutants is suppressed by the salt composition and concentration in the culture medium (GAREN 1961; SEKIGUCHI 1966; see SINGER, SHINEDLING and GOLD 1983 for a discussion of the salt effect). The similarity found here raises the possibility that this sequence mediates DNA binding of the rIIA protein in response to changes in the ionic composition of the culture medium.

Since the rIIA protein is strongly associated with the bacterial inner membrane, does it contain the hydrophobic α-helical regions associated with integral membrane proteins? Using GES analysis for identifying hydrophobic α-helices, it was found that the most hydrophobic 17-amino-acid sequence in rIIA had a hydrophobicity of 0.85 kcal/amino acid (ENGELMAN, STEITZ and GOLDMAN 1986). This makes it unlikely that rIIA is an integral membrane protein.

We have carried out extensive computer comparisons between the rIIA protein sequence and the following sequence banks: NBRF, PASTEUR, and our T4 protein sequences bank (168 protein sequences derived from the genes and ORFs contained in the 95-kbp of sequence in our possession). Although short regions of similarity are found with a number of proteins in these banks, no similarity extended over a large proportion of the rIIA protein. One region of short similarity deserves mention. A strong similarity is seen between amino acids 155–181 of the methyl-accepting chemotaxis protein I of *E. coli*, tsr (BOYD, KENDALL and SIMON 1983), and the motif centered on the duplicated EI-(R or T)-GYVAYM sequence in rIIA (Figure 6) . This reinforces the idea that the rIIA protein plays a role in reacting to the external environment of T4 infected cells. This same motif is also found in the region of amino acids 222 to 236 of the T4 protein 63 (RNA ligase).

The *rIIA.1* ORF: This ORF is a hydrophilic, somewhat basic peptide containing no proline, cysteine, or tryptophan (Table 2). Secondary structure analysis, by either the method of BIOU *et al.* (1988) or of GABORIAUD *et al.* (1987), predicts this peptide to be almost entirely α-helical (Figures 3b and 4b). Extensive comparative searches reveal short similarities to a number of *E. coli* proteins. The greatest similarity (60% with 4 gaps) is to a region between amino acids 279–349 of citrate synthase of *E. coli* (see Figure 8). Less impressive, but perhaps more intriguing, is a similarity to the VirG protein of *Agrobacterium tumefaciens.* The *VirG* locus codes for a protein necessary for virulence, and its sequence shows it to be homologous to the OmpR group of protein activators which respond to the cells' environment. In fact, the region of similarity of the VirG protein with rIIA.1 lies just downstream (amino acids 102–169; data not shown) of the presumed helix-turn-helix region of this protein (amino acids 67–93) with apparent homology to the helix-turn-helix region of rIIA.

Mutants in the Ala region of the *rIIA* gene: The unexpected finding of a small ORF between genes *60* and *rIIA* leads to the question of whether some mutants in the original *rII* collection might not be in this ORF. We have sequenced three *rIIA* mutants from the most amino-terminal portion of the *rIIA* genetic map (Ala, BENZER 1961). As shown in Figure 9, all of these mutations are in the beginning of the *rIIA* gene. In all of these mutants the sequence of the *rIIA.1* ORF was identical to the wild-type sequence. We conclude that the BENZER map starts with the *rIIA* coding sequence. The mutant *AP80* is particularly interesting

```
(a)  rIIA  protein.

                                                  50                                               100
MIITTEKETILGNGSKSKAFSITASPKVFKILSSDLYTNKIRAVVRELITNMIDAHALNGNPEKFIIQVPGRLDPRFVCRDFGPGMSDFDIQGDDNSPGL
CECEHCCCEEHCCCCCCCEEEEECCCHEEEEHHHHCCCHHHHHHHHHHHHHHHHHHHHCCCCCCHEEEHCCCCCCCCEEECCCCCCCCCCECCCCCCCCCCE
1212122133134555421111135311234312112111234454544323343113455411123124554542232344555343222355555552
                                                 150                                               200
YNSYFSSSKAESNDFIGGFGLGSKSPFSYTDTFSITSYHKGEIRGYVAYMDGDGPQIKPTFVKEMGPDDKTGIEIVVPVEEKDFRNFAYEVSYIMRPFKD
ECCCEECCCCCCCCEEEEEEECCCCCEEEEEEEEEEEECCCHHEEEEEECCCCCCCCCCCCHHCCCCCCCCCEEEEECCCHCCHHHHHHHHHHHHCCCCC
3221313423343332112115444521332344332342411113332245555553311131355554441333222111112121322333221122
                                                 250                                               300
LAIINGLDREIDYFPDFDDYYGVNPERYWPDRGGLYAIYGGIVYPIDGVIRDRNWLSIRNEVNYIKFPMGSLDIAPSREALSLDDRTRKNIIERVKELSE
HHHHCCCCCCCCECCCCCCCCCCCCCCCCCCCCEEEECCEEECCCEEEECCCEECECCCCCCECCCCCCCCCCCCCHCHCCHHHHHHHHHHHHHHHHHH
2333232222232334333223354322255343133223223132212223331211122132122335433354314131213211213444443313
                                                 350                                               400
KAFNEDVKRFKESTSPRHTYRELMKMGYSARDYMISNSVKFTTKNLSYKKMQSMFEPDSKLCNAGVVYEVNLDPRLKRIKQSHETSAVASSYRLFGINTT
HHHHHHHHHHHHCCCCCCCCHHHHHHCCCCCCEEECCCCCCCCCCCCCHCCHHHHHCCCCCCCCCCCEEEEEECCHHHHHHCCCCCCCCCCCEEEEEEECCC
4232313332114454443111113143344211314322122222123111113343312353112313341133111232221111223232122331
                                                 450                                               500
KINIVIDNIKNRVNIVRGLARALDDSEFNNTLNIHHNERLLFINPEVESQIDLLPDIMAMFESDEVNIHYLSEIEALVKSYIPKVVKSKAPRPKAATAFK
CHHHHHHHHHCCHHHHHHHHHHCCCHHHHHHHHHHHHHHHHHHHEEECCHHHCCHHHCCHHHHHHCCCHHHHHHHHHHHHHHHHHHCCHHHCCCCCCHHHHHHHH
1111123222332323222213331133211211313112211512221123321244433123113323455555555443221112444413344434
                                                 550                                               600
FEIKDGRWEKRNYLRLTSEADEITGYVAYMHRSDIFSMDGTTSLCHPSMNILIRMANLIGINEFYVIRPLLQKKVKELGQCQCIFEALRDLYVDAFDDVD
HHHHCCHHHHCCECEEEECCCHHHHHHHHEEECCECEECCCCCCCCCCCCCHHHHHHHHHHCCCCHHHHHHHHHHHHHHHHHCCHHHHHHHHHHHHHHHHHCCCC
3431231331221121132221212331213323223244423344551333222131342311112234331111234212445554343332212234
                                                 650                                               700
YDKYVGYSSSAKRYIDKIIKYPELDFMMKYFSIDEVSEEYTRLANMVSSLQGVYFNGGKDTIGHDIWTVTNLFDVLSNNASKNSDKMVAEFTKKFRIVSD
CCCCEEECCCHHHHHHHHCCCCHHHHHHHHCCHHHHHHHHHHHHHHHHHHHCCEEECCCCCCECEECEEEHHHHHHHCCCCCCCCHHHHHHHHHHHHEEHH
1312221333112222111233133333313411122444434444432123334355332354312233333321333343332434453232333311
                      725
FIGYRNSLSDDEVSQIAKTMKALAA
HHCCCCCCCCHHHHHHHHHHHHHCC
2222234453124445554444532


(b)  rIIA.1   protein.

                                                  50               67
MKSYKVNLELFDKAVHREYRIIQRFFDMGEAEEFKTRFKDIRDKIQSDTATKDELLEVAEVIKRNMN
HHHCHHCHHHHHHHHHHHHHCEHHHHHCHCHHHHHHHHHHHHHHHHHHHCCCHHHHHHHHHHHHHHHHHCHC
2111221444434342223112112211414444443331312321322212344334554433213
```

FIGURE 3.—Secondary structure prediction of rIIA and rIIA.1 proteins, by the COMBINE methods. (a) rIIA protein. The COMBINE method (BIOU *et al.* 1988) used to predict the secondary structure of the rIIA protein is a combination of three complementary secondary structure prediction methods: homolog prediction (LEVIN *et al.* 1986), GORIII prediction (GIBRAT *et al.* 1987) and the bit pattern prediction method for helix and β-strand structures (BIOU *et al.* 1988) . We show here (from a longer listing output), for each group of four lines: the amino acid number scale; the amino acid sequence; the result of the COMBINE prediction method for each amino acid in one of three states: α-helix (H; helical), β-strand (E; extended) and aperiodic structure (C; coil); the confidence scale index expressing the reliability of the prediction (this numerical value increases from 1 to 5 with an increasing probability to find a particular amino acid residue in a particular state). (b) rIIA.1 protein. Same as in (a) for the rIIA.1 protein.

because it is an ATG → ATA mutation in the initiation codon of the *rIIA* gene. An analogous mutation in the *rIIB* gene, *HD263*, has been shown to be temperature-sensitive for rIIB protein synthesis. The ATG → ATA mutation in the initiation codon of *rIIA* does not lead to a temperature-sensitive defect. The plating efficiency of T4 *AP80* is about 10^{-3} on a λ-lysogen (compared to a nonlysogen) at 20°, at 37° and at 42° (data not shown).

DISCUSSION

We have completed the analysis of the DNA sequence between genes *60* and *rIIB*, and have found a new ORF just upstream of *rIIA*. The complete sequence of this region allows us to make some remarks on the large number of *rIIA* mutants collected during the last 40 yr. First of all, one of the most striking aspects of BENZER's genetic map of spontaneous *rII* mutations is the existence of hot spots (BENZER 1961). Mutation *131* in *rIIA* and *117* in *rIIB* are extraordinarily overrepresented in his collection; mutant *114* in *rIIB* is less hot, but is still remarkable for its frequency. GOLD and his collaborators (PRIBNOW *et al.* 1981; SINGER, SHINEDLING and GOLD 1983) have shown that these three hot spots correspond to runs of six consecutive A:T bp (all, in fact, with the A's in the RNA-like strand). They found no other runs of six consecutive A:T base pairs in the 873-bp fragment that they sequenced. In the *rIIA* gene, site *131* is by far the "hottest spot, and it is also the only run of six consecutive A:T base pairs. We find, however, 12 runs of 5 consecutive A:T base pairs in the *rIIA* gene (10 with A's in the RNA-like strand). Mutations at hot spots in the *rII* genes are thought to arise by slippage of these A:T base pairs when they are traversed by the replication apparatus. Slippage generates frame-

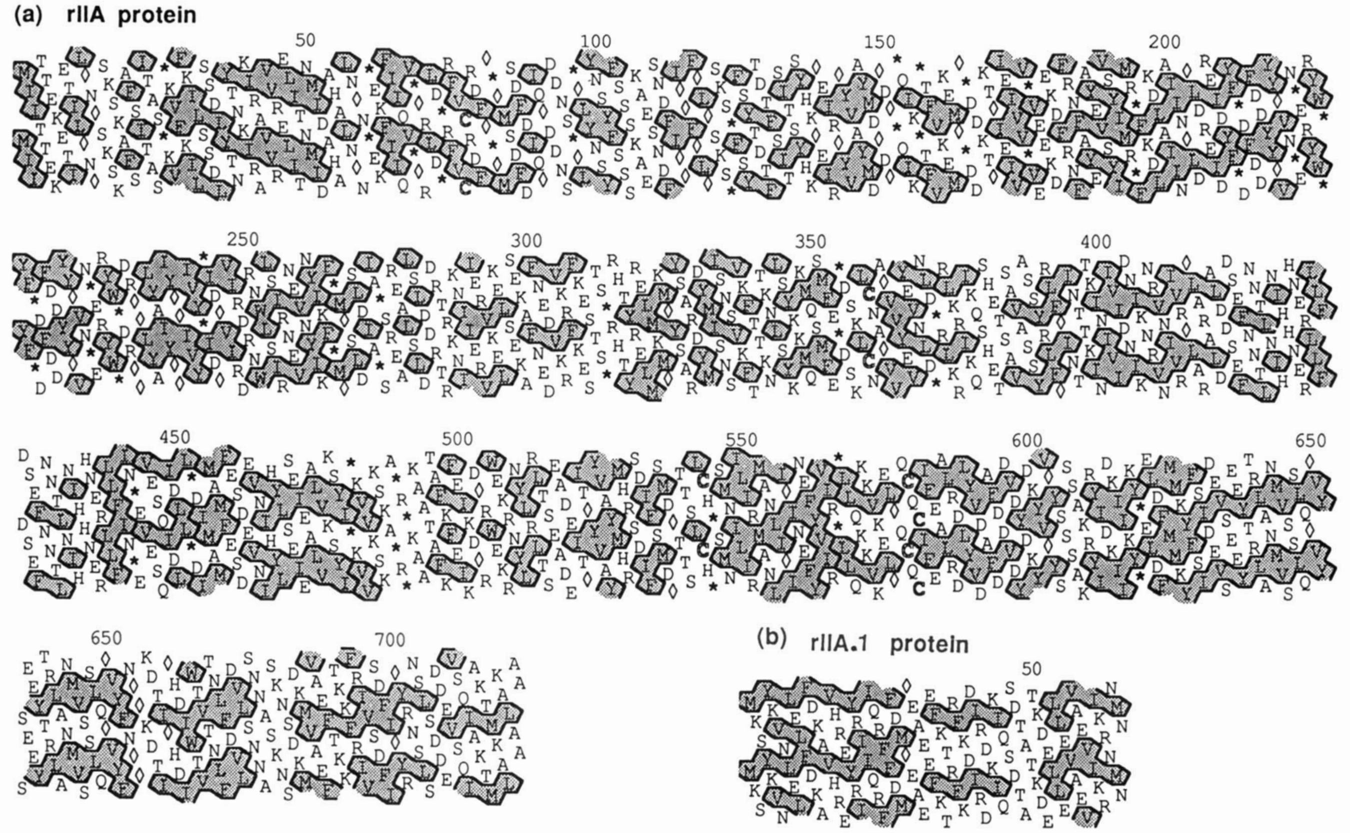

FIGURE 4.—Hydrophobic cluster analysis of rIIA and rIIA.1 proteins. (a) rIIA protein. The rIIA protein has been analyzed with the HCA program from the DNAid$^+$ package (DARDEL and BENSOUSSAN 1988) based on the hydrophobic cluster analysis algorithm of GABORIAUD *et al.* (1987) . The HCA method was originally designed for comparing and aligning amino acid sequences from distantly related proteins (*i.e.*, proteins which cannot be aligned using classical methods of sequence comparison but which can be with the HCA method because they fold into similar three-dimensional structures). The figure shown above is based on a representation of the *rII*A amino acid sequence in the manner of an α-helical three-dimensional pattern expanded at the surface of a cylinder (3.6 amino acids per turn, parallel to the generator of the cylinder). The cylinder is cut along this axis, unrolled onto a two-dimensional surface and then duplicated. Sets of adjacent hydrophobic residues (F, I, L, M, V, W, Y and A or C when they are in a hydrophobic environment) on the surface have have been arranged in hydrophobic clusters. Three special symbols are used: * for prolines considered as breakers of these clusters, ◊ for glycines and ℂ for cysteines. To avoid disruption in the drawing, the end of each line 1, 2 and 3 overlaps, over 18 amino acids, the beginning, respectively, of the lines 2, 3 and 4. (b) rIIA.1 protein. Same as in (a) for the rIIA.1 protein.

shift mutations, either to 7 (unstable) or to 5 (stable) A:T base pairs. The physical-chemical differences between $(dA{:}dT)_5$ and $(dA{:}dT)_6$ do not seem to explain this gigantic difference in mutation frequencies and PRIBNOW *et al.* (1981) have already shown that there must be a context to this effect. Whatever mechanism is at work here, it generates an impressive differential response in replication fidelity to the addition of one dA:dT base pair.

The Ala segment of the *rII* genetic map defines the leftmost region of the *rIIA* gene (BENZER 1961) . The mutations in the Ala segment recombine with every *rII* deletion in the BENZER collection except the deletion *r1272.* Three of these Ala mutations map in the NH_2-terminal portion of the *rIIA* gene (*AP80* in the initiator codon of the protein). We guess that if mutants had arisen in the *rIIA.1* ORF, they would not have been detected as *rIIA* mutants and, in fact, would probably not be viable. Our reasoning is as follows: many *rII* deletions have a right end in the nonessential D region just downstream of gene *rIIB* (BENZER 1961). Only one deletion of the original collection goes as far as the NH_2 terminus of *rIIA*. This suggests that there is a barrier either to deletion formation or to deletion viability upstream of the *rIIA* gene. Until now, it has always been thought that this barrier was the essential gene *60*. If, as thought, deletion formation is strongly dependent on direct repeats of DNA sequences (PRIBNOW *et al.* 1981; SINGER, SHINEDLING and GOLD 1983), there are a large number of potential sites in gene *rIIA.1*. The longest are one 11-bp sequence and five 9-bp sequences in *rIIA.1* which are directly repeated in either *rIIA* or *rIIB* (data not shown). It seems likely, then, that the deletion asymmetry arises simply because *rIIA.1* codes for an essential protein in T4 development.

The codon usage in gene *rIIA.1* is also interesting; it suggests that the gene has evolved to use at least some of the T4 coded tRNAs for its translation. As can be seen in Table 3, the codon usage of *rIIA.1*

Gene	Sequence					G+C
	−12	−4	1	7	34	
rIIA	AUAUGAAUUAAU**GAGG**A	AAUU	*AUG*	AUUA	UCACCACUGAAAAAGAAACAAUUCUUG	0
44	UAAAACUUGAAU**GAGG**A	AAUU	*AUG*	AUUA	CUGUAAAUGAAAAAGAACACAUUCUUG	0
rpbA	UAUUAUGACUAA**AGGU**G	UAUU	*AUG*	ACUA	AAAUUACUGUGAAUUAUACUGUUGAUG	1
dexA	UGAUUUAGC**GAGG**AAAA	UUUA	*AUG*	UUUG	AUUUUAUUAUAGAUUUUGAAACAAUGG	1
dexA.1	AUCUUUAU**GAGG**CGAUU	AUUA	*AUG*	AUUG	AAUUAAGUUGGUACCAGUUUAAAUCUC	1
rIIB	UGCGGCCUAAUA**AGGA**A	AAUU	*AUG*	UACA	AUAUUAAAUGCCUGACCAAAAACGAAC	1
45	AUUUGAAUUGA**AGGA**AA	UUAC	*AUG*	AAAC	UGUCUAAAGAUACUACUGCUCUGCUUA	2
alc	ACAUAACAU**GAGG**ACUU	UAUG	*AUG*	GAUU	UACAACUUAUUACUACUGAAAUGGUCG	2
62	GCGAAAUGCAGU**GGA**AG	UGAU	*AUG*	AGCU	UAUUUAAAGAUGAUAUUCAAUUAAACG	3
52	AUUCACUAGUAU**GGU**AA	AUUU	*AUG*	CAAC	UGAAUAAUCGCGAUUUAAAAAGUAUCA	2
regA	AACUAGCAUU**GGA**AUGG	UAAA	*AUG*	AUUG	AAAUUACUCUUAAAAAACCUGAAGAUU	1
1	AUUAAAUUU**GAGG**AGAA	ACAC	*AUG*	AAAC	UAAUCUUUUUAAGCGGUGUAAAGCGUA	3

FIGURE 5.—The nucleotide sequence surrounding the AUG initiator, and the sensitivity of some T4 genes to the translational repression mediated by the T4 *regA* regulator. The sequences at the beginning of 11 T4 genes for which the sensitivity to regA protein is known, are aligned under the analogous *rIIA* sequence. Sequences are aligned around the first AUG in the coding region. The nucleotides are numbered with the A from the initiator as the +1 reference. Around the initiator, the two tetranucleotides have been separated for legibility. The nucleotides that are identical at a given position (*i.e.*, without deletion or insertion) in genes *rIIA* and 44 are underlined. For all other genes, the nucleotides located in the area between position −4 and position +7, which are identical to those in the *rIIA-44* sequence, have been underlined. The number of (G + C)'s between positions −4 and +7 are shown on the right. The SHINE-DALGARNO sequences for each gene are shown in bold.

```
gprIIA  481-  YIPKVVKSKAPRPKAATAFKFEIKDGRWEKRNYLRLTSEADEITGYVAYMHRSDIFSMDGTTSLCHPSMNILIRMANLI   -559
              |      ||| |      |    |        |   |    |||||||||                |     ||  | |
gprIIA  101-  YNSYFSSSKAESNDFIGGFGLGSKSPFSYTDTFSITSYHKGEIRGYVAYMDGDGPQIKPTFVKEMGPDDKTGIEIVVPV   -179

gprIIA  139-                                      HKGE.IRGYVAYM.DG                              -152
                                                  |  | | |||| | ||
gp63    222-                                      NA.ENIEGYVAVMKDG                              -236

gprIIA  134-                                  SITSYHKG.EIRGYVAYMDGDGPQIKPT                      -160
                                              | |    | | | |||||         |
gptsr   155-                                  RPRDIRNGFE.KQYVAYMEQNDRLHDIA                      -181
```

FIGURE 6.—The *rIIA* gene has possibly evolved by gene duplication. The two 79-amino-acid sequence fragments of rIIA protein starting at amino acid 101 and 481 are aligned with no insertions or deletions; perfect matches are represented in bold (top). In the same manner, similar regions from other proteins (*E. coli* tsr protein and T4 gene *63* product) are aligned with the repeated motif of rIIA protein. Alignments have been obtained with the BESTFIT program using the matrix of RISLER (see MATERIALS AND METHODS). Some gaps have been included in order to maintain the one-to-one correspondence in the alignment.

approximates more closely that of gene *63* than that of *rIIA*, *rIIB* or *E. coli* proteins. Most significant is the AGA arginine codon, which is used frequently in genes *63* and *rIIA.1* but not at all in *rIIB*, nor in highly expressed *E. coli* proteins.What is the significance of the identity of the translation initiator regions of genes *rIIA* and *44*? As we have mentioned, these proteins are the two whose synthesis is most sensitive to translational repression by the T4 regA protein. We think that the identical sequences between −12 and +7 define a strong *regA* repression site. One or all of the parameters discussed in RESULTS and shown in Figure 5 must contribute to the weakening of the *regA* effect on other sites. The significance of the quasi-identity of these two genes between +1 and +34 could involve other regulatory mechanisms; alternatively, this identity could define a protein domain shared by these two proteins.

CAMBELL and GOLD (1982) have suggested a model for *regA* physiology which may be pertinent. The model states that the real purpose of the *regA* gene is to regulate DNA synthesis; *regA*-sensitive RNAs, in this model, all code for proteins involved in a "supra-replisome." The primary ligand of regA protein would be a nucleic acid directly involved in DNA synthesis (RNA primers, for example). *regA*-sensitive RNAs would be secondary ligands used only after the primary ligand is saturated. If this model were correct, the sequence data presented here would imply that rIIA is a component of this "supra-replisome." This is consistent with the complete arrest of DNA synthesis

```
  Source of
Helix-Turn-Helix                         Sequences
   Motif
__________________            _______________________________________

ompR consensus             LTEKBPDLVVLDLNLPGMDGLELLKRL
                           |      ||||  || ||||| ||| |
rIIA protein          544- LCHPSMNILIRMANLIGINEFYVIRPL  -570
                                      ||  ||| |  | | |
repressors consensus                 EVAQKLGVSQSTVSRFI
```

FIGURE 7.—Amino acid alignment among rIIA protein, the ompR consensus, and the helix-turn-helix motif of certain prokaryotic repressors. The region located between amino acids 544 and 570 of the rIIA protein is aligned with two different consensus sequences: the ompR consensus derived from the putative regions of the nine proteins (see Figure 4 of IKENAKA *et al.* 1988) involved in sensory systems that share common features with osmoregulation, and the helix-turn-helix consensus derived from the 21 protein sequences (see Figure 12 of PABO and SAUER 1984) which are proved or assumed to be DNA-binding proteins. Alignments were done using PROFILE and PROFILEGAP programs (see MATERIALS AND METHODS) . Perfect matches are shown in bold.

```
rIIA.1   protein      1-MKSYKVNLELFDKAVHREYRIIQRFFDMGE..AEEFKTR...FKDIRDKIQSDTATKDELLEVA.EVIKRNMN  -66
                        | | |   ||| ||  ||  |  |||||      ||  |   |||    |  | |||||||||  ||    ||
citrate  synthase   300-ISSVKHIPEFFRRAKDKNDSF..RLMGFGHRVYKNYDPRATVMRETCHEVLKELGTKDDLLEVAMELENIALN  -348
```

FIGURE 8.— Similarity between the rIIA.1 protein and the citrate synthase from *E. coli*. All experimental details are as in Figure 6.

```
Mutation     Position                    Sequences
_________    ________    _______________________________________________

wild type                ATG ATT ATC ACC ACT GAA AAA GAA ACA ATT CTT GGT

AP80             3       ..A ... ... ... ... ... ... ... ... ... ... ...
AP129           32       ... ... ... ... ... ... ... ... ... ... .C. ...
F120            35       ... ... ... ... ... ... ... ... ... ... ... .A.

                          M   I   I   T   T   E   K   E   T   I   L   G
                          I   .   .   .   .   .   .   .   .   .   P   D
```

FIGURE 9.— Sequence analysis of three mutants in the Ala region of *rIIA*. The three mutants (*AP80, AP129* and *F120*) isolated originally by S. BENZER and located in the Ala region of *rIIA*, were sequenced by primer-extension of the r2A3 oligodeoxynucleotide, using as template T4 RNA isolated 5 min after infection at 30°. The nucleotide sequence of the beginning of the wild-type *rIIA* gene is shown at the top. For each mutant, only the nucleotide replacing the corresponding wild-type nucleotide is shown in bold. The amino acids deduced from these sequences are shown at the bottom.

TABLE 3

T4 rare codon usage

Amino acid	Codon	*rIIA*	*rIIA.1*	*rIIB*	*63*[a]	*E. coli* L[b]	*E. coli* H[c]
Gly	GGA	0.17	1.00	0.08	0.37	0.11	0.005
Arg	AGA	0.07	0.33	0.00	0.44	0.12	0.00
Ile	ATT	0.76	0.80	0.95	0.60	0.50	0.17
Thr	ACA	0.42	0.00	0.24	0.15	0.11	0.04
Leu	TTA	0.29	0.25	0.00	0.32	0.15	0.02
Ser	TCA	0.28	0.05	0.24	0.36	0.16	0.02
Gln	CAA	0.45	1.00	0.82	0.50	0.36	0.14
Pro	CCA	0.34	0.00	0.33	0.44	0.23	0.15

[a] T4 gene *63* (RAND and GAIT 1984).

[b] Codon usage for poorly expressed *E. coli* genes are from UWGCG (see MATERIALS AND METHODS, and GRANTHAM *et al.* 1981).

[c] Same as [b] except for highly expressed genes.

seen when *rIIA* mutants infect *rex*$^+$ bacteria (GAREN 1961; SEKIGUCHI 1966).

Although the amino acid sequence of the rIIA protein does not suggest an overall homology to any known protein, there are sequence elements which may suggest how, if not why, it functions. The apparent homology between the DNA-binding domain of the ompR family of proteins and a motif in rIIA suggests that the rIIA protein may bind to DNA in response to some change in the ionic composition of the environment. Since the rIIA protein does not seem to be membrane spanning, even though it is closely associated with the inner membrane of infected cells (TAKACS and ROSENBUSCH 1975), it may play a role quite analogous to this group of proteins. This idea is reinforced by the repeated motif in rIIA which is similar to a region of the tsr protein. The tsr motif is adjacent to, but not in, the presumed membrane spanning region of this protein. It would not be unreasonable for rIIA to be a protein with two elements that respond to some ionic component in the external medium, and that the concentration of these ions determines whether or not rIIA detaches from the inner membrane to affix itself to some sequence on T4 DNA. Alternatively, rIIA may simultaneously bind to DNA and the inner membrane and the response to changes in the medium may lead to dissociation of one of these contacts. If there really were an analogy to the ompR group of proteins, one wonders what protein would play the role of the envZ part of this system. EnvZ is suggested to be an integral membrane protein which is the site of attachment (and detachment) of ompR. It is envZ that senses directly the osmolarity of the external medium and transmits this information to ompR on the inner surface of the membrane. Could rIIB be the envZ analog which fixes rIIA to the membrane? Hydrophobic cluster analysis of envZ predicts a membrane spanning region whereas the analysis for rIIB is less clear, although there is a potential 20 amino acid hydrophobic α-helix in the middle of this protein (data not shown).

We are grateful to LARRY GOLD and JOHN F. PULITZER for helpful discussions and criticisms during the course of this work and during the preparation of the manuscript. We thank WAI MUN HUANG for communicating her unpublished results, VALÉRIE BIOU and JEAN GARNIER for their help in the determination of rIIA and rIIA.1 protein structures, and JEAN-LUC POPOT for the GES analysis of rIIA and rIIA.1 proteins. The research reported here was supported by research grants from the CNRS (Centre National de la Recherche Scientifique) and INSERM (Institut National de la Santé et de la Recherche Medicale).

LITERATURE CITED

BENZER, S., 1959 On the topology of the genetic fine structure. Proc. Natl. Acad. Sci. USA **45:** 1607–1620.

BENZER, S., 1961 On the topography of the genetic fine structure. Proc. Natl. Acad. Sci. USA **47:** 403–415.

BIOU, V., J. F. GIBRAT, J. LEVIN, B. ROBSON and J. GARNIER, 1988 Secondary structure prediction: combination of three different methods. Protein Eng. **2:** 185–191.

BOLIVAR, F., R. L. RODRIGUEZ, M. C. BETLACH and H. W. BOYER, 1977 Construction and characterization of new cloning vehicles. I. Ampicillin-resistant derivatives of the plasmid pMB9. Gene **2:** 75–93.

BOYD, A., K. KENDALL and M. I. SIMON, 1983 Structure of the serine chemoreceptor in *Escherichia coli.* Nature **201:** 623–626.

BRODY, E. N., D. RABUSSAY and D. H. HALL, 1983 Regulation of transcription of prereplicative genes, pp. 174–183 in *Bacteriopnage T4*, edited by C. K. MATHEWS, E. M. KUTTER, G. MOSIG and P. B. BERGET. AMERICAN SOCIETY FOR MICROBIOLOGY, WASHINGTON. D.C.

CAMPBELL, K., and L. GOLD, 1982 Construction of the bacteriophage T4 replicatin machine: regulation of synthesis of component proteins, pp. 69–81 in *Interaction of Translational and Transcriptional Controls in the Regulation of Gene Expression*, edited by M. GRUNBERG-MANAGO and B. SAFER. Elsevier, New York.

CASADABAN, M. J., and S. N. COHEN, 1980 Analysis of gene control signals by DNA fusion and cloning in *Escherichia coli.* J. Mol. Biol. **138:** 179–207.

CRICK, F. H. C., L. BARNETT, S. BRENNER and R. J. WATTS-TOBIN, 1961 General nature of the genetic code for proteins. Nature **192:** 1227–1232.

DAEGELEN, P., and E. N. BRODY, 1976 Early bacteriophage T4 transcription. A diffusible product controls *rIIA* and *rIIB* RNA synthesis. J. Mol. Biol. **103:** 127–142.

DAEGELEN, P., and E. N. BRODY, 1990 The *rIIA* gene of bacteriophage T4. II. Regulation of its messenger RNA synthesis. Genetics **125:** 249–260.

DARDEL, F., and P. BENSOUSSAN, 1988 A Macintosh full screen editor featuring a built in regular expression interpreter for the search of specific patterns in biological sequences using finite state automata. Comput. Appl. Biosci. **4:** 483–486.

DAYHOFF, M. O., W. C. BARKER and L. T. HUNT, 1983 Establishing homologies in protein sequences. Methods Enzymol. **91:** 524–545.

DEVEREUX, J., P. HAEBERLI and O. SMITHIES, 1984 A comprehensive set of sequence analysis programs for the VAX. Nucl. Acids Res. **12:** 387–395.

DRETZEN, G., M. BELLARD, P. SASSONE-CORSI and P. CHAMBON, 1981 A reliable method for the recovery of DNA fragments from agarose and acrylamide gels. Anal. Biochem. **112:** 295–298.

ENGELMAN, D. M., T. A. STEITZ and A. GOLDMAN, 1986 Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. Annu. Rev. Biophys. Biophys. Chem. **15:** 321–353.

ENNIS, H. L., and K. D. KIEVITT, 1973 Association of the rIIA protein with bacterial membrane. Proc. Natl. Acad. Sci. USA **70:** 1468–1472.

GABORIAUD, C., V. BISSERY, T. BENCHETRIT and J. P. MORNON, 1987 Hydrophobic cluster analysis: an efficient new way to compare and analyze amino acid sequences. FEBS Lett. **224:** 149–155.

GAREN, A., 1961 Physiological effects of *rII* mutations in bacteriophage T4. Virology **14:** 151–163.

GRANTHAM, R., C. GAUTIER, M. GOUY, M. JACOBZONE and R. MERCIER, 1981 Codon catalog usage is a genome strategy modulated for gene expressivity. Nucl. Acids Res. **9:** r43–r74.

HUANG, W. H., 1986 The 52-protein subunit of T4 DNA polymerase is homologous to the gyrA-protein of gyrase. Nucl. Acids Res. **14:** 7379–7390.

HUANG, W. M., S. Z. AO, S. CASJENS, R. ORLANDI, R. ZEIKUS, R. WEISS, D. WINGE and M. FANG, 1988 A persistent untranslated sequence within bacteriophage T4 DNA topoisomerase gene *60.* Science **239:** 1005–1012.

HUANG, W. M., and J. M. BUCHANAN, 1974 Synergistic interactions of T4 early proteins concerned with their binding to DNA. Proc. Natl. Acad. Sci. USA **71:** 2226–2230.

IKENAKA, K., K. TSUNG, D. E. COMEAU and M. INOUYE, 1988 A dominant mutation in *Escherichia coli OmpR* lies within a domain which is highly conserved in a large family of bacterial regulatory proteins. Mol. Gen. Genet. **211:** 538–540.

KARAM, J. D., and M. G. BOWLES, 1974 Mutation to overproduction of bacteriophage T4 genes products. J. Virol. **13:** 428–438.

LIPMAN, D. J., and W. R. PEARSON, 1985 Rapid and sensitive protein similarity searches. Science **227:** 1435–1441.

MANOIL, C., N. SINHA and B. ALBERTS, 1977 Intracellular DNA-protein complexes from bacteriophage T4-infected cells isolated by a rapid two-step procedure. J. Biol. Chem. **252:** 2734–2741.

MATZ, K., M. SCHMANDT and G. N. GUSSIN, 1982 The *rex* gene of bacteriophage λ is really two genes. Genetics **102:** 319–327.

MCPHEETERS, D. S., A. CHRISTENSEN, E. T. YOUNG, G. STORMO and L. GOLD, 1986 Translational regulation of expression of the bacteriophage T4 lysozyme gene. Nucleic Acids Res. **14:** 5813–5826.

MESSING, J., and J. VIEIRA, 1982 A new pair of M13 vectors for selecting either DNA strand of double-digest restrictions fragments. Gene **19:** 269–276.

MILLER, E. S., J. KARAM, M. DAWSON, M. TROJANOWSKA, P. GAUSS and L. GOLD, 1987 Translational repression: biological activity of plasmid-encoded bacteriophage T4 regA protein. J. Mol. Biol. **194:** 397–410.

NEEDLEMAN, S. B., and C. D. WUNSCH, 1970 A general method applicable to the search for similarities in the amino-acid sequence of two proteins. J. Mol. Biol. **48:** 443–453.

NORRANDER, J., T. KEMPE and J. MESSING, 1983 Construction of improved M13 vectors using oligodeoxynucleotide directed mutagenesis. Gene **26:** 101–106.

PABO, C. O., and R. T. SAUER, 1984 Protein-DNA recognition. Annu. Rev. Biochem. **53:** 293–321.

PRIBNOW, D., D. C. SIGURDSON, L. GOLD, B. S. SINGER and C. NAPOLI, 1981 *rII* cistrons of bacteriophage T4. DNA sequence around the intercistronic divide and positions of genetic landmarks. J. Mol. Biol. **149:** 337–376.

RAND, K. N., and M. J. GAIT, 1984 Sequence and cloning of bacteriophage T4 gene *63* encoding RNA ligase and tail fibre attachment activities. EMBO J. **3:** 397–402.

RIPLEY, L. S., J. S. DUBINS, J. G. DEBOER, D. M. DE MARINI, A. M. BOGERD and K. N. KREUZER, 1988 Hotspot sites for acridine-induced frameshift mutations in bacteriophage T4 correspond to sites of action of the T4 type II topoisomerase. J. Mol. Biol. **200:** 665–680.

RISLER, J. L., M. O. DELORME, H. DELACROIX and A. HENAUT, 1988 Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix. J. Mol. Biol. **204:** 1019–1029.

SANGER, F., S. NICKLEN and A. R. COULSON, 1977 DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74:** 5463–5467.

SEKIGUCHI, M., 1966 Studies on the physiological defect in *rII* mutants of bacteriophage T4. J. Mol. Biol. **16:** 503–522.

SELZER, G., A. BOLLE, H. KRISCH and R. EPSTEIN, 1978 Construction and properties of recombinant plasmids containing the *rII* genes of bacteriophage T4. Mol. Gen. Genet. **159:** 301–309.

SELZER, G., D. BELIN, A. BOLLE, G. VAN HOUWE, T. MATTSON and R. EPSTEIN, 1981 *In vivo* expression of the *rII* region of bacteriophage T4 present in chimeric plasmids. Mol. Gen. Genet. **183:** 505–513.

SHINEDLING, S. T., L. T. WALKER and L. GOLD, 1986 Cloning the complete *rIIB* gene of bacteriophage T4 and some observations concerning its middle promoteurs. J. Virol. **60:** 787–792.

SHINEDLING, S. T., M. GAYLE, D. PRIBNOW and L. GOLD, 1987a Mutations affecting translation of the bacteriophage T4 *rIIB* gene cloned in *Escherichia coli.* Mol. Gen. Genet. **207:** 224–232.

SHINEDLING, S. T., B. S. SINGER, M. GAYLE, D. PRIBNOW, E. JARVIS, B. EDGAR and L. GOLD, 1987b Sequences and studies of bacteriophage T4 *rII* mutants. J. Mol. Biol. **195:** 471–480.

SINGER, B. S., S. T. SHINEDLING and L. GOLD, 1983 The *rII* genes: a history and a prospectus, pp. 327–333 in *Bacteriophage T4,* edited by C. K. MATHEWS, E. M. KUTTER, G. MOSIG and P. B. BERGET. American Society for Microbiology, Washington D.C.

SMITH, T. F., and M. S. WATERMAN, 1981 Comparison of biosequences. Adv. Appl. Math. **2:** 482–489.

SUGINO, A., and J. W. DRAKE, 1984 Modulation of mutation rates in bacteriophage T4 by a base-pair change a dozen nucleotides removed. J. Mol. Biol. **176:** 239–249.

TAKACS, B. J., and J. P. ROSENBUSCH, 1975 Modification of *Escherichia coli* membranes in the prereplicative phase of T4 infection. Specificity of association and quantification of bound phage proteins. J. Biol. Chem. **250:** 2339–2350.

UZAN, M., R. FAVRE and E. N. BRODY, 1988 A nuclease that cuts specifically in the ribosome binding site of some T4 mRNAs. Proc. Natl. Acad. Sci. USA **103:** 8895–8899.

UZAN, M., Y. D'AUBENTON-CARAFA, R. FAVRE, V. DE FRANCISCIS and E. N. BRODY, 1985 The T4 mot protein functions as part of a pre-replicative DNA-protein complex. J. Biol. Chem. **260:** 633–639.

WEBSTER, K. R., H. Y. ADARI and E. K. SPICER, 1989 Bacteriophage T4 RegA protein binds to the Shine-Dalgarno region of gene 44 mRNA. Nucleic Acids Res. **17:** 10047–10068.

WEINTRAUB, S. B., and F. R. FRANKEL, 1972 Identification of the T4*rIIB* gene product as a membrane protein. J. Mol. Biol. **70:** 589–615.

WIBERG, J. S., and J. D. KARAM, 1983 Translational regulation in T4 phage development, pp. 193–201 in *Bacteriophage T4,* edited by C. K. MATHEWS, E. M. KUTTER, G. MOSIG and P. B. BERGET. American Society for Microbiology, Washington D.C.

WINTER, R. B., L. MORRISSEY, P. GAUSS, L. GOLD, T. HSU and J. KARAM, 1987 Bacteriophage T4 regA protein binds to mRNAs and prevents translation initiation. Proc. Natl. Acad. Sci. USA **84:** 7822–7826.

Communicating editor: J. W. DRAKE